Predictive Hidden Markov Model Selection for Speech Recognition

Jen-Tzung Chien, *Senior Member, IEEE,* and Sadaoki Furui, *Fellow, IEEE*

***Abstract*—This paper surveys a series of model selection approaches and presents a novel *predictive information criterion* (PIC) for hidden Markov model (HMM) selection. The approximate Bayesian using Viterbi approach is applied for PIC selection of the best HMMs providing the largest prediction information for generalization of future data. When the perturbation of HMM parameters is expressed by a product of *conjugate prior densities*, the segmental prediction information is derived at the frame level without Laplacian integral approximation. In particular, a multivariate t distribution is attained to characterize the prediction information corresponding to HMM mean vector and precision matrix. When performing model selection in tree structure HMMs, we develop a top-down *prior/posterior propagation* algorithm for estimation of structural hyperparameters. The prediction information is determined so as to choose the best HMM tree model. Different from maximum likelihood (ML) and minimum description length (MDL) selection criteria, the parameters of PIC chosen HMMs are computed via *maximum a posteriori* estimation. In the evaluation of continuous speech recognition using decision tree HMMs, the PIC criterion outperforms ML and MDL criteria in building a compact tree structure with moderate tree size and higher recognition rate.**

***Index Terms*—Approximate Bayesian, decision tree state tying, model selection, multivariate t distribution, predictive information criterion, prior/posterior propagation, speech recognition.**

I. Introduction

MODEL selection is a major problem in signal processing where the model parameters and their number are unknown and therefore must be estimated. To achieve robust data modeling, it is necessary to precisely estimate the underlying parameters of a stochastic model and properly determine the size of model at the same time. The estimated models are then tested for robustness against the *underestimation* or *overestimation* dilemma. Sometimes, this model selection problem is referred as model identification [1], model regularization [22], model order estimation [23], or stochastic model complexity [26]. In this study, we aim to develop a novel predictive information criterion to estimate hidden Markov models (HMMs) and simultaneously select the proper size of HMMs to represent the observed data. HMM model selection is not only useful for speech recognition but also for general data clustering/modeling applications.

For applications to speech recognition, many research topics involve model selection problem. For example, when constructing speech HMMs, we should assign the number of states in an HMM as well as the number of mixture components in an HMM state. It is feasible to apply model selection criterion to determine these numbers from the observed data. In the study of speaker clustering [24], a hierarchy of HMMs was built for individual speaker cluster. Each cluster contained HMMs corresponding to a set of similar speakers. When performing speaker-independent (SI) speech recognition, the closest cluster of HMMs were chosen for recognition. Similarly, the noise-cluster HMMs were estimated to improve the performance of noisy speech recognition [34]. How to decide the cluster number and population was identified as a model selection problem. Further, in the research of speaker adaptation, we endeavored to adapt the existing SI HMM parameters to a new speaker. It was necessary to share adaptation data for different distribution identities and perform structural adaptation [30]. Dynamic model sharing should be established [5]. Even in language modeling research, the model selection problem presented itself in category n-gram models, where the conditional probability of a current word given its word history was approximated using its word category [31]. Specifying the word categories and the cluster numbers was important. In general, a tree model was useful to control the degree of model sharing [5], [28] although *there was no guarantee of tree optimization.* How to determine the suitable tree level and sharing population was a challenging model selection problem. In this study of HMM decision tree state tying, all observation frames corresponding to a context-independent phonetic unit are collected and split according to the phonetic questions of their contexts. It is important to choose the best split question and validate whether the split should be terminated or not. The complexity of the decision tree is determined so that the tied context-dependent HMM parameters are properly estimated [3], [4], [7], [29].

In previous studies, model selection problems using HMMs were solved empirically without evaluating the fitness between the observed data and the estimated model. The robustness of speech recognition cannot be guaranteed. To prevent subjective judgment, MDL and BIC model selection methods were explored for decision tree state tying [8], [29]. This paper presented a predictive information criterion (PIC) to select and estimate HMM parameters where the model regularization was incorporated in decision tree construction. We combined the *predictive Bayesian* [17], [21] and *structural Bayesian* [28] approaches to develop a new PIC model selection for HMMs. The segmental predictive information of HMM parameters

Manuscript received August 22, 2003; revised March 1, 2004. This work was supported in part by the National Science Council, Taiwan, R.O.C., under Contract NSC92-2213-E006-059. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jerome R. Bellegarda.

J.-T. Chien is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan, R.O.C. (e-mail: jtchien@mail.ncku.edu.tw).

S. Furui is with the Department of Computer Science, Tokyo Institute of Technology, Tokyo 152-8552, Japan (e-mail: furui@cs.titech.ac.jp).

Digital Object Identifier 10.1109/TSA.2005.845810

were derived without Laplacian integral approximation. In this study, we modeled the uncertainty of HMM parameters using a product of conjugate prior densities for two reasons. One was to develop the top-down prior/posterior propagation algorithm for computation of prediction information. The other was to estimate the HMM parameters via maximum *a posteriori* (MAP) theory [13]. To that purpose, we selected the best model and estimated the parameters for speech recognition based on decision tree HMMs. PIC criterion was found to be effective to cluster the context-dependent speech frames into compact groups. The HMM parameters were properly estimated for continuous Mandarin speech recognition. In the following Section, a series of model selection approaches are surveyed. The PIC model selection is presented and compared with previous approaches. Section III focuses on the formulation of segmental prediction information for HMMs. The calculation of hierarchical hyperparameters and the interpretation of model selection are addressed. In Section IV, the experiments on decision tree based acoustic modeling are evaluated in terms of recognition rate, HMM state number and processing time. Finally, Section V describes the conclusions drawn from this study.

II. Model Selection Approaches

The model selection problem has been widely discussed in the literature of statistics, information theory, neural network and signal processing. This problem aims at selecting a parametrical model M_m with distribution $f(X|\lambda_m, \mathrm{k}_m)$ for the observed data sequence $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ and trying to estimate the vector parameter $\lambda_m = \{\lambda_1, \ldots, \lambda_{\mathrm{k}_m}\}$, where the number of parameters k_m is also to be estimated. Traditionally, the maximum likelihood (ML) model selection leans toward choosing the highest possible dimension, which leads to overestimation and an overlarge model [27]. Also, the likelihood function is sensitive to the small parameter variations around the true values [1]. Several universal approaches were presented to penalize the overlarge model and de-emphasize the sensitive likelihood function.

A. AIC, BIC, MDL, and NPC Criteria

Akaike [1] presented the Akaike's information criterion (AIC) for model selection. This pioneering work adopted the *estimate of the mean log likelihood*

$$S(h\colon f(\cdot|M_m)) = E_X[\log f(X|M_m)] = \int h(X)\log f(X|M_m)dX \quad (1)$$

as a criterion of fitness of the structural model $f(X|M_m)$ to the true distribution $h(X)$ of X. From the possible models $\mathrm{M} = \{M_m\}$ containing parameters $\Lambda = \{\lambda_m\}$ and their numbers $\mathrm{K} = \{\mathrm{k}_m\}$, AIC criterion selects the model $M_{\mathrm{AIC}} = (\lambda_{\mathrm{AIC}}, \mathrm{k}_{\mathrm{AIC}})$ by

$$\begin{aligned} M_{\mathrm{AIC}} &= \arg\min_{M_m \in \mathrm{M}} \mathrm{AIC}(M_m) \\ &= \arg\min_{\lambda_m \in \Lambda, \mathrm{k}_m \in \mathrm{K}} [-\log f(X|\lambda_m, \mathrm{k}_m) + \mathrm{k}_m]. \end{aligned} \quad (2)$$

The selected parameter λ_{AIC} is an ML estimate λ_m^{ML} of parameter λ_m. AIC criterion is analogous to minimizing the entropy $-S(h\colon f(\cdot|M_m))$.

Schwarz [27] resolved the problem from a Bayesian perspective and proposed the Bayesian information criterion (BIC) for model selection. By considering *a priori* model probability $P(M_m)$ and *a priori* parameter distribution $g(\lambda|M_m)$, the logarithm of the integral of *a posteriori* distribution

$$\log f(X, M_m) = \log \int P(M_m) f(X|\lambda, M_m) dg(\lambda|M_m) \quad (3)$$

is maximized to select the most probable model. BIC selection is performed in accordance with

$$\begin{aligned} M_{\mathrm{BIC}} &= \arg\max_{M_m \in \mathrm{M}} \mathrm{BIC}(M_m) \\ &= \arg\max_{\lambda_m \in \Lambda, \mathrm{k}_m \in \mathrm{K}} \left[\log f(X|\lambda_m, \mathrm{k}_m) - \frac{1}{2}\mathrm{k}_m \log n\right]. \end{aligned} \quad (4)$$

Both AIC and BIC were derived by assuming that the observation data come from a Koopman-Darmois exponential distribution family [27]. The only difference between (2) and (4) is due to the second term playing the role of *penalty* for selecting the high dimensions. BIC can easily choose a lower-dimensional model compared to AIC. If the second term is neglected, the BIC criterion is simplified to a *ML model selection* where only likelihood function affects the selected models. A tunable penalization parameter was merged in BIC for acoustic decision tree state tying [8].

Rissanen [26] found an interesting relation between estimation and coding, from which he was able to exploit the minimum description length (MDL) selection criterion. From the data coding viewpoints, MDL was designed to find the minimum number of bits required to describe the observation data. When formulating MDL, the real-valued parameters were converted to integers by dividing them by their precision. The prior probability was determined using the universal prior for *integers* and optimizing the precision. This MDL approach encoded each component of parameter λ_m by $(1/2)\log n$ bits and allocated the observation X by $-\log f(X|\lambda_m^{\mathrm{ML}}, \mathrm{k}_m)$ bits. Although MDL and BIC initialize from different aspects, they come up with the same formula

$$\mathrm{MDL}(M_m) = -\log f(X|\lambda_m^{\mathrm{ML}}, \mathrm{k}_m) + \frac{1}{2}p \cdot \mathrm{k}_m \cdot \log n \quad (5)$$

where a penalization factor p is incorporated to control model complexity [8]. Again, the selected parameter λ_{MDL} is an ML estimate λ_m^{ML}. Shinoda and Watanabe [29] applied MDL for acoustic decision tree construction. MDL was employed to unsupervised learning of mixture models [12]. It gave good initialization in expectation-maximization (EM) algorithm for parameter estimation.

Merhav [23] presented a model order estimator based on Neyman-Pearson hypotheses testing criterion (NPC). He derived order estimator $\mathrm{k}_{\mathrm{NPC}}$ by minimizing the underestimation probability $P(\mathrm{k}_{\mathrm{NPC}} < \mathrm{k})$ under the constraint that the overestimation probability $P(\mathrm{k}_{\mathrm{NPC}} > \mathrm{k})$ decays faster than $2^{-\varepsilon n}$ for $\varepsilon > 0$. Assuming elements of X are distributed with Koopman-Darmois exponential densities, the estimator is obtained by

$$\mathrm{k}_{\mathrm{NPC}} = \min_{\mathrm{k}_m \in \mathrm{K}} \left\{ \mathrm{k}_m : \frac{1}{n} \log \frac{f(X \,|\lambda_0^{\mathrm{ML}}, \mathrm{k}_0)}{f(X \,|\lambda_m^{\mathrm{ML}}, \mathrm{k}_m)} < \varepsilon \right\}. \quad (6)$$

Here, k_0 is an *a priori* known upper bound for model order. $(\lambda_0^{\mathrm{ML}}, \lambda_m^{\mathrm{ML}})$ are ML estimates with dimensions $(\mathrm{k}_0, \mathrm{k}_m)$. Parameter ε controls the trade-off between the overestimation and underestimation probabilities. In the area of neural network, it is crucial to develop *regularization theory* to select salient synaptic weights and build a feedforward network model [15]. The Bayesian inference approach has been used as one solution for the regularization problem [22].

B. Predictive Information Criterion (PIC)

In general, the criteria of AIC, BIC, MDL and NPC are universal and applicable to coding, estimation, prediction, adaptation and pattern recognition. Most implementation procedures have been carried out for *integer data* in *coding/compression* system [26]. The penalization was presumed identical for each parameter component. The ML estimates λ_m^{ML} were first calculated for different models $\mathrm{M} = \{M_m\}$. The most probable model was selected by testing whether its parameter number k_m was too large or too small. These methods focused on finding the model order estimate $\hat{\mathrm{k}}$ based on *ML parameter* λ_m^{ML}. They used the *Taylor expansion* of $\log f(X \,|\lambda_m, \mathrm{k}_m)$ around λ_m^{ML} to express its uncertainty due to parameter λ_m. A Fisher information (Hessian) matrix was calculated to fulfill the *Laplacian integral approximation*. The model penalty was explicitly represented using the parameter number k_m. Instead of using λ_m^{ML} and k_m, we present a new model selection where the model complexity is embedded in the prior density of parameter $g(\lambda \,|M_m)$ and the parameter uncertainty of posterior density $f(\lambda \,|X, M_m)$. The model structure and the type of model parameters are represented through the *prediction information*. Related works are described below.

Geisser and Eddy [14] addressed a predictive approach to model selection where the *predictive density* was maximized for selection. The predictive sample reuse (PSR) criterion was presented for structural model selection. Djuric and Kay [10] used the predictive density as a criterion for model selection. A normal linear regression data modeling was considered to calculate the predictive density. The cross validation principle was employed to yield the consistent model order estimates [11]. MacKay [22] adopted the *Bayesian evidence* to penalize the over-complex model. Based on Bayesian philosophy, he selected the model bearing the best *prediction* capability. The selected model had good *generalization* to predict future data occurrence. The prior density of model parameter $g(\lambda \,|M_m)$ acted as a regularizer. Instead of using ML estimate λ_m^{ML}, we will present the maximum *a posteriori* (MAP) parameter estimate λ_m^{MAP} for the selected model. Specially, we adopt an information-theoretic criterion, called the *predictive information criterion* (PIC), which is expressed by the logarithm of the predictive distribution

$$\begin{aligned}\mathrm{PIC}(M_m)& \\ &= \log f(X \,|M_m) = \log \int f(X|\lambda, \varphi_m) g(\lambda \,|\varphi_m) d\lambda \\ &\cong \underbrace{\log f(X|\lambda_m^{\mathrm{MAP}}, \varphi_m)}_{\text{log likelihood}} + \underbrace{\log g(\lambda_m^{\mathrm{MAP}}|\varphi_m) + \log \Delta\lambda}_{\text{model regularization}}. \quad (7)\end{aligned}$$

Taking the logarithm embodies the *prediction information* we gain from the observation data X. For ease of interpretation, we use the integral approximation [22]. The first term represents the log likelihood given the most probable parameters. The model regularization is accordingly controlled by the *prior density* $g(\lambda \,|\varphi_m)$ as well as the *parameter uncertainty* $\Delta\lambda$ of the posterior likelihood $f(\lambda \,|X, \varphi_m)$. Considering the case of uniform prior $g(\lambda \,|\varphi_m)$ and posterior $f(\lambda \,|X, \varphi_m)$, the complex models are endowed with greater parameter varieties, which result in smaller quantity of prior density $g(\lambda \,|\varphi_m)$. Also, the complex model provides better model fitting and higher posterior likelihood. Equivalently, the parameter uncertainty $\Delta\lambda$ is reduced. For these reasons, the complex models can be autonomously penalized according to PIC criterion. Similar to using penalization factor p in MDL criterion, we may merge a forgetting factor ρ in the exponent of the second term $g(\lambda \,|\varphi_m)$ of (7) to adjust the effect of prior density [16] in the PIC criterion. The forgetting factor has a value in the range of $0 < \rho \leq 1$. We neglect the expression of ρ in the following formulation.

AIC, BIC, MDL, NPC, and PIC approaches aim at generalizing likelihood for model selection. There are similarities and dissimilarities between different criteria. When we look at the original formulas of BIC and MDL, it is interesting to see them arising from a similar mathematical form. Both criteria were incorporated with the prior density $g(\lambda \,|M_m)$ for integration over parameter λ. BIC criterion is related to PIC by

$$\mathrm{BIC}(M_m) = \log f(X, M_m) = \log P(M_m) + \mathrm{PIC}(M_m). \quad (8)$$

BIC formulation in (4) was obtained when the observations were modeled using Koopman-Darmois exponential density and the integral was carried out through Taylor expansion. Interestingly, the model selection can be viewed as a pattern classification problem where the most probable model is selected via maximizing *a posteriori* density $f(M_m \,|X)$. The MAP model selection is equivalent to BIC criterion because

$$\begin{aligned}M_{\mathrm{MAP}} &= \arg \max_{M_m \in M} f(M_m \,|X) = \arg \max_{M_m \in M} \log f(X, M_m) \\ &= M_{\mathrm{BIC}}. \quad (9)\end{aligned}$$

Similar to BIC and MDL, PIC involves an integral in prediction distribution, which is often intractable. In [2], a stochastic approximation scheme, called variational Bayes (VB), was employed in calculating the integral through maximizing the free energy. This energy approximated the joint posterior distribution over model parameters.

The novelties of this paper are clarified as follows. The proposed PIC model selection is specially exploited for *continuous-density HMMs* with *multivariate real-valued observation data.* The problem of building compact decision trees is tackled. Instead of approximating the integrals in predictive density using VB or through Taylor expansion, this paper presents an exact solution to finding prediction information of individual HMM parameters for each frame. The parameters are estimated via MAP rather than ML theory adopted in AIC, BIC, MDL and NPC selection criteria. By incorporating the conjugate priors, the structural hyperparameters are derived for PIC calculation and MAP estimation. Consistently, the prediction information, parameter estimation and hyperparameter evolution are initialized from the Bayesian theory. In what follows, we address the PIC model selection for the framework of HMMs.

III. PIC for HMM Selection

In [6], [17], [21], the Bayesian predictive classification (BPC) was developed for robust decision using the predictive distribution. The BPC-based speech recognizer was robust to perturbation of HMM parameters due to estimation error, noise inference, etc. The uncertainties of HMMs were represented by a prior density. The integral was achieved using Laplacian approximation [17]. Herein, we apply the approximate Bayesian method in [17], [21] for resolving *model selection problem instead of classification problem.* We are concerned with how to determine suitable size of HMM parameters to fit the observed data. A novel PIC criterion is discovered for state-clustered decision tree construction.

A. PIC Formulation for HMMs

In the context of HMMs, the prediction information is calculated by merging the unobserved state and mixture component sequences (s, l). Jiang *et al.* [21] presented a contributive work for calculating the Bayesian predictive density of HMMs. Referring to Jiang's work, we determine the prediction information via Viterbi approximation. Namely, we decode the best sequences $(\hat{\mathrm{s}}, \hat{\mathrm{l}})$ for observation X through the Viterbi algorithm. The *segmental prediction information* of HMMs is computed by

$$\begin{aligned}\mathrm{PIC}(M_m) &= \log \int \sum_{\mathrm{s,l}} f(X, \mathrm{s}, \mathrm{l} \,|\lambda, \varphi_m) g(\lambda \,|\varphi_m) d\lambda \\ &\cong \log \int f(X, \hat{\mathrm{s}}, \hat{\mathrm{l}} \,|\lambda, \varphi_m) g(\lambda \,|\varphi_m) d\lambda \qquad (10)\end{aligned}$$

where the summation of likelihood of all possible sequences (s, l) is approximated by the single likelihood of the best sequences $(\hat{\mathrm{s}}, \hat{\mathrm{l}})$. The continuous-density HMM parameters $\lambda = (\pi, A, \theta)$ consist of the initial state probabilities $\pi = \{\pi_i\}$, state transition probabilities $A = \{a_{ij}\}$ and mixture Gaussian parameters $\theta = \{\omega_{ik}, m_{ik}, r_{ik}\}$ including mixture weights, mean vectors and precision matrices for states i and mixture components k. The state observation density of d-dimensional x_t has the form

$$\begin{aligned}f(\mathrm{x}_t \,|\theta_i) &= \sum_k \omega_{ik} f(\mathrm{x}_t \,|m_{ik}, r_{ik}) \\ &\propto \sum_k \omega_{ik} |r_{ik}|^{1/2} \exp\left[-\frac{1}{2}(\mathrm{x}_t - m_{ik})^T r_{ik} (\mathrm{x}_t - m_{ik})\right]. \qquad (11)\end{aligned}$$

ML estimation of parameters λ can be easily found in [25]. This paper aims at selecting the optimal model M_{PIC} producing the largest prediction information. The integral in (10) is calculated over the randomness of those HMM parameters $\lambda = \{\pi_{\hat{s}_1}, a_{\hat{s}_t \hat{s}_{t+1}}, \omega_{\hat{s}_t \hat{l}_t}, m_{\hat{s}_t \hat{l}_t}, r_{\hat{s}_t \hat{l}_t}\}$ corresponding to the optimal state and mixture component sequences $(\mathrm{s}, \mathrm{l}) = \{\hat{s}_t, \hat{l}_t\}$. If we assume that four sets of parameters $\{\pi_{\hat{s}_1}\}$, $\{a_{\hat{s}_t \hat{s}_{t+1}}\}$, $\{\omega_{\hat{s}_t \hat{l}_t}\}$ and $\{m_{\hat{s}_t \hat{l}_t}; r_{\hat{s}_t \hat{l}_t}\}$ are independent at each frame, the prediction information is decomposed as shown in (12) at the bottom of the page, which is a summation of the corresponding prediction information $\mathrm{PIC}(\pi_{\hat{s}_1})$, $\mathrm{PIC}(a_{\hat{s}_t \hat{s}_{t+1}})$, $\mathrm{PIC}(\omega_{\hat{s}_t \hat{l}_t})$ and $\mathrm{PIC}(m_{\hat{s}_t \hat{l}_t}, r_{\hat{s}_t \hat{l}_t})$. The prediction information $\mathrm{PIC}(m_{\hat{s}_t \hat{l}_t}, r_{\hat{s}_t \hat{l}_t})$ involves a double integral operation. Hereafter, we formulate the prediction information for $\pi_{\hat{s}_1} = \pi_i$, $a_{\hat{s}_t \hat{s}_{t+1}} = a_{ij}$, $\omega_{\hat{s}_t \hat{l}_t} = \omega_{ik}$, $m_{\hat{s}_t \hat{l}_t} = m_{ik}$ and $r_{\hat{s}_t \hat{l}_t} = r_{ik}$.

To solve (12), the prior density should be specified beforehand. The choice of prior density is crucial in PIC calculation. The vague or diffuse distribution [10], [14] and the uniform distribution [21] served as the priors to calculate the predictive density. Nevertheless, it is attractive to adopt the *conjugate prior* where the prior and the pooled posterior densities belong to the same distribution family [9]. With this property, the incremental learning algorithm was constructed for speaker adaptation [16]. A conjugate prior approach to Bayesian model selection was exploited for nonlinear regression models [32]. Using HMMs, it is appropriate to use Dirichlet density as the conjugate prior for probability parameters π_i, a_{ij} and ω_{ik}, i.e., $g(\pi_i \,|\eta_i) \propto \pi_i^{\eta_i - 1}$, $g(a_{ij} \,|\eta_{ij}) \propto a_{ij}^{\eta_{ij} - 1}$ and $g(\omega_{ik} \,|\nu_{ik}) \propto \omega_{ik}^{\nu_{ik} - 1}$, because of the conditions $\sum_i \pi_i = 1$, $\sum_j a_{ij} = 1$ and $\sum_k \omega_{ik} = 1$. The

$$\begin{aligned}\mathrm{PIC}(M_m) = \log\Bigg\{ &\int \pi_{\hat{s}_1} g(\pi_{\hat{s}_1} \,|\varphi_m) d\pi_{\hat{s}_1} \times \prod_t \Bigg[\int a_{\hat{s}_t \hat{s}_{t+1}} g(a_{\hat{s}_t \hat{s}_{t+1}} \,|\varphi_m) d a_{\hat{s}_t \hat{s}_{t+1}} \\ &\cdot \int \omega_{\hat{s}_t \hat{l}_t} g(\omega_{\hat{s}_t \hat{l}_t} \,|\varphi_m) d\omega_{\hat{s}_t \hat{l}_t} \cdot \int \left(\int f\left(\mathrm{x}_t \,\middle|\, m_{\hat{s}_t \hat{l}_t}, r_{\hat{s}_t \hat{l}_t}\right) g\left(m_{\hat{s}_t \hat{l}_t} \,\middle|\, r_{\hat{s}_t \hat{l}_t}, \varphi_m\right) d m_{\hat{s}_t \hat{l}_t} \right) g(r_{\hat{s}_t \hat{l}_t} \,|\varphi_m) d r_{\hat{s}_t \hat{l}_t} \Bigg] \Bigg\} \\ = \mathrm{PIC}(\pi_{\hat{s}_1}) + &\sum_t \left(\mathrm{PIC}(a_{\hat{s}_t \hat{s}_{t+1}}) + \mathrm{PIC}(\omega_{\hat{s}_t \hat{l}_t}) + \mathrm{PIC}(m_{\hat{s}_t \hat{l}_t}, r_{\hat{s}_t \hat{l}_t}) \right) \qquad (12)\end{aligned}$$

normal-Wishart density serves as the conjugate prior for HMM mean and precision parameters [9], [13]

$$g(m_{ik}, r_{ik} \,|\tau_{ik}, \mu_{ik}, \alpha_{ik}, u_{ik}) \propto |r_{ik}|^{(\alpha_{ik}-d)/2} \times \exp\left[-\frac{\tau_{ik}}{2}(m_{ik}-\mu_{ik})^T r_{ik}(m_{ik}-\mu_{ik})\right]\exp\left[-\frac{1}{2}\mathrm{tr}(u_{ik}r_{ik})\right] \tag{13}$$

where hyperparameters $\eta_i > 0$, $\eta_{ij} > 0$, $\nu_{ik} > 0$, $\tau_{ik} > 0$, $\alpha_{ik} > d - 1$, μ_{ik} is a $d\times 1$ vector and u_{ik} is a $d \times d$ positive definite matrix. The predictive densities of parameters π_i, a_{ij} and ω_{ik} correspond to the means of Dirichlet densities [9]. The prediction information is obtained by

$$\begin{aligned}\mathrm{PIC}(\pi_i) &= \log \int \pi_i g(\pi_i \,|\eta_i)d\pi_i = \log E(\pi_i) \\ &= \log \eta_i - \log \sum_i \eta_i \end{aligned}\tag{14}$$

$$\mathrm{PIC}(a_{ij}) = \log \eta_{ij} - \log \sum_j \eta_{ij} \tag{15}$$

$$\mathrm{PIC}(\omega_{ik}) = \log \nu_{ik} - \log \sum_{\mathrm{k}} \nu_{ik}. \tag{16}$$

To find the prediction information $\mathrm{PIC}(m_{ik}, r_{ik})$ for multivariate mean m_{ik} and precision r_{ik}, we determine the marginal density $g(r_{ik}\,|\varphi_m)$ and the conditional density $g(m_{ik}\,|r_{ik}, \varphi_m)$, which are proportional to the Wishart and Gaussian density functions, respectively [6]. The inner integral in $\mathrm{PIC}(m_{ik}, r_{ik})$ is derived by

$$\sqrt{\frac{\tau_{ik}}{\tau_{ik}+1}}\,|r_{ik}|^{1/2} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\frac{\tau_{ik}}{\tau_{ik}+1}(\mathbf{x}_t - \mu_{ik})(\mathbf{x}_t - \mu_{ik})^T r_{ik}\right]\right\} \tag{17}$$

using the property of the integral of an arranged Gaussian function being unity. The outer integral turns out to

$$(\tau_{ik}+1)^{-1/2}\,|U|^{-(\alpha_{ik}+1)/2} \times \int |U|^{(\alpha_{ik}+1)/2}\,|r_{ik}|^{(\alpha_{ik}-d)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(Ur_{ik})\right]dr_{ik} \tag{18}$$

with notation $U = u_{ik} + (\tau_{ik}+1)^{-1}\tau_{ik}(\mathbf{x}_t - \mu_{ik})(\mathbf{x}_t - \mu_{ik})^T$. The integral is done over a Wishart density and has the result of unity. We finally derive the predictive density of (m_{ik}, r_{ik}) having the form of *d-dimensional multivariate t distribution* with $\alpha_{ik} - d + 1$ degrees of freedom, location vector μ_{ik} and precision matrix $(\alpha_{ik} - d + 1)\tau_{ik}(\tau_{ik}+1)^{-1}u_{ik}^{-1}$ [6]. The resulting prediction information $\mathrm{PIC}(m_{ik}, r_{ik})$ is given by

$$-\frac{1}{2}\left\{\log(\tau_{ik}+1) + (\alpha_{ik}+1) \times \left[\log|u_{ik}| + \log\left(1 + \frac{\tau_{ik}}{\tau_{ik}+1}(\mathbf{x}_t - \mu_{ik})^T u_{ik}^{-1}(\mathbf{x}_t - \mu_{ik})\right)\right]\right\}. \tag{19}$$

Generally, the t distribution has a flatter shape with thicker tails as compared to a Gaussian distribution. With larger variance, the selected models are robust to parameter perturbation and insufficient data. The merit of PIC is due to the *incorporation of increasing variance* for model selection. Also, it is noted that the conjugate priors of HMM parameters using Dirichlet and normal-Wishart densities are not only helpful to *obtain the closed-form prediction information* at each frame, but also make it feasible to *calculate the MAP estimate of HMM parameters* λ_m^{MAP}. The number of parameters k_m is determined from λ_m^{MAP}, accordingly. Readers may refer to [13] for MAP formulation of HMMs. We may represent the model either by $M_m = \varphi_m$ or $M_m = (\lambda_m^{\mathrm{MAP}}, \mathrm{k}_m)$. In general pattern recognition, we require the estimated parameters λ_m^{MAP} to compute the likelihood function of test data matched with various patterns with parameters λ_m^{MAP}.

B. Estimation of Hyperparameters for Tree Structure HMMs

When adopting PIC decision for model selection, we need to determine the hyperparameters φ_m for candidate models M_m. Generally, different models built using the common data X could be characterized via a tree structure, e.g., decision tree state tying. The models locating in a tree layer reflect a certain degree of parameter sharing. Higher tree nodes share parameters for wider data range. The hyperparameters of a smaller cluster in layer κ are estimated using the corresponding data X_κ and the hyperparameters of the broader cluster in layer $\kappa - 1$. Consequently, the problem of model selection is equivalent to *selecting a tree cut from the tree structure* as illustrated in Fig. 1. Each tree cut Γ_m represents a specific data partition $X = X_1 \cup \cdots \cup X_{\kappa_m}$ for model M_m. Larger cluster number κ_m (or parameter dimension k_m) is inherent in complex models. Simple models have smaller κ_m. The overestimation and underestimation dilemma happens accordingly. We intend to evaluate the prediction information for different tree cuts. The hyperparameters $\hat{\varphi}_m = (\hat{\varphi}_1, \ldots, \hat{\varphi}_{\kappa_m})$ along the tree cut $\hat{\Gamma}_m$ producing the largest prediction information are selected. The prediction information is accumulated for all nodes in the tree cut. Finally, the hyperparameters $\hat{\varphi}_m$ or the resulting MAP parameters λ_m^{MAP} are determined to characterize the observation data X.

However, estimating hyperparameters for structural HMMs is tricky. In a decision tree structure, the observations $X_{\kappa-1}$ of a tree node in layer $\kappa - 1$ are split into the observation subset X_κ in layer κ. Let the hyperparameters of tree nodes in layers $\kappa - 1$ and κ be denoted by $\varphi_{\kappa-1}$ and φ_κ, respectively. Hyperparameters φ_κ include $\{\eta_i^\kappa, \eta_{ij}^\kappa, \nu_{ik}^\kappa, \tau_{ik}^\kappa, \alpha_{ik}^\kappa, \mu_{ik}^\kappa, u_{ik}^\kappa\}$. Due to the attractive property of conjugate prior, we can establish the *top-down prior/posterior propagation algorithm* for estimation of structural hyperparameters. The underlying concept is similar to structural Bayes method discovered by Shinoda and Lee [28]. In [28] and [30], the hierarchical priors were presented to estimate the MAP parameters for speaker adaptation. The priors were used to tackle the data sparseness problem in adaptation. Herein, we are engaged in estimating hyperparameters for measuring model complexity and resolving model selection problem. The proposed algorithm aims to determine the hyperparameters through *calculation of the posterior density.* The posterior density $f(\lambda\,|X_\kappa, \varphi_{\kappa-1})$ of tree node in layer κ is

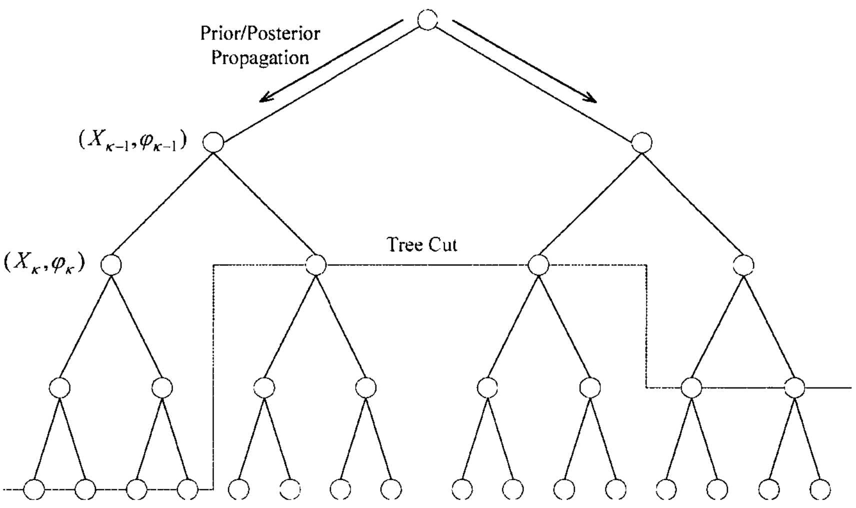

Prior/Posterior Propagation
$(X_{\kappa-1},\varphi_{\kappa-1})$
$(X_\kappa,\varphi_\kappa)$
Tree Cut

Fig. 1. Tree structure of HMMs illustrating the prior/posterior propagation in top-down manner. A tree cut is determined for model selection.

formulated by applying the corresponding observations $X_\kappa = \{\mathbf{x}_t^\kappa\}$ and the prior density $g(\lambda\,|\varphi_{\kappa-1})$ of its father node in layer $\kappa - 1$

$$\begin{aligned} f(\lambda\,|X_\kappa, \varphi_{\kappa-1}) &\propto \sum_{\mathbf{s},\mathbf{l}} f(X_\kappa, \mathbf{s}, \mathbf{l}\,|\lambda\,, \varphi_{\kappa-1}) g(\lambda\,|\varphi_{\kappa-1}) \\ &\cong f(X_\kappa, \hat{\mathbf{s}}_\kappa, \hat{\mathbf{l}}_\kappa\,|\lambda, \varphi_{\kappa-1}) g(\lambda\,|\varphi_{\kappa-1}) \\ &\equiv g(\lambda\,|\varphi_\kappa). \end{aligned} \tag{20}$$

Again, the Viterbi approach is employed to obtain the formulation. When incorporating conjugate prior $g(\lambda\,|\varphi_{\kappa-1})$ with hyperparameters $\varphi_{\kappa-1}$, the resulting posterior density $f(\lambda\,|X_\kappa, \varphi_{\kappa-1})$ belongs to the same distribution family, which can be expressed by a new prior density $g(\lambda\,|\varphi_\kappa)$ with new hyperparameters φ_κ. By repeating this procedure, the hyperparameters φ_κ of all tree nodes are estimated in a top-down manner. Here, the prior density $g(\lambda\,|\varphi_{\kappa-1})$ is represented as the product of Dirichlet and normal-Wishart densities. The derived posteriori density $f(\lambda\,|X_\kappa, \varphi_{\kappa-1})$ has the same joint density $g(\lambda\,|\varphi_\kappa)$ with new hyperparameters $\varphi_\kappa = \{\eta_i^\kappa, \eta_{ij}^\kappa, \nu_{ik}^\kappa, \tau_{ik}^\kappa, \alpha_{ik}^\kappa, \mu_{ik}^\kappa, u_{ik}^\kappa\}$ [13]

$$\eta_i^\kappa = \eta_i^{\kappa-1} + \gamma_1^\kappa(i) \tag{21}$$

$$\eta_{ij}^\kappa = \eta_{ij}^{\kappa-1} + \sum_t \gamma_t^\kappa(i, j) \tag{22}$$

$$\nu_{ik}^\kappa = \nu_{ik}^{\kappa-1} + \zeta_{ik}^\kappa \tag{23}$$

$$\tau_{ik}^\kappa = \tau_{ik}^{\kappa-1} + \zeta_{ik}^\kappa \tag{24}$$

$$\alpha_{ik}^\kappa = \alpha_{ik}^{\kappa-1} + \zeta_{ik}^\kappa \tag{25}$$

$$\mu_{ik}^\kappa = \frac{\tau_{ik}^{\kappa-1}\mu_{ik}^{\kappa-1} + \zeta_{ik}^\kappa \overline{\mathbf{x}}_{ik}^\kappa}{\tau_{ik}^{\kappa-1} + \zeta_{ik}^\kappa} \tag{26}$$

$$u_{ik}^\kappa = u_{ik}^{\kappa-1} + S_{ik}^\kappa + \frac{\tau_{ik}^{\kappa-1}\zeta_{ik}^\kappa}{\tau_{ik}^{\kappa-1} + \zeta_{ik}^\kappa} (\overline{\mathbf{x}}_{ik}^\kappa - \mu_{ik}^{\kappa-1})(\overline{\mathbf{x}}_{ik}^\kappa - \mu_{ik}^{\kappa-1})^T \tag{27}$$

where $\delta(\cdot)$ is Kronecker delta function, $\gamma_1^\kappa(i) = \delta(\hat{s}_1^\kappa - i)$, $\gamma_t^\kappa(i, j) = \delta(\hat{s}_t^\kappa - i)\delta(\hat{s}_{t+1}^\kappa - j)$, $\zeta_{ik}^\kappa = \sum_t \zeta_t^\kappa(i, k) = \sum_t \delta(\hat{s}_t^\kappa - i)\delta(\hat{l}_t^\kappa - k)$ and

$$\overline{\mathbf{x}}_{ik}^\kappa = \frac{\sum\limits_t \zeta_t^\kappa(i, k)\mathbf{x}_t^\kappa}{\zeta_{ik}^\kappa} \tag{28}$$

$$S_{ik}^\kappa = \sum_t \zeta_t^\kappa(i, k)(\mathbf{x}_t^\kappa - \overline{\mathbf{x}}_{ik}^\kappa)(\mathbf{x}_t^\kappa - \overline{\mathbf{x}}_{ik}^\kappa)^T. \tag{29}$$

When the hyperparameters of each tree node are estimated, the prediction information PIC(M_m) of all tree cuts Γ_m is determined. The best PIC model M_{PIC} can be selected. The HMM mean vectors and covariance matrices in tree node κ are obtained through MAP estimation, i.e., $m_{ik,\mathrm{MAP}} = \mu_{ik}^\kappa$ and $r_{ik,\mathrm{MAP}}^{-1} = (\alpha_{ik}^\kappa - d)^{-1} u_{ik}^\kappa$ [13], [16].

C. Interpretation of PIC Model Selection

We further interpret how PIC criterion is employed for model selection. Conceptually, PIC chooses the model carrying the largest prediction capability. The derived predictive density of HMM mean vector and precision matrix has a form of multivariate t distribution. Fig. 2 shows an example of PIC selection between simple and complex models illustrated by univariate t distributions. A univariate t distribution with α degrees of freedom, location parameter μ and precision τ is expressed by

$$\begin{aligned} f(x\,|M) &= t(\alpha, \mu, \tau) \\ &= \frac{\tau^{1/2}\Gamma\left[\frac{(\alpha+1)}{2}\right]}{(\alpha\pi)^{1/2}\Gamma\left(\frac{\alpha}{2}\right)} \left[1 + \frac{\tau}{\alpha}(x - \mu)^2\right]^{-(\alpha+1)/2} \end{aligned} \tag{30}$$

where $\alpha > 0$, $\tau > 0$ and $\Gamma(\cdot)$ is a gamma function [9]. Generally, simple model M_s has a smaller number of parameters than complex model M_c. Without loss of generality, the predictive densities of simple and complex models are assigned with a single t distribution $f(x\,|M_s) = t(5, 0, 1)$, and a mixture of equally weighted t distributions $f(x\,|M_c) = 0.5 \cdot t(5, 2, 2) + 0.5 \cdot t(5, -2, 1)$, respectively. A simple model M_s makes only a small range of predictions but shapes steep distribution $f(x\,|M_s)$ in range R_s. Conversely, the complex model M_c has more free parameters to

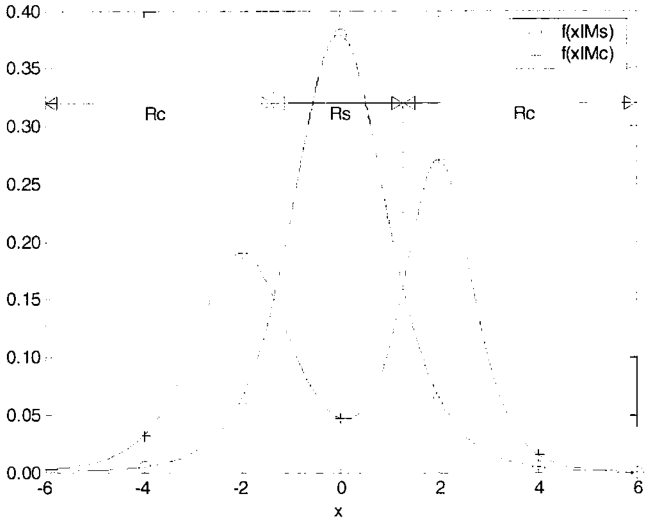


Fig. 2. PIC selection between simple and complex models is interpreted. Predictive densities of simple and complex models are respectively specified by a t distribution, $f(x\,|M_s) = t(5,0,1)$, and a mixture of t distributions $f(x\,|M_c) = 0.5 \cdot t(5,2,2) + 0.5 \cdot t(5,-2,1)$.

cover a greater variety of observation randomness. However, the flatter distribution shape makes the density $f(x\,|M_c)$ smaller for a wide range of data occurrence. The range of complex model excluding R_s is denoted by R_c. In case of data falling in wider range R_c, complex model M_c is preferably chosen. But, when data are located in smaller range R_s, the less powerful model M_s becomes a more probable model.

IV. Experiments

The proposed PIC criterion is a general model selection approach to different data modeling. This paper highlights the model selection problem for HMM framework. Except deriving the prediction information, we establish the relation between HMM parameter estimation and model selection from a Bayesian viewpoint. To investigate the effectiveness of PIC, we conduct a series of continuous speech recognition experiments where PIC is applied to construct HMM decision trees.

A. Decision Tree Construction Acted as a Model Selection Problem

Decision tree state tying is a top-down clustering mechanism. The state observation data of a phonetic unit are successively split according to questions about their context dependencies [3]. The tied HMM state parameters are estimated using clustering data. Importantly, the fitting of observed data to HMM parameters using decision tree involves the model selection problem. When performing a node split, we require the objective criterion to select the best question to separate data $X_{\kappa-1}$ of a tree node into X_κ^y and X_κ^n corresponding to the answers of "Yes" and "No", respectively. Whether the split is continued or terminated depends on the calculated information measures of $X_{\kappa-1}$, X_κ^y and X_κ^n. ML [33] and MDL [29] are popular for goodness-of-fit evaluation in decision tree construction. We implemented ML and MDL for a comparative study. Using ML or MDL, we choose the best question for node split which can attain the largest improvement in log likelihood or description length. In case of no increase in log likelihood or decrease in description length, $\Delta_{\mathrm{ML}} < 0$ or $\Delta_{\mathrm{MDL}} > 0$, the split should be terminated. To prevent building overlarge decision trees using ML criterion, the increasing amount of log likelihood T_Δ and the floor number of observation frames in tree node T_n are used in place of termination condition $\Delta_{\mathrm{ML}} < 0$. Using MDL, the penalization factor p is tuned in addition to the condition $\Delta_{\mathrm{MDL}} > 0$. In this manner, the optimal tree cut $\hat{\Gamma}_m$ is determined when the splits are terminated in all branches. A good model selection criterion is crucial for node split as well as termination evaluation. In this study, we *adopt the PIC criterion to select the best split question*, which achieves the largest gain in prediction information Δ_{PIC}. In the case of $\Delta_{\mathrm{PIC}} > 0$, a twofold complex model using X_κ^y and X_κ^n is better than a simpler model using $X_{\kappa-1}$. The split is continuously applied to child data X_κ^y and X_κ^n. Conversely, when $\Delta_{\mathrm{PIC}} < 0$, the simple model is selected and the split is terminated. The compact decision tree models are selected accordingly. Using PIC, we adopt the forgetting factor ρ in addition to the termination condition $\Delta_{\mathrm{PIC}} < 0$.

B. Experimental Setup and Implementation Issues

The benchmark speech database MAT-400 was used for building HMM decision trees. MAT-400 contained the *M*andarin speech *A*cross *T*aiwan (MAT) for 400 speakers (216 males and 184 females) recorded over telephone networks. We sampled 20 800 utterances with 12.3 hours covering isolated syllables (MATDB-3), command words (MATDB-4) and continuous speech (MATDB-5), and hence generated speaker-independent HMMs. The benchmark test telephone speech (Test500) consisted of 500 sentences of 15 males and 15 females, which were different from those of MAT-400. Test500 contained 4754 syllables serving as the common evaluation set for Mandarin speech recognition [7]. All utterances were sampled at 8 kHz with 16-bit resolution. Each frame was characterized by 12 Mel-frequency cepstral coefficients (MFCCs), 12 delta MFCCs, one delta log energy and one delta-delta log energy. The sentence-based cepstral mean subtraction was applied for telephone channel normalization. Without considering the tonal information, there were 408 confusing base Mandarin syllables covering 22 initials (consonants) and 38 finals (vowels). The sub-syllable HMM units were specified. The initial and final of a syllable were modeled via three and six HMM states, respectively. The null initial was modeled by two states for the syllable containing only final part. Five states were used to characterize the pre-silence, post-silence and between-syllable silence. The intra- and inter-syllable dependencies were modeled through right context-dependent (RCD) initials and RCD finals, respectively. RCD initials were group-dependent of 38 finals. We built 94 RCD initials, 2400 RCD finals and 7 group-dependent null initials for continuous speech recognition [7]. Due to lack of training data, only the three rear states were context dependent for RCD finals. We empirically chose the mixture component number according to the number of frames in a state. Without decision tree state tying, we generated 7615 HMM states. This model set was too large to estimate reliable parameters. To resolve the overestimation problem, we prepared 31 consonant phonetic questions and built 38 decision trees for state tying of Mandarin finals instead of using 2400 RCD finals. The syllable bigram was applied for base syllable

TABLE I
PERFORMANCE COMPARISON OF BASELINE CI INITIAL/CI FINAL AND CD INITIAL/CI FINAL WITHOUT DECISION TREE CONSTRUCTION

	CI Initial, CI Final	CD Initial, CI Final
Number of trained states	375	502
Syllable recognition rate (%)	*48.9*	*51*

decoding of 500 test utterances. The resulting perplexity 85.2 was measured with syllable vocabulary size being 408. The syllable recognition rates (%) and the number of trained HMM states were reported for evaluation. The computation costs in training as well as recognition sessions were evaluated by executing the proposed algorithms on Sun Workstation with model Ultra 10. For comparison, we realized HMM modeling using 22 context-independent (CI) initials and 38 CI finals. The resulting syllable recognition rate was 48.9% as listed in Table I. The number of trained HMM states was only 375. In CI system, we used eight states to model the syllable with only final. When we applied 94 RCD initials and 38 CI finals, the syllable recognition rate 51% was obtained using 502 HMM states. In both baseline cases, the trained HMM parameters were underestimated.

To achieve robust estimation of context-dependent HMM parameters, the model selection approach was crucial to perform decision tree state tying. In PIC implementation, the HMM precision r_{ik} and hyperparameter u_{ik} were assumed to be diagonal matrices. The PIC criterion was evaluated only for HMM mean vector m_{ik} and precision matrix r_{ik}. Evaluation of other HMM parameters was neglected. The PIC procedure for decision tree construction is described as follows.

1) Perform Viterbi alignment and collect the data X_1 for each CI HMM state. Set $\kappa = 1$ for this root node. Use X_1 to estimate the initial hyperparameters φ_1.
2) For each nonleaf node, find the best split question producing the largest PIC gain Δ_{PIC}. If $\Delta_{\text{PIC}} > 0$, split the data $X_{\kappa-1}$ into two child nodes X_κ^y and X_κ^n for "Yes" and "No" answers, respectively. Set X_κ^y and X_κ^n as nonleaf nodes. Go to step 3. If $\Delta_{\text{PIC}} < 0$, stop splitting and set $X_{\kappa-1}$ as leaf node. Go to step 4.
3) Use child node data X_κ and father node hyperparameters $\varphi_{\kappa-1}$ and apply (24)–(29) to estimate hyperparameters $\varphi_\kappa = \{\tau_{ik}^\kappa, \alpha_{ik}^\kappa, \mu_{ik}^\kappa, u_{ik}^\kappa\}$ for child node X_κ. Go to step 2.
4) When all splitting processes are terminated, a decision tree is constructed. Estimate MAP parameters for all leaf nodes. Finally, HMM parameters λ_m^{MAP} are estimated and selected.

Here, the initial hyperparameters φ_1 are estimated by referring to the mechanism [16] using the data X_1 of a CI phonetic unit. The ML estimation method is employed to obtain $\tau_{ik}^1 = \zeta_{ik}, \alpha_{ik}^1 = d + \zeta_{ik}, \mu_{ik}^1 = \overline{\mathbf{x}}_{ik} = m_{ik,\text{ML}}$ and $u_{ik}^1 = \zeta_{ik} \cdot r_{ik,\text{ML}}^{-1}$. To examine the effect of hyperparameters, we also carry out PIC without top down prior/posterior propagation algorithm. The prediction information of all tree nodes is determined using fixed and common hyperparameters, i.e., $\{\tau_{ik}^\kappa, \alpha_{ik}^\kappa, \mu_{ik}^\kappa, u_{ik}^\kappa\} = \{\zeta_{ik}, d + \zeta_{ik}, m_{ik,\text{ML}}, \zeta_{ik} \cdot r_{ik,\text{ML}}^{-1}\}$ for all κ.

TABLE II
PERFORMANCE COMPARISON OF DECISION TREE CONSTRUCTION USING ML, MDL, AND PIC CRITERIA UNDER DIFFERENT VALUES OF CONTROL PARAMETERS T_n, T_Δ, p AND ρ. PIC REALIZATIONS USING COMMON AND STRUCTURAL HYPERPARAMETERS ARE COMPARED

ML	T_n	50	150	250	350	550
	Number of trained states	5142	3988	**2020**	1539	996
	Syllable recognition rate (%)	*54.6*	*55.5*	***56***	*55.2*	*53.7*
ML	T_Δ ($T_n = 250$)	0	200	400	600	1000
	Number of trained states	2020	**1832**	1507	1097	688
	Syllable recognition rate (%)	*56*	***56.7***	*55*	*53.5*	*52.4*
MDL	p	0.5	1	1.5	2	5
	Number of trained states	2853	1898	**1720**	1267	710
	Syllable recognition rate (%)	*57*	*57.7*	***58***	*56.4*	*54.9*
PIC (common)	ρ	0.7	0.8	0.9	0.95	1
	Number of trained states	2383	2121	**1870**	1709	1587
	Syllable recognition rate (%)	*56.3*	*57.1*	***58.3***	*57.9*	*57.5*
PIC (structural)	ρ	0.7	0.8	0.9	0.95	1
	Number of trained states	2510	2253	1903	**1795**	1620
	Syllable recognition rate (%)	*57.4*	*59.5*	*60.4*	***60.9***	*59.8*

C. Performance Comparison of ML and MDL Using Different Thresholds

When ML and MDL model selection criteria are employed in HMM decision tree state tying, we are investigating the recognition performance versus different control parameters. The node split questions with the largest differences in log likelihood Δ_{ML} and description length Δ_{MDL} are chosen. The ML based split is stopped according to the increasing amount of log likelihood T_Δ and the floor number of observation frames T_n. Using MDL, the penalization factor p is adopted to control the stop condition. In Table II, we list the syllable recognition rates and the numbers of trained HMM states for two cases of ML based decision trees. In the first case, different thresholds T_n are examined under $T_\Delta = 0$. We can see that the number of trained states is reduced from 7615 without decision trees to 5142 with decision trees in case of $T_n = 50$. The syllable recognition rate 54.6% is obtained. When we modify the threshold by $T_n = 250$, the recognition rate is improved to 56% with smaller HMM parameter number 2020. This illustrates the importance of applying decision trees in reducing the amount of context-dependent HMM parameters. If the threshold T_n is further increased, smaller states are trained but with lower recognition rates. In the second set of ML evaluation, we fix the threshold $T_n = 250$ and adjust another stop condition T_Δ. We find that the best recognition rate is slightly raised to 56.7% under $T_\Delta = 200$. Interestingly, the trained HMM parameters are simplified to 1832 states. The higher the threshold T_Δ is, the smaller the size of HMM parameters attained. Here, the result of 56.7% is the best performance of ML when we search on a two-dimensional space of T_n and T_Δ. With regard to MDL model selection, the stop condition $\Delta_{\mathrm{MDL}} > 0$ with tunable penalization factor p is applied to control the tradeoff between speech recognition rate and HMM parameter size. The best recognition rate 58% is achieved under $p = 1.5$. More attractively, the number of trained states is reduced to 1720. It is obvious that MDL consistently obtains a smaller number of states and higher recognition rates than ML. MDL is superior to ML for model selection. Compared with the baseline cases, the recognition results using ML and MDL based decision trees are significantly improved. But, the disadvantage is to induce a larger amount of HMM parameters and higher computation costs.

D. Mutual Information Criterion for Decision Tree Construction

When performing node splitting for decision tree construction, we select the best question to separate data $X_{\kappa-1}$ into the most incoherent subsets X_κ^y and X_κ^n. We choose one model or two models. MDL and PIC model selection methods consider only one preset model complexity rather than different models with different complexities. It is more reasonable to apply the mutual information criterion to examine data homogeneity for optimal node splitting. The minimization of mutual information was feasible to find *statistically independent* models [20]. This property is desirable for decision tree splitting. To fulfill mutual information criterion, we need to calculate the weighted entropies for the variables in $X_{\kappa-1}$, X_κ^y and X_κ^n. The resulting delta entropy was used to merge the most similar pair of *discrete* HMM densities in [18]. Also, this criterion was applied for optimal node splitting in decision tree construction of semi-continuous HMMs [19]. Herein, the decision trees of *continuous* HMMs using Gaussian densities are built. The entropy decrease due to a node splitting is given by

$$\Delta_{\mathrm{Entropy}} = \zeta_\kappa^y H(X_\kappa^y) + \zeta_\kappa^n H(X_\kappa^n) - \zeta_{\kappa-1} H(X_{\kappa-1}) \quad (31)$$

where ζ_κ^y, ζ_κ^n and $\zeta_{\kappa-1}$ are the observation frame numbers serving as weighting factors and the entropy of Gaussian density of father node $X_{\kappa-1} \propto N(m_{\kappa-1}, r_{\kappa-1})$ is [15]

$$H(X_{\kappa-1}) = \frac{1}{2}\left[d + d\log 2\pi - \log|r_{\kappa-1}|\right]. \quad (32)$$

The delta entropy has the form of

$$\begin{aligned}\Delta_{\mathrm{Entropy}} &= \frac{1}{2}\left[\zeta_{\kappa-1}\log|r_{\kappa-1}| - \zeta_\kappa^y \log|r_\kappa^y| - \zeta_\kappa^n \log|r_\kappa^n|\right] \\ &= -\Delta_{\mathrm{ML}} \quad (33)\end{aligned}$$

which is equivalent to the negative of maximum likelihood criterion [33]. The node splitting using maximal delta entropy has the same realization as that using maximal delta log likelihood. We have investigated the performance of ML decision trees.

E. Effectiveness of PIC versus Different Hyperparameters and Forgetting Factors

Further, the proposed PIC model selection is evaluated in terms of syllable recognition rate and number of trained states. During decision tree construction, the data clustering is performed according to the question providing the largest gain of prediction information Δ_{PIC}. The clustering is terminated when $\Delta_{\mathrm{PIC}} < 0$ accompanied by different forgetting factors ρ. The forgetting factor adapts the effect of prior parameters in derived prediction information. The case of $\rho = 1$ means no forgetting. Prior statistics is entirely kept for model selection. When forgetting factor ρ approaches zero, the model complexity penalization is almost de-emphasized. Also, we evaluate two PIC realizations using fixed common hyperparameters and proposed structural hyperparameters. This evaluation aims to know the goodness of structural hyperparameters for PIC model selection. As shown in Table II, the PIC realizations using structural hyperparameters are much better than those using common hyperparameters for different forgetting factors. Using common hyperparameters, the best syllable recognition rate 58.3% is obtained when $\rho = 0.9$. However, PIC can achieve syllable recognition rate as high as 60.9% when applying structural hyperparameters and $\rho = 0.95$. Hereafter, we only report the results of PIC using structural hyperparameters. Fig. 3 displays the syllable recognition rates using PIC model selection versus different factors ρ. The results of CI Initial/CI Final, CD Initial/CI Final, ML with $T_\Delta = 200$ and MDL with $p = 1.5$ are included for comparison. In case of a very small forgetting factor $\rho = 0.001$, PIC attains a recognition rate of 53.6%. The original PIC without applying forgetting factor, i.e., $\rho = 1$, had a recognition rate of 59.8%. For the cases of forgetting factor larger than 0.7, PIC outperforms other methods in term of recognition performance. Also, Table II indicates that the best recognition rate of PIC attains a slightly larger parameter size than that of MDL. Nevertheless, PIC is advantageous because of its good performance in recognition rate as well as suitable model size.

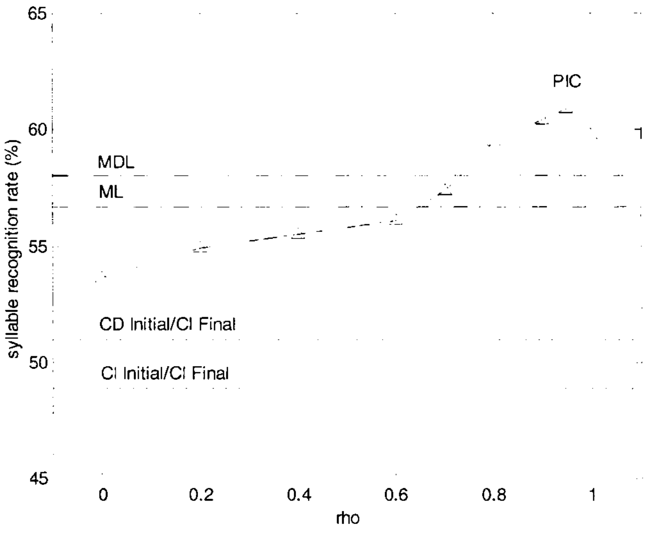


Fig. 3. PIC based decision tree construction is evaluated under different forgetting factors ρ. Syllable recognition rates of CI Initial/CI Final, CD Initial/CI Final, ML, and MDL based decision tree construction are compared.

F. Performance Comparison of Training and Recognition Time Costs

Finally, we evaluate different model selection approaches in terms of training and recognition time costs. The results are provided in Fig. 4, which illustrates the overall performance com-

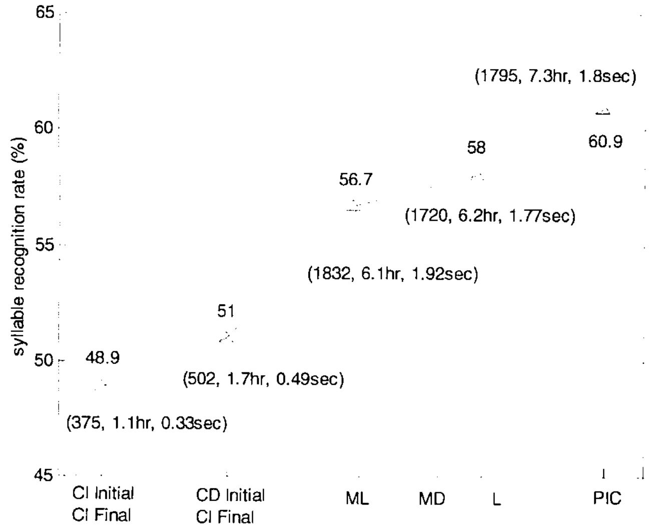


Fig. 4. Performances of CI Initial/CI Final, CD Initial/CI Final and the best decision tree construction using ML, MDL and PIC model selection criteria are compared. The numbers in brackets denote the corresponding state number, training hour and recognition second.

parison. We compare the recognition rate, number of trained states, training time (in hours) and recognition time (in seconds) for CI Initial/CI Final, CD Initial/CI Final, ML, MDL, and PIC model selection. The training time means the time of executing an EM iteration of HMM training. The recognition time per sentence is averaged over 500 test sentences. We can see that using context-dependent decision trees takes much more training cost than the baseline system. The training and recognition times of the baseline system are less than two hours and a half second, respectively. In the construction of decision trees, the training overhead arises from the larger number of HMM states, the question selection operation and the split termination evaluation. The question selection for data clustering is the most time-consuming work. However, the recognition time is mainly affected by the size of HMM states. We find that PIC requires the highest training cost due to the extensive computation of evolutionary hyperparameters and HMM prediction information. MDL attains the desirable computation costs in spite of tunable control parameters. Overall, the proposed PIC model selection is promising for HMM decision tree construction because of its moderate recognition time and storage requirement of HMM parameters.

V. Conclusion

We surveyed several important model selection approaches using AIC, BIC, MDL and NPC. A new PIC model selection using approximate Bayesian is proposed to generalize the likelihood criterion and resolve the overestimation and underestimation dilemmas for HMM data modeling. The similarities and dissimilarities between PIC and other criteria were described. Using PIC, the model complexity was autonomously controlled according to the prediction information. The model with the largest prediction information was retrieved. Theoretically, the selected model provided the best generalization for future data occurrences. To realize PIC for HMMs, we applied Viterbi approach and characterized the statistics of real-valued multivariate HMM parameters by the conjugate prior densities. The formulated prediction information had exact solution at the frame level without Laplacian integral approximation. A multivariate t distribution was obtained to express the prediction information due to HMM mean vector and precision matrix. When the most likely model was selected, the corresponding HMM parameters were determined via MAP estimate. Using the property of conjugate prior, we also constructed a top-down prior/posterior propagation algorithm to calculate the structural hyperparameters for HMM decision trees. These structural hyperparameters were essential for evaluating prediction information and retrieving compact decision trees. The prediction information, parameter estimation and hyperparameter evolution for HMMs were analytically addressed based on Bayesian learning viewpoints. We also addressed the relationship between ML and mutual information criteria for decision tree construction of HMMs. In the experiments, the proposed PIC achieved the highest speech recognition rate with moderate number of HMM states compared to ML and MDL criteria. PIC using structural hyperparameters outperformed that using shared hyperparameters. PIC spent higher processing times in training and recognition sessions. In the future, we will investigate whether the estimated structural hyperparameters are proper for Bayesian model regularization. The PIC model selection will be further expanded for other data modeling problems and pattern recognition applications. A straightforward extension of PIC for constructing compact Gaussian mixture models will be developed for speaker recognition. We will also improve mutual information criterion for data homogeneity evaluation in decision tree state tying.

Acknowledgment

The authors acknowledge Dr. J. R. Bellegarda and the anonymous reviewers for providing valuable comments, which considerably improved the quality of this paper. We also thank helpful comments from Prof. K. Shinoda at Tokyo Institute of Technology.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[2] H. Attias, "Inferring parameters and structure of latent variable methods by variational Bayes," in *Proc. 15th Conf. Uncertainty Artificial Intelligence*, 1999, pp. 21–30.

[3] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision trees for phonological rules in continuous speech," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 1991, pp. 185–188.

[4] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process.*, vol. 2, 1998, pp. 645–648.

[5] J.-T. Chien and J.-C. Junqua, "Unsupervised hierarchical adaptation using reliable selection of cluster-dependent parameters," *Speech Commun.*, vol. 30, no. 4, pp. 235–253, Apr. 2000.

[6] J.-T. Chien and G.-H. Liao, "Transformation-based Bayesian predictive classification using online prior evolution," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 399–410, May 2001.

[7] J.-T. Chien, C.-H. Huang, and S.-J. Chen, "Compact decision trees with cluster validity for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 1, 2002, pp. 873–876.

[8] W. Chou and W. Reichl, "Decision tree state tying based on penalized Bayesian information criterion," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 1, 1999, pp. 345–348.

[9] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[10] P. M. Djuric and S. M. Kay, "Predictive probability as a criterion for model selection," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 5, 1990, pp. 2415–2418.

[11] ——, "Model selection based on Bayesian predictive densities and multiple data records," *IEEE Trans. Signal Process.*, vol. 42, no. 7, pp. 1685–1699, Jul. 1994.

[12] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[13] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 291–298, Apr. 1994.

[14] S. Geisser and W. F. Eddy, "A predictive approach to model selection," *J. Amer. Statist. Assoc.*, vol. 74, no. 365, pp. 153–160, Mar. 1979.

[15] S. Haykin, *Neural Networks – A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[16] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 161–172, Mar. 1997.

[17] Q. Huo, H. Jiang, and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 1997, pp. 1547–1550.

[18] M.-Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, pp. 414–420, Oct. 1993.

[19] M.-Y. Hwang, X. Huang, and F. A. Alleva, "Predicting unseen triphones with senones," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 412–419, Nov. 1996.

[20] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[21] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on a Bayesian prediction approach," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 426–440, Jul. 1999.

[22] D. J. C. Mackay, "Bayesian interpolation," *Neural Comput.*, vol. 4, pp. 405–447, 1992.

[23] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inf. Theory*, vol. 35, pp. 1109–1114, Sep. 1989.

[24] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 71–77, Jan. 1998.

[25] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[26] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[27] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[28] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 276–287, Mar. 2001.

[29] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. Eur. Conf. Speech Commun. Technol.*, vol. 1, 1997, pp. 99–102.

[30] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum *a posteriori* linear regression for fast HMM adaptation," *Comput. Speech Lang.*, vol. 16, pp. 5–24, 2002.

[31] J. P. Ueberla, "Domain adaptation with clustered language models," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 1997, pp. 807–810.

[32] J.-P. Vila, V. Wagner, and P. Neveu, "Bayesian nonlinear model selection and neural networks: A conjugate prior approach," *IEEE Trans. Neural Networks*, vol. 11, no. 2, pp. 265–278, Mar. 2000.

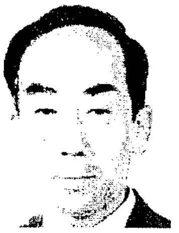
[33] S. J. Young, J. J. Odell, and P. Woodland, "Tree-based state-tying for high accuracy acoustic modeling," in *Proc. ARPA Workshop Human Language Technol.*, 1994, pp. 286–291.

[34] Z. Zhang and S. Furui, "Piecewise-linear transformation-based HMM adaptation for noisy speech," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, 2001.

Jen-Tzung Chien (S'97–A'98–M'99–SM'04) was born in Taipei, Taiwan, R.O.C., on August 31, 1967. He received the Ph.D. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 1997.

He was an Instructor in the Department of Electrical Engineering, NTHU, in 1995. In 1997, he joined the Department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, where he was an Assistant Professor. Since 2001, he has been an Associate Professor in NCKU. He was the Visiting Researchers at the Speech Technology Laboratory, Panasonic Technologies, Inc., Santa Barbara, CA, in the summer of 1998, the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in the summer of 2002, and the Institute of Information Science, Academia Sinica, Taipei, Taiwan, in the summers of 2003 and 2004. His research interests include statistical pattern recognition, speech recognition, speaker adaptation, language modeling, face recognition/verification, biometrics, document classification, multimedia information retrieval and multimodal human-computer interaction. He serves on the board of the *International Journal of Speech Technology* and a guest editor of the *Special Issue on Chinese Spoken Language Technology*.

Dr. Chien is member of the IEEE Signal Processing Society, the International Speech Communication Association, the Chinese Image Processing and Pattern Recognition Society, and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). He serves on the board and chairs the Special Interest Group on Spoken Language Processing for ACLCLP. He was listed in *Who's Who in the World*, 2002, and awarded the Outstanding Young Investigator Award (Ta-You Wu Memorial Award) from National Science Council, Taiwan, in 2003, and the Research Award for Junior Research Investigators from Academia Sinica, Taiwan, in 2004.



Sadaoki Furui (F'93) is currently a Professor with the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 350 published articles. From 1978 to 1979, he served on the staff of the Acoustics Research Department of Bell Laboratories, Murray Hill, NJ, as a visiting researcher working on speaker verification. He is the author of *Digital Speech Processing, Synthesis, and Recognition* (New York: Marcel Dekker, 1989, revised 2000) in English, *Digital Speech Processing* (Kanagawa, Japan: Tokai University Press, 1985) in Japanese, *Acoustics and Speech Processing* (Tokyo, Japan: Kindai-Kagaku-Sha, 1992) in Japanese, and *Speech Information Processing* (Tokyo, Japan: Morikita, 1998) in Japanese. He edited *Advances in Speech Signal Processing* (New York: Marcel Dekker, 1992) with M. M. Sondhi. He has translated *Japanese Fundamentals of Speech Recognition* authored by Drs. L. R. Rabiner and B.-H. Juang (Tokyo, Japan: NTT Advanced Technology, 1995) and *Vector Quantization and Signal Compression* authored by A. Gersho and R. M. Gray (Tokyo, Japan: Corona-sha, 1998). He has served as Editor-in-Chief of both the *Journal of Speech Communication* and the *Transactions of the IEICE*.

Dr. Furui is a Fellow of the Acoustical Society of America and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE). He was President of the Acoustical Society of Japan (ASJ) from 2001 to 2003 and is currently President of the International Speech Communication Association (ISCA) and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He was a Board of Governor of the IEEE Signal Processing Society from 2001 to 2003. He has served on the IEEE Technical Committees on Speech and MMSP and on numerous IEEE conference organizing committees. He is an Editorial Board member of *Speech Communication, the Journal of Computer Speech and Language* and the *Journal of Digital Signal Processing*. He has received the Yonezawa Prize and the Paper Awards from the IEICE (1975, 1988, 1993, and 2003) and the Sato Paper Award from the ASJ (1985 and 1987). He has received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He has received the Technical Achievement Award and the Book Award from the IEICE (1990 and 2003). He has also received the Mira Paul Memorial Award from the AFECT, India (2001). In 1993, he served as an IEEE SPS Distinguished Lecturer.