

Predictive Hidden Markov Model Selection for Decision Tree State Tying

Jen-Tzung Chien^a and Sadaoki Furui^b

^a Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC

^b Department of Computer Science, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, Japan

jtchien@mail.ncku.edu.tw & furui@cs.titech.ac.jp

Abstract

This paper presents a novel *predictive information criterion* (PIC) for hidden Markov model (HMM) selection. The PIC criterion is exploited to select the best HMMs, which provide the largest prediction information for generalization of future data. When the randomness of HMM parameters is expressed by a product of *conjugate prior densities*, the prediction information is derived without 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 HMM selection in tree structure HMMs, we develop a top-down *prior/posterior propagation* algorithm for estimation of structural hyperparameters. The prediction information is accordingly determined so as to choose the best HMM tree model. The parameters of chosen HMMs can be rapidly computed via maximum *a posteriori* (MAP) estimation. In the evaluation of continuous speech recognition using decision tree HMMs, the PIC model selection criterion performs better than conventional maximum likelihood and minimum description length criteria in building a compact tree structure with moderate tree size and higher recognition rate.

1. Introduction

Model selection is a major problem in signal processing where the model parameters and their number are unknown. To achieve robust data modeling, it is necessary to precisely estimate the underlying parameters of a stochastic model and properly determine the size/order/dimension of the model at the same time. The estimated models are then tested for robustness against the *underestimation* or *overestimation* dilemma by applying future observation data. In this study, we aim to develop a new predictive information criterion to estimate HMMs and simultaneously select the proper size of HMMs to represent the observed data. This approach is not only useful for speech recognition but also for general data clustering/modeling. For speech recognition applications, the construction of HMM decision trees involve the model selection problems. In HMM 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 [2]. How to determine the suitable tree level from observed data is a challenging topic.

Usually, model selection problems using HMMs are 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

judgments, we explored a predictive information criterion (PIC) to select and estimate HMMs where the model complexity penalties are properly incorporated. The underestimation and overestimation conditions were alleviated. Using PIC, the penalization was performed by means of investigating the prediction/generalization information in tree structure HMMs. The predictive densities of HMM parameters were derived without Taylor's approximation of integrals. In this study, we modeled the uncertainty of HMM parameters using a product of conjugate prior densities for two purposes. One is to develop the top-down prior/posterior propagation algorithm for computation of prediction information. The other is to estimate the HMM parameters via MAP theory [4]. To that purpose, we selected the best model set and estimated the model parameters for general pattern recognition. In this paper, we examined the proposed algorithm by carrying out experiments on speech recognition based on decision tree models. PIC criterion was found to be effective to cluster the context-dependent speech frames into compact groups.

2. Model Selection Approaches

The model selection problem aims at selecting a parametrical model M_m with distribution $f(X|\lambda_m, k_m)$ for the observed data sequence $X = \{x_1, \dots, x_n\}$ and trying to estimate the vector parameter $\lambda_m = \{\lambda_1, \dots, \lambda_{k_m}\}$, 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 [8].

2.1. AIC, BIC and MDL Criteria

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

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

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

$$M_{AIC} = \arg \min_{\lambda_m \in \Lambda, k_m \in K} [-\log f(X|\lambda_m, k_m) + k_m]. \quad (2)$$

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

Schwarz [8] resolved the problem from a Bayesian perspective and proposed the Bayesian information criterion (BIC) for model selection. By considering *a priori* model

This work was partially completed while Prof. J.-T. Chien visited Tokyo Institute of Technology.

probability $P(M_m)$ and *a priori* parameter distribution $g(\lambda|M_m)$, the logarithm of the integral of *a posteriori* distribution

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

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

$$M_{\text{BIC}} = \arg \max_{\lambda_m \in \Lambda, k_m \in K} [\log f(X|\lambda_m, k_m) - \frac{1}{2} k_m \log n]. \quad (4)$$

Both AIC and BIC were derived by assuming that the observation data come from a Koopman-Darmois exponential distribution family [8]. The only difference between (2) and (4) is due to the second term playing the role of *penalty* for selecting the high dimensions. If the second term is neglected, the BIC criterion is simplified to a *ML model selection*.

Rissanen [7] found an interesting relation between estimation and coding, from which he was able to exploit the minimum description length (MDL) selection criterion. From the information theoretic and 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 $-\log f(X|\lambda_m^{\text{ML}}, k_m)$ bits. These description lengths were derived to achieve the lower bound on the expected code length for a set of probabilistic sources $\{f(X|M_m)\}$. Although MDL and BIC initialize from different aspects, they come up with the same formula as given in (4).

2.2. Predictive Information Criterion

In general, the criteria of AIC, BIC and MDL are universal and applicable to coding, estimation, prediction and pattern recognition. Most implementation procedures have been carried out for *integer data* in *compression* system [7]. The penalization was presumed identical for each parameter component. In the procedure, the ML estimates λ_m^{ML} were first calculated for different models $M = \{M_m\}$. They used the *Taylor expansion* of $\log f(X|\lambda_m, k_m)$ around λ_m^{ML} to express its uncertainty due to parameter λ_m . A Fisher information matrix was calculated to fulfill the *Laplacian integral approximation*. The model penalty was explicitly represented using parameter number k_m . Instead of using λ_m^{ML} and k_m , we present a new model selection where the model complexity is controlled according to the prior density of model parameter $g(\lambda|M_m)$. It turns out that the hyperparameters φ_m of prior distribution $g(\lambda|\varphi_m)$ were used to represent model $M_m = \varphi_m$.

Unlike other methods 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 predictive distribution

$$\text{PIC}(M_m) = \log f(X|M_m) = \log \int f(X|\lambda, \varphi_m) g(\lambda|\varphi_m) d\lambda. \quad (5)$$

Taking the logarithm embodies the *prediction information* we gain from the observation data X . The integral over the entire parameter space provides a meaningful way to *penalize complex models with greater parameter varieties*.

The novelties of this paper are 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 integrals using Taylor expansion, this paper presents an exact PIC approach to selection of HMMs, which is attractive for many pattern recognition applications, e.g. speech recognition, image classification, document retrieval, etc. The HMM parameters are calculated via MAP rather than ML estimate. By adopting the conjugate priors, the structural hyperparameters are derived. Consistently, the prediction information, parameter estimation and hyperparameter evolution are initialized from the Bayesian theory.

3. PIC HMM Selection for Decision Tree State Tying

3.1. PIC Formulation for HMMs

In the context of HMMs, the prediction information is calculated by merging the unobserved state and mixture component sequences (\mathbf{s}, \mathbf{l}) . The best sequences $(\hat{\mathbf{s}}, \hat{\mathbf{l}})$ are decoded through the Viterbi algorithm to approximate the prediction information by $\log \int f(X, \hat{\mathbf{s}}, \hat{\mathbf{l}}|\lambda, \varphi_m) g(\lambda|\varphi_m) d\lambda$. The HMM parameters λ consist of the initial state probabilities $\{\pi_i\}$, transition probabilities $\{a_{ij}\}$ and d -variate mixture Gaussian parameters $\{\omega_{ik}, m_{ik}, r_{ik}\}$ including weights, mean vectors and precision matrices for states i and mixture components k . This paper aims at selecting the optimal model M_{PIC} where the resulting hyperparameters φ_{PIC} produce the largest prediction information. If we assume that $\{\pi_i\}$, $\{a_{ij}\}$, $\{\omega_{ik}\}$ and $\{m_{ik}, r_{ik}\}$ are independent, the prediction information $\text{PIC}(M_m)$ is decomposed by

$$\begin{aligned} & \log \int \pi_{\hat{s}_1} g(\pi_{\hat{s}_1}|\varphi_m) d\pi_{\hat{s}_1} + \sum_t \left[\log \int a_{\hat{s}_{t-1}\hat{s}_t} g(a_{\hat{s}_{t-1}\hat{s}_t}|\varphi_m) da_{\hat{s}_{t-1}\hat{s}_t} \right. \\ & \quad \left. + \log \int \omega_{\hat{s}_t\hat{l}_t} g(\omega_{\hat{s}_t\hat{l}_t}|\varphi_m) d\omega_{\hat{s}_t\hat{l}_t} \right. \\ & \quad \left. + \log \int \int f(\mathbf{x}_t | m_{\hat{s}_t\hat{l}_t}, r_{\hat{s}_t\hat{l}_t}) g(m_{\hat{s}_t\hat{l}_t}, r_{\hat{s}_t\hat{l}_t} | \varphi_m) dm_{\hat{s}_t\hat{l}_t} dr_{\hat{s}_t\hat{l}_t} \right]. \quad (6) \end{aligned}$$

The choice of prior density is crucial in PIC calculation. It is appropriate to use Dirichlet density as the conjugate prior for probability parameters, 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 constraints $\sum_i \pi_i = 1$, $\sum_j a_{ij} = 1$ and $\sum_k \omega_{ik} = 1$. The normal-Wishart density serves as the prior for HMM mean and precision parameters [4]

$$\begin{aligned} & g(m_{ik}, r_{ik} | \tau_{ik}, \mu_{ik}, \alpha_{ik}, u_{ik}) \propto |r_{ik}|^{(\alpha_{ik}-d)/2} \\ & \quad \times \exp \left[-\frac{\tau_{ik}}{2} (m_{ik} - \mu_{ik})^T r_{ik} (m_{ik} - \mu_{ik}) \right] \exp \left[-\frac{1}{2} \text{tr}(u_{ik} r_{ik}) \right], \quad (7) \end{aligned}$$

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

$$\text{PIC}(\pi_i) = \log \int \pi_i g(\pi_i | \eta_i) d\pi_i = \log \eta_i - \log \sum_i \eta_i, \quad (8)$$

$$\text{PIC}(a_{ij}) = \log \eta_{ij} - \log \sum_j \eta_{ij}, \quad (9)$$

$$\text{PIC}(\omega_{ik}) = \log v_{ik} - \log \sum_k v_{ik}. \quad (10)$$

The predictive density of (m_{ik}, r_{ik}) can be derived in a form of d -dimensional multivariate t distribution with $\alpha_{ik} - d + 1$ degrees of freedom, location vector μ_{ik} and precision matrix $(\alpha_{ik} - d + 1)\tau_{ik}(\tau_{ik} + 1)^{-1}u_{ik}^{-1}$ [3]. The resulting prediction information $\text{PIC}(m_{ik}, r_{ik})$ is given by

$$-\frac{1}{2} \left[\log(\tau_{ik} + 1) + (\alpha_{ik} + 1) \left[\log |u_{ik}| + \log \left(1 + \frac{\tau_{ik}}{\tau_{ik} + 1} (\mathbf{x}_t - \mu_{ik})^T u_{ik}^{-1} (\mathbf{x}_t - \mu_{ik}) \right) \right] \right]. \quad (11)$$

Therefore, we can express the total prediction information by

$$\text{PIC}(\pi_{\hat{s}_i}) + \sum_i (\text{PIC}(a_{\hat{s}_i, \hat{s}_i}) + \text{PIC}(\omega_{\hat{s}_i, \hat{s}_i}) + \text{PIC}(m_{\hat{s}_i, \hat{s}_i}, r_{\hat{s}_i, \hat{s}_i})). \quad (12)$$

Generally, the variance of t distribution is larger 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 solutions to prediction information*, but also make it feasible to *calculate the MAP estimate of HMM parameters* λ_m^{MAP} [4]. At the same time, the number of parameters k_m is determined from λ_m^{MAP} . We may represent the model by $M_m = (\lambda_m^{\text{MAP}}, k_m)$. Different from AIC, BIC and MDL, we adopt MAP parameters λ_m^{MAP} instead of ML parameters λ_m^{ML} .

3.2. Estimation of Structural Hyperparameters

Applying PIC model selection, we need to determine the hyperparameters φ_m for different models M_m . Hyperparameters are critical to conduct effective model comparison. Generally, different models built using the common data X could be characterized via a tree structure. The models locating in a tree layer reflect a certain degree of parameter sharing. During tree construction, the observations are grouped into several cluster scatters and cluster numbers. The hyperparameters of a smaller cluster κ are estimated using the corresponding data X_κ and the hyperparameters extracted from the broader cluster. The problem of model selection is equivalent to *selecting a tree cut from the tree structure* as illustrated in Fig. 1. The tree cut Γ_m represents a specific data partition $X = X_1 \cup \dots \cup X_{\kappa_m}$ for model M_m . Larger cluster number κ_m is inherent to complex models. Simple models have smaller κ_m . We intend to evaluate the prediction information for different tree cuts. The hyperparameters $\hat{\varphi}_m = (\hat{\varphi}_1, \dots, \hat{\varphi}_{\kappa_m})$ along the tree cut $\hat{\Gamma}_m$ producing the largest prediction information are selected.

In top-down clustering, 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 in layers $\kappa-1$ and κ be respectively denoted by $\varphi_{\kappa-1}$ and φ_κ . Here, we establish the *top-down prior/posterior propagation algorithm* for estimation of structural hyperparameters. The proposed algorithm aims to determine the hyperparameters through *calculation of posterior density*. The posterior density $f(\lambda | X_\kappa, \varphi_{\kappa-1})$ of tree node in layer κ is 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 $\kappa-1$

$$\begin{aligned} f(\lambda | X_\kappa, \varphi_{\kappa-1}) &\equiv \sum_{\mathbf{s}_1} f(X_\kappa, \mathbf{s}_1 | \lambda, \varphi_{\kappa-1}) g(\lambda | \varphi_{\kappa-1}) \\ &\equiv f(X_\kappa, \hat{\mathbf{s}}_\kappa, \hat{\mathbf{I}}_\kappa | \lambda, \varphi_{\kappa-1}) g(\lambda | \varphi_{\kappa-1}) \equiv g(\lambda | \varphi_\kappa) \end{aligned} \quad (13)$$

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. When $g(\lambda | \varphi_{\kappa-1})$ is a product of Dirichlet and normal-Wishart densities, the derived $f(\lambda | X_\kappa, \varphi_{\kappa-1})$ has the same density $g(\lambda | \varphi_\kappa)$ with $\varphi_\kappa = \{\eta_i^\kappa, \eta_{ij}^\kappa, v_{ik}^\kappa, \tau_{ik}^\kappa, \alpha_{ik}^\kappa, \mu_{ik}^\kappa, u_{ik}^\kappa\}$. Reader may refer [4] for the formulation.

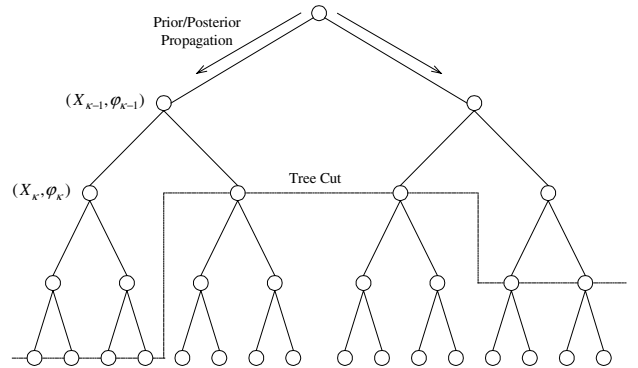


Fig. 1. Prior/posterior densities propagate in top-down manner

3.3. Decision Tree State Tying Used in a Model Selection Problem

When constructing HMM decision trees, the state observation data of a phonetic unit are successively split according to questions about their context dependencies [2]. 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 and MDL are popular choices for goodness-of-fit evaluation in decision tree construction [9]. 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, $\Delta_{\text{ML}} < 0$ or $\Delta_{\text{MDL}} > 0$, the split should be terminated.

The optimal tree cut $\hat{\Gamma}_m$ is finally 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 splitting question, which achieves the largest gain in prediction information Δ_{PIC} . In the case of $\Delta_{PIC} > 0$, a complex model using X_k^y and X_k^n is better than a simpler model using X_{k-1} . The split is continuously applied to child data X_k^y and X_k^n . Conversely, when $\Delta_{PIC} < 0$, the simple model is selected and the split is stopped. The compact decision tree models are selected accordingly.

4. Experiments

4.1. Experimental Setup

The benchmark speech database MAT was used for building HMM decision trees. We sampled speech material from MAT-400 containing Mandarin speech Across Taiwan (MAT) for 400 speakers recorded over telephone networks. The test telephone speech (Test500) consisted of 500 sentences of 15 males and 15 females. Test500 contained 4754 highly confusing syllables serving as the benchmark evaluation set for Mandarin speech recognition. Syllable language models were used. The initial (consonant) and final (vowel) of a Mandarin syllable were modeled. There were totally 408 syllables covering 22 initials and 38 finals. The intra and inter syllable dependencies were respectively modeled through right context-dependent (RCD) initials and RCD finals. Without decision tree state tying, we generated 10k 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 RCD finals. The syllable recognition rates and the number of trained HMM states were reported for evaluation. For comparison, we realized HMM modeling using context-independent (CI) initials/CI finals and CD initials/CI finals. These baseline results without decision trees are shown in Fig. 2.

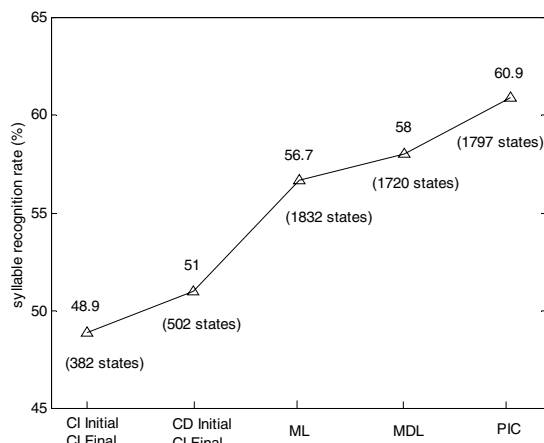


Fig. 2. Recognition rates and state numbers for different methods

4.2. Performance Comparison of ML, MDL and PIC

When ML, MDL and PIC are employed in decision tree construction, some parameters are incorporated to control the stop condition of node splitting. Using ML, the control parameters contain the increasing amount of log likelihood and the floor number of observation frames in tree nodes. Using MDL, a penalization factor is merged to tune the effect of model complexity. Using PIC, we introduce a weighting factor to adjust the effect of prior parameters in derived prediction information. Figure 2 displays the recognition results of ML,

MDL and PIC with comparable numbers of HMM states. State numbers are denoted in brackets. We can see that number of context-dependent states is greatly reduced when applying ML decision tree state tying. These results are significantly better than those of baseline results. However, it is obvious that MDL obtains a smaller number of states and higher recognition rates than ML. When applying PIC, the best syllable recognition rate 60.9% is achieved. The number of states of PIC is slightly larger than that of MDL. Overall, PIC is promising because of its good performance in recognition rate as well as suitable model size.

5. Conclusion

We surveyed several model selection approaches using AIC, BIC and MDL. A new PIC model selection is proposed to resolve the overestimation and underestimation dilemmas for HMM data modeling. Using PIC, the model with the largest prediction information was retrieved. The selected model provided the best generalization for data occurrences. To realize PIC in HMM data modeling, we properly characterized the statistics of real-valued multivariate HMM parameters by the conjugate prior densities. The formulated prediction information had exact solution without using Laplacian integral approximation. A multivariate t distribution was obtained to express the prediction information due to HMM mean vector and precision matrix. The corresponding HMM parameters were rapidly determined via MAP estimate. We also constructed a top-down prior/posterior propagation algorithm to calculate the structural hyperparameters of HMMs for evaluating prediction information. The prediction information, parameter estimation and hyperparameter evolution for HMMs were analytically addressed based on Bayesian learning viewpoints. In the experiments for building HMM decision trees, the proposed PIC achieved the highest speech recognition rate with moderate number of HMM states compared to ML and MDL criteria. In the future, the PIC model selection will be expanded for other data modeling problems and pattern recognition applications.

6. References

- [1] H. Akaike, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, December 1974.
- [2] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M. A. Picheny, "Decision trees for phonological rules in continuous speech", *ICASSP*, pp. 185-188, 1991.
- [3] J.-T. Chien and G.-H. Liao, "Transformation-based Bayesian predictive classification using online prior evolution", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 399-410, May 2001.
- [4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291-298, April 1994.
- [5] S. Geisser and W. F. Eddy, "A predictive approach to model selection", *Journal of the American Statistical Association*, vol. 74, no. 365, pp. 153-160, March 1979.
- [6] D. J. C. Mackay, "Bayesian interpolation", *Neural Computation*, vol. 4, pp. 405-447, 1992.
- [7] J. Rissanen, "Modeling by shortest data description", *Automatica*, vol. 14, pp. 465-471, 1978.
- [8] G. Schwarz, "Estimating the dimension of a model", *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464, 1978.
- [9] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition", *EUROSPEECH*, vol. 1, pp. 99-102, 1997.