7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

CONSTRUCTING SHARED-STATE HIDDEN MARKOV MODELS BASED ON A BAYESIAN APPROACH

Shinji Watanabe[1], *Yasuhiro Minami*[1,2], *Atsushi Nakamura*[1], *Naonori Ueda*[1]

Nippon Telegraph and Telephone Corporation
NTT Communication Science Laboratories 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan [1]
NTT Cyber Space Laboratories 1-1, Hikari-no-Oka, Yokosuka-shi, Kanagawa, Japan [2]
{watanabe,minami,ats,ueda}@cslab.kecl.ntt.co.jp

ABSTRACT

In this paper, we propose a method for constructing shared-state triphone HMMs (SST-HMMs) within a practical Bayesian framework. In our method, Bayesian model selection criterion is derived for SST-HMM based on the *Variational Bayesian* approach. The appropriate phonetic decision tree structure of SST-HMM is found by using the criterion according to a given data set. This criterion, unlike the conventional MDL criterion, is applicable even in the case of insufficient amounts of data. We conduct two experiments on speaker independent word recognition in order to prove the effectiveness of the proposed method. The first experiment demonstrates that the Bayesian approach is valid for determining the tree structure. The second experiment demonstrates that the Bayesian criterion can design SST-HMMs with higher recognition performance than the MDL criterion when dealing with small amounts of data.

1. INTRODUCTION

Shared-state triphone HMMs (SST-HMMs) have been widely used for speech recognition to deal with the problem of data insufficiency [1] [2]. An SST-HMM can be constructed by successively partitioning states based on phonetic decision tree clustering. The important problem to be faced when building an SST-HMM is to find a way of adaptively optimizing the tree structure (equivalent to the number of states) according to available training data. Namely, maintaining the balance between model complexity and training data size is quite important for obtaining high generalization performance.

The maximum likelihood (ML) is inappropriate as a model selection criterion since ML increases monotonically as the number of states increases. Some heuristic thresholding is therefore necessary to terminate the partitioning. To solve this problem, the minimum description length (MDL) criterion has been employed to determine the tree structure of SST-HMMs [3][4]. However, since the MDL is based on an asymptotic assumption, it is invalid when the number of training data is small because of the failure of the assumption. Clearly, computationally heavy cross-validation procedures also become unreliable when the sample is small.

To overcome the problem associated with the ML approach, we present a Bayesian criterion based on the *Variational Bayesian* (VB) approach [5][6]. More specifically, we derive an objective function consisting of distributions over model parameters and a structure of SST-HMM within the VB framework. Maximizing the objective function with respect to the number of states leads to the maximization of the *variational posterior distribution* over the model structure. Therefore, maximizing the objective function of a given data set, we can find not only the model parameters, but also the tree structure of SST-HMMs in the MAP sense. Since we do not use an asymptotic assumption when deriving the criterion, (unlike the MDL criterion), it is available even in the case of small amounts of data.

2. BAYESIAN SELECTION OF MODEL STRUCTURE

Let $\boldsymbol{O}$ be a set of training data and $p(\boldsymbol{O}|\theta_j)$ be a probability distribution with a set of parameters θ_j for label j. The ML approach estimates θ_j. On the other hand, the Bayesian approach estimates true posterior distributions $p(\theta_j|\boldsymbol{O})$ for θ_j. In the Bayesian approach, once $p(\theta_j|\boldsymbol{O})$ is estimated, the class label for unknown data $\boldsymbol{x}$ is determined by:

$$j = \arg\max_{j'} \int p(\boldsymbol{x}|\theta_{j'})p(\theta_{j'}|\boldsymbol{O})d\theta_{j'}. \quad (1)$$

As shown in Eq.(1), the Bayesian approach suppresses the problem of over-fitting the training data by integrating out the parameters. Moreover, by regarding model structure M as a random variable, we can introduce a true posterior distribution $p(M|\boldsymbol{O})$ for a model structure M. By maximizing $p(M|\boldsymbol{O})$ with respect to M, we can estimate the optimal model structure.

The true posterior distributions, however, are not easy to obtain since the expectations and the integrals are difficult to calculate. Recently, Waterhouse *et al.* [6] proposed a *Variational Bayesian* (VB) approach for obtaining the analytic form of posterior distributions by using the variational approximation technique. In the VB approach, Ueda and Ghahramani [5] introduced the functional:

$$\mathcal{F}_M[q(\theta_j|\boldsymbol{O},M)] = \left\langle \log \frac{p(\boldsymbol{O}|\theta_j,M)p(\theta_j|M)}{q(\theta_j|\boldsymbol{O},M)} \right\rangle_{q(\theta_j|\boldsymbol{O},M)} \quad (2)$$

$\langle u(y) \rangle_{p(y)}$ represents the expectation of $u(y)$ with respect to the distribution $p(y)$. Here, $p(\theta_j|M)$ is prior for θ_j and is given by a designer. $q(\theta_j|\boldsymbol{O},M)$ is *variational posterior distribution* to be estimated and is an approximation of the true posterior distribution. The optimal posterior distributions for a fixed M can be obtained analytically by maximizing $\mathcal{F}_M$ with respect to $q(\theta_j|\boldsymbol{O},M)$. Moreover, by using $\mathcal{F}_M$, the following monotonic property is obtained [5]:

$$\mathcal{F}_{M'} \geq \mathcal{F}_M \Rightarrow q(M'|\boldsymbol{O}) \geq q(M|\boldsymbol{O}). \quad (3)$$

Here, $q(M|\boldsymbol{O})$ denotes the variational posterior distribution over the model structure M. This indicates that by maximizing $\mathcal{F}_M$ with respect to not only $q(\theta_j|\boldsymbol{O},M)$, but also M, we can obtain the optimal parameter distributions and optimal model structure simultaneously. In this paper, we apply this objective function to the selection of an acoustic model structure.

3. PHONETIC DECISION TREES USING MDL CRITERION

We use phonetic decision tree clustering [1] as a basic framework for constructing shared-states in a set of triphone HMMs. A phonetic decision tree is a kind of binary tree, at each node (m) of which a phonetic "Yes/No" question is attached.

Let $\Omega(m)$ denotes a set of states that a tree node m holds. We start with only a root node $(m = 0)$ which holds a set of all triphone HMM states $\Omega(0)$ for an identical center phoneme. The set of triphone states is then split into two sets, $\Omega(m_Y)$ and $\Omega(m_N)$, which are held by two new nodes, m_Y and m_N, respectively. This occurs as a result of answers to a phonetic question, which inquires if the preceding phoneme is a vowel, if the following phoneme is a nasal, etc. A particular question ξ is chosen so that the partition is the optimal of all the possibilities. We continue this splitting successively for every new set of the states, and this produces a binary tree, each leaf node of which holds a clustered set of triphone states. The states belonging to the same cluster are merged into a single state. A set of triphones is thus represented by a set of shared-state HMMs. An HMM, which represents a phonetic segment, usually consists of a linear sequence of three or four states. A decision tree is produced specifically for each state in the sequence, and the trees are independent of each other.

In order to choose a question properly at each split, the ML criterion is most commonly applied. When a node m is split into a Yes node (m_Y^ξ) and a No node (m_N^ξ), the question ξ for splitting the node is chosen from the question set Ξ as follows:

$$\xi = \arg\max_{\xi' \in \Xi} \Delta\mathcal{L}_{\xi'}. \quad (4)$$

$\Delta\mathcal{L}_\xi$ is the gain in total likelihood when a node is split by ξ. The question is chosen to maximize the gain in total likelihood by the splitting.

Now, we describe how to calculate $\Delta\mathcal{L}_\xi$. We assume the following conditions.

- The state assignment while splitting is fixed.
- A single Gaussian for one state is used.
- A diagonal covariance matrix is used.
- Contributions of the transition probabilities to the likelihood are ignored.

By using the conditions, a likelihood can be obtained as a simple form without iterative calculation, and the conditions reduce calculation time.

Let $\boldsymbol{O}(i) = \{O_t(i) \in \mathcal{R}^D : t = 1, ..., T(i)\}$ be a training data set assigned to a state i. $T(i)$ denotes the total number of frames for state i. D is the dimensionality of the feature vector. A log likelihood for the data set assigned to a set of states Ω can be expressed as:

$$\mathcal{L}_\Omega = -\frac{1}{2}\left(T_\Omega \log\left(\left|\hat{\Sigma}_\Omega\right|\right)\right). \quad (5)$$

Note that the terms which do not contribute to $\Delta\mathcal{L}_\xi$ are neglected. $T_\Omega (\equiv \sum_{i\in\Omega} T(i))$ denotes the total number of frames in a data set assigned to Ω. $\hat{\Sigma}_\Omega$ is an ML estimate of a covariance matrix for a data set assigned to Ω. From Eq.(5), $\Delta\mathcal{L}_\xi$ can be obtained by:

$$\Delta\mathcal{L}_\xi = \mathcal{L}_{\Omega(m_Y^\xi)} + \mathcal{L}_{\Omega(m_N^\xi)} - \mathcal{L}_{\Omega(m)}. \quad (6)$$

The value of $\Delta\mathcal{L}_\xi$ is positive for any split, which causes nodes to continue splitting without a stop criterion. In order to stop the splitting, some heuristic thresholding is used. However, there is no valid way to determine the threshold value. Therefore, the value of the threshold is decided experimentally. On the other hand, an objective function (MDL function) using the MDL criterion instead of Eq.(5) can stop the splitting and needs a smaller heuristic threshold. With the MDL function, Eq.(6) is represented as [3]:

$$\Delta\mathcal{L}_\xi^{MDL} = \Delta\mathcal{L}_\xi - \lambda D \log T_{\Omega(0)}. \quad (7)$$

λ is the tuning parameter for MDL. By using Eq.(7) instead of $\Delta\mathcal{L}_\xi$ in Eq.(4) and stopping the node splitting when the condition $\Delta\mathcal{L}_\xi^{MDL} \leq 0$ is satisfied, the model structure which is optimized locally by using the MDL criterion can be obtained. Note that the term $D \log T_{\Omega(0)}$ in Eq.(7) is regarded as a penalty term added to Eq.(6), and is dependent on the total frame number $T_{\Omega(0)}$ of the training data. That is, a model structure using the MDL criterion is selected according to the training data

The MDL criterion is applicable only in the case of large amounts of data since the MDL criterion is derived by using the asymptotic property. Therefore, in the case of insufficient amounts of data, the model structure obtained by using the MDL criterion is not necessarily correct.

4. PHONETIC DECISION TREES USING BAYESIAN CRITERION

The Bayesian approach solves the problem of insufficient amounts of data. In this section we apply the objective function given by Eq.(2) to phonetic decision trees. We assume the same conditions as in Section 3.

The prior distributions for a mean vector $\boldsymbol{\mu}_\Omega$ and a covariance matrix Σ_Ω are assumed to be the following normal-gamma distribution:

$$p(\boldsymbol{\mu}_\Omega, \Sigma_\Omega) = \mathcal{N}(\boldsymbol{\mu}_\Omega | \boldsymbol{\nu}_0, \xi_0^{-1}\Sigma_\Omega) \prod_{d=1}^{D} \mathcal{G}(\Sigma_{\Omega,d}^{-1} | \eta_0, R_{0,d}), \quad (8)$$

where $\{\xi_0, \boldsymbol{\nu}_0, \eta_0, R_0\}(\equiv \psi_0)$ is a set of hyper parameter constants. Note that ξ_0 and η_0 are a scalar, $\boldsymbol{\nu}_0$ is a vector of dimension D, and R_0 is a $D \times D$ diagonal matrix. $\Sigma_{\Omega,d}$ denotes the diagonal element of Σ_Ω in row d of column d.

From the log likelihood (5) and the prior distributions (8), the variational posterior distributions for a mean vector $\boldsymbol{\mu}_\Omega$ and a covariance matrix Σ_Ω can be obtained as:

$$\begin{aligned} & q(\boldsymbol{\mu}_\Omega, \Sigma_\Omega | \boldsymbol{O}) \\ & \quad = \mathcal{N}(\boldsymbol{\mu}_\Omega | \boldsymbol{\nu}_\Omega, \xi_\Omega^{-1}\Sigma_\Omega) \prod_{d=1}^{D} \mathcal{G}(\Sigma_{\Omega,d}^{-1} | \eta_\Omega, R_{\Omega,d}), \quad (9) \end{aligned}$$

where $\{\xi_\Omega, \boldsymbol{\nu}_\Omega, \eta_\Omega, R_\Omega\}(\equiv \psi_\Omega)$ is a set of parameters defined by:

$$\begin{aligned} \boldsymbol{\nu}_\Omega &= \frac{\xi_0 \boldsymbol{\nu}_0 + \sum_{i \in \Omega} \sum_{t=1}^{T(i)} \boldsymbol{O}_t(i)}{\xi_0 + \sum_{i \in \Omega} T(i)} && (10) \\ \xi_\Omega &= \xi_0 + \sum_{i \in \Omega} T(i) && (11) \\ \eta_\Omega &= \eta_0 + \sum_{i \in \Omega} T(i) && (12) \\ R_{\Omega,d} &= R_{0,d} + \xi_0 (\nu_{0,d} - \nu_{\Omega,d})^2 + \\ & \quad + \sum_{i \in \Omega} \sum_{t=1}^{T(i)} (O_{t,d}(i) - \nu_{\Omega,d})^2. && (13) \end{aligned}$$

ψ_Ω is composed of ψ_0 and a function of a training data set assigned to Ω. Note that the variational posterior distributions (9) are the same as the true posterior distributions since there is no latent variable in the model structure assuming the conditions in Section 3. It follows that when $T(i) \to 0$, the posterior distributions are mainly influenced by the prior distributions while when $T(i) \to \infty$, they are mainly influenced by the training data. By using Eqs.(2) and (10)-(13), we can derive the gain in the objective function when m is split into m_Y^ξ, m_N^ξ by a question ξ,

$$\begin{aligned} \Delta\mathcal{F}_\xi &= f(\psi_{\Omega(m_Y^\xi)}) + f(\psi_{\Omega(m_N^\xi)}) - \\ & \qquad - f(\psi_{\Omega(m)}) - f(\psi_0). && (14) \end{aligned}$$

Here f is defined by:

$$\begin{aligned} f(\psi) &= -\frac{D}{2} \log \xi - \frac{\eta}{2} \log |R| + D \log \Gamma \left(\frac{\eta}{2} \right) \\ \text{for} & \quad \psi = \left\{ \psi_0, \psi_{\Omega(m)}, \psi_{\Omega(m_Y^\xi)}, \psi_{\Omega(m_Y^\xi)} \right\}, && (15) \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function. By stopping the node splitting when the condition $\Delta\mathcal{F}_\xi \leq 0$ is satisfied, a model structure based on Bayesian approach can be obtained.

5. EXPERIMENTS

We conducted two experiments in order to prove the effectiveness of our proposed Bayesian model selection. The first experiment was designed to verify how accurately the Bayesian criterion works when choosing the phonetic question and stopping the growth of the trees. The second experiment was designed to compare how the Bayesian and MDL criteria work with variable amounts of training data.

5.1. Validity of Bayesian Criterion

We performed the first experiment under the conditions, shown in Tables 1 and 2.

Table 1. Acoustic Conditions

Sampling Rate	16 kHz (Quantization 16 bit)
Feature Vector	12 - order MFCC with Δ MFCC
Window	Hamming
Frame Size/Shift	25/10 ms

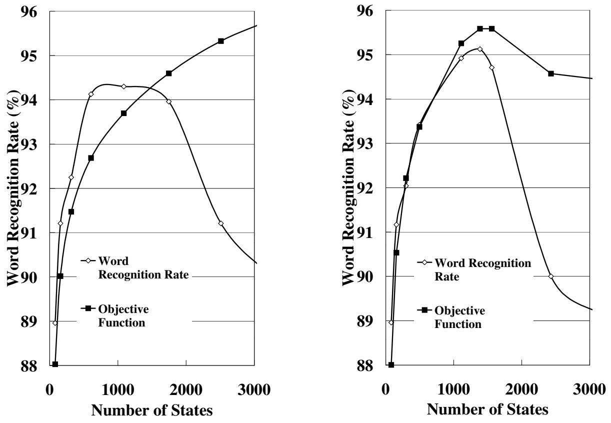
Table 2. Prepared HMM

# of States	3 (Left to Right)
# of Phoneme Categories	27
Output Distribution	Single Gaussian

The total number of root nodes (81) was obtained by multiplying the number of phoneme categories (27) with the number of HMM states (3). The parameter set ψ_0 for the prior distributions in each root node was determined from the average and the variance of each root node. The training was divided into two steps. In the first step, the model selection described in Section 3 or 4 was carried out. In the second step, the parameters of the current model structure which was obtained in the first step were re-estimated. We also kept a single Gaussian per state in the second step in order to evaluate the model structure obtained in the first step. The training and recognition data used in these experiments are shown in Table 3.

Table 3. Training and Recognition Data (A)

Training Data	ASJ Continuous Speech DB
Recognition Data	JCSD City (100) × Male 25



(a) ML

(b) Bayesian

Fig. 1. Objective Functions and Word Recognition Rates.

The training data consisted of about 3200 sentences from 30 males and the recognition data consisted of 100 words per male. Figure 1(a)/(b) shows the total value of the likelihood/Bayesian objective function and the word recognition rate for each number of states. In Figure 1, both objective function values were suitably normalized. The total value of the likelihood always increased as the number of states increased by splitting nodes, while the word recognition rate decreased gradually due to the over-training effect, as shown in Figure 1(a). On the other hand, a similar tendency between the total values of the Bayesian objective function and the word recognition rates was confirmed, as shown in Figure 1(b). Moreover, the objective function and the recognition rate in Figure 1(b) had its the maximum value at almost the same number of states. This indicates that our objective function was valid for determining tree structures.

5.2. Comparison between Bayesian and MDL Criteria

We compared the Bayesian approach with the MDL approach for variable amounts of training data. The acoustic conditions and prepared HMM were based on the same conditions as those listed in Section 5.1. The training and recognition data used in these experiments consisted of 100 words per male and are shown in Table 4. The number of speakers in the training data was varied between 3 and 40. We adopted the parameter $\lambda = 2$ in Eq.(7) from [3] [4].

The experimental results, as shown in Table 5, confirmed that the model selection using our Bayesian criterion resulted in better word recognition rates compared with that using the MDL criterion, especially in the case of small amounts of training data.

Table 4. Training and Recognition Data (B)

Training Data	JCSD City (100) × Male 3 $\sim$ 40
Recognition Data	JCSD City (100) × Male 25

Table 5. MDL and Bayesian

# of Speakers	MDL $\lambda = 2$		Bayesian	
	# of States	WRR (%)	# of States	WRR (%)
3	164	91.0	179	92.0
4	191	93.0	208	93.9
5	202	94.0	220	95.0
10	268	98.5	295	98.4
20	387	98.9	415	99.1
30	475	99.1	501	99.2
40	538	99.2	587	99.3

6. SUMMARY

In this paper, a method for constructing shared-state triphone HMMs using the Bayesian criterion has been described and applied to phonetic decision tree clustering. The results on speaker independent isolated word recognition have demonstrated the validity of the selection of a model structure using the Bayesian criterion. The results have also showed that the Bayesian criterion is superior to the MDL criterion as regards model selection, especially when dealing with small amounts of data.

7. REFERENCES

[1] J. J. Odell, “The Use of Context in Large Vocabulary Speech Recognition,” PhD thesis, Cambridge University, (1995).

[2] M. Ostendorf, H. Singer, “HMM Topology Design Using Maximum Likelihood Successive State Splitting,” Computer Speech and Language, 11, pages 17-41, (1997).

[3] K. Shinoda, T. Watanabe, “Acoustic Modeling Based on the MDL Principle for Speech Recognition,” *Proc. EuroSpeech’97*, vol. 1, pp. 99-102, (1997).

[4] W. Chou, W. Reichl, “Decision Tree State Tying Based on Penalized Bayesian Information Criterion,” *Proc. ICASSP’99*, vol. 1, pp. 345-348, (1999).

[5] N. Ueda, Z. Ghahramani, “Optimal Model Inference for Bayesian Mixture of Experts,” *Proc. NNSP’00*, pp. 145-154, (2000).

[6] S. Waterhouse, D. MacKay, T. Robinson, “Bayesian Methods for Mixture of Experts,” *Proc. NIPS8*, pp. 351-357, (1995).