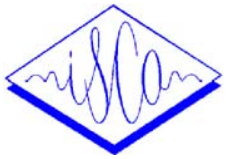

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MODEL SELECTION IN ACOUSTIC MODELING

S. S. Chen & R. A. Gopinath

IBM T.J. Watson Research Center, P.O. Box 218
Yorktown Heights, NY 10598, USA
Email: {schen,rameshg}@watson.ibm.com

10.21437/Eurospeech.1999-274x

ABSTRACT

Recently several classes of models have been suggested for use in continuous density HMMs for speech recognition. This paper proposes to choose both the model type and model size (number of parameters) by optimizing the Bayesian information criterion. Specifically we apply this to Gaussian mixture density estimation to determine both the number of Gaussians and the covariance structure of each Gaussian, and decision tree clustering of HMM states. A numerical algorithm similar to the EM algorithm for mixture density estimation is proposed for optimizing BIC.

1. INTRODUCTION

Most automatic speech recognition systems consist of acoustic models and language models. The acoustic models typically utilizes mixtures of diagonal Gaussians to describe the acoustic features. Recently a number of alternative acoustic models have been suggested:

- Covariance Modeling. Techniques such as semi-tied covariance [5, 7] and factor analysis [11, 6] have been suggested to explicitly model the correlation among acoustic dimensions. In this paper, we will also consider the so-called covariance-selection models [4, 9] which explicitly specify the structure of the inverse covariance matrix.
- non-Gaussian distributions. Instead of using mixtures of Gaussian components, power-exponential distributions [8] and Richter distributions [8] have been proposed to model the non-Gaussian nature of the acoustic features.

In this paper, we concentrate on the problem of covariance modeling. We consider the following two questions:

- How many Gaussian components should be used?
- For each Gaussian component, what is the appropriate covariance structure? The structures which we consider are diagonal covariance, factor analyzed covariance with m factors, full covariance and covariance-selection models. The covariance structure can be shared at various level, such as global level, phone level, HMM state level and Gaussian level.

We propose to identify the both the number of Gaussian components and the structure of each Gaussian components by optimizing the Bayesian information criterion (BIC) [12], a model selection criterion in the statistics literature. To overcome the high computational complexity, we propose a modified EM algorithm in which the M-step optimizes a penalized version of the so-called Q-function.

This paper is organized as follows. In section 2, we describe a variety of covariance structures. In section 3, we introduce the BIC criterion and the modified EM algorithm. In section 4, we apply our BIC approach in choosing the number of HMM states and choosing the number of diagonal Gaussian components for each HMM state in the standard acoustic models; we present experiments on 1997 DARPA Hub4 evaluation task.

2. COVARIANCE MODELING

From computational, storage and/or data-sparsity considerations, in several applications such as speech recognition, the covariances of the Gaussians are usually constrained to be either diagonal or block diagonal. For example in acoustic modeling for speech recognition, the state-of-the-art systems typical have about 100K diagonal Gaussians. In this section we review several other variants of covariance modeling.

2.1. Semi-Tied Covariance

The full covariance structure has rarely been used for acoustic modeling because of the overhead of huge number of parameters. However, the semi-tied covariance has been used with much success [5, 7]. Here each covariance Σ is of the form ADA^T, where D is diagonal and A is an invertible linear transform which is shared over a set of Gaussian components. Compared with diagonal Gaussians, semi-tied covariance models are able to boost the likelihood with very small number of extra parameters. It was reported in [5, 7] that most of the gain came from the global level semi-tied covariance; increasing the number of semi-tied variances does not significantly improve the recognition accuracy, and can sometimes degrade the performance. In this paper, we will assume the global semi-tied covariance in our model; we will not consider the issue of model selection on the number of semi-tied covariances.

2.2. Factor-analyzed Covariance

In factor analysis, each covariance is of the form

$$\Sigma_{p\times p} = \Lambda_{m\times p}\Lambda_{p\times m}^T + D_{p\times p}$$

where Λ is typically a matrix with fewer columns than rows and D is a diagonal matrix. The matrix Λ can be viewed

as a factor that models the off-diagonal terms in the covariance matrix. As the number of factors m varies, one gets covariance structures with various degrees of complexity: when $m = 0$, this corresponds to diagonal covariance; when $m = p$, this amounts to full covariance. For Gaussian mixture models with factor-analyzed covariances where the numbers of factors are pre-determined, the EM algorithm can be used to obtain the the maximum likelihood parameter estimates [11]. [11] also suggested parameter estimation by the MCE criterion. [11] reported improved results in digits recognition by using factor-analyzed covariances instead of diagonal covariances.

Typically the number of factors m is very small, and the increase in the number of parameters is small. We point out that by linear algebra

$$(\Lambda\Lambda^T + D)^{-1} = D^{-1} - D^{-1}\Lambda(I + \Lambda^T D^{-1}\Lambda)^{-1}\Lambda^T D^{-1}$$

we have the following inversion

$$\Sigma^{-1} = D^{-1} - \Phi^T\Phi$$

where $\Phi_{p\times m} = D^{-1}\Lambda(I+\Lambda^T D^{-1}\Lambda)^{-\frac{1}{2}}$. Thus by storing Φ, the acoustic likelihood can be computed efficiently.

2.3. Covariance-selection models

Yet another possibility of covariance modeling is to use a sparsity structure on the covariance that is derived from the data. However, sparse covariances do not necessarily lead to computational and/or storage advantages since ultimately the Gaussian likelihoods are evaluated as a quadratic form determined by the inverse covariance. As such it has been recognized [4, 9] that it is advantageous to model the so called "concentration" (inverse covariance) matrix with a sparsity structure. This paper considers the following models for the covariance of each Gaussian in a mixture: Σ^{-1} is a sparse matrix with a specific sparsity structure. Typically this structure is inferred from sample covariance estimates by some form of thresholding of the sample concentration matrix entries. When the sparsity structure is known, an iterative proportional scaling algorithm can be used to infer the the ML values of the concentration matrix. For Gaussian mixture models with covariance-selection where the sparsity structures is known, the EM algorithm can be used to obtain ML parameter estimates.

3. MODEL SELECTION VIA THE BAYESIAN INFORMATION CRITERION

In the problem of covariance modeling, we have to determine both the covariance structure and the covariance complexity. In this section, we will describe the Bayesian information criterion (BIC); we will present a modified EM algorithm to optimize the BIC for covariance modeling.

3.1. Bayesian Information Criterion

The problem of model identification is to choose one among a set of candidate models to describe a given data set. We often have candidates of a series of models with different number of parameters. It is evident that when the number of parameters in the model is increased, the likelihood of the training data is also increased; however, when the number of parameters is too large, this might cause the problem of overtraining. Several criteria for model selection have been introduced in the statistics literature, ranging from non-parametric methods such as cross-validation, to parametric methods such as the Bayesian Information Criterion (BIC) [12].

BIC is a likelihood criterion penalized by the model complexity: the number of parameters in the model. In detail, let $\mathcal{X} = \{x_i : i = 1, \cdots, N\}$ be the data set we are modeling; let $\mathcal{M} = \{M_i : i = 1, \cdots, K\}$ be the candidates of desired parametric models. Assuming we maximize the likelihood function separately for each model M, obtaining the maximum likelihood, say $L(\mathcal{X}, M)$. Denote $\#(M)$ as the number of parameters in the model M. The BIC criterion is defined as:

$$BIC(M) = \log L(\mathcal{X}, M) - \lambda\frac{1}{2}\#(M) \times \log(N) \qquad (1)$$

where the penalty weight $\lambda = 1$. The BIC procedure selects the model for which the BIC criterion is maximized. This procedure can be derived as a large-sample version of Bayes procedures for the case of independent, identically distributed observations and linear models [12].

The BIC criterion is well-known in the statistics literature; it has been widely used for model identification in statistical modeling, time series, linear regression, etc. Recently, BIC has been successfully used for automatic audio segmentation [3]. It is commonly known in the engineering literature as the minimum description length (MDL). BIC is closely related to other penalized likelihood criterions such as AIC [1]. One can vary the penalty weight λ in (1), although BIC is typically defined as $\lambda = 1$.

3.2. Choosing the number of mixture components

We first consider the problem of choosing the number of Gaussian mixture components where both the type and the complexity of the covariances have been pre-determined. Here we assume diagonal covariances, although our approach also applied for other types of covariances.

One problem in Gaussian mixture modeling is how to choose the number of Gaussians. It is well-known that too few Gaussians does not give sufficient model complexity whereas too many leads to overtraining. Our goal here is to adaptively choose the number of Gaussians according to the underlying complexity of the HMM state.

A common heuristic solution of this problem is the thresholding method. According to the number of samples belonging to the HMM state in the training data, one choose the number of Gaussians proportionally.

Here we propose to choose the number of Gaussians by optimizing the BIC criterion for each HMM state:

$$\hat{n} = \arg\max BIC(\text{mixture with } n \text{ diagonal Gaussians}).$$

Figure 3.2 illustrates how this procedure works for a particular HMM state. The horizontal axis represents the number of Gaussians. The vertical axis represents the log-likelihood in Panel (a) and the BIC value in Panel (b). Clearly as the number of Gaussians increases, the likelihood always improves, whereas the BIC value first increases then declines. The BIC value is optimized at $n = 27$.

We point out that here the search over the number of components in the BIC optimization can be efficiently implemented by utilizing the bisection method, since typically

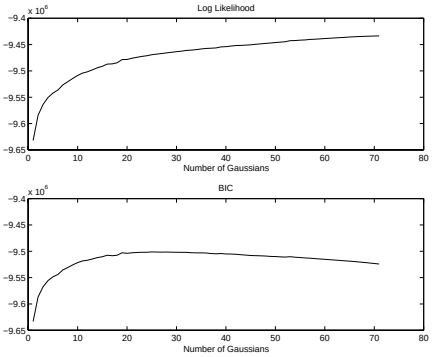

Figure 1: Choosing the number of Gaussians by maximizing the BIC criterion

the BIC curve is a concave function. We can start with the one component model, and successively obtain models with $2, 4, \ldots, 2^l$ components, until the BIC value starts to drop. Then the optimal number of components belongs to $[2^{l-2}2^l]$, and can be fond efficiently using the bisection method.

3.3. Choosing the number of HMM states

Most state-of-the-art speech recognition systems utilize context-dependent HMM states which are obtained by growing decision trees with phonetic questions. Usually the number of HMM states is determined by a heuristic thresholding scheme: the top-down decision tree growing process is terminated if the best split does not increase the likelihood by a threshold.

Here we apply the BIC criterion to determine the number of HMM states. Let $\mathcal{X} = \{x_i \in \mathcal{R}^p : i = 1, \cdots, N\}$ be the training data set. Top-down decision tree methods start with all data samples as the root node, then successively split according to the best question. Let $\mathcal{S} = \{s_1, \ldots, s_k\}$ be the current nodes; suppose s_k are the best candidate for splitting by the best question into new nodes s_{k+1} and s_{k+1}. Thus we are comparing the current clustering tree $\mathcal{S}$ with a new clustering tree $\mathcal{S}' = \{s_1 \ldots, s_{k-1}, s_{k+1}, s_{k+2}\}$. We model each node s_i as a diagonal Gaussian $N(\mu_i, \Sigma_i)$. It is clear that the increase of the BIC value by splitting s_k into s_{k+1} and s_{k+2} is

$$-n \log |\Sigma_k| + n_1 \log |\Sigma_{k+1}| + n_2 \log |\Sigma_{k+2}| + \lambda p \log(N) \quad (2)$$

where n_k is sample size of node k and Σ is the diagonal covariance matrix of node k.

Our BIC termination procedure is that two nodes should not be merged if (2) is negative. Since the BIC value is increased at each split, we are searching for an "optimal" clustering tree by optimizing the BIC criterion in a greedy fashion.

Note that we merely use our criterion (2) for termination. It is possible to use our criterion (2) as the distance measure in the bottom-up process. However, in many applications, it is probably better to use more sophisticated distance measures. It is also clear that our criterion can be applied to bottom-up tree methods.

3.4. A modified EM algorithm for covariance modeling

Our goal is to determine the number of Gaussian components, the types and the complexity of the covariances for each Gaussian by optimizing the BIC criterion. This optimization is challenging because of the huge model space we have to search through. Here we propose a modified EM algorithm which can possibly alleviate this difficulty.

Let $\mathcal{X} = \{x_i \in \mathcal{R}^p : i = 1, \cdots, N\}$ be the training data associated with a particular HMM state. Let $\sum_{k=1}^{n} \pi_k p(x, \mu_k, \Sigma_k)$ be the Gaussian mixture model with n components. The types of covariances we consider are the factor-analyzed covariances and the covariance selection models; the complexity corresponds to the number of factors or the number of entries in the inverse covariance matrix. For a particular number of Gaussian components n, we select the optimal covariance type and complexity for each Σ_k by the following modified EM algorithm.

- E-step. Compute the posterior probabilities

$$\gamma(i,k) = \frac{\pi_k p(x_i, \mu_k, \Sigma_k)}{\sum_j \pi_j p(x_i, \mu_j, \Sigma_j)}$$

- M-step. Identify the covariance matrix Σ_k by maximizing the penalized likelihood of the k-th "soft" cluster defined by $\{\gamma(i,k)\}_{i=1}^{n}$. For factor-analyzed covariances with n factors, we compute the maximum likelihood via EM [10]. For covariance selection models with, we first compute the sample inverse covariance matrix and locate the top m entries, then compute the maximum likelihood via the iterative proportional scaling algorithm [9]. We finally choose estimates (μ_k, Σ_k) with the highest BIC value.

We point out the Both the search within the M-step and the search over the number of components n can be implemented efficiently using the bisection approach discussed in section 3.2. The convergence of this algorithm is still under investigation.

4. EXPERIMENTS

In this section, we present experiments applying our BIC-based techniques on the 1997 DARPA broadcast news evaluation task. The IBM speech recognizer [2] was used in all our experiments.

4.1. Choosing the number of mixture components

We conducted experiments comparing the BIC approach in section 3.2 with the heuristic thresholding method. We designed a system by the thresholding method which had 90K Gaussians. By choosing the penalty weight $\lambda = 1$, we obtained a system which had roughly 90K Gaussians using the BIC method. Figure 4.1 plots each HMM state by the its training sample size and its number of Gaussians determined by the BIC procedure. Notice that a certain state belonging to F-2 and a certain state belonging to AO-2 are indicated in the figure. They both had roughly the same number of samples. It is interesting that the BIC procedure chose about 25 Gaussians for the state belonging to F-2 whereas about 105 Gaussians for the state belonging to AO-2. In fact, we found out that most of the "upper" states, which have big angles from the horizontal axis if

connected with the origin, are mostly vowels; most of the "lower" states, which have small angles from the horizontal axis if connected with the origin, are most consonants. This shows that the BIC procedure indeed tends to choose more Gaussians for more complex states. Table 1 shows that the system built by the BIC procedure outperformed the system built by the thresholding method, by 0.8% absolute. In fact, in our experiments, we have observed consistently that compared with the thresholding method, the BIC approach can produce systems achieving reduced error rate with the same number of Gaussians, or produce systems achieving the same error rate but with smaller number of Gaussians. Here the test set was subsampled from the the 1997 DARPA broadcast news evaluation task.

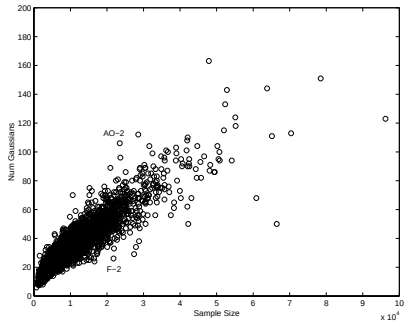

Figure 2: The BIC procedure tends to assign more Gaussians to more complex states

By varying the penalty weight λ in the BIC criterion, we can obtain systems with various numbers of Gaussians. As we decreased λ, we obtained systems with increasing numbers of Gaussians. As indicated in Table 1, the recognition accuracy dropped. We decided to use the 289K system as our baseline.

	# Gaussians	All	Prep	Spon
Standard	90K	26.0	11.9	23.5
$\lambda = 1.00$	90K	25.2	11.6	23.1
$\lambda = 0.80$	135	24.7	11.2	21.2
$\lambda = 0.65$	178	24.2	10.7	21.5
$\lambda = 0.54$	237	23.8	10.7	21.6
$\lambda = 0.45$	289	23.5	10.5	21.5

Table 1: Choosing the number of Gaussians

4.2. Choosing the number of HMM states

We designed decision trees using both the heuristic thresholding scheme and the BIC scheme. In both cases, we obtained $\approx$ 3800 HMM states; we then constructed acoustic models with $\approx 150K$ diagonal Gaussians where the number of Gaussians was determined by the heuristic thresholding scheme. Table 2 shows the recognition results on the 1997 evaluation set. Overall, the BIC approach gained about 0.3% absolute; we see improvements on both the prepared speech and the spontaneous speech. We believe further improvements can be obtained if the number of Gaussians is determined by the BIC method.

	All	Prep	Spon
Threshold	18.7	11.8	18.8
BIC	18.4	11.4	18.5

Table 2: Choosing the number of HMM states

5. CONCLUSION

We applied the Bayesian information criterion for choosing the number of Gaussians and the number of HMM states. We proposed a modified EM algorithm for general covariance modeling. We have not obtained results on general covariance modeling at the time the paper was written.

6. REFERENCES

[1] H. Akaike, "A new look at the statistical identification model", IEEE Trans. Auto. Control, vol 19, pp 716-723, 1974.

[2] S.S. Chen et al, "Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News", Proc. ICASSP, 1999.

[3] S. Chen et al, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", Proc. of DARPA Speech Recognition Workshop, Feb 8–11, Lansdowne VA, 1998.

[4] A. P. Dempster, "Covariance Selection", Biometrics, vol 28, 1972

[5] R.A. Gopinath, "Constrained Maximum Likelihood Modeling with Gaussian Distributions," Proc. of DARPA Speech Recognition Workshop, Feb 8–11, Lansdowne VA, 1998.

[6] R.A. Gopinath, "Factor Analysis Invariant to Linear transforms for Data", Proc. ICSLP, Sydney, Australia, Dec 1998.

[7] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov Models", IEEE Trans. Speech Audio Processing, Vol 7, to appear, 1999.

[8] M. J. F. Gales and P. A. Olsen, "Tail Distribution Modeling Using the Richter and Power Exponential Distributions," submitted to EuroSpeech 99.

[9] S. Lauritzen Graphical Models, Oxford University Press, 1996.

[10] D.B. Rubin and D.T. Thayer, "More on EM for factor analysis", Psychometrika, vol 48, 1983.

[11] L. Saul and M. Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition", Submitted to IEEE Trans. Speech Audio Proc, 1997.

[12] G. Schwarz, "Estimating the dimension of a model", Annals of Statistics, vol. 6, pp 461-464, 1978.