The Application of Bayesian Information Criterion in Acoustic Model Refinement

Jáchym Kolář, Luděk Müller
Department of Cybernetics, University of West Bohemia, Univerzitní 22,
Plzeň 306 14, Czech Republic, jachym@kky.zcu.cz, muller@kky.zcu.cz

Abstract: Automatic speech recognition (ASR) systems usually consist of an acoustic model and a language model. This paper describes a technique of an efficient deployment of the acoustic model parameters. The acoustic model typically utilizes Continuous Density Hidden Markov Models (CDHMM). The output probability of a particular CDHMM state is represented by a Gaussian mixture density with a diagonal covariance structure. Usually, the output probability density function of each CDHMM state contains the same number of mixture components although a different number of components in individual states may yield more accurate recognition results, especially for low-resource ASR systems. The central idea is to assign more components to states where it is effective and less components to states where the increasing number of components is not warranting a significantly better description of the training data. The number of mixture components for a particular CDHMM state is chosen by optimizing the Bayesian Information Criterion (BIC).

I. INTRODUCTION

Automatic speech recognition (ASR) systems usually consist of an acoustic model and a language model. The acoustic model typically utilizes Continuous Density Hidden Markov Models (CDHMM). CDHMM state output probability is commonly represented by a Gaussian mixture density with a diagonal covariance structure. In this paper, we concentrate on the problem of determining an appropriate number of mixture components. Usually, the output probability density function of each CDHMM state contains the same number of mixture components although a different number of components in individual states may yield more accurate recognition results.

The model selection problem is to choose one model from a set of candidate models to describe a given training data. The candidate models are models with a different number of parameters. It is evident that when the number of parameters is increased, the likelihood of the training data is also increased. But when the number of parameters is too large, the problem of overtraining may appear. It means that the training data are fitted too closely and the model does not generalize well. The performance of the model is then excellent on the training set but not on other data. On the other hand, when the number of parameters is too small, the model will not adequately represent the data. A natural way to find the balance between these two extremes is the use of the Bayesian Information Criterion (BIC).

II. MODEL ORDER ESTIMATE

The maximum-likelihood (ML) method is an efficient method for estimating parameter vectors when the dimension of the parameter space is fixed. But how to choose an appropriate dimension of the parameter space? The right choice is very important since models with too few parameters will not adequately represent the training data, whereas models with too many parameters might cause the problem of overtraining. The aim is to find a balance between these two extremes. A couple of criteria for model size selection have been introduced in the statistics literature, ranging from non-parametric methods such as cross-validation to parametric methods as the Akaike Information Criterion [1] or the Bayesian Information Criterion.

In the model selection problem, we have to choose one model m among a set of candidate models (hypotheses). The probability of a specific model given by the observed data X can be by using the Bayes' relation written as

$$p(m \mid X) = \frac{p(X \mid m)\, p(m)}{p(X)} \qquad (1)$$

where $p(m)$ is the prior probability reflecting our prior belief in the specific model. The model is typically defined by a set of parameters denoted by θ, so that we set up a generative model density $p(X|m,\theta)$. Thus, we obtain the following relation

$$p(X \mid m) = \int_{\theta} p(X,\theta \mid m) d\theta = \int_{\theta} p(X \mid \theta, m) p(\theta \mid m) d\theta \qquad (2)$$

where $p(\theta|m)$ carries a possible prior belief on the level of parameters. The integral in (2) is often too complicated to be evaluated analytically. A number of various approximations have been proposed, here we use the BIC approximation which has been introduced for the first time by G. Schwarz in 1978 [2]. This method approximates the integral by a Gaussian in the vicinity of parameters θ^* that maximizes the integrant. With this approximation, we get

$$\log p(X,m) \approx \log p(X \mid \theta^*, m) - \frac{d}{2} \log N \qquad (3)$$

where d is the dimension of a parametric model and N is the number of training cases. A detailed inference of BIC can be found in [3]. The BIC criterion has often been used for model identification in statistical modeling, time series, linear regression, automatic audio segmentation etc. [4, 5].

III. ACOUSTIC MODEL REFINEMENT

In CDHMM based speech recognition, it is assumed that the sequence of observed speech vectors is generated by a finite state machine which changes its state every time unit. Each time that a state is entered a speech vector is generated from the state's output probability density. To each speech unit (e.g. monophone or triphone) is assigned just one CDHMM, typically with 3 emitting states. CDHMM state output probability is represented by a Gaussian mixture density with a diagonal covariance structure. The output distribution is then defined as

$$b(o) = \sum_{m=1}^{M} c_m N(\mathrm{o} \mid \mu_{\mathrm{m}}, Q_m) \qquad (4)$$

where M is the number of mixture components, o is the observed vector, c_m is the weight of m-th component, and $N(o|\mu_m, Q_m)$ is the multidimensional Gaussian density with the mean vector μ_m and the diagonal covariance matrix Q_m. For c_m it holds

$$\sum_{m=1}^{M} c_m = 1. \qquad (5)$$

The mixture model is here a parsimonious representation of a non-standard output density. An illustration of a Gaussian mixture density and its components is shown in Fig. 1. A zero cepstral coefficient distribution of a particular state serves as an example there.

Now we will assume an application of the BIC to the output density of CDHMM states. We concentrate on the problem of determining an appropriate number of mixture components. Usually, the output probability density function of each CDHMM state contains the same number of mixture components although a different number of components in individual states may yield more accurate recognition results. The central idea is to assign more components to states where it is effective and less components to states where the increasing number of components is not warranting a significantly better description of the training data. Thus, BIC should tend to choose more components for the states representing more complex sounds and vice versa less components for the states representing less complex sounds. The parameters of the whole acoustic model are then efficiently deployed.

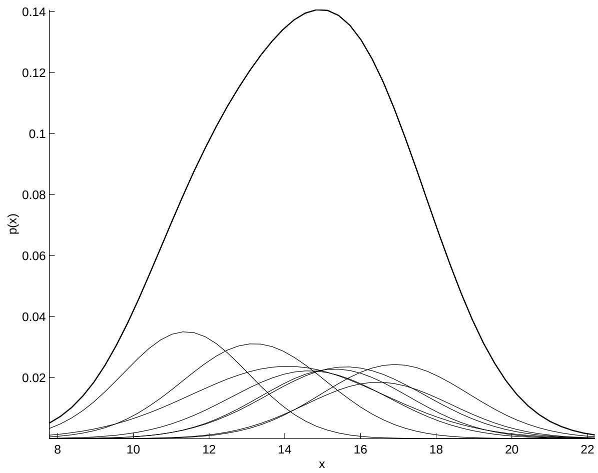

Fig. 1: An example of a Gaussian mixture density and its components

Let m be an acoustic model containing n_g Gaussians with diagonal covariance matrixes and K the dimensionality of the training data. Then, the total number of parameters needed to describe the model is

$$d = (2K + 1) \cdot n_g \,. \tag{6}$$

Let X be the training data set comprising N samples, and let $p(X|m)$ be the training data likelihood. With this notation, the BIC approximation in (3) can be rewritten as

$$BIC(X, m) = \log p(X \mid m) - \lambda \frac{d \log N}{2} \tag{7}$$

where parameter λ >0 is arbitrary chosen by a system designer. Rigidly taken, λ = 1 is set in (3), but the possibility of varying λ allows us to affect the overall model size. This fact is very important in many cases and will be mentioned later. The greater value of λ is chosen, the smaller model we get.

The aim is to choose a model m that maximizes $BIC(X,m)$. Note that the size of the model is exponentially penalized, so the large models can be selected only if they considerably better describe the training data. We can discuss two distinct cases of the BIC application. The maximal number of parameters that the ASR system can support is either limited or not. The first case arises when we are designing a low-resource system (e.g. ASR for mobile phones, PDA etc.) [6]. In the resource-constrained system, model size has significant economic and energetic consequences. A large model requires more non-volatile storage than a small one, and its associated computations usually require more processor cycles and runtime memory. The limited maximal number of parameters is here suboptimal, so we choose such value of λ at which the total number of parameters is equal to the maximal allowed number. In the latter case we can test different values of λ and determine that one that maximizes recognition accuracy [7].

We applied the BIC criterion on triphone models with shared states. However, the resembling strategy could be applied on any models that we use (e.g. monophones, biphones etc.) We searched for the BIC-optimal triphone models with shared states using a following strategy. We trained sets of triphone models with a fixed number of mixture components assigned to each state and stored them. Subsequently, we computed training data likelihood $p(m|X)$ for each state of each set by the forced alignment. Then we were able to easily determine BIC maximizing n_g for each state of the triphone set. Triphone models with a varying number of components were consequently retrained allowing a variable alignment.

This procedure is illustrated for a particular state in Fig. 2. The second emitting state of the model of *L_B-d+a* triphone serves as an example there. A maximal number of components is set to 32. The horizontal axes represent number of components. The vertical axes represent the training data log-likelihood (in the top part) and the BIC value (in the bottom part). As the number of mixture components increases, the log-likelihood improves too, whereas the BIC value first increases and then decreases. In this example, the optimal value of the BIC criterion was reached at n_g = 18.

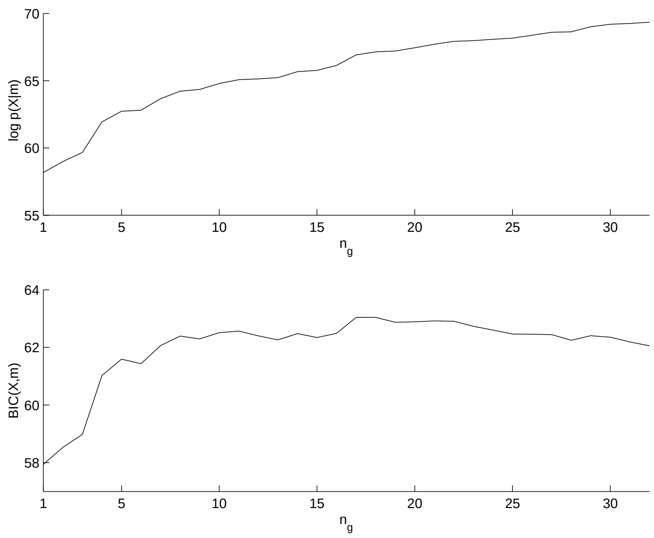

Fig. 2: An example of choosing a BIC optimal number of mixture components

IV. EXPERIMENTAL RESULTS

The method was tested on a subset of the Czech Read-Speech Corpus (UWB_S01) recorded at University of West Bohemia [8]. The used subset of the corpus consists of 40 phonetically balanced sentences that were read by 105 different speakers (66 males and 39 females). Thus, the used dataset altogether contains 4200 records. All waveforms were parametrized by the PLP method with 13 cepstral coefficients with additional delta and delta-delta coefficients. The whole dataset was divided into a training set (100 speakers) and a testing set (5 speakers not appearing in the training set). The vocabulary comprising 679 items was used both for the training and the testing. So we revolve a speaker independent system with a medium vocabulary. To evaluate only the impact of the acoustic model, no language model was used during the testing. The HTK speech recognizer [9] was used in all experiments. Recognition accuracy was used as an evaluation metric. It is defined as

$$Acc = \frac{N - D - S - I}{N} \cdot 100\% \tag{8}$$

where N is the total number of labels in a transcript file, D is the number of deletions, S is the number of substitutions, and I is the number of insertions.

We made several experiments to evaluate the impact of the acoustic model refinement on the performance of an ASR system. The value of λ was on each occasion chosen so that the BIC-refined system had the same total number of Gaussians as a corresponding baseline system. The systems having assigned a constant number of mixture components to each state were chosen as those baseline systems. We tested systems with 5, 6, 8, 10, and 13 mixture components. Results are shown in Table 1 where *Avg* n_g denotes the average number of components per state, *BIC Acc* the accuracy after applying BIC, *BL Acc* the baseline accuracy, and *Imp* the absolute accuracy improvement. As it is possible to see, a slight recognition accuracy improvement was achieved. A more significant improvement was reached when the number of components per state was 5 and 6. This case corresponds to a low-resource system with a limited number of parameters.

TABLE 1. A comparison of the recognition accuracy of a baseline and the BIC-refined system

λ	*Avg* n_g	*BIC Acc* [%]	*BL Acc* [%]	*Imp* [%]
0.0027	5	79.07	77.52	1.55
0.0022	6	79.40	77.93	1.47
0.0016	8	79.29	78.97	0.32
0.0012	10	79.69	79.26	0.43
0.0008	13	79.58	79.33	0.25

V. CONCLUSION

In this paper we have described the application of the Bayesian Information Criterion in an ASR acoustic model refinement. By optimizing BIC, the overall acoustic model parameters are efficiently deployed between individual states. This yields a slight recognition accuracy improvement. By varying the penalizing parameter λ we are able to influence overall model size, so we can generate a superior model at a given fixed size. This is convenient in building a low-resource system since the model size has relevant economic consequences. A more significant recognition accuracy improvement was achieved for the case of a low-resource system.

VI. ACKNOWLEDGEMENT

Support for this work was provided by the Ministry of Education of the Czech Republic, project No. MSM234200004.

REFERENCES

[1] H. Akaike: “A new look at the statistical identification model”, IEEE Trans. Automatic Control, Vol. 19, pp. 719 – 723, 1974
[2] G. Schwarz: “Estimating the Dimension of a Model”, Annals of Statistics, Vol. 6, pp.461 – 464, 1978
[3] A. Lanterman: “Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Order Estimation”, International Statistical Review, Vol. 69, No. 2, August 2001, pp. 185 – 212
[4] S. Chen, R. Gopinath: “Model Selection in Acoustic Modeling”, Proc. EUROSPEECH99, Budapest, Hungary,1999
[5] L.K. Hansen, J. Larsen, T. Kolenda: “Blind Detection of Independent Dynamic Components“ Proc. of ICASSP'2001, Salt Lake City, USA, SAM-P8.10, Vol. 5, 2001
[6] S. Deligne, E. Eide, R. Gopinath, D. Kanevsky, B. Maison, P. Olsen, H. Printz, J. Sedivy: “Low – Resource Speech Recognition of 500 – Word Vocabularies”, Proc. EUROSPEECH 2001 Scandinavia, Aalborg, Denmark, 2001
[7] S. Chen, E. Eide, M. Gales, R. Gopinath, D. Kanevsky, P. Olsen: “Automatic Transcription of Broadcast News”, IBM T.J. Watson Research Center, Yorktown Heights, USA, 2001
[8] J. Psutka, V. Radová , L. Müller, J. Matoušek, P. Ircing, D. Graff: “Large Broadcast News and Read Speech Corpora of Spoken Czech”, Proc. EUROSPEECH 2001 Scandinavia, pp. 2067 – 2070, Aalborg, Denmark, 2001
[9] S. Young et al.: “The HTK Book (for HTK Version 3.1)”, Cambridge University, available at http://htk.eng.cam.ac.uk/, 2002