Gaussian Elimination Algorithm for HMM Complexity Reduction in Continuous Speech Recognition Systems

Glauco F. G. Yared, Fábio Violaro and Lívio C. Sousa

Department of Telecommunications
State University of Campinas, Campinas-SP, Brazil
{glauco,fabio,livio}@decom.fee.unicamp.br

Abstract

Nowadays, HMM-based speech recognition systems are used in many real time processing applications, from cell phones to automobile automation. In this context, one important aspect to be considered is the HMM model size, which directly determines the computational load. So, in order to make the system practical, it is interesting to optimize the HMM model size constrained to a minimum acceptable recognition performance. Furthermore, topology optimization is also important for reliable parameter estimation. Previous works in this area have used likelihood measures in order to obtain models with a better compromise between acoustic resolution and robustness. This work presents a new approach based on a Gaussian Importance Measure (GIM) used in the Gaussian Elimination Algorithm (GEA) for determining the more suitable HMM complexity. The results are compared to the classical Bayesian Information Criterion.

1. Introduction

In the last years, applications requiring speech recognition techniques have grown considerably, varying from speaker independent to speaker adapted systems, in low and high noise environments. Moreover, the speech recognition systems have been designed to achieve both high processing performance and accuracy. In order to achieve such aims, the model size is an important aspect to be analyzed as it is directly related to the processing load and to the model classification capability. The model size is related to the number of Gaussian components used in each mixture. So, this work presents a well-known approach and a new one used to determine the number of Gaussian components of each state model: the first is the classical Bayesian Information Criterion (BIC) and the second uses the new Gaussian Importance Measure (GIM) in the Gaussian Elimination Algorithm (GEA).

Briefly speaking, the training process is performed by means of the Baum-Welch algorithm, then the model selection is used to choose the more appropriate HMM topology and the HTK [1] is used for the recognition task. The system performances are compared in terms of word recognition rate.

2. Background

The statistical modeling problem has some points that do not depend on the specific task for which the model was designed. One classical problem that must be overcome is the over-parameterization [2] which may occur with large models, that is, models with an excessive number of parameters. In general, such models present low training error rate, due to the high flexibility, but the performance, with a test database, is almost always unsatisfactory. On the other hand, models with insufficient number of parameters can not even be trained. At this point there is a trade-off between robustness and trainability, which must be accomplished in order to obtain a high performance system. In the speech recognition context, the word recognition rate is used as a performance measure and the total number of Gaussian components as the model size.

Another important issue to be considered is that a reliable parameter estimation is achieved only when there is enough data available, otherwise they are poorly estimated [3]. As the training database normally have a different number of data samples of each phone, it is reasonable to expected that the number of data samples is also a limiting factor for increasing the number of clusters in a given phone model.

Thus the idea of using different number of components for each mixture is firstly supported by the fact that each acoustic unit has a certain number of samples in the database. Depending on the amount of data samples available, the acoustic resolution of the HMM models should be increased or decreased in order to perform a reliable parameter estimation. In addition, the complexity of the acoustic distribution boundary also determines the number of components required for correct modeling different classes.

There are also some practical arguments to support the idea of determining a varying number of Gaussian components per state. The computational cost is directly related to the number of Gaussian components present in the system. As a consequence, the number of operations and the memory requirements increase with the number of components.

The theoretical and practical reasons presented above are the main concepts which support the idea of obtaining acoustic models with varying number of Gaussian components per state.

Briefly, section 3 describes the speech recognition system and the database used in the experiments. Then, the classical BIC method for determining the more suitable model size is summarized in section 4. Section 5 presents the new Gaussian Elimination Algorithm (GEA) and section 6 shows the experimental results. Finally, the conclusions are given in section 7.

3. Speech Recognition System and Database

A brazilian Portuguese small vocabulary database (700 different words) [4], with a set of 200 different sentences, spoken by 40 speakers (20 male and 20 female) was used in the experiments. In the experiments, 1200 utterances were used for training and 400 for testing [4].

A basic maximum likelihood (ML) training system was im-

plemented in order to obtain three state left-right continuous density HMM models. As 36 different context-independent phones (including silence) are used, 108 multidimensional Gaussian mixtures, with a variable or fixed number of components in each mixture, are present in the system. In addition, each Gaussian component is represented in a 39-dimensional space (12 mel cepstral coefficients, 1 log-energy parameter and their first and second order derivatives), using diagonal covariance matrix. The recognition task is performed by the HTK [1] using a Back-off bigram language model.

4. Bayesian Information Criterion

The BIC have been widely used within statistical modeling in many different areas for model structure selection. The main concept that supports the BIC is the principle of parsimony, that is, the selected model should be that with the smallest complexity and the highest capacity of modeling the training data. This can be observed directly from Equation (1)

$$BIC\left(M_l^j\right) = \sum_{t=1}^{N_j} \log P\left(o_t^j | M_l^j\right) - \lambda \frac{\nu_l^j}{2} \log N_j, \qquad (1)$$

where M_l^j is the candidate model "l" of state "j", N_j is the number of data samples of state "j", o_t^j is the t-th data sample of state "j", ν_l^j is the number of free parameters present in M_l^j and the parameter λ controls the penalization term.

According to such criterion, the selected model is the one over all models that gives the highest BIC value. As can be noted, the topology of each state model is selected despite of the other existing states. However, there are some proposed modifications in the BIC in order to take into acount all existing states during the topology selection [5].

5. Gaussian Elimination Algorithm (GEA)

Previous works in this area have used likelihood measures in model topology selection criterions [5]. The present work defines a Gaussian Importance Measure (GIM) which is first used in a new discriminative Gaussian selection algorithm and then in a Euclidian distance based Gaussian selection procedure.

5.1. Discriminative Determination of Model Complexity

The proposed discriminative method for determining the more suitable number of Gaussians per state differs from previous works in this area [5, 6, 7], in the sense that the algorithm starts from a well-trained system and indicates which Gaussians should be eliminated using the new GIM, instead of likelihood measures. All multidimensional Gaussians $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are represented in a 39-dimensional acoustic space, and the probability density function is given by Equation (2)

$$f(\mathbf{O_t}) = \frac{1}{(2\pi)^{\dim/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\mu)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mu)/2}, \qquad (2)$$

where $|\boldsymbol{\Sigma}|$ is the determinant of covariance matrix. If the parameters are statistically independent (diagonal covariance matrix), then the pdf can be written as in Equation (3)

$$f(\mathbf{O_t}) = \prod_{d=1}^{\dim} \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\left[(x_d-\mu_d)^2/2\sigma_d^2\right]}. \qquad (3)$$

Futhermore, the contribution of each sample to the GIM, along each acoustic dimension, is given by the areas indicated in Figure 1(a) and 1(b).

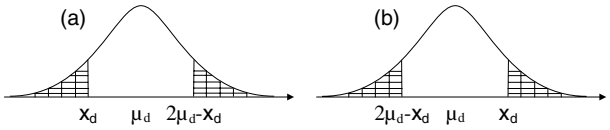

Figure 1: *Contribution of each sample to the GIM. (a) for* $x_d \leqslant \mu_d$. *(b) for* $x_d > \mu_d$.

For each Gaussian, the contribution along all dimensions to the GIM can be calculated by Equation (4)

$$GIM\,(O_t)^{(i;j;s)} = \prod_{d=1}^{\dim} \left(1 - \left[\frac{2}{\sqrt{\pi}} \int_0^{||Z_d||} e^{-(z_d)^2} dz_d\right]\right), \qquad (4)$$

where "dim" is the feature vector dimension, $z_d = \frac{x_d - \mu_{dij}}{\sqrt{2}\sigma_{dij}}$, and $\mathbf{O_t} = (x_1, x_2, \ldots, x_{\dim})$ is the feature vector. The values μ_{dij} and σ_{dij} correspond to the mean and standard deviation respectively, along dimension "d", of Gaussian "i" that belongs to state "j".

The GIM of Gaussian "i", that belongs to state "j", is calculated for every data sample of state "s", so that the mean value of GIM can be obtained in respect to each existing state, as shown in Equation (5) below

$$P_{GIM}^{(i;j;s)} = \frac{\sum_{t=1}^{N_s} GIM\,(O_t)^{(i;j;s)}}{N_s}, \qquad (5)$$

where N_s is the number of data samples of state "s".

In order to calculate the GIM value, it is required a segmented database, since Equation (4) uses labeled frames. The segmentation is obtained by means of a Viterbi alignment against the correct transcription using the best available system.

On a previous work [8], a likelihood based measure was used to compute contribution of each data sample to the importance of each Gaussian. The new proposed measure (GIM) is based on the probability of the feature samples being out of the interval

$$\mu_d - ||x_d - \mu_d|| < x < \mu_d + ||x_d - \mu_d||.$$

As can be noted, the closer the feature sample is to the Gaussian mean value, the higher is the contribution to the GIM along the analyzed dimension.

The $P_{GIM}^{(i;j;s)}$ can be used as a measure of the importance of each Gaussian in respect to each existing state. In this manner, a discriminative Gaussian selection method can be implemented. Thus, the main objective is to maximize the discriminative relation so that each final model gives the highest $P_{GIM}^{(i;j;s)}$ for the correct patterns and gives the smallest $P_{GIM}^{(i;j;s)}$ for the wrong patterns at the same time. This can be done by the maximization of Equation (6) below

$$DC^{(j)} = \frac{\left[\sum_{i=1}^{M_j} P_{GIM}^{(i;j;j)}\right]^K}{\left[\sum_{s\neq j}^{N} \sum_{i=1}^{M_j} P_{GIM}^{(i;j;s)}\right] \Big/ N-1} \qquad (6)$$

where K is the rigour exponent, M_j is the number of Gaussians in state "j" and N is the total number of states. If the logarithm of the Discriminative Constant (DC) is taken, the resulting expression in Equation (7) is similar to Equation (1), in the sense that the first term measures the capacity of modeling the correct patterns and the second is a penalty term

$$\log DC^{(j)} = K \log \sum_{i=1}^{M_j} P_{GIM}^{(i;j;j)} - \log \frac{\sum_{s \neq j}^{N} \sum_{i=1}^{M_j} P_{GIM}^{(i;j;s)}}{N-1}. \quad (7)$$

However, it differs in the sense that the penalty term in Equation (1) only considers aspects inherent to the analyzed model, while the penalty term in Equation (7) takes into acount aspects from models of all system states. In a wide sense, the DC relation also can be thought as a signal to noise relation.

The main idea is to eliminate Gaussians from a well trained model with a fixed number of Gaussian components per state and observe the new DC value obtained. The Discriminative Constant (DC) given above may increase or decrease depending on the relevance of the eliminated Gaussian. In this manner, the rigor exponent plays an important role in the Gaussian selection, since it makes the discriminative criterion DC more restrictive, that is, the greater the exponent the more rigorous the criterion is and therefore less Gaussians are eliminated.

The procedure described above is applied for each HMM state model. Once the discriminative Gaussian selection is finished, the resulting model is trained by means of the Baum-Welch algorithm.

It is also important to observe that the discriminative algorithm only detects Gaussians that converged to the wrong distributions during training. However, there may still exist exceeding Gaussians after applying the discriminative algorithm. Although these Gaussians are not detected by the discriminative algorithm, they must be discarded, since this can be performed without degradation of the model capacity of classification. In this manner, an additional distance based algorithm is applied to clean the exceeding Gaussians.

5.2. Gaussian Selection Based on Euclidian Distance Measure

The HMM models obtained after training by means of the Baum-Welch algorithm, frequently have redundant components, that is, Gaussians that converged to almost the same position in the acoustic space and that give practically the same contribution for classification.

A distance threshold is necessary for determining groups of components that should be replaced by just one. In addition, a different threshold should be used for Gaussians in the boundaries and for Gaussians in the central part of the distribution. Therefore, at this point it is necessary to determine which Gaussians are in the boundary or in the central part of the distribution, and it is also required to find a threshold for each of them.

Initially, the Euclidian distance is calculated between every Gaussians for a specific state, and the exceeding components are replaced by the Gaussian with the highest determinant of the covariance matrix.

After that, the $P_{GIM}^{(i;j;s)}$ was used in order to give an indication of which Gaussians are in the boundaries, and also to give a different weight for these Gaussians and those in the central part of the distribution. Therefore, a modified Euclidian distance measure was defined in Equation (8)

$$M_{d_{xy}} = \frac{d_{xy}}{\frac{\sum_{i=x,y} P_{GIM}^{(i;j;j)}}{\sum_{i=x,y} P_{GIM}^{(i;j;j)} + \sum_{s \neq j} \sum_{i=x,y} P_{GIM}^{(i;j;s)}}}, \quad (8)$$

where d_{xy} is given by

$$d_{xy} = \sqrt{(\mu_x - \mu_y) \cdot (\mu_x - \mu_y)^T}, \quad (9)$$

and μ_x and μ_y are the mean vector of Gaussian components x and y respectively.

6. Experiments

The experiments were performed using the speech recognition system and database previously described. The first results were obtained for a fixed number of Gaussian components per state (baseline systems), which are shown in Figure 2

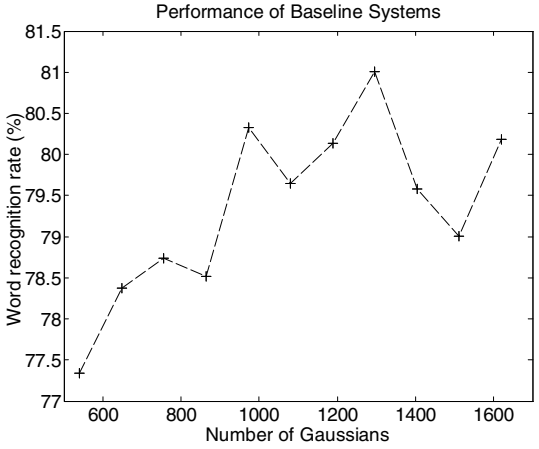

Figure 2: *Baseline systems containing a fixed number of Gaussian components per state (from 5 to 15 Gaussians per state). The HTK was used for the recongition task.*

As can be observed, the system with 1296 Gaussians (12 Gaussians per state) gives the highest recognition performance. The strategy adopted in this work is to find the more appropriate compromise between the size and performance starting from a well trained system containing a fixed number of Gaussians per state. Moreover, the final topology obtained after applying the proposed method has 3 states per phone model and a varying number of Gaussian components per state.

The first step is to apply the discriminative algorithm in order to eliminate Gaussians that are modeling the wrong distributions. The tests have shown that the rigour exponent "K" must be determined so that the reduced systems obtained from the discriminative algorithm should present at least the same perfomance as the original baseline system. The rigour exponent used in the experiments was 10^5.

The second step is to apply the distance based algorithm to the reduced system obtained from the discriminative method. The idea is to clean the exceeding Gaussians which are not detected in the first step.

Finally, the reduced systems from the GEA are evaluated and compared to the baseline one and to the systems obtained

from the BIC method (using the HTK in the recognition task). The results are shown in Figure 3.

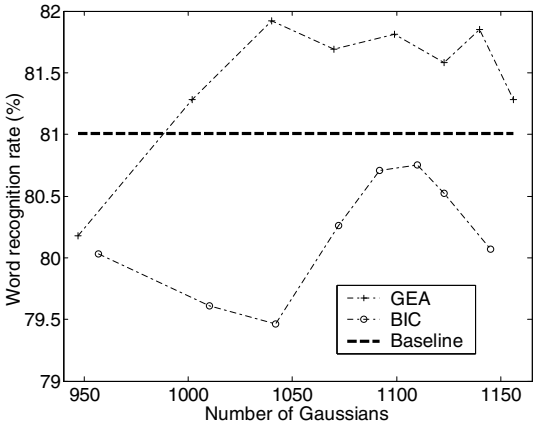

Figure 3: *The reduced systems obtained from the Gaussian Elimination Algorithm ("+") are compared to those obtained from the BIC ("o") and to the baseline one ("- -").*

The results shown in Figure 3 are summarized in Tables 1 and 2.

Table 1: Comparison between systems obtained from the GEA (k=10^5) and the baseline system containing 12 Gaussians per state.

Distance threshold ($M_{d_{xy}}$)	Reduced model size	Economy of Gaussians (%)	Recog. rate (%)	Differ. in perform. (%)
7	947	26.9	80.18	-0.83
6	1002	22.7	81.28	+0.27
5.5	1040	19.7	81.92	+0.91
5	1070	17.4	81.69	+0.68
4.5	1099	15.2	81.81	+0.80
4	1123	13.3	81.58	+0.57
3.5	1140	15.3	81.85	+0.84
0	1156	12.0	81.28	+0.27

Table 2: Comparison between systems obtained from the BIC and the baseline system containing 12 Gaussians per state.

regular. paramet. (λ)	Reduced model size	Economy of Gaussians (%)	Recog. rate (%)	Differ. in perform. (%)
0.3	957	26.2	80.03	-0.98
0.2	1010	22.1	79.61	-1.40
0.15	1042	19.6	79.46	-1.55
0.1	1072	17.3	80.26	-0.75
0.07	1092	15.7	80.71	-0.30
0.05	1110	14.4	80.75	-0.26
0.03	1123	13.3	80.52	-0.49
0.01	1145	11.7	80.07	-0.94

As can be noted, the reduced systems obtained from the GEA algorithm clearly outperform the baseline system and those obtained from the BIC method. The reason may rely on the fact that the GEA takes in acount information from models of all states during analysis, while the BIC uses only inherent information of the analyzed state model. In addition, the rigour exponent K used in GEA controls the degree of model degradation during the elimination process, while the parameter λ used in BIC only controls the penalization due to model size.

It is important to remark that the GEA may be applied to more complex systems (using context-dependent phone models), using a large vocabulary, for HMM complexity optmization, since there is no restriction in the formulation. Moreover, the same concepts may be extended for optimizing HMM complexity in any pattern recognition problem.

7. Conclusion

The GEA proposed here is able to determine HMM models with a better compromise between size and robustness, increasing system recognition performance while using less parameters (around 0.9% when compared to the original baseline system, with an economy of 19.7%). In addition, it outperforms the classical BIC when comparing the performance of the reduced systems. So, the next work is to apply the GEA during the process of building tree based tyed states context-dependent phone models using a large vocabulary database.

8. Acknowledgement

We would like to thank the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the research funding.

9. References

[1] *The HTK Book*, Cambridge University Engineering Departament, 2002.

[2] L. A. Aguirre, *An Introduction to Systems Identification - Linear and Non-linear Techniques Applied to Real Systems (in portuguese)*. Editora UFMG, 2000.

[3] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[4] C. A. Ynoguti, "Continuous speech recognition using hidden markov models (in portuguese)," Ph.D. dissertation, State University of Campinas, 1999.

[5] A. Biem, "Model selection criterion for classification: Application to hmm topology optimization," in *7-th International Conference on Document Analysis and Recognition (ICDAR'03)*, 2003.

[6] Y. Gao, E.-E. Jan, M. Padmanabhan, and M. Picheny, "Hmm training based on quality measurement," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1999.

[7] M. Padmanabhan and L. R. Bahl, "Model complexity adaptation using a discriminant measure," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 205–208, 2000.

[8] G. F. G. Yared and F. Violaro, "Finding the more suitable hmm size in continuous speech recognition systems," in *International Information and Telecommunications Technologies Symposium*, 2004.