STATE DIVERGENCE-BASED DETERMINATION OF THE NUMBER OF GAUSSIAN COMPONENTS OF EACH STATE IN HMM

Xiao-Bing Li *Ren-Hua Wang*

Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei, Anhui, 230027, China
lixiaobing@ustc.edu, rhw@ustc.edu.cn

ABSTRACT

A new, state divergence-based algorithm is proposed in this paper to determine the number of Gaussian components of each state in continuous density HMM by maximizing the between-state divergence. The unscented transform based approximation of the Kullback-Leibler divergence is adopted to measure the between-state model divergence to direct the determination. Due to the advantage of being more discriminative, the proposed approach can lead to more compact HMM. Our experimental evaluation shows that compared with the conventional Bayesian Information Criterion based determination (which is better than the uniform determination), the presented method can reduce the total number of Gaussian components to about 63%, while it results in almost negligible degradation of the recognition performance.

1. INTRODUCTION

Modelling the state by a mixture of Gaussian components is a common technique in current state-of-the-art, continuous density HMM-based speech recognition systems. In order to estimate the model parameters using the Expectation Maximization (EM) algorithm [1], the number of Gaussian components for each state should be specified firstly. In general, two methods are used to determine the number of Gaussian components: (1) same number of components is used across each state; (2) for each state, the number of components is determined proportionally to the corresponding state occupancy.

Though with an acceptable recognition performance, the models resulted from the two determination approaches are suboptimal and are usually at a price of large number of redundant Gaussian components. In other words, large storage and computation resources are required. Furthermore different states depend on different phonetic context and may contribute unequally to the classifier's classification error, as has been reported in [2], so the two approaches are also unfair to the states (i.e. the "non-aggressive" states in [2]) that need more Gaussian components for their more contribution.

Therefore, many alternative determination approaches to assigning the number of components in Gaussian mixture models have been proposed, e.g., Bayesian Information Criterion (BIC) [3] or Minimum Description Length (MDL) [4] based determination [5][6][7], state quality measurement [2], agglomerative clustering based determination [8], and leave-one-out likelihood based method [9]. Among them, BIC-based determination has been reported to give a much better performance than other methods.

However, in the BIC-based approach, the determination for each state is given separately, and the competitive relationship between states is not considered. In this paper, we propose a new algorithm to determine the number of components of each state. The algorithm gives the determination by maximizing between-state divergence given the total number of Gaussian components. As the competitive relationship between states is considered, this approach has the advantage of more discrimination.

In our approach, the symmetric Kullback-Leibler divergence (KLD) [10], an information theoretic measure of the distortion (distance) between two probability density functions, is adopted with its unscented transform [11] based approximation [12] to measure the between-state model divergence. The Gaussian components are allocated successively to a state where maximum increment of the between-state model divergence is obtained. Therefore, we can determine the number of Gaussian components of each state at any operating point (measured by the total number of Gaussians used) while the between-state model divergence is maximized.

The rest of the paper is organized as follows. In Section 2, the Bayesian Information Criterion and its corresponding determination are briefly reviewed. An overview of Kullback-Leibler divergence and the unscented transform based approximation are given in Section 3. Section 4 proposes the algorithms for generating the competitive states and the divergence-based determination of the number of Gaussian components. Database, experimental setups and results are presented in Section 5. Conclusion is given in Section 6.

2. BAYESIAN INFORMATION CRITERION BASED DETERMINATION

Well known as a model selection criterion in the statistics literature, Bayesian Information Criterion [3], is a likelihood criterion penalized by the number of parameters in the model. It is defined as:

$$BIC(\theta) = log\, P(X|\theta) - \frac{\lambda}{2}\#(\theta) log\, N \tag{1}$$

where $X = \{x_i, i = 1, \cdots, N\}$ is the data set, θ is the model, $log\, P(X|\theta)$ is the log likelihood of X given θ, and $\#(\theta)$ denotes the number of parameters in θ. The parameter λ is the penalty weight.

Fig. 1 plots the BIC values against the number of Gaussian components. By increasing the number of model parameters, the BIC value first increases then declines. The number of Gaussian components of a state S is specified at the point (In Fig. 1, it is 20.) with the maximum BIC value.

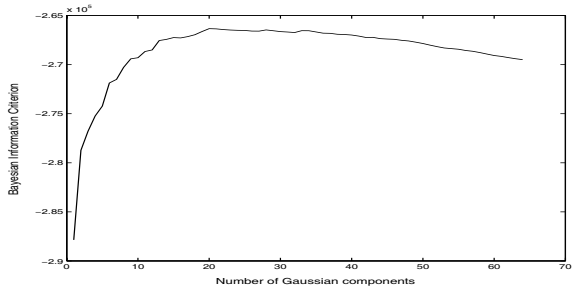

Fig. 1. Bayesian Information Criterion value vs. number of Gaussian components

3. KULLBACK-LEIBLER DIVERGENCE AND UNSCENTED TRANSFORM BASED APPROXIMATION

The Kullback-Leibler divergence [10], is a distortion measure for measuring (dis)similarity between two given probability density functions, f and g. It is defined as:

$$d(f\|g) = \int f(x) log \frac{f(x)}{g(x)}\, dx \tag{2}$$

In classification and discrimination problems, a symmetrized version of two asymmetric KLDs, also known as the Jeffrey divergence, is widely used.

$$J(f,\, g) = d(f\|g) + d(g\|f) \tag{3}$$

There is no closed-form analytical expression for the KLD between two multivariate Gaussian mixture models. It can be approximated by Monte-Carlo simulation techniques as:

$$d(f\|g) \approx \frac{1}{N}\sum_{n=1}^{N} log \frac{f(x_n)}{g(x_n)} \tag{4}$$

where $x_1, \cdots, x_N$ are randomly sampled from $f(x)$. However the computation burden is somewhat heavy. An efficient, unscented transform [11] based approximation, which can significantly decrease the computation complexity, was presented in [12]. For two D-dimensional Gaussian mixture distributions:

$$f = \sum_{m=1}^{M_f} c_{f,m} N(\mu_{f,m}, \Sigma_{f,m})$$
$$g = \sum_{m=1}^{M_g} c_{g,m} N(\mu_{g,m}, \Sigma_{g,m})$$

by using only a small number of points, which are deterministicly sampled from $f(x)$, the KLD between them can be approximated as:

$$d(f\|g) \approx \frac{1}{2D}\sum_{m=1}^{M_f} c_{f,m} \sum_{d=1}^{2D} log\, g(x_{m,d}) \tag{5}$$

where:

$$x_{m,d} = \mu_{f,m} + (\sqrt{D\Sigma_{f,m}})_d \quad d = 1, \cdots, D$$
$$x_{m,d+D} = \mu_{f,m} - (\sqrt{D\Sigma_{f,m}})_d \quad d = 1, \cdots, D$$

and $(\sqrt{\Sigma})_d$ refers to the d-th column of the matrix square root of Σ.

4. STATE DIVERGENCE-BASED DETERMINATION OF THE NUMBER OF GAUSSIAN COMPONENTS

Given a HMM Λ with N_S states $(S_1, \cdots, S_{N_S})$, we define the between-state model divergence as:

$$D(\Lambda) = \frac{1}{2}\sum_{i=1}^{N_S} D(S_i) \tag{6}$$

where

$$D(S_i) = \sum_{j=1,\, j\neq i}^{N_S} J(S_i,\, S_j) \tag{7}$$

is the total divergence between state S_i and all other states. Our divergence-based determination is thus to obtain a model with a maximum between-state model divergence, given the total number of parameters.

However, a HMM usually has a large number of states (i.e., with a large N_S). The computation of equation (7) for all the states is somewhat prohibitive. We therefore resort to

a computationally tractable, approximate solution by calculating the divergence between the state and its corresponding top N_c ($N_c \ll N_S$) competitive states. Then equation (7) can be rewritten as:

$$D(S_i) \approx \sum_{j=1}^{N_c} J(S_i,\, S_j) \tag{8}$$

Therefore, how to get the top N_c competitive states for each state and how to efficiently allocate Gaussian components are the two key points of our state divergence-based determination. In the following we'll describe them respectively.

4.1. Competitive States

In our study, a data-driven method was used to obtain the corresponding top N_c competitive states of each state by the following procedure:

1. Use a well-trained model to align the training data (or a randomly selected subset of the full training set) to the frame-level with the correct transcription;
2. For each state, find its top N_f competitive states for each frame aligned to it, and increase these competitive states' appearance number counters with the corresponding factors weighted by relative likelihood;
3. Sort the competitive state list for each state according to the number of appearance in a descending order;
4. For each state, the N_c states in the front of the sorted competitive state list is selected as the top N_c competitive states of it.

4.2. Divergence-based Determination

Our goal is to maximize the between-state model divergence with a given total number of Gaussians. Thus, adding one extra Gaussian component in state S_i, the resulted increment of the between-state model divergence

$$\begin{aligned} \Delta D(\Lambda) &= \Delta D(S_i) \\ &= D(S_i(m+1)) - D(S_i(m)) \end{aligned} \tag{9}$$

(where m is the number of Gaussian components of state S_i) is adopted as the indicative measure to perform the divergence-based Gaussian component allocation method.

The divergence-based determination method searches for the state to allocate successively one extra Gaussian component. The state for allocating the extra Gaussian component is chosen, based upon the maximum increment of the between-state divergence. It is a greedy search algorithm.

The method starts with allocating one Gaussian component in each state and the corresponding between-state divergence is computed. Then one extra Gaussian component is tentatively assigned to each state in turn and a corresponding increment of the divergence is computed. The state that yields the maximum increment of the between-state model divergence is assigned with one more Gaussian component and the procedure then repeats itself. The algorithm stops when the total number of Gaussian components reaches a preassigned limit.

5. EXPERIMENTAL RESULTS

TiDigits, a speaker independent, connected digit utterances database, was used to test our method. The speech signal was recorded from various regions of the United States. It contains 12,549 strings for training and 12,547 strings for testing. The digits string has a random length from 1 to 7. Each digit was modelled by a 10-state, whole-word based HMM. And a 3-state silence model and a 1-state short pause model were added. The features were the conventional 39-dimensional MFCCs (12 static MFCCs, log energy, and their first- and second-order time derivatives).

Fig. 2 gives the recognition performance (in Word Error Rate (WER)) curves against the average number of Gaussian components per state for fixed, uniform determination (same number of Gaussians per state), the BIC-based and the divergence-based determinations. It shows that, with comparable total number of Gaussians, both the BIC-based and our proposed determinations give much better recognition performance than the fixed, uniform determination. We also found that using our divergence-based determination, the resulted model with on average 12 components per state has almost the same performance as the model resulted by the BIC-based determination with on average about 19 Gaussians per state. That is to say, compared with the BIC-based determination, our proposed method successfully reduced the total number of Gaussians by about 37% with almost the same recognition performance.

6. CONCLUSIONS

In this paper, we propose a new method to determine the number of Gaussian components of each state in HMM-based speech recognition system. Different from the fixed determination or the Bayesian Information Criterion-based determination, we use the between-state model divergence as the indicative measure to determine the number of Gaussians. The Kullback-Leibler divergence is used with its unscented transform based approximation to measure the between-state divergence. The algorithm is performed successively, one Gaussian at a time, by searching over all the states. Significant improvement of the recognition performance was found in our experiments. A smaller size model can be obtained with almost the same recognition performance.

For the time reason, we only tested our approach on the TiDigits database. Its efficiency has been proved on this sim-

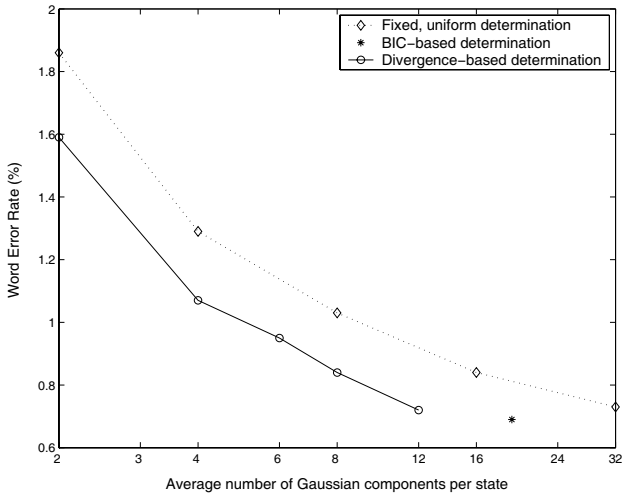

Fig. 2. Recognition performance with different determination methods vs. average number of Gaussian components per state

ple system. In the near future, we'll test our method on large vocabulary continuous speech recognition tasks to see whether its efficiency is consistently maintained on more complex systems.

7. ACKNOWLEDGEMENT

The authors are grateful to Dr. Frank K. Soong at Microsoft Research Asia, Beijing, China, for his valuable comments and suggestions.

8. REFERENCES

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 1-38, 1977.

[2] Y. Gao, E.E. Jan, M. Padmanabhan, and M. Picheny, "HMM Training Based on Quality Measurement", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 129-132, 1999.

[3] G. Schwarz, "Estimating the Dimension of A Model", *Annals of Statistics*, Vol. 6, pp. 461-464, 1978.

[4] J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length", *Annals of Statistics*, Vol. 11, pp. 416-431, 1983.

[5] S.S. Chen, and R.A. Gopinath, "Model Selection in Acoustic Modeling", *Proc. European Conference on Speech Communication and Technology*, Vol. 3, pp. 1087-1090, 1999.

[6] Y. Jia, Y. Yan, and B. Yuan, "Dynamic Threshold Setting via Bayesian Information Criterion (BIC) in HMM training", *Proc. International Conference on Spoken Language Processing*, PAd(14, 15)-M3-10, 2000.

[7] H. Tenmoto, M. Kudo, and M. Shimbo, "Determination of the Number of Components Based on Class Separability in Mixture-based Classifiers", *Proc. International Conference on Knowledge-Based Intelligent Information Engineering Systems*, pp. 439-442, 1999.

[8] S. Medasani, and R. Krishnapuram, "Determination of the Number of Components in Gaussian Mixtures Using Agglomerative Clustering", *Proc. IEEE International Conference on Neural Networks*, Vol. 3, pp. 1412-1417, 1997.

[9] J.P. Hoffbeck, and D. Landgrebe, "A Method for Estimating the Number of Components in Normal Mixture Density Function", *Proc. IEEE International Symposium on Geoscience and Remote Sensing*, Vol. 4, pp. 1675-1677, 2000.

[10] S. Kullback, *Information Theory and Statistics*, Dover Publications, 1997.

[11] S. Julier, and J.K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions", *Technical Report*, RRG, Department of Engineering Science, University of Oxford, 1996.

[12] J. Goldberger, S. Gordon, and H. Greenspan, "An Efficient Image Similarity Measure based on Approximations of KL-Divergence Between Two Gaussian Mixtures", *Proc. IEEE International Conference on Computer Vision*, Vo. 1, pp. 487-493, 2003.