Minimum Error Rate Training for Designing Tree-Structured Probability Density Function

Wu Chou

Bell Laboratories
Rm. 2D-526, 600 Mountain Avenue
Murray Hill, New Jersey, NJ 07974, U.S.A.
Email: wuchou@research.bell-labs.com

ABSTRACT

In this paper, we propose a signal prototype classification and evaluation framework in acoustic modeling. Based on this framework, a new tree-structured likelihood function is derived. It uses a designated cluster kernel f_m^C for signal prototype classification and a designated cluster kernel f_m^L for likelihood evaluation of outlier or tail events of the cluster. A minimum classification error (MCE) rate training approach is described for designing tree-structured likelihood function. Experimental results indicate that the new tree-structured likelihood function significantly improves the acoustic resolution of the model. It has a more significant speedup in decoding than the one obtained from the conventional approach.

1. INTRODUCTION

In general, a speech recognition system based on continuous density mixture Gaussian *pdf*s often demonstrates a better speech recognition performance than its discrete counter parts. However, one of the problems associated with using mixture Gaussian *pdf*s in speech recognition is its computational cost. Evaluating mixture Gaussian *pdf* in speech recognition often consumes significant amount of CPU time. This problem becomes even more acute when the HMM model set contains a large number of different Gaussian *pdf*s, as typical in detailed context dependent acoustic modeling.

In this paper, we describe a general signal prototype classification and likelihood evaluation paradigm in acoustic modeling. In this approach, the signal space formed by observation vectors is partitioned into subspaces, and in each subspace, the signal becomes more homogeneous with less impurities[6]. Therefore, it can be modeled by a suitable mixture density function often with very few *pdf* kernels. This approach integrates signal prototype classification with mixture density modeling, providing a new dimension in modeling various acoustic units in speech recognition. Based on this framework, a new type of tree-structured likelihood function is derived and a minimum classification error rate approach is described for designing its cluster kernels.

For a mixture Gaussian density, the likelihood of observing O at a given state is

$$l(O) = \sum_{m=1}^{M} \epsilon_m G_m(O, \mu_m, \sigma_m^2) \qquad (1)$$

where G_m is multi-dimensional Gaussian *pdf* kernel, μ_m and σ_m^2 are mean and variance vectors. In order to reduce the cost of evaluating mixture Gaussian *pdf*s in speech recognition, several schemes were proposed in the past with different level of success[1][2]. From the point of view of the signal prototype classification and likelihood evaluation framework, these schemes introduce certain signal prototype classification procedures for a given set of Gaussian *pdf* kernels in the mixture. For an input observation vector, Gaussian *pdf*s in the mixture density are classified to be either active or non-active. For Gaussian *pdf*s in the non-active class, the current observation vector prototype corresponds to an outlier event. The likelihood value of the observation vector from these Gaussian *pdf*s will be small and insignificant. Therefore, using a table look-up or a grand Gaussian kernel to approximate its value becomes reasonable. In decoding, only Gaussian *pdf*s in the active class are fully evaluated. This results in a significant savings in computational cost.

One approach of signal prototype classification is based on VQ clustering of Gaussian *pdf*s in the model set. In this approach, the input observation vector is first classified to its nearest codeword using a weighted Mahalanobis type distance[2]. Each codeword in the codebook is related to a subset of Gaussian *pdf*s, and only the Gaussian *pdf*s which belong to the nearest codeword of the observation vector are marked active.

This approach is simple and easy to implement. However, using a Mahalanobis type distance to decide an outlier event for Gaussian *pdf*s is, in many cases, not very accurate. Tree-structured likelihood function is another approach[1]. It uses a decision tree based method for signal prototype classification. In our experiments, it often offers better speed and performance trade-offs than VQ based approach.

2. TREE-STRUCTURED LIKELIHOOD FUNCTION

In tree-structured likelihood function, each Gaussian kernel is organized as a node in a decision tree. The Gaussian *pdf*s which occur in the original set of HMMs, $\{N1[.], \cdots, N_k[.], \cdots N_K[.]\}$, form the base root nodes of the tree. The Gaussian at the higher level tree node is called a cluster *pdf*, because it corresponds to the set of its next level child node Gaussian *pdf*s. The cluster *pdf* is usually obtained by approximating the mixture of all child *pdf*s in the cluster using a single Gaussian *pdf*. Fig 1. illustrates the structure of the tree structured likelihood function.

In recognition, a top down tree classification procedure is performed. The top level tree nodes are calculated first to determine active nodes at the top level according to certain selection criterion. One typical selection criterion is based on top N rule. Tree nodes are marked active if their cluster *pdf* likelihood scores are among the top N scores at that level. This process is propagated down to every level of the tree. At each level, nodes with active parents are calculated, and among them, active nodes are determined. In decoding, the base layer Gaussian *pdf* is evaluated only if its immediate parent is active. Otherwise, its value is approximated by its closest active cluster (ancestor) *pdf*. Typically, the number of tree levels is less than three, and the number of nodes at each level follows roughly a square root law.

One approach of designing tree-structured *pdf* is based on a divergence distance measure using a standard K-means procedure[1]. The divergence distance measure between two *pdf*s f_k, f_m in this approach is

$$d(f_k, f_m) = D(f_k \mid\mid f_m) + D(f_m \mid\mid f_k), \qquad (2)$$

where $D(p \mid\mid q)$ is the Kulllback-Leibler (K-L) distance or relative entropy between two distributions[5]. For Gaussian *pdf*s with diagonal covariance matrix, the right hand side of (2) becomes

$$\sum_i \left(\frac{\sigma_k^2(i) + \Delta_{km}^2(i)}{2\sigma_m^2(i)} + \frac{\sigma_m^2(i) + \Delta_{km}^2(i)}{2\sigma_k^2(i)} - 1\right), \qquad (3)$$

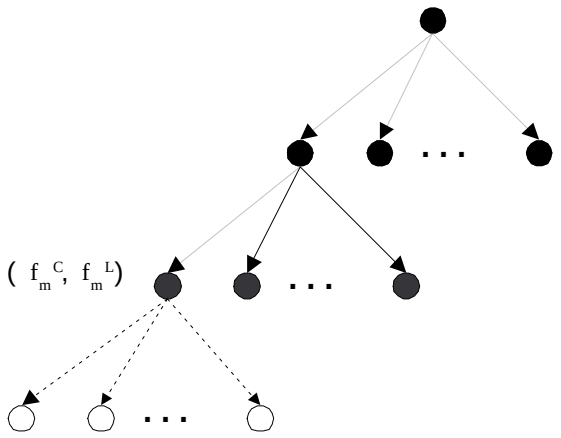

Figure 1: A layered decoding network architecture.

where $\Delta_{km}^2(i) = (\mu_k(i) - \mu_m(i))^2$, $\mu_k(i)$, $\mu_m(k)$ are i-th components of the mean vectors of f_k, f_m respectively, and $\sigma_k^2(i)$, $\sigma_m^2(k)$ are i-th diagonal components of the covariance matrices of f_k, f_m.

The clustering algorithm in this design is done by approximating the Gaussian *pdf*s in the cluster by a single cluster Gaussian *pdf*, assuming that child Gaussians in the cluster form one grand mixture distribution. The mean and variance of the cluster *pdf* f_m are estimated as follows:

$$\mu_m(i) = \frac{1}{K}\sum_{k=1}^{K}\int x_k(i) f_k(i) dx = \frac{1}{K}\sum_{k=1}^{K}\mu_k(i); \qquad (4)$$

$$\begin{aligned}\sigma_m^2 &= \frac{1}{K}\sum_{k=1}^{K}\int (x_k(i) - \mu_m(i))^2 f_k(i) dx \\ &= \frac{1}{K}\left[\sum_{k=1}^{K}\int x_k^2(i) f_k(i) dx - K\mu_m^2(i)\right] \\ &= \frac{1}{K}\left[\sum_{k=1}^{K}\sigma_k^2(i) + \sum_{k=1}^{K}\mu_k^2(i) - K\mu_m^2(i)\right] \qquad (5)\end{aligned}$$

The K-means estimation procedure of tree-structured likelihood function consists the following steps:

1. Initialization of cluster *pdf*s.
2. Determine cluster membership using the distance function of (2).
3. Re-estimate the cluster *pdf*s using (4) and (5).

Iterate steps 2 and 3 until the improvements in global distortion between each iteration becomes less than a specified threshold δ.

3. A CLASSIFICATION AND EVALUATION APPROACH

One of the important issues in tree-structured likelihood function is how to improve its acoustic resolution while reducing the number of Gaussian *pdf*s to evaluate. This is critical, because the loss of acoustic resolution due to likelihood function approximations can lead to a significant increase in the number of active nodes in the search beam. Additional savings in likelihood computation can be offset by a much heavier cost on the search and recognition performance. In particular, when high resolution acoustic model is used, the percentage of Gaussian *pdf*s actually evaluated in the search beam can be already well below 10%. This imposes a very stringent constraint on the quality of the tree-structured likelihood function.

In tree-structured likelihood function, the original approach is based on one single cluster kernel to do both signal prototype classification and likelihood evaluation. This can create problem and is a cause of losing acoustic resolution in tree-structured likelihood function. One way to see this is from the signal prototype classification and evaluation framework. In tree-structured *pdf* function, classification kernel is to represent the inliers of the cluster. However, the likelihood value of the cluster kernel is used in likelihood evaluation only when the observation vector is an outlier to the cluster. This is because if the observation vector is an inlier to the cluster, it will activate the child nodes in the cluster and be evaluated by more accurate lower layer kernels. Therefore, likelihood evaluation at cluster level is for outlier or tail events of the cluster, whereas the classification kernel is to determine if the observation vector is an inlier to the cluster.

In order to solve this problem, we propose a new tree-structured likelihood function, in which a designated cluster kernel f_m^C is used for signal prototype classification and a separate designated cluster kernel f_m^L is used for likelihood evaluation. The classification kernel f_m^C is for inlier signal prototype detection and the likelihood evaluation kernel f_m^L is for evaluating outliers or tail events of the cluster. It is not difficult to visualize that f_m^C and f_m^L can have very different characteristics. The cost of evaluating f_m^L in decoding is almost negligible. This is because f_m^L is for outlier events and does not need to be evaluated in advance. It can be evaluated on demand and be cached. Moreover, f_m^L is evaluated only if a *pdf* in the mixture distribution of an active node is not active. The model enhancement due to f_m^L often leads to a significant reduction of the number of active nodes in the search beam and a much better recognition performance.

4. MCE TRAINING OF TREE-STRUCTURED *PDF*

Designing likelihood evaluation kernel f_m^L for outliers or tail events is in general a very difficult problem. In this paper, we propose an approach in which cluster kernel f_m^L estimation is embedded into the minimum classification error (MCE) rate training[4]. In particular, the string error rate based MCE approach can be applied to designing f_m^L in tree-structured *pdf*. The loss function in this approach is constructed through the following four steps.

(1) Discriminant function in minimum string error rate training is defined as

$$g(O, S_k, \Lambda) = \log f(O, \Theta_{\bar{M}_{S_k}}, S_k \mid \Lambda), \tag{6}$$

where S_k is the k-th best string, $\Theta_{\bar{M}_{S_k}}$ is the optimal path (state sequence) of the k-th string given the model set Λ, and $\log f(O, \Theta_{\bar{M}_{S_k}}, S_k \mid \Lambda)$ is the related log-likelihood score on the optimal path of the k-th string.

For the correct string S_{lex}, the discriminant function is given by

$$g(O, S_{lex}, \Lambda) = \log f(O, \Theta_{\bar{M}_{S_{lex}}}, S_{lex} \mid \Lambda), \tag{7}$$

where S_{lex} is the correct string, $\Theta_{\bar{M}_{S_{lex}}}$ is the optimal alignment path and $\log f(O, \Theta_{\bar{M}_{S_{lex}}}, S_{lex} \mid \Lambda)$ is the corresponding log-likelihood score.

(2) Misclassification measure in minimum string error rate training is defined as

$$d(O, \Lambda) = -g(O, S_{lex}, \Lambda) + \log\{\frac{1}{N-1} \sum_{S_k \neq S_{lex}} e^{g(O, S_k, \Lambda)\eta}\}^{\frac{1}{\eta}}. \tag{8}$$

(3) Loss function in minimum string error rate training is defined as

$$l(O, \Lambda) = \frac{1}{1 + e^{-\gamma d(O, \Lambda)}}, \tag{9}$$

where γ is a positive constant, which controls the slope of the sigmoid function.

(4) The expected loss which is associated with the string error rate is given by

$$L(\Lambda) = E_O[l(O, \Lambda)]. \tag{10}$$

The likelihood evaluation kernel f_m^L occurs explicitly in the string likelihood function $\log f(O, \Theta_{\bar{M}_S}, S \mid \Lambda)$. However, in order to reconstruct the local likelihood function in training, the activation history of the classification cluster kernel f_m^C needs to be saved.

5. EXPERIMENTAL RESULTS

The proposed approach was applied in several applications. In order to illustrate the issue, we focus on one application of recognizing long digit strings from a relatively noisy environment. The acoustic model used in this study was a high resolution, inter-word context dependent model set with 274 context dependent acoustic model units [3]. This model is of very high acoustic resolution. The inter-word context dependency is explicitly modeled and each digit model was represented as a context dependent graph with 12 fan-in heads, one body and 12 fan-out tails.

The speech used in the test were utterances of 16 digit strings collected from the telephone network with an average duration of 11 seconds/utterance. Table 2. illustrates the performance comparison for various approaches based on 100 test sentences. In VQ-assisted approach, an 128 entry codebook with $\Theta = 1.5$ was used[2]. The tree_baseline was a two level tree-structured *pdf* obtained from the original divergence distortion measure approach described in *Section 2*. It has 10 first level cluster nodes and 80 second level cluster nodes. The root layer has 6700 Gaussian *pdf*s which occurs in the original model. A top (5, 10) rule is used in recognition. The active nodes at the first level of the tree were among the top 5 scoring nodes and the active nodes at the second level were among the top 10 scoring nodes whose parents were active. Tree_new_gpd was based on the new tree-structured *pdf* derived in *Section 3* and trained using MCE approach. It had the same number of tree nodes as the tree_baseline, but there were two cluster *pdf*s (f_m^C, f_m^L) on each tree node as illustrated in Fig 1. Same top (5, 10) rule was used in recognition. The VQ based approach suffered a significant loss of recognition performance and the decoding speed was also heavily penalized. The tree_baseline gave a reasonable performance. The new tree-structured *pdf* showed a more significant speedup and a much better recognition performance. The use of the designated likelihood evaluation kernel has a positive impact on both speed and performance.

	WdErr	Run Time	Speed factor
Baseline	1.3%	1305s	1
VQ assisted	3.4%	1153.7s	1.13
Tree_baseline	1.6%	1066s	1.22
Tree_new_gpd	1.4%	872.2s	1.50

Table 1: Comparisons between different approaches

One thing needs to mention is that the baseline HMMs was of very high resolution. With beam search, only 9.2% Gaussian *pdf*s were actually evaluated in decoding. The percentage of Gaussian *pdf*s evaluated in Tree_baseline was about 4.5%, but the average number of active nodes in the search beam was increased by a significant 33%. In Tree_new_gpd, the percentage of Gaussian *pdf*s evaluated in decoding was about 4%, but the number of active nodes in the search beam was much reduced. This is a strong indication of a much improved acoustic resolution in the new tree-structured likelihood function, and the speedup gain is more than doubled from 22% to 50% which is on top of an already very effective beam search strategy.

6. SUMMARY

In this paper, we propose a signal prototype classification and evaluation framework in acoustic modeling. Based on this framework, a new tree-structured likelihood function is derived. It uses a designated cluster kernel f_m^C for signal prototype classification and a designated cluster kernel f_m^L for likelihood evaluation of outlier or tail events of the cluster. A minimum classification error (MCE) rate training approach is described for designing tree-structured likelihood function. Experimental results indicate that the new tree-structured likelihood function significantly improves the acoustic resolution of the model. It has much smaller performance degradation and a more significant speedup in decoding than the one obtained from the conventional approach.

Acknowledgment

The author would like to thank Dr. S. Moon, Dr. E. Burhke, Dr. C.-H. Lee and Dr. B.-H. Juang for their help and discussion.

REFERENCES

[1] T. Watanabe, K. Shinoda, K. Takagi and E. Yamada "Speech Recognition Using Tree-Structured Probability Density Function",*Proc. ICSLP-94*

[2] E. Bocchieri "Vector Quantization for the Efficient Computation of the Continuous Density Function", *Proc. ICASSP-93, pp 692–695 (1993)*, pp 380–385.

[3] C.-H. Lee, W. Chou, B.-H. Juang, "Context dependent acoustic modeling for connected digit recognition", 1993 ASA Fall meeting, Denver, Oct 93.

[4] W. Chou, C.-H. Lee and B.-H. Juang, "Minimum error rate training of hidden Markov models based on the N-best string models", *Proc. ICASSP'93*, Vol. 1, pp. 652–655

[5] T. Cover and J. Thomas "Information Theory", *John Wiely & Sons Inc, 1991.*

[6] L. Breiman, J. Friedman, R. Olshen and C. Stone "Classification and Regression Trees" *Wadsorth & Brooks/Cole Advanced Books & Software, 1984*