CLUSTERING OF BOOTSTRAPPED ACOUSTIC MODEL WITH FULL COVARIANCE

Xin Chen[1], Xiaodong Cui[2], Jian Xue[2], Peder Olsen[2], John Hershey[3], Bowen Zhou[2] and Yunxin Zhao[1]*

Department of Computer Science, University of Missouri, Columbia, MO, 65211 USA[1]
IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598[2]
Mitsubishi Electric Research Laboratories, Cambridge, MA, 02139, USA[3]
Emails: {xck82,zhaoy}@mail.missouri.edu[1], {cuix, jxue, pederao, zhou}@us.ibm.com[2], hershey@merl.com[3]

ABSTRACT

HMM-based acoustic models built from bootstrap are generally very large, especially when full covariance matrices are used for Gaussians. Therefore, clustering is needed to compact the acoustic model to a reasonable size for practical applications. This paper discusses and investigates multiple distance measurements and algorithms for the clustering. The distance measurements include Entropy, KL, Bhattacharyya, Chernoff and their weighted versions. For clustering algorithms, besides conventional greedy bottom-up, algorithms such as N-Best distance Refinement (NBR), K-step Look-Ahead (KLA), Breadth-First Searched (BFS) best path are proposed. A two-pass optimization approach is also proposed to improve the model structure. Experiments in the Bootstrap and Restructuring (BSRS) framework on Pashto show that the discussed clustering approach can lead to better quality of the restructured model. It also shows that final acoustic model that is diagonalized from the full covariance yields good improvement over BSRS model directly with diagonal model and yields significant improvement over the conventional diagonal model.

Index Terms— Acoustic modeling, Clustering, Full covariance, K-step lookahead, Global search optimization

1. INTRODUCTION

Acoustic modeling of under-resourced languages such as Farsi, Dari and Pashto suffers from limited training data which are both difficult and expensive to collect. To overcome the sparsity problem in training data, bootstrap modeling was proposed to achieve robust acoustic modeling [1][2][3]. With significant improvements in word accuracy performance, the aggregated bootstrapped acoustic models have a large number of Gaussian components in the Gaussian Mixture Model (GMM) in each state with parametric redundancy. Computing acoustic likelihood scores for such a large model is expensive, especially when full covariance modeling is used for GMM.

There are many existing works for compacting the acoustic model to a desirable size and eliminating the redundancy without significantly degrading model quality. In [4], a weighted distance has been investigated. In [5], Bayesian Information Criterion (BIC) was used to determine the optimal GMM structure for clustering. In [6], compacting models by clustering the Gaussian components in the random forest based acoustic model was investigated. Most of these proposed methods are based on greedy agglomerative clustering scheme.

In this paper, we investigate the problem of clustering bootstrapped acoustic models. Several distance measures are evaluated, including Entropy change, KL divergence, Bhattacharyya distance, and Chernoff distance. We also consider their weighted forms as suggested in [4]. Several clustering algorithms are proposed to improve the clustering quality, including an Entropy based N-Best distance Refinement method (NBR) to improve the speed for distance computation as well as to implicitly impose mixture component weights to improve clustering quality. To overcome the deficiency of the greedy approach we also propose to use global optimization techniques such as K-step look ahead (KLA), Breadth First Search (BFS) in models clustering. In addition, a two-pass clustering algorithm is proposed to improve the structure of GMMs. We evaluate the proposed distance measurements and clustering algorithms in acoustic modeling with full covariance Gaussians. The experiments on the speech recognition task have shown that the compact acoustic model obtained using our proposed methods improved the recognition accuracy over the conventional methods. As an important step in Bootstrap and Restructuring (BSRS) training [2], we also evaluate the proposed methods on a BSRS full covariance to diagonal scheme. This full covariance BSRS shows good improvements over its diagonal covariance BSRS counterpart.

The reminder of the paper is organized as follows. Section 2 describes distance measures for clustering. Section 3 discusses the proposed algorithms for clustering. Experimental results are presented in section 4. A summary and possible future extension are given in section 5.

2. DISTANCE MEASURES FOR GAUSSIAN CLUSTERING

In this section, we discuss several distances to measure the dissimilarity between each pair of Gaussian components. These distances include KL divergence, entropy change, Bhattacharyya distance and Bayes error.

2.1. KL Divergence

The KL divergence is commonly applied to measure the dissimilarity between two distributions [8]. Given the two distributions, $f_1(x)$ and $f_2(x)$, the KL divergence is defined as

$$D_{\text{kl}}(f_1 \parallel f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx \quad (1)$$

Since the definition in Eq.1 is asymmetric, a symmetric version defined in Eq.2 is used.

$$D_{\text{kls}}(f_1 \parallel f_2) = \int \left[f_1(x) \log \frac{f_1(x)}{f_2(x)} + f_2(x) \log \frac{f_2(x)}{f_1(x)} \right] dx \quad (2)$$

*This work was conducted when the first author was doing his summer intern at IBM T.J. Watson Research Center.

2.2. Entropy change

The entropy criterion (ENT) measures the change of entropy when two distributions are merged. If the two distributions are Gaussians, then the change of entropy is computed as

$$D_{\text{ent}}(f_1 \parallel f_2) = (w_1 + w_2)\log|\Sigma| - w_1 \log|\Sigma_1| - w_2 \log|\Sigma_2| \quad (3)$$

2.3. Bayes Error, Bhattacharyya and Chernoff Distances

The Bayes error measures the overlap of two distributions. Given the two distributions, $f_1(x)$ and $f_2(x)$, the Bayes error is defined as

$$D_{\text{bayes}}(f_1 \parallel f_2) = \int \min(f_1(x), f_2(x))dx \quad (4)$$

There is no closed-form expression for the Bayes error even when $f_1(x)$ and $f_2(x)$ are multivariate Gaussians. However, it can be shown to be bounded from above by the Chernoff function:

$$C_s(f_1 \parallel f_2) = C(s) = \int f_1(x)^s f_2(x)^{1-s} dx, \quad 0 \le s \le 1 \quad (5)$$

The Chernoff function in Eq.8 is a general function describing the distance between two distributions and it has close relationships with some other well-known distribution distances.

Bhattacharyya (Bha) distance is defined as

$$D_{\text{bha}}(f_1 \parallel f_2) = \int \sqrt{f_1(x) f_2(x)}dx \quad (6)$$

which is a special case of the Chernoff function with $s = 0.5$. It has an analytic form when f_1 and f_2 are Gaussian distributions

$$-\log D_{\text{bha}}(f_1 \parallel f_2) = \frac{1}{8}(\mu_1 - \mu_2)^T \left[\frac{1}{2}(\Sigma_1^{-1} + \Sigma_2^{-1})\right](\mu_1 - \mu_2) + \frac{1}{2}\log\left(\frac{\frac{1}{2}|\Sigma_1 + \Sigma_2|}{|\Sigma_1|^{(1/2)} + |\Sigma_2|^{(1/2)}}\right) \quad (7)$$

Based on the Chernoff function, the Chernoff distance is defined to be the minimum of Eq.8 with respect to s:

$$D_{\text{chern}}(f_1 \parallel f_2) = \min_{0 \le s \le 1} \int f_1(x)^s f_2(x)^{1-s} dx \quad (8)$$

When both $f_1(x)$ and $f_2(x)$ are Gaussians, which is the case in this work, the Chernoff distance can be computed via Newton-Raphson algorithm

$$s_{k+1} = s_k - \frac{c'(s)}{c''(s)} \quad (9)$$

where $c(s) = \log C(s)$ which is a convex function of s. Therefore convergence is guaranteed. A reasonable starting point is $s = 0.5$ which amounts to the Bhattacharyya distance. The details of formula derivation to obtain the Chernoff distance are elaborated in [2]. Note that when both $f_1(x)$ and $f_2(x)$ are Gaussians, which is the case in this work, the Chernoff distance demonstrated in Eq.5 has an analytic solution

$$-\log D_{\text{chern}}(f_1 \parallel f_2) = \frac{1}{2}(s(1-s))(\mu_1 - \mu_2)^T \left[\frac{1}{2}((1-s)\Sigma_1^{-1} + s\Sigma_2^{-1})\right](\mu_1 - \mu_2) + \frac{1}{2}\log\left(\frac{|(1-s)\Sigma_1 + s\Sigma_2|}{|\Sigma_1|^{(1-s)} + |\Sigma_2|^{(s)}}\right)$$

Therefore it is also possible to use a line search with adaptive step sizes to compute the Chernoff distance. In our pilot experiment this method is two times slower than the Newton-Raphson algorithm method with similar accuracy level. However, the line search does not need to compute the derivatives of $c(s)$ in the Newton-Raphson algorithm.

2.4. Weighted Distances

It is noted that the KL, Bhattacharyya and Chernoff distances defined above do not include the weights of Gaussians. It is observed in [4] that weighted distances gave better performance than non-weighted distances. Therefore, it is of interest to evaluate weighted distance in clustering Gaussian components. Take Bhattacharyya distance as an example. Given the two distributions, $f_1(x)$ and $f_2(x)$, suppose the weight for $f_1(x)$ is w_1 and the weight for $f_2(x)$ is w_2, the weighted Bhattacharyya is defined as

$$D_{\text{bhat}}(f_1 \parallel f_2) = \int \sqrt{w_1 f_1(x) w_2 f_2(x)}dx \quad (10)$$

3. GAUSSIAN COMPONENT CLUSTERING ALGORITHMS

In this section, we discuss the proposed clustering algorithms. Suppose one state in the bootstrapped model has a total of M Gaussians. Our task is to cluster the M Gaussian components into N Gaussian components. The most widely used algorithm is to perform agglomerative clustering for each GMM state, i.e. calculate the distances between all of the Gaussian mixture component pairs and merge the closest one, so that the change between the GMMs before and after the merge are minimized. Repeat this process for M-N times to construct a Gaussian component binary tree in the bottom-up direction. This process can be illustrated in the following formula.

$$D(f, g) = \sum_{i=1}^{M-N} Distance(f_a^i, f_b^i) \quad (11)$$

In the greedy algorithm, for the step i, we find the minimum distance pairs f_a^i and f_b^i. However, our optimization target is to find the minimum distance change $D(f, g)$, which is also called the global optimization. In this case, for each step i, we are trying to look for pairs f_a^i and f_b^i that will lead to the minimum of $D(f, g)$.

3.1. Entropy based N-Best Distance Refinement (NBR)

Both Chernoff distance and KL distance are computationally heavy, especially for full covariance Gaussians in a bootstrapped large GMM. It is noted that the computation of Entropy distance is fast (20 times faster than Chernoff distance computation), and the clustering quality is also very good as evaluated in pilot experiments. The proposed NBR idea is using fast distance criterion like entropy to first select N-best combining candidates, and use the slow but better distance criteria such as Chernoff or KL to refine the distance for the selected N candidate pairs. Since mixture weights are explicitly used in the entropy criterion, a potential advantage with the NBR approach is that the weights are implicitly used for non-weighted distances such as KL, Bhattacharyya and Chernoff. In addition, NBR also provides more than one view of distance to potentially take advantage of complementary information from two distances.

3.2. K-step Lookahead (KLA)

It is well known that the greedy approach of combining candidates based on the smallest distance in the current clustering stage only leads to a local optimization. Our target is global optimization that minimizes $D(f, g)$ of total distance change between the Gaussian mixture models f and g. In [7], KLA is applied on a global optimize phonetic decision tree based triphone clustering. In this work, we hope global optimization can lead to better clustering than greedy local optimization, and KLA is therefore used to obtain a global optimized solution.

The example in Fig.1 illustrats that KLA no longer chooses the best combining candidate in the step1. A better solution will be obtained in the second step.

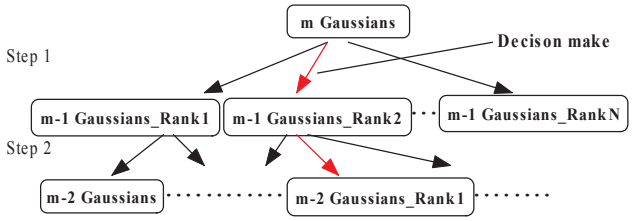

Fig.1 K-step Look Ahead for global optimization

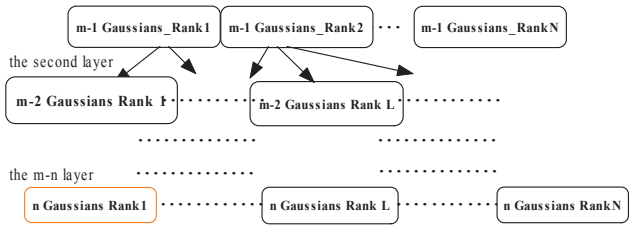

Fig.2 BFS for global optimization

3.3. Breadth-Fist Search (BFS) Global optimization

Searching the best combining path is another global optimization approach. It is computationally intensive. With a proper beam and pruning set up, the amount of computation can be reduced and a sub-optimal global solution can be obtained. Since acoustic model is trained off-line, and computational resource is becoming rich, it is worth a try to apply search methods in the task of Gaussian component clustering.

The clustering process using the Breadth First Search (BFS) algorithm is illustrated in Fig.2. At the first layer, we have one copy of the current mixture components. For each subsequent layer, we have the combining candidates from all copies of the previous layer. We extend from the previous layer with all possible candidates, and keep top-N best copies in the subsequent layer, where N is the beam width for this layer. Finally, we have N combining candidates at the last layer. We can therefore pick the best one among the N candidates.

Beam setup remains a question for this algorithm. A large beam will lead to slow speed, whereas a small beam will keep only limited candidates and the possible best path might be missed. Intuitively, at the beginning steps, we can keep the beam small, as the difference between the combining candidates is small. For the last few steps, the differences become bigger and we shall make the beam larger to avoid missing the potential best path. Based on this property, the beam width can be increased with the clustering process.

3.4. Two-Pass model structure optimization

For conventional one-pass clustering algorithm, each clustered state shall have $S_i^{new} = S_i \cdot \frac{N}{M}$ Gaussian mixtures, so that the total number of Gaussians for the acoustic model is N. However, fixing the compression rate to $\frac{N}{M}$ in each state is not the best option. Intuitively, some states may need more mixture components to represent a more complex distribution while some states just need less due to their simpler distribution. To deal with this issue, we proposed a two-pass approach to globally optimize the model structure.

While fixing the total number of Gaussians to N, each state can have different compression rate. In the first pass, we can use Bayesian Information Criterion (BIC)[5] or other similar criteria to decide the specific number of Gaussian mixture components for each state. Following the information in the first pass, we conduct clustering in the second pass. Let $S = s_1, s_2, ...s_k$ be the current k cluster GMM, suppose we combine s_1, s_2 to s_1'. then we have $S' = s_1', ...s_k$. The change from S to S' if measured with BIC [5] is

$$\begin{aligned} &- n \cdot \log|\Sigma| + n_1 \cdot \log|\Sigma_1| + n_2 \cdot \log|\Sigma_2| \\ &+ N \cdot (d + 0.5 \cdot d(d+1)) \end{aligned} \tag{12}$$

Where $n = n_1 + n_2$ is sample size of the merged node and N is defined as log data set that we are modeling. Note that the first part is similar to the entropy change, and the second part $N(d + 0.5 \cdot d(d+1))$ is a term that is not sensitive in this task and therefore entropy change can be used as a threshold measurement criterion.

In the first pass, we get a series of entropy values for the candidates from $(S_i \cdot \frac{N}{M}) - K, .., S_i \cdot \frac{N}{M}, .., (S_i \cdot \frac{N}{M}) + K$. In this way, we keep 2K candidate mixture-size and their corresponding entropy values. We therefore conduct a search to choose a threshold that makes the total number of Gaussians to N. In the second pass, we will cluster the Gaussians in each state to the number we obtained from the first pass. By keeping the same N Gaussian mixture components, the model structure is improved over the conventional one-pass approach.

4. EXPERIMENTS

Experiments were carried out on Pashto, one of the two major languages spoken in Afghanistan. The data was collected and transcribed by DARPA under the Transtac project. There were 135 hours of training data and 10 hours of test data. The feature space was constructed by splicing 9 frames of 24 dimensional PLP features and projecting down to a 40 dimensional space via LDA. Context-dependent quinphone states were tied by decision tree. A trigram language model with 1.2M n-grams was used for test, with a dictionary of 30K words. For acoustic modeling, full covariance Gaussian densities were used. The original bootstrapped acoustic model was constructed from 15 bootstrapped models with each subset covering 70% of original data. The bootstrapped model had 6K quinphone states with in total of 1.8M Gaussians. Two compression cases were investigated in this work: 100K Gaussians (1/18 of original model) and 50K Gaussians (1/36 of original model). The baseline method was using a certain distance criterion to find the minimum distance pair of Gaussian mixture components and combine the pair, repeat this process multiple times until we obtained the designed compact model.

4.1. Experimental Result on NBR

By normalizing the time cost for computing ENT as 1.0X, the time used for KL and Chernoff distance is presented in Table 1. It is noted that after using NBR the speed improvement is significant.

	KL	Chernoff
Baseline	6.5X	24.4X
NBR	1.2X	1.9X

Table 1. Investigation on speed improvement for NBR

Fig.3 shows that the NBR has improved the baseline in terms of WER too. It also outperformed the entropy distance produced compact model. Fig.4 evaluated on the Bhattacharyya distance shows that using weighted distance can improve the quality of clustering over non-weighted distance, especially when the compression rate is high. Implicitly imposing mixture weights from the NBR approach outperformed the weighted distance.

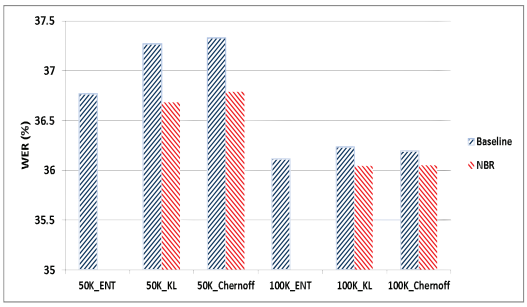

Fig.3 The effects of NBR method on WER

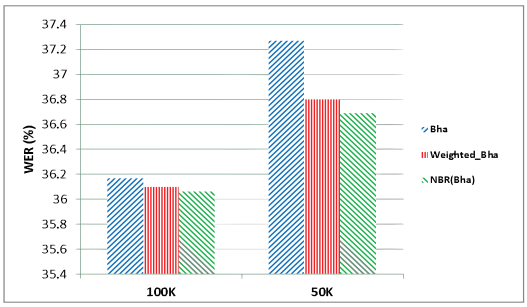

Fig.4 The effects of weighted Bha criteria and NBR method on WER

4.2. Experimental Results on Global Optimization

We used Two-Step Look Ahead (2S_LA). The beam was empirically set for BFS, the first 1/3 layers used beam 2, the middle 1/3 layers used beam 4, the last 1/3 layers used beam 8.

	Baseline	2S_LA	BFS
100K_KL	36.23	36.11	36.14
100K_ENT	36.11	36.08	36.08
50K_ENT	36.77	36.81	36.60

Table 2. WERs on proposed global clustering algorithms

As showed in Table 2, the global optimization solutions outperformed the local optimize except in the 50K_ENT task with the 2S_LA method. We also measured the global distance change in Eq.11 to evaluate the performance of the proposed methods. Entropy change criterion was used, the greedy approach on state 0 had an overall entropy change of 3336.80. The 2S_LA had a very small improvement of 0.03 comparing with the greedy approach. The proposed BFS approach had an overall entropy change of 3299.04. This was a relatively large improvement over 2S_LA approach as well as local optimization which showed the proposed algorithm was effective in finding the globally optimized clustering.

4.3. Experimental Results on Two-pass Model Structure Refinement

From Table 3, we can see that the case of 100K with two-pass produced improvement over the conventional one-pass 100K model, showing that the proposed methods were able to improve the model structure and had a positive effect on acoustic model quality. In the two-pass 100K experiment, The NBR method for the Chernoff distance generated model yielded 35.98% in WER, which was the best result obtained for the 1/18 compression rate performance.

100K test	Baseline	NBR
Two-Pass_KL	36.18	36.02
Two-Pass_ENT	36.04	–
Two-Pass_Che	–	35.98

Table 3. Experiment results on two-pass model structure refine

ML model	Conventional	BSRS_diag	BSRS_full2diag
WER(%)	39.6	38.5	38.1

Table 4. Experiment results on conventional, BSRS_diag and BSRS_full model

4.4. Experimental Results on Diagonal Covariance Model

the BSRS_full2diag represented that full covariance BSRS acoustic model was diagonalized and diagonal model was used for speech recognition. The detail of this framework was discussed in [2]. The proposed clustering algorithm performed on full covariance Gaussian components was one essential step in the training pipeline of BSRS_full2diag model, where all the models were trained with Maximum Likelihood (ML) with 100K Gaussian mixtures. The baseline model in WER was 39.6%, the BSRS model with diagonal training was 38.5%, while the BSRS model trained with full2diag had a WER of 38.1%, which was a 0.4% improvement comparing to BSRS_diag.

5. SUMMARY

In this work, we have investigated entropy, KL, Bhattacharyya, and Chernoff distance measures and proposed a fast distance computation method NBR, global optimization methods KLA and BFS, global structure optimization two-pass method of clustering, to improve the performance of compacting Gaussian mixture models. The evaluated results on full covariance Gaussian using the proposed methods have shown improvements over the conventional methods. Future extensions include combining several proposed methods in this work, and extend these methods to other similar tasks.

6. REFERENCES

[1] X. Cui, J. Xue, et. al.,"Acoustic modeling with bootstrap and restructuring for low resourced languages," Proc.Interspeech, pp.291-294, 2010.

[2] X. Cui, et. al., "Acoustic modeling with bootstrap and restructuring", IBM Technical Report.

[3] X. Chen and Y. Zhao,"Data sampling based ensemble acoustic modelling," Proc. ICASSP, pp.3805-3808, 2009

[4] Ogawa, A. and Takahashi, S. ,"Weighted distance measures for efficient reduction of Gaussian mixture components in HMM-based acoustic model," Proc. ICASSP, pp.4173-4176, 2008.

[5] Scott Chen, Gopalakrishnan, P.S.,"Clustering via the Bayesian information criterion with applications in speech recognition," Proc. ICASSP, pp.645-648, 1998.

[6] J. Xue and Y. Zhao,"Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," IEEE Trans. ASLP, vol.16, no. 3, pp. 519-528, 2008.

[7] J. Xue and Y. Zhao,"Novel lookahead decision tree state tying for acoustic modeling," Proc. ICASSP, pp.1133-1136, 2007.

[8] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models," Proc. ICASSP, 2007.