OPTIMAL CLUSTERING AND NON-UNIFORM ALLOCATION OF GAUSSIAN KERNELS IN SCALAR DIMENSION FOR HMM COMPRESSION

Xiao-Bing Li[1,2], *Frank K. Soong*[1], *Tor André Myrvoll*[1], *Ren-Hua Wang*[2]

[1]ATR Spoken Language Translation Research Labs, Kyoto, Japan
[2]University of Science and Technology of China, China
{xiaobing.li, frank.soong}@atr.jp, myrvoll@iet.ntnu.no, rhw@ustc.edu.cn

ABSTRACT

We propose an algorithm for optimal clustering and non-uniform allocation of Gaussian Kernels in scalar (feature) dimension to compress complex, Gaussian mixture-based, continuous density HMMs into computationally efficient, small footprint models. The symmetric Kullback-Leibler divergence (KLD) is used as the universal distortion measure and it is minimized in both kernel clustering and allocation procedures. The algorithm was tested on the Resource Management (RM) database. The original context-dependent HMMs can be compressed to any resolution, measured by the total number of clustered scalar kernel components. Good trade-offs between the recognition performance and model complexities have been obtained; HMM can be compressed to 15-20% of the original model size, which needs 1-5% of multiplication/division operations, and results in almost negligible recognition performance degradation.

1. INTRODUCTION

Current state-of-the-art, context-dependent, continuous density HMM-based, large vocabulary speech recognition system can deliver a fairly decent recognition performance but usually at a price of large memory for its storage and high computation complexities in computing local log-likelihoods and dynamic programming search. It poses a research challenge to come up with HMMs of smaller footprints while maintaining the high performance of a much larger model. The problem can be approached from two forefronts: (1) training parsimonious models of high discrimination, e.g., minimum classification error (MCE) [1] or variational Bayesian (VB) [2] training; (2) compressing a given high resolution (hence high performance) model into a smaller one, hopefully without compromising the recognition performance. In this study we concentrate on the 2nd approach to continuous, Gaussian mixture-based HMM model compression. Similar attempts have been taken along the idea of HMM model compression before, e.g., feature-level parameter tying [3], subspace distribution clustering [4], or divergence based vector quantized variances [5]. However, there are some missing or incomplete parts in the previous attempts on how to find the optimal centroid of clustered kernels for a chosen information-theoretic distortion measure and how to allocate kernels efficiently across different feature subspaces or dimensions.

The symmetric Kullback-Leibler divergence [6], an information-theoretic measure of inter-distribution distortion, is chosen in this study for measuring (dis)similarity between two given Gaussian probability density functions (*pdf*'s). It has been shown that the optimal centroid of a Gaussian *pdf* cluster can be computed through a fast iterative procedure [7].

In the feature-level parameter tying or the subspace HMM clustering, same number of kernels was used across each dimension or subspace. But it is fairly well known that features in different dimensions or subspaces can have unequal discriminations, e.g., [8]. Enlightened by the rate-distortion theory which has been well exploited for assigning bits non-uniformly to different LPC parameters to minimize distortion at a given bit rate, e.g., [9], we propose an non-uniform kernel allocation algorithm. The same symmetric KLD is used to measure the model precision and kernels are allocated successively to a feature dimension where maximum KLD reduction is obtained. Via this non-uniform kernel allocation algorithm, we can compress the original HMM into any size (measured by the total number of Gaussian kernels used) while the total KLD between the original and the compressed HMMs is minimized.

The rest of the paper is organized as follows. In Section 2, an overview of the symmetric KLD and the corresponding optimal centroid of clustered multivariate, Gaussian *pdf*'s, especially for the diagonal covariance case, are given. In Section 3, an algorithm for non-uniform kernel allocation is proposed. Database, experimental setups and results are presented in Section 4. In Section 5, a conclusion is given.

2. KULLBACK-LEIBLER DIVERGENCE AND CORRESPONDING OPTIMAL CENTROID

The symmetric Kullback-Leibler divergence or the Jeffrey's divergence [6], a distortion measure for measuring (dis)similarity between two given *pdf*'s, f and g, is defined as:

$$d(f,g) = \int f \log \frac{f}{g}\, dx + \int g \log \frac{g}{f}\, dx \qquad (1)$$

T.A. Myrvoll, a visiting researcher at ATR SLT Labs when this work was done, on leave from the Dept. of Elec. and Telecom., NTNU.

This measure is a symmetrized version of two asymmetric KLD, or the first and second terms in equation (1). The optimal centroid probability distribution, f_c, of a cluster of N distributions is obtained by minimizing the total KLD, between the cluster centroid *pdf* and all *pdf*'s in the cluster, as:

$$f_c = \arg\min_{f_c'} \sum_{n=1}^{N} d(f_c', f_n) \tag{2}$$

For multivariate Gaussian distributions, a closed form of the symmetric KLD in equation (1) is:

$$d(f, g) = \frac{1}{2} trace\{(R_f^{-1} + R_g^{-1})(\mu_f - \mu_g)(\mu_f - \mu_g)^T + R_f R_g^{-1} + R_g R_f^{-1} - 2I\} \tag{3}$$

where μ and R are the mean and covariance of the corresponding Gaussian distribution, respectively.

The optimal centroid of multivariate Gaussians can be obtained by solving a set of Riccati matrix equations [7]. For the special case of diagonal covariance, the i-th dimension mean and variance of the centroid, μ_{ci} and σ_{ci}^2, can be computed iteratively by alternating eqs. (4) and (5) as:

$$\mu_{ci} = \frac{\sum_{n=1}^{N} (\sigma_{ci}^{-2} + \sigma_{ni}^{-2}) \mu_{ni}}{\sum_{n=1}^{N} (\sigma_{ci}^{-2} + \sigma_{ni}^{-2})} \tag{4}$$

$$\sigma_{ci}^2 = \frac{\sum_{n=1}^{N} \sigma_{ni}^2 + (\mu_{ci} - \mu_{ni})^2}{\sum_{n=1}^{N} \sigma_{ni}^{-2}} \tag{5}$$

It has been shown that the overall KLD is a convex function of both mean and covariance of the centroid Gaussian *pdf*. Experimentally, we have also found that only few iterations are needed for a centroid to converge to its optimum [7].

Fig. 1 depicts the 7,070 Gaussian kernels of a context-dependent HMM (will be described in Section 4) in feature dimension C_1 and they are clustered into 4, 8 and 16 centroid kernels optimally. The fidelity of representing the original 7,070 kernels by their nearest centroids improves monotonically when the number of centroids increases from 4 to 16.

For the case of diagonal covariance, which is assumed for this study, the multivariate Gaussian *pdf* can be written as a product of all its scalar, statistically independent components. Consequently, the corresponding multivariate KLD is linearly additive in terms of its scalar components. This property forms the foundation of why our non-uniform kernel allocation algorithm, which will be presented in the next section, can be decomposed into a scalar search in feature dimension.

3. NON-UNIFORM KERNEL ALLOCATION FOR HMM COMPRESSION

Our non-uniform kernel allocation algorithm searches for the feature dimension to allocate successively one extra centroid kernel (from the set generated in the optimal clustering procedure) to a centroid kernel subset. The dimension for allocating the extra kernel is chosen, based upon the maximum reduction of total KLD distortions. The KLD is computed by measuring the distortion between the kernels used in the unquantized, original HMM and their nearest neighbors (in the centroid kernel subset). It is searched component by component in scalar feature dimensions because KLD distortion between two multivariate Gaussian kernels with diagonal covariances is linearly additive in its scalar components. It is still, by all means, a greedy search algorithm. But as it will be shown later, despite its greedy nature, this algorithm gives virtually the same result as an (M, L) search where a much larger M (retained candidates) is kept in each search cycle.

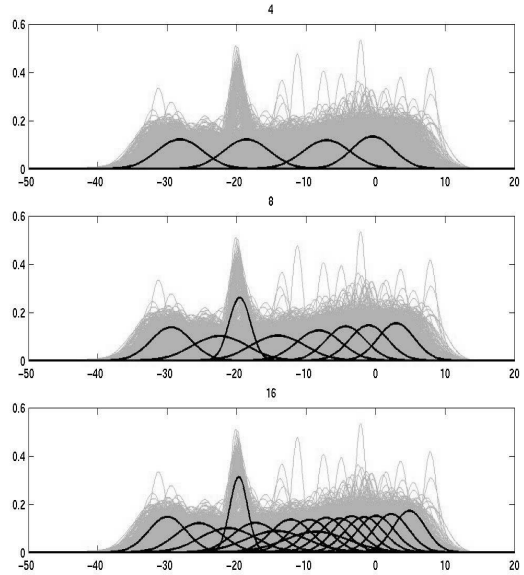

Fig. 1. 7,070 distributions in feature dimension C_1 are clustered into 4, 8 and 16 centroid kernels by LBG clustering

Table 1. Pseudocode of non-uniform kernel allocation method

for i = 1 … D			
	starting with 1 kernel per dimension: $k_i = 1$		
compute current KLD: Q_{new}			
successively allocate one extra kernel until $Q_{new} < Threshold$			
	$Q_{old} = Q_{new}$ $\Delta Q = 0$		
	for i = 1 … D		
		add 1 kernel in the i-th dimension: k_i ++	
		compute current KLD: Q_{new}	
		if (Q_{old} - Q_{new}) > ΔQ	
			$\Delta Q = Q_{old} - Q_{new}$ $dim = i$ select the dimension
		remove the allocated kernel in the i-th dimension: k_i --	
	allocate one extra kernel in the dimension of maximum KLD reduction: k_{dim} ++		
	$Q_{new} = Q_{old} - \Delta Q$		

Strictly speaking, the distortion measure for HMM compression should be the KLD between the state *pdf* of the original model and the "quantized" state *pdf*, parameterized by the selected kernels in the compressed centroid subset. However, the state output *pdf* is in general a mixture of Gaussians, there is no closed-form for computing the corresponding KLD. The numerical computation of KLD in a multi-dimensional space is somewhat prohibitive for a complex HMM. We therefore resort to a computationally tractable, approximate solution by measuring the KLD between the Gaussian kernels used in the original HMMs and their nearest neighbors in a subset of successively allocated centroids.

The non-uniform kernel allocation method starts with allocating one kernel in each dimension and the corresponding KLD is computed. Then one extra kernel is tentatively assigned in each dimension in turn and a corresponding reduction of KLD is computed. The feature dimension that yields the maximum reduction of KLD is assigned with one more kernel and the procedure then repeats itself. The algorithm stops when the total KLD becomes less than a given threshold or optionally, the total number of kernels reaches a preassigned limit. The pseudocode of the non-uniform kernel allocation is given in Table 1.

4. EXPERIMENTAL RESULTS

Our optimal clustering and non-uniform kernel allocation method was tested on the DARPA 991-word Resource Management (RM) database. The standard SI-109 training data set of 3,990 utterances was used for training the HMMs. The CMU 48 phone set was used to create a context-dependent (CD) model, with 1,414 tied states and a mixture of 5 Gaussians per state. Altogether, there are 7,070 kernels used in the CD HMM. The features are the conventional 39-dimension MFCCs (12 static MFCCs, log energy, and their first- and second-order time derivatives). The Sep92 test set was used for evaluation with the standard word-pair grammar of perplexity 60. The baseline performance of the original, uncompressed HMM is 7.35% word error rate (WER).

4.1. Clustering and allocation results

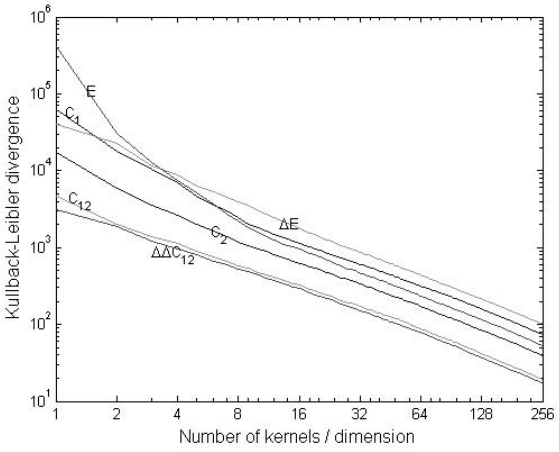

Fig. 2. KLD vs. number of clustered centroid kernels

Fig. 2 shows the kernel clustering performance where corresponding KLD is plotted against the number of kernels for the features of C_1, C_2, C_{12}, E, ΔE and $\Delta\Delta C_{12}$. Except at smaller number of kernels (i.e., lower rate), the log-log plot shows relatively straight, parallel lines of rate-distortion curves where different features exhibit different intersecting points. The rate distortion curves of other features show similar trends.

Based on the component rate-distortion curves of different features, the non-uniform kernel allocation algorithm generates a composite rate-distortion curve, which is plotted in Fig. 3 where KLD is shown against average kernels per dimension. The non-uniform allocation yields a better rate-distortion curve than that of fixed, uniform kernel allocation, which is also plotted (broken line) in the same figure. Fig. 4 gives the recognition performance curves against the average number of kernels per dimension for both fixed and adaptive, non-uniform kernel allocations. As expected, with comparable number of kernels, the non-uniform allocation generally gives better recognition performance than the fixed allocation, due to its lower KLD. We also found that our greedy algorithm gives virtually the same results as an (M, L) search with M = 200 and L = 39.

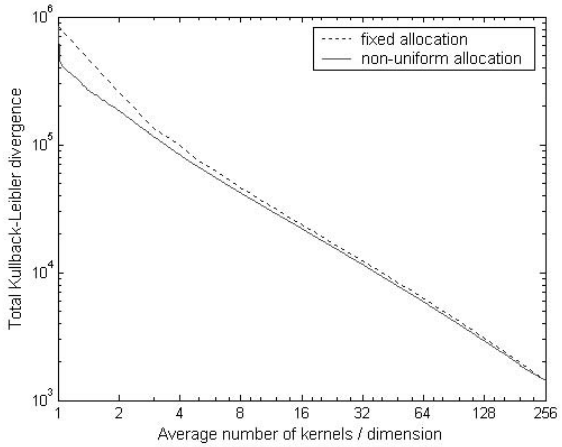

Fig. 3. Total KLD vs. average number of kernels per dimension

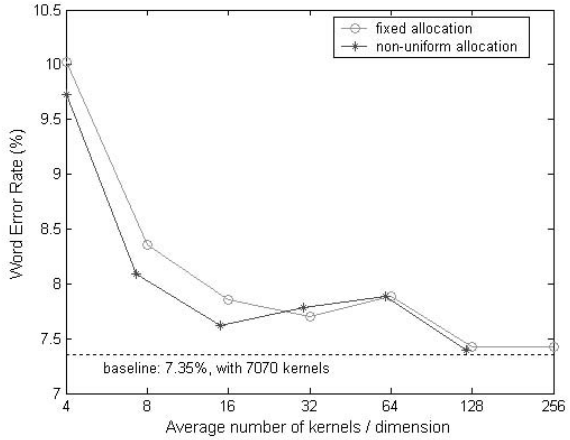

Fig. 4. Recognition performance vs. average number of kernels per dimension

4.2. Experimental Results

In Figs. 5 and 6, the recognition performance (WER) is plotted vs. memory storage and computation (multiplication/division only) for different compression ratios, respectively. For the storage, one byte (256 possibilities) is used for encoding the index of each kernel used in both fixed and adaptive, non-uniform kernel allocation. For the computation, the log likelihood for the *j*-th state is calculated as follows:

$$\log b_j(\vec{o}_t) = \log \sum_{m=1}^{M} c_{jm} \exp\left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log\sigma_{jmi}^2 - \sum_{i=1}^{D}\frac{(o_{ti}-\mu_{jmi})^2}{2\sigma_{jmi}^2} \right] \quad (6)$$

As Gaussian kernels are clustered and shared in the scalar dimension, the third term inside the square bracket, $(o_{ti}-\mu_{jmi})^2/2\sigma_{jmi}^2$, is pre-computed and stored in a small table shared among all output *pdf*'s. For the fixed and the non-uniform kernel allocations, the addition/subtraction requirement is similar, about 50% of the original HMM's complexity. Therefore, we only plot the computation ratios for multiplication/division only. From the two figures, both the fixed and the non-uniform allocation yield significant savings of storage and computations. With comparable computation and memory resources, the non-uniform allocation generally gives better performance than the fixed allocation, especially for larger compression ratios (i.e., less kernels).

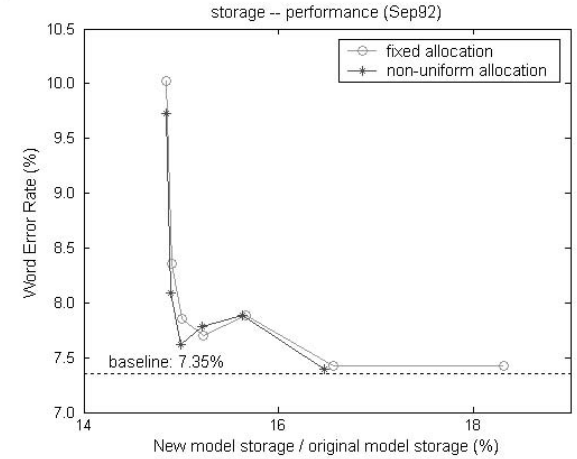

Fig. 5. Storage vs. recognition performance

5. CONCLUSIONS

We propose an HMM model compression algorithm for optimal clustering and non-uniform allocation of Gaussian kernels in HMM feature (scalar) dimension. The symmetric KLD is used as the universal distortion measure for both kernel clustering and allocation. Non-uniform kernel allocation in model compression is performed successively, one kernel at a time, by searching over all feature dimensions. Computationally efficient and small footprint, compact HMMs can be custom made at any operating point along the rate-distortion curve. Tested on the RM database, we found the original, context-dependent phone HMMs can be compressed to 15-20% of its original size and 1-5% of the original multiplication/division operations, with almost negligible recognition performance degradation.

6. ACKNOWLEDGEMENTS

This research was supported in part by the National Institute of Information and Communications Technology.

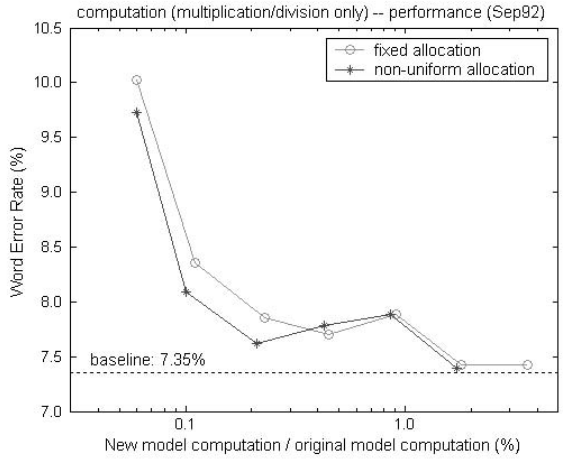

Fig. 6. Multiplication/division vs. recognition performance

7. REFERENCES

[1] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, pp. 257-265, May 1997.

[2] H. Attias, "A Variational Bayesian Framework for Graphical Models", In S.A. Solla, T.K. Leen, and K. Muller, editors, *Advances in Neural Information Processing Systems 12*, pp. 49-52, Cambridge, MA, 2000, MIT Press.

[3] S. Takahashi, and S. Sagayama, "Four-Level Tied-Structure for Efficient Representation of Acoustic Modeling", *Proc. ICASSP*, pp. 520-523, 1995.

[4] E.L. Bocchieri, and K.W. Mak, "Subspace Distribution Clustering Hidden Markov Model", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 3, pp. 264-275, March 2001.

[5] J. Kim, R. Haimi-Cohen, and F.K. Soong, "Hidden Markov Models with Divergence based Vector Quantized Variances", *Proc. ICASSP*, pp. 125-128, 1999.

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

[7] T.A. Myrvoll, and F.K. Soong, "Optimal Clustering of Multivariate Normal Distributions using Divergence and its Application to HMM Adaptation", *Proc. ICASSP*, pp. 552-555, 2003.

[8] E.L. Bocchieri, and J.G. Wilpon, "Discriminative Analysis for Feature Reduction in Automatic Speech Recognition", *Proc. ICASSP*, pp. 501-504, 1992.

[9] F.K. Soong, and B.-H. Juang, "Optimal Quantization of LSP Parameters", *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 1, pp. 15-24, January 1993.