Maximum Likelihood Clustering of Gaussians for Speech Recognition

A. Kannan, M. Ostendorf, and J. R. Rohlicek

Abstract—**This correspondence describes a method for clustering multivariate Gaussian distributions using a maximum likelihood criterion. We point out possible applications of model clustering, and then use the approach to determine classes of shared covariances for context modeling in speech recognition, achieving an order of magnitude reduction in the number of covariance parameters, with no loss in recognition performance.**

I. Introduction

Distribution clustering is an important tool in statistical modeling. In speech recognition in particular, distribution clustering can be used to reduce the number of context-dependent models (which enables robust parameter estimates and reduces recognition computation and storage costs), to provide an initial estimate of component distributions for mixture models, and to group similar models for building an initial "fast-match" function in large vocabulary recognition. This work describes a new method of distribution clustering that handles continuous observations and is consistent with a maximum likelihood (ML) parameter estimation criterion.

Two important issues associated with clustering distributions include the general method (agglomerative or divisive hierarchical methods versus K-means clustering) and the clustering criterion or objective function. Initial work in clustering models for speech recognition used variations on agglomerative clustering [1], [2]. Subsequent work explored divisive clustering methods, using linguistically motivated questions (partitioning functions) about phonetic context for splitting the data [3], [4]. An advantage of this divisive clustering approach, as Lee *et al.* point out [3], is that conditioning contexts unseen in training can be easily mapped to a cluster that provides a robust but detailed model. For this reason, our work uses divisive clustering, although the similarity criterion we propose could easily be applied to agglomerative clustering as well.

The second issue in clustering is the choice of a clustering criterion or objective function. One possibility is to use a measure of distribution similarity, such as information divergence (see [5] for the hidden Markov model (HMM)) or the chi-squared-like measure used in [1] for Gaussian distributions. However, such similarity measures tend to be more useful for agglomerative clustering than for divisive clustering, because agglomerative clustering does not require the computation of a centroid associated with the similarity measure and the centroid is difficult to define for these criteria. In addition, similarity measures on distributions may not faithfully represent the similarity of the data from which the distributions were estimated if distribution assumptions were inaccurate or parameters were estimated from sparse data. Other objective functions proposed for distribution clustering include likelihood ratios for discrete observations [2], [4] and for Gaussian distributions [6]. (The likelihood ratio in [2] was shown to be equivalent to the entropy measure used in [3], and the likelihood ratio in [4] was reduced to a similar measure.) The likelihood ratio criterion represents the relative probability of a set of data using one versus two models, and its use in divisive clustering guarantees an increase in the likelihood of the data. Thus the likelihood ratio criterion has the advantage that it is consistent with the objective of maximum likelihood parameter estimation, that is if the clustered distributions and not just the cluster definition are used in the model (the distinction between our use of clustering in estimating the model parameters and the use of clustering in [4] to determine regions of parameter tying).

This work investigates the use of a likelihood ratio criterion in divisive clustering for context-dependent modeling in speech recognition, extending the work of Gish *et al.* [6] which looked at agglomerative clustering for speaker segmentation and identification. We describe general methods for clustering data to determine appropriate multivariate Gaussian models under different parameter tying conditions (tying all Gaussian parameters versus only covariances), and then present experiments in clustering covariances, specifically for estimating unimodal Gaussian distributions that represent regions of a phoneme segment as used in the stochastic segment model (SSM). (The region-dependent distributions in the SSM are analogous to state-dependent observation distributions in an HMM.) Note that for the speech recognition application, the question of whether to cluster on the phone level or sub-phone level arises. Like [7] but unlike most other reported work, the experiments here focus on the sub-phone level, though the general method is also applicable to phone-level clustering. The results show that the number of covariance parameters can be reduced by more than a factor of ten through clustering, with no loss in recognition performance.

II. Clustering Paradigm

The clustering algorithm is a binary tree growing procedure, similar to decision tree design [8], that successively partitions the observations (splits a node in the tree), at each step minimizing a splitting criterion over a pre-determined set of allowable binary partitions. For each allowable binary partition of the data, we evaluate a likelihood ratio to choose between one of two hypotheses:

- H_0: the observations were generated from one distribution (that corresponds to the ML estimate for the parent node), and
- H_1: the observations were generated from two different distributions (that correspond to the ML estimates for the child nodes).

The likelihood ratio, λ, is defined as the ratio of the likelihood of the observations being generated from one distribution (H_0) to the likelihood of the observations in the partition being generated from two different distributions (H_1). For Gaussians, λ can be expressed as a product of the quantities λ_{COV} and λ_{MEAN} [6], which are expressed in terms of the sufficient statistics of the observation sets

$$\lambda_{\mathrm{MEAN}} = \left(1 + \frac{n_l n_r}{n^2}(\hat{\mu}_l - \hat{\mu}_r)^t W^{-1}(\hat{\mu}_l - \hat{\mu}_r)\right)^{\frac{-n}{2}} \quad (1)$$

$$\lambda_{\mathrm{COV}} = \left(\frac{|\hat{\Sigma}_l|^{\alpha}|\hat{\Sigma}_r|^{1-\alpha}}{|W|}\right)^{\frac{n}{2}} \quad (2)$$

where n_l and n_r are the number of observations in the left and right child nodes with $n = n_l + n_r$, $\hat{\mu}_l$ and $\hat{\mu}_r$ are the sample means of the left and right nodes, $\hat{\Sigma}_l$ and $\hat{\Sigma}_r$ are the sample covariances

Manuscript received April 12, 1993; revised September 14, 1993. This work was supported by NSF and ARPA under NSF grant number IRI-8902124, and by ARPA and ONR under ONR grant number N00014-92-J-1778.

A. Kannan and M. Ostendorf are with the ECS Department, Boston University, Boston, MA 02215 USA.

J. R. Rohlicek is with Bolt, Beranek, and Newman, Inc., Cambridge, MA 02138 USA

IEEE Log Number 9400751.

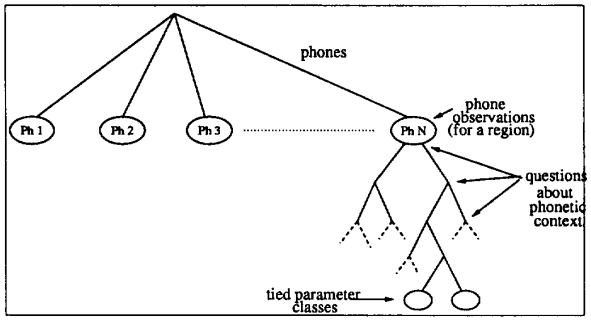

Fig. 1. Illustration of divisive clustering.

associated with the left and right nodes, $\alpha = \frac{n_l}{n}$, and W is the frequency weighted tied covariance, viz., $W = \frac{n_l}{n}\hat{\Sigma}_l + \frac{n_r}{n}\hat{\Sigma}_r$.

There are different variations on clustering with the likelihood ratio criterion, corresponding to different hypothesis tests on the candidate partition of a node. If the clustering is to determine whether the observations in two sets share a common covariance, then the increase in log likelihood is given by $-\log \lambda_{\mathrm{COV}}$. Alternatively, if the hypothesis test is over the complete distributions, then the increase in likelihood due to the partition is $-(\log \lambda_{\mathrm{COV}} + \log \lambda_{\mathrm{MEAN}})$. Finally, if the goal is to cluster distribution means assuming a common covariance, then the likelihood ratio criterion is $-\log \lambda_{\mathrm{MEAN}}$. (Note that the mean clustering case would require a hybrid divisive plus K-means clustering to guarantee increase in likelihood, since the common covariance is defined as the sample covariance of the parent node.) Derivations leading to these different cases can be found in [9].

Divisive clustering involves growing a binary tree using a greedy algorithm for maximizing the likelihood of the data. For each terminal node in the tree, we evaluate the increase in likelihood for all binary partitions allowed, and then split the terminal node with the partition that results in the largest increase in likelihood. The tree is grown until there are no more splits that result in valid child nodes. Here, it is assumed that valid terminal nodes must have more than T_c observations, where T_c is an empirically determined threshold to indicate that a reliable covariance can be estimated for that node (we use $T_c = 250$, for vector dimension 29). The full tree can be used for the set of clustered models as in the experiments described here, or alternatively, one could use tree pruning techniques (e.g., [10]) to determine the appropriate number of distributions.

This technique to cluster Gaussians can be used for clustering context-dependent acoustic models in speech recognition. In this work, we cluster region-specific distributions across triphones, where a triphone is a phone conditioned on the phone label of its left and right neighbor. In other words, divisive clustering is performed independently on the observations that correspond to each region of the center phone, with the goal of finding classes of triphones that can share a common covariance. The partitions used to test the likelihood ratio are found by asking linguistically motivated questions related to features such as the place and manner of articulation of the immediate left and right neighboring phones of the triphone. Of course, the conditioning contexts could potentially include a larger window of neighbors [4] or information such as lexical stress. Only simple questions (i.e., questions about one variable) are used in this implementation; a method for designing trees with compound questions is described in [3]. The clustering framework is illustrated in Fig. 1.

When the tree is grown, each terminal node has a set of observations associated with it that map to a set of region-specific triphone distributions. The partition of observations directly implies a partition of triphones, since the allowable questions refer to the left and right neighboring phone labels. Each node is associated with a covariance, which is an unbiased estimate of the tied covariance for the constituent distributions computed by taking a weighted average of the separate triphone-dependent covariances. During recognition, all distributions associated with a terminal node share this covariance. Although it would have been possible to cluster means as well, we simply used the triphone-dependent means and backed off to combined left- and right-context-dependent means when necessary due to insufficient triphone training data. In these experiments, clustering of means was not warranted, since for full-covariance Gaussian distributions it contributes only marginally to parameter reduction and could cause a degradation in recognition accuracy. However, clustering of the means does become important if, instead of full-covariance Gaussians, diagonal-covariance Gaussians are used to model observation distributions, or if there is only a small amount of training data available.

III. Experiments

We evaluated the effects of clustering triphones for the stochastic segment model (SSM). The SSM, first introduced in [11], represents a variable-duration observation sequence $Y = [y_1, \cdots, y_L]$ of random length L (a segment) using a model for each phone α consisting of (1) a family of joint density functions (one for every observation length), and (2) a collection of mappings that specify the particular density function for a given observation length. Typically, the model assumes that segments are described by a fixed-length sequence of locally time-invariant regions (or regions of tied distribution parameters). A deterministic mapping specifies which region corresponds to each observation vector.

The specific SSM version used here [12], [13] assumes that frames within a segment are conditionally independent given the segment length. In this case, the probability of a segment given phone α is the product of the probability of each observation y_i and the probability of its (known) duration L

$$p(Y|\alpha) = p(Y, L|\alpha) = p(L|\alpha) \prod_{i=1}^{L} p(y_i|\alpha, T_L(i))$$

where the distribution used corresponds to region $T_L(i)$. The distributions associated with a region j, $p(y|\alpha, j)$, are multivariate Gaussians. In this work, the phone length distribution $p(L|\alpha)$ is a smoothed relative frequency estimate. The function T_L is a deterministic mapping of the L-long observation to the m regions in the model, and here T_L is linear in time for the entire segment.

To reduce the computational costs associated with a segment-based model, which has a much higher effective search space than an HMM, we use the N-best rescoring formalism for continuous word recognition [14]. In this formalism, one recognition system produces the top N hypotheses for an utterance, the hypotheses rescored by other knowledge sources, and the different scores are combined to rerank the hypotheses. In addition to reducing computation for the SSM (by reducing the search space), the N-best rescoring paradigm provides a mechanism for integrating very different types of knowledge sources, though this aspect is not explored here. For these experiments, the initial list of candidate sentences were generated using BBN's BYBLOS system and then rescored by the SSM. The BYBLOS system is an HMM-based system that uses tied Gaussian mixtures and context-dependent models including cross-word triphones [15]. Once the N-best list is rescored by the SSM, it is reordered according to a linear combination of the SSM log acoustic score, the number of words in the sentence (insertion penalty) and the number of phonemes in the sentence. We estimate the set of weights

in the linear combination that minimizes average word error in the top ranking hypotheses [16].

Both the SSM and the HMM use gender-dependent acoustic models, and the gender used by the SSM in recognition is determined by the BBN system. The SSM system uses frame-based observations of spectral features, including 14 mel-warped cepstra and their first differences, plus the first difference of log energy. Each phone-sized segment model uses $m = 8$ multivariate (full), unimodal Gaussian distributions, assuming frames are conditionally independent given the segment length.

Results are reported on the speaker-independent Resource Management task (continuous speech, 991 word vocabulary). The SSM models are trained on the SI-109, 3990 utterance SI training set. In these experiments, we use $N = 20$ for the N-best list. The correct sentence is included in this list about 98% of the time by the Byblos system, using the word-pair grammar. The February 89 speaker-independent (SI) test set was used to estimate gender-independent weights that were then used to combine scores for the evaluation test set (October 89). Recognition performance was computed as the word error rate based on the top ranking hypotheses after rescoring.

The performance of our system on the October 89 test set was 5.0% for the base-line triphone system [13], 4.9% for the system using clustering with the full likelihood criterion and 4.9% when clustering with only the λ_{COV} likelihood criterion. The corresponding numbers for the February 89 development test set were 4.6%, 4.2% and 4.1%, respectively. (For reference, the recognition performance achieved by combining the SSM scores with HMM scores was 3.3% and 2.5% word error for the October 89 and February 89 test sets, respectively, using clustering with the λ_{COV} criterion.) Although the performance differences on the October 89 test set are not significant, the performance on the development test set provides some evidence that clustering with the theoretically appropriate λ_{COV} criterion is also a good choice in terms of recognition performance. The results are consistent with those reported by others, in that the main benefit of clustering is a reduction in model complexity rather than an improvement in performance. In these experiments, we reduced the number of covariance parameters required by more than a factor of ten with no loss in recognition performance, and further reduction may be possible. In preliminary experiments on the much larger Wall Street Journal corpus we have observed similar results. Since the covariance parameters are the dominating factor in computation and storage costs, this represents a significant overall reduction.

IV. Conclusion

In summary, we have described a divisive clustering paradigm for multivariate Gaussians based on a likelihood ratio test. In the context of speech recognition, we use the clustering formalism to determine classes of triphones over which to tie covariances in the SSM, finding that we can reduce the number of covariances by more than a factor of ten without any loss in recognition performance. This method will be useful for any pattern recognition problem where features are modeled using Gaussian distributions, including HMM's. In particular, this approach to clustering may also be useful for providing initial estimates of components in tied-mixtures, determining classes of like models for designing fast initial search procedures in large vocabulary recognition, or determining model topology as in successive state splitting [17]. The algorithms described here apply only to unimodal Gaussians, however, and further analytical development is needed to extend the likelihood ratio criterion to clustering of Gaussian mixtures (used in many successful HMM systems) rather than the individual mixture components.

References

[1] D. B. Paul and E. A. Martin, "Speaker stress-resistant continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 283–286.

[2] K.-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, April 1990, pp. 599–609.

[3] K.-F. Lee *et al.*, "Allophone clustering for continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990 pp. 749–752.

[4] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Context dependent modeling of phones in continuous speech using decision trees," in *Proc. DARPA Speech, Natural Language Workshop*, Feb. 1991, pp. 264–269.

[5] B.-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.*, vol. 64, no. 2, pp. 391–408, 1985.

[6] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, May 1991, pp. 873–876.

[7] M.-Y. Hwang and X. Huang, "Subphonetic modeling for speech recognition," in *Proc. DARPA Speech, Natural Language Workshop*, Feb. 1992, pp. 174–179.

[8] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks/Cole, 1984.

[9] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1984, pp. 404–450.

[10] P. Chou, T. Lookabaugh, and R. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Inform. Theory*, pp. 299–315, Mar. 1989.

[11] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1857–1869, Dec. 1989.

[12] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous word recognition based on the stochastic segment model," *Proceedings of the 1992 DARPA Workshop on Artificial Neural Networks and Continuous Speech Recognition*.

[13] O. Kimball, M. Ostendorf and I. Bechwati, "Context modeling with the stochastic segment model," *IEEE Transactions Signal Processing*, pp. 1584–1587, June 1992.

[14] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz and J. R. Rohlicek, "Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses," *Proc. DARPA Speech and Natural Language Workshop*, pp. 83–87, February 1991.

[15] F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway and R. Schwartz, "BYBLOS speech recognition benchmark results," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 77–82, February 1991.

[16] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Weight estimation for N-best rescoring," in *Proc. DARPA Speech, Natural Language Workshop*, Feb. 1992, pp. 455–456.

[17] J. Takami and S. Sagayama, "A successive splitting algorithm for efficient allophone modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992, pp. 573–576.