Speaker Recognition with Mixtures of Gaussians with Sparse Regression Matrices

Constantinos Boulis
University of Washington,
Electrical Engineering dept., SSLI Lab
boulis@ee.washington.edu

Abstract

When estimating a mixture of Gaussians there are usually two choices for the covariance type of each Gaussian component. Either diagonal or full covariance. Imposing a structure though may be restrictive and lead to degraded performance and/or increased computations. In this work, several criteria to estimate the structure of regression matrices of a mixture of Gaussians are introduced and evaluated. Most of the criteria attempt to estimate a discriminative structure, which is suited for classification tasks. Results are reported on the 1996 NIST speaker recognition task and performance is compared with structural EM, a well-known, non-discriminative, structure-finding algorithm.

1 Introduction

Most state-of-the-art systems in speech and speaker recognition use mixtures of Gaussians when fitting a probability distribution to data. Reasons for this choice are the easily implementable estimation formulas and the modeling power of mixtures of Gaussians. For example, a mixture of diagonal Gaussians can still model dependencies on the global level. An established practice when applying mixtures of Gaussians is to use either full or diagonal covariances. However, imposing a structure may not be optimum and a more general methodology should allow for joint estimation of both the structure and parameter values.[1].

The first question we have to answer is what type of structure we want to estimate. For mixtures of Gaussians there are three choices. Covariances, inverse covariances or regression matrices. For all cases, we can see as selecting a structure by introducing zeros in the respective matrix. The three structures are distinctively different and zeros in one matrix do not, in general, map to zeros in another matrix. For example, we can have sparse covariance but full inverse covariance or sparse inverse covariance and full regression matrix.

There are no clear theoretical reasons why one choice of structure is more suitable than others. However, introducing zeros in the inverse covariance can be seen as deleting arcs in an Undirected Graphical Model (UGM) where each node represents each dimension of a single Gaussian (Bilmes, 2000). Similarly, introducing zeros in the regression matrix can be seen as deleting arcs in a Directed Graphical Model (DGM). There is a rich body of work on structure learning for UGM and DGM and therefore the view of a mixture of Gaussians as a mixture of DGM or UGM may be advantageous. Under the DGM framework, the problem of Gaussian parameter estimation can be cast as a problem of estimating linear regression coefficients. Since the specific problem of selecting features for linear regression has been encountered in different fields in the past, we adopt the view of a mixture of Gaussians as a mixture of DGM.

In (Bilmes, 2000), the problem of introducing zeros in regression matrices of a mixture of Gaussians was presented. The approach taken was to set to zero the pairs with the lowest mutual information, i.e. $b^m_{i,j} = 0 \iff I(X_i, X_j) \approx 0$, where m is the Gaussian index and $b_{i,j}$ is the (i, j) element of regression matrix B. The approach was tested for the task of speech recognition in a limited vocabulary corpus and was shown to offer the same performance with the mixture of full-covariance Gaussians with 30% less parameters. full covariances. One issue with the work in (Bilmes, 2000) is that the structure-estimation criterion that was used was not discriminative. For classification tasks, like speaker or speech recognition, discriminative parameter estimation approaches achieve better performance than generative ones, but are in general hard to estimate especially for a high number of classes. In this work, a number of discriminative structure-estimation criteria tailored for the task of speaker recognition are introduced. We avoid the complexities of discriminative parameter estimation by estimating a discriminative structure and then applying generative parameter estimation techniques. Thus, overall the models attempt to model the discriminability between classes without the numerical and implementation diffi-

[1]Here, we describe the Maximum Likelihood estimation methodology for both structure and parameters. One alternative is Bayesian estimation.

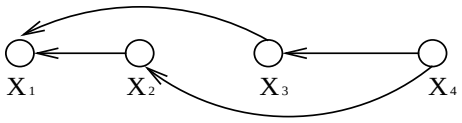

Figure 1: A Directed Graphical Model

culties that such techniques have. A comparison of the new discriminative structure-estimation criteria with the structural EM algorithm is also presented.

This paper is structured as follows. In section 2, the view of a Gaussian as a directed graphical model is presented. In section 3, discriminative and generative structure-estimation criteria for the task of speaker recognition are detailed, along with a description of the structural EM algorithm. In section 4, the application task is described and the experiments are presented. Finally, in section 5, a summary and possible connections of this work with the speaker adaptation problem are discussed.

2 Gaussians as Directed Graphical Models

Suppose that we have a mixture of M Gaussians:

$$p(x) = \sum_{m}^{M} p(z = m) N(x; \mu_m, \Sigma_m) \quad (1)$$

It is known from linear algebra that any square matrix A can be decomposed as $A = LDU$, where L is a lower triangular matrix, D is a diagonal matrix and U is an upper triangular matrix. In the special case where A is also symmetric and positive definite the decomposition becomes $A = U^T DU$ where U is an upper triangular matrix with ones in the main diagonal. Therefore we can write $U = I - B$ with $b_{ij} = 0$ if $i >= j$.

The exponent of the Gaussian function can now be written as (Bilmes, 2000):

$$(\tilde{x} - B\tilde{x})^T D(\tilde{x} - B\tilde{x}) \quad (2)$$

where $\tilde{x} = x - \mu$. The i-th element of $(\tilde{x} - B\tilde{x})$ can be written as:

$$\tilde{x}_i - \sum_{k=i+1}^{V} b_{i,k}\tilde{x}_k \quad (3)$$

with V being the dimensionality of each vector. Equation 3 shows that the problem of Gaussian parameter estimation can be casted as a linear regression problem. Regression schemes can be represented as Directed Graphical Models. In fact, the multivariate Gaussian can be represented as a DGM as shown in Figure 1. Absent arcs represent zeros in the regression matrix. For example the B matrix in Figure 1 would have $b_{1,4} = b_{2,3} = 0$.

We can use the EM algorithm to estimate the parameters of a mixture of Gaussian $\theta = [\mu_m B_m D_m]$. This formulation offers a number of advantages over the traditional formulation with means and covariances. First, it avoids inversion of matrices and instead solves $V + 1$ linear systems of d equations each where $d = 1 : V + 1$. If the number of components and dimensionality of input vectors are high, as it is usually the case in speech recognition applications, the amount of computations saved can be important. Second, the number of computations scale down with the number of regression coefficients set to zero. This is not true in the traditional formulation because introducing zeros in the covariance matrix may result in a non-positive definite matrix and iterative techniques should be used to guarantee consistency (Dempster, 1972). Third, for the traditional formulation, adapting a mixture of non-diagonal Gaussians with linear transformations leads to objective functions that cannot be maximized analytically. Instead, iterative maximization techniques, such as gradient descent, are used. With the new formulation even with arbitrary Gaussians, closed-form update equations are possible. Finally, the new formulation offers flexibility in tying mechanisms. Regression matrices B_m and variances D_m can be tied in different ways, for example all the components can share the same regression matrix but estimate a different variance diagonal matrix for each component. Similar schemes were found to be succesful for speech recognition (Gales, 1999) and this formulation can provide a model that can extend such tying schemes. The advantages of the new formulation are summarized in Table 1.

3 Structure Learning

In general, structure learning in DGM is an NP-hard problem even when all the variables are observed (Chickering et al., 1994). Our case is further complicated by the fact that we have a hidden variable (the Gaussian index). Optimum structure-finding algorithms, like the one in (Meila and Jordan, 2000) assume a mixture of trees and therefore are making limiting assumptions about the space of possible structures. In this paper, no prior assumptions about the space of possible structures are made but this leads to absence of guarantee for an optimum structure. Two approaches for structure-learning are introduced.

The first approach is to learn a discriminative structure, i.e. a structure that can discriminate between classes even though the parameters are estimated in an ML fashion. The algorithm starts from the fully connected model and deletes arcs, i.e. sets $b_m^{i,j} = 0 \; \forall \; m = 1 : M$ (M is the number of Gaussian components in a mixture). After setting regression coefficients to zero, maximum likelihood parameter estimation of the sparse mixture is employed. A number of different structure-estimation criteria were tested for the speaker recognition task (at the right of each equation a shorthand for each criterion is defined):

	$\theta = [\mu_m, B_m, D_m]$	$\theta = [\mu_m, \Sigma_m]$
Computations	• Each component m=1:M requires solution of V+1 linear systems of d equations each, d=1:V+1. • Computations scale down with the number of regression coefficients set to zero.	• Each component m=1:M requires an inversion of a rank V matrix. • Iterative techniques must be employed for the sparse case.
Adaptation	Easy EM equations for estimating a linear transformation of a mixture of arbitrary Gaussians.	Gradient descent techniques for the M-step of EM algorithm for estimating a linear transformation of a mixture of arbitrary Gaussians.
Tying	More flexible tying mechanisms across components, i.e. components can share B but estimate D_m.	Components can share the entire Σ.

Table 1: Comparison between viewing a mixture of Gaussians as a mixture of DGMs and the traditional representation

$$\mathbf{MI} = I(X_i; X_j | target\ speaker\ s) \tag{4}$$

$$\mathbf{DMIimp} = I(X_i; X_j | target\ speaker\ s) - (1/N) \sum_n I(X_i; X_j | impostor\ n) \tag{5}$$

$$\mathbf{MIimp} = \sum_n I(X_i; X_j | impostor\ n) \tag{6}$$

$$\mathbf{DMIconf} = I(X_i; X_j | target\ speaker\ s) - I(X_i; X_j | target\ speaker\ k) \tag{7}$$

where $I(X_i; X_j)$ is the mutual information between elements X_i and X_j of input vector X. The mutual informations are estimated by first fitting a mixture of 30 diagonal Gaussians and then applying the methods described in (Bilmes, 1999). All but **MI** and **MIimp** are discriminative criteria and all are based on finding the pairs (i, j) with the lowest values and zeroing the respective regression coefficients, for every component of the mixture. **MIimp** assigns the same speaker-independent structure for all speakers. For **DMIconf** target speaker k is the most confusable for target speaker s in terms of hits, i.e. when the truth is s, speaker k fires more than any other target speaker. We can see that different criteria aim at different goals. **MI** attempts to avoid the overfitting problem by zeroing regression coefficients between least marginally dependent feature elements. **DMIimp** attempts to discriminate against impostors, **MIimp** attempts to build a speaker-independent structure which will be more robustly estimated since there are more data to estimate the mutual informations and **DMIconf** attempts to discriminate against the most confusable target speaker. The most confusable target speaker k for a given target speaker s should be determined from an independent held-out set.

There are three main drawbacks that are shared by all of the above criteria. First, they are limited by the fact that all Gaussians will have the same structure. Second, since we are estimating sparse regression matrices, it is known that the absence of an arc is equivalent to conditional independencies, yet the above criteria can only test for marginal independencies. Third, we introduce another free parameter (the number of regression coefficients to be set to zero) which can be determined from a held-out set but will require time consuming trial and error techniques. Nevertheless, they may lead to better discrimination between speakers.

The second approach we followed was one based on an ML fashion which may not be optimum for classification tasks, but can assign a different structure for each component. We used the structural EM (Friedman, 1997), (Thiesson et al., 1998) and adopt it for the case of mixtures of Gaussians. Structural EM is an algorithm that generalizes on the EM algorithm by searching in the combined space of structure and parameters. One approach to the problem of structure finding would be to start from the full model, evaluate every possible combination of arc removals in every Gaussian, and pick the ones with the least decrease in likelihood. Unfortunately, this approach can be very expensive since every time we remove an arc on one of the Gaussians we have to re-estimate all the parameters, so the EM algorithm must be used for

each combination. Therefore, this approach *alternates* parameter search with structure search and can be very expensive even if we follow greedy approaches. On the other hand, structural EM *interleaves* parameter search with structure search. Instead of following the sequence $Estep \rightarrow Mstep \rightarrow$ structure search, structural EM follows $Estep \rightarrow$ structure search $\rightarrow Mstep$. By treating expected data as observed data, the scoring of likelihood decomposes and therefore local changes do not influence the likelihood on other parameters. In essence, structural EM has the same core idea as standard EM. If M is the structure, Θ are the parameters and n is the iteration index, then the naive approach would be to do:

$$\{M^n, \Theta^n\} \rightarrow \{M^{n+1}, \Theta^{n+1}\} \quad (8)$$

On the other hand, structural EM follows the sequence:

$$\{M^n, \Theta^n\} \rightarrow \{M^{n+1}, \Theta^n\} \rightarrow \{M^{n+1}, \Theta^{n+1}\} \quad (9)$$

If we replace M with H, i.e. the hidden variables or sufficient statistics, we will recognize the sequence of steps as the standard EM algorithm. For a more thorough discussion of structural EM, the reader is referred to (Friedman, 1997). The paper in (Friedman, 1997) has a general discussion on the structural EM algorithm for an arbitrary graphical model. In this paper, we introduced a greedy pruning algorithm with step size K for mixtures of Gaussians. The algorithm is summarized in Table 2. One thing to note about the scoring criterion is that it is local, i.e. zeroing regression coefficient m, i, j will not involve computations on other parameters.

4 Experiments

We evaluated our approach in the male subset of the 1996 NIST speaker recognition task (Przybocki and Martin, 1998). The problem can be described as following. Given 21 target speakers, perform 21 binary classifications (one for each target speaker) for each one of the test sentences. Each one of the binary classifications is a YES if the sentence belongs to the target speaker and NO otherwise. Under this setting, one sentence may be decided to have been generated by more than one speaker, in which case there will be at least one false alarm. Also, some of the test sentences were spoken by non-target speakers (impostors) in which case the correct answer would be 21 NO. All speakers are male and the data are from the Switchboard database (Godfrey et al., 1992). There are approximately 2 minutes of training data for each target speaker. All the training data for a speaker come from the same session and the testing data come from different sessions, but from the same handset type and phone number (matched conditions). The algorithms were evaluated on sentence sizes of three and thirty seconds. The features are 20-dimensional MFCC vectors, cepstrum mean normalized and with all silences and pauses removed. In the test data there are impostors who don't appear in the training data and may be of different gender than the target speakers.

A mixture of Gaussians is trained on each one of the target speakers. For impostor modeling, a separate model is estimated for each gender. There are 43 impostors for each gender, each impostor with 2 minutes of speech. Same-gender speakers are pooled together and a mixture of 100 diagonal Gaussians is estimated on each pool. Impostor models remained fixed for all the experiments reported in this work. During testing and because some of the impostors are of different gender than the target speakers, each test sentence is evaluated against both impostor models and the one with the highest log-likelihood is chosen. For each test sentence the log-likelihood of each target speaker's model is subtracted from the log-likelihood of the best impostor model. A decision for YES is made if the difference of the log-likelihoods is above a threshold. Although in real operation of the system the thresholds are parameters that need to be estimated from the training data, in this evaluation the thresholds are optimized for the current test set. Therefore the results reported should be viewed as a best case scenario, but are nevertheless useful for comparing different approaches.

The metric used in all experiments was Equal Error Rate (EER). EER is defined as the point where the probability of false alarms is equal to the probability of missed detections. Standard NIST software tools were used for the evalution of the algorithms (Martin et al., 1997).

It should be noted that the number of components per Gaussian is kept the same for all speakers. A scheme that allowed for different number of Gaussians per speaker did not show any gains. Also, the number of components is optimized on the test set which will not be the case in the real operation of the system. However, since there are only a few discrete values for the number of components and EER was not particularly sensitive to that parameter, we do not view this as a major problem.

Table 3 shows the EER obtained for different baseline systems. Each cell contains two EER numbers, the left is for 30-second test utterances and the right for 3-second. For the **Diagonal** case 35 components were used, while for the **full** case 12 components were used. The **Random** case corresponds to randomly zeroing 10% of the regression coefficients of a mixture of 16 components. This particular combination of number of parameters pruned and number of components was shown to provide the best results for a subset of the test set. All structure-finding experiments are with the same number of components and percent of regression coefficients pruned.

Table 4 shows the EER obtained for different baseline

Algorithm: Finding both structure and parameter values using structural EM

Start with the full model for a given number of Gaussians

while (number of pruned regression coefficients $< T$)
$\mathbf{E-step}$: Collect sufficient statistics for given structure, i.e, $\gamma_m(n) = p(z_n = m|x_n, M_{old})$

$\mathbf{StructureSearch}$: Remove one arc from a Gaussian at a time, i.e. set $b^m_{i,j} = 0$.
The score associated with zeroing a single regression coefficient is.
$Score_{m,i,j} = 2D^i_m b^m_{i,j} \sum_n^N \gamma_m(n)\tilde{x}^j_{n,m}(\tilde{x}^i_{n,m} - B^i_m \tilde{x}_{n,m}) + D^i_m (b^m_{i,j})^2 \sum_n^N \gamma_m(n)\tilde{x}^j_{n,m}$
Order coefficients in ascending order of score. P is the set of the first K coefficients.
Set the new structure M_{new} as $M_{new} = M_{old} \backslash \{P\}$.

$\mathbf{M-step}$: Calculate the new parameters given M_{new}.
This step can be followed by a number of EM iterations to obtain better parameter values.
end

Table 2: The Structural EM algorithm for a mixture of Gaussians

Full	Diagonal	Random
6.3/10.3	5.6/9.0	6.3/10.3

Table 3: Baseline EER, left number is for 30-second test utterances and right number for 3-second

sparse structures. **SEM** is structural EM. The first column is zeroing the pairs with the minimum values of the corresponding criterion and the second column is zeroing the pairs with the maximum values. The second column is more of a consistency check. If the min entry of criterion A is lower than the min entry of criterion B then the max entry of criterion A should be higher than the max entry of criterion B. For the structural EM, pruning step sizes of 50 and 100 were tested and no difference was observed.

	min	max
MI	6.3/10.0	6.6/10.6
DMIimp	5.9/9.0	6.6/10.3
MIimp	6.3/10.3	6.3/10.3
DMIconf	5.9/9.3	6.6/10.3
SEM	6.3/9.6	6.6/10.3

Table 4: EER for different sparse structures, left number is for 30 second test utterances and right number for 3-second.

From Table 4 we can see improved results from the full-covariance case but results are not better than the diagonal-covariance case. All criteria appear to perform similarly. Table 4 also shows that zeroing the regression coefficients with the maximum of each criterion function does not lead to systems with much different performance. Also from Table 3 we can see that randomly zeroing regression coefficients performs approximately the same as taking the minimum or maximum. These numbers, seem to suggest that the structure of a mixture of Gaussians is not a critical issue for speaker recognition, at least with the current structure-estimation criteria used.

5 Summary-Future work

In this work the problem of estimating sparse regression matrices of mixtures of Gaussians was addressed. Different structure-estimation criteria were evaluated, both discriminative and generative. The general problem of finding the optimum structure of a mixture of Gaussians has direct applications in speaker identification as well as speech recognition.

Interesting connections can be drawn with Maximum Likelihood Linear Regression (MLLR) speaker adaptation (Leggetter and Woodland, 1995). Not surprisingly, the estimation equations for the regression matrix bare resemblance with the MLLR equations. However, researchers have thus far barely looked into the problem of structure-finding for speaker adaptation, focusing mostly on parameter adaptation. An interesting new topic for speaker adaptation could be joint structure and parameter adaptation.

References

J. Bilmes. 1999. *Natural Statistical Models for Automatic Speech Recognition*. Ph.D. thesis, U.C. Berkeley, Dept. of EECS.

J. Bilmes. 2000. Factored sparse inverse covariance matrices. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

D.M. Chickering, D. Geiger, and D.E. Heckerman. 1994. Learning bayesian networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft Research.

A. P. Dempster. 1972. Covariance selection. *Journal of Biometrics*, 28:157–75.

N. Friedman. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proc. 14th International Conference on Machine Learning (ICML)*, pages 125–133.

M. Gales. 1999. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7:272–281.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research development. In *Proceedings of ICASSP*, pages 517–520.

C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. 1997. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech '97*, pages 1895–1898.

M. Meila and M. I. Jordan. 2000. Learning with mixtures of trees. *Journal of Machine Learning Research*, 1:1–48.

M. Przybocki and A. Martin. 1998. NIST speaker recognition evaluations:1996-2001. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 331–335.

B. Thiesson, C. Meek, D. Chickering, and D. Heckerman. 1998. Learning mixtures of dag models. Technical Report MSR-TR-97-30, Microsoft Research.