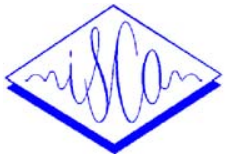

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

GMM SAMPLE STATISTIC LOG-LIKELIHOODS FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Michael Schmidt *John Golden* *Herbert Gish*

BBN Systems and Technologies
70 Fawcett St., Cambridge, MA 02138 USA
mschmidt@bbn.com

ABSTRACT

A novel approach to scoring Gaussian mixture models is presented. Feature vectors are assigned to the individual Gaussians making up the model and log-likelihoods of the separate Gaussians are computed and summed. Furthermore, the log-likelihoods of the individual Gaussians can be decomposed into sample weight, mean, and covariance log-likelihoods. Correlation likelihoods can also be computed. The results of the various systems are comparable on text-independent speaker recognition experiments despite the fact that the models and scoring are all quite different. By decomposing log-likelihoods of models into various sample statistic log-likelihoods, it is possible to diagnose which part of the model has the greatest discriminative power, whether the location of the Gaussians or their shapes.

10.21437/Eurospeech.1997-288

1. Introduction

Text-independent systems using Gaussian Mixture Models (GMMs) to model speakers are state-of-the-art [1]. In such systems, the candidate speaker with the maximum log-likelihood is identified, alternatively, if a log-likelihood is above a certain threshold the speaker is verified. In this article, generalizations to the GMM likelihood computation are presented. In particular, cepstra are (probabilistically) mapped to component mixture model Gaussians. Log-ikelihoods for each mixture model Gaussian are then computed separately and combined. Not only is the match between individual Gaussians in training and testing measured, the scores are further decomposed into GMM weights, means, and covariance log-likelihoods. Being able to determine the components of the GMM which have the greatest discriminative power is useful as a diagnostic tool. Furthermore, the various Gaussian statistic log-likelihoods can be weighted and combined, giving an increased flexibility in scoring. Experimental results on Switchboard are presented, showing results are comparable with standard log-likelihood computation systems.

2. Standard GMM Likelihood

Given a GMM, $\sum_{i=1}^{N} N(\cdot;\mu_i,\Sigma_i)$, the formula for computing the log-likelihood of a collection $X = \{x_1,\ldots,x_j,\ldots,x_n\}$ of test frames is

$$\ell(X;\alpha_i,\mu_i,\Sigma_i) = \sum_{j=1}^{n} \log\Big[\sum_{i=1}^{N} \alpha_i N(x_j;\mu_i,\Sigma_j)\Big]. \quad (1)$$

The log-likelihood of each frame is computed individually and then the log-likelihoods for all frames are summed.

3. Individual Gaussian Log-Likelihood Sums.

We now propose an alternative scoring algorithm based on assigning test frames to individual model mixture components. Frames are assigned in a probabilistic manner according to the posterior probabilities c_{ij} :

$$c_{ij} = P(i|x_j) = \frac{\alpha_i N(x_j;\mu_i,\Sigma_i)}{\sum_k \alpha_k N(x_j;\mu_k,\Sigma_k).} \quad (2)$$

The idea now is to see how well the collection of the points assigned to a particular Gaussian matches the Gaussian: The log-likelihood of the assigned collection is computed given the single Gaussian. These single Gaussian log-likelihoods are then summed to get a score for the mixture model. Note that this new pseudo-log-likelihood is not equivalent to the standard GMM log-likelihood. If the mixture model Gaussian densities did not overlap (which we know not to be the case, otherwise the assignments would be "hard,"), then the two log-likelihood scores would be equal. However, it is believable that frames are assigned "mostly" to a few Gaussians, in which case the new score may be a reasonable approximation to the original.

The log-likelihood of data given a p-dimensional Gaussian model can be expressed in terms of the sample mean $\overline{x}$ and covariance S of the data:

$$\ell(X;\mu,\Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S) - \frac{n}{2}(\overline{x}-\mu)'\Sigma^{-1}(\overline{x}-\mu). \quad (3)$$

The sum of individual Gaussian log-likelihood scores is

$$\ell^*(X;\alpha_i,\mu_i,\Sigma_i) = \sum_i^N \tilde{\alpha}_i n \Big[\log(\alpha_i) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\mathrm{tr}(\Sigma_i^{-1}S_i) - \frac{1}{2}(\overline{x}_i-\mu_i)'\Sigma_i^{-1}(\overline{x}_i-\mu_i)\Big], \quad (4)$$

where the sample weights $\tilde{\alpha}_i$, means $\overline{x}_i$ and covariances S_i are computed from the "assigned" cepstra for each mixture Gaussian:

$$\tilde{\alpha}_i = \frac{\sum_j c_{ij}}{\sum_i \sum_j c_{ij}}, \quad (5)$$

$$\overline{x}_i = \frac{\sum_j c_{ij} x_j}{\sum_j c_{ij}}, \quad (6)$$

$$S_i = \frac{\sum_j c_{ij}(x_j-\overline{x}_i)(x_j-\overline{x}_i)'}{\sum_j c_{ij}}. \quad (7)$$

The expression in (4) is referred to as a pseudo-log-likelihood.

4. GMM Sample Statistic Log-Likelihood Score

This section extends ideas from [2] to GMMs. The likelihood of data given a Gaussian can be roughly expressed as the match of the sample mean to the model mean plus the match of the sample covariance to the model covariance. The distributions of the sample means, covariances and weights are well known [3]. In fact, the weight, mean and sample covariance pseudo-log-likelihoods modulo a few terms can be expressed as follows:

$$\sum_i^N \tilde{\alpha}_i n \log(\alpha_i), \quad (8)$$

$$-\sum_i^N \Big[\frac{1}{2}\log|\Sigma_i| + \frac{\tilde{\alpha}_i n}{2}(\overline{x}_i-\mu_i)'\Sigma_i^{-1}(\overline{x}_i-\mu_i)\Big], \quad (9)$$

and

$$-\sum_i^N \Big[\frac{\tilde{\alpha}_i n - 1}{2}\log|\Sigma_i| + \frac{\tilde{\alpha}_i n}{2}\mathrm{tr}(\Sigma_i^{-1}S_i)\Big] \quad (10)$$

Speakers can now be identified based on weights, means and covariances separately.

5. Correlation Log-Likelihoods

This approach also allows for the introduction of new sample statistic pseudo-log-likelihoods. In [4] sample correlation log-likelihoods were shown to give a gain when using single Gaussian models, especially when the handsets change between training and testing. Correlation log-likelihoods are a measure of the match between the orientation of the model and sample covariances. Specifically, in training only covariance eigenvectors are retained. Let E denote a matrix of eigenvectors for a model covariance Σ. Let S denote the sample covariance matrix. The sample correlation matrix R^* is computed in a rotated coordinate system where the eigenvectors of E are the coordinate axis:

$$R^* = D^{-1/2}(E'SE)D^{-1/2} \quad (11)$$

where D is the diagonal matrix of $E'SE$, so that variances are scaled to 1 in the rotated space. The log-likehood of R^* is then

$$\ell(R^*;E) = k_1 \log|R^*| + k_2 \quad (12)$$

for some k_1 and k_2 constant for each Gaussian. The correlation sample statistic pseudo-log-likelihoods for the GMMs are computed by summing the individual sample correlation log-likelihoods as above.

6. Experimental Results

Experimental results are reported for a subset of the NIST '96 Speaker Recognition Evaluation. (All conversations for the evaluation were taken from the Switchboard corpus, a corpus of conversational speech over the telephone.) Specifically, results are presented for the 21 male targets given two minutes of training speech from a single conversation. Scores are reported for 324 half-minute target tests. (The original evaluation included 628 target tests, but these tests included two cuts from each conversation and so are were not independent.) Each target test is also used as an impostor test for the 20 targets not speaking in that test. In total there are 324 target scores and 20x324 male impostor scores. Though the subset contains only a limited number of targets (and impostors), the experiments should give a good indication of the relative power of each of the scoring algorithms presented.

Results are presented in terms of Equal Error Rates (EERs): The accept/reject threshold is set so that the number of False Accepts (FAs) equals the number of False Rejects (FRs). In addition, results are separated according to whether target tests used the same or different phone number as in the target's training. Roughly half (159/324) of the target tests are on the same handset as training.

All systems presented employ as input 19 mel-warped cepstra and 19 derivative cepstra computed every 10ms using 20ms Hamming windows. Target log-likelihood scores are normalized by the average log-likelihood of ten highest scoring cohorts out of a collection of 85 cohorts.

Baseline Result. The equal error rates for the baseline Gaussian mixture model system as well as for full covariance systems and the new sum of separate Gaussian log-likelihoods system are reported in Table 1. Using 128 mixture terms (rather than 64 or 256) results in the best performance. Note that results on the mismatched phone number tests are significantly worse than performance in the matched case.

2m train 30s Male Test	EER Same	EER Diff
128 Diag	.075	.214
24 Full	.074	.268
24 Full, 0 Mean	.072	.274
24 Full, Sum Indiv	.068	.268

Table 1. Diag. covariances vs. full.

Full Covariance GMM Scoring. Comparing the results using full covariance models (Table 1), the 24 term mixture models give comparable performance to baseline at least on same channel tests. The number of model parameters using 24 full covariance Gaussians (18734) is roughly double the number of parameters used in the baseline system (9766.) Results using 24 terms with full covariances are better than results using 12 and 48 terms, but perhaps could be further improved by using between 12 and 24 or between 24 and 48 terms.

In addition to modeling speakers with full covariance models, modeling targets using zero mean Gaussians was also tried. All 24 Gaussians now are centered at the origin. Interestingly, results do not degrade.

Individual Gaussian Log-Likelihood Sums Results. Also in Table 1 are results for the system which sums the log-likelihoods of the individual mixture model Gaussians. Recall that feature vectors are assigned to Gaussians probabilistically. There is no degradation in performance using the novel log-likelihood score.

2m train 30s Male Test	EER Same	EER Diff
128 Diag	.075	.214
24 Full, Weight	.635	.573
24 Full, Mean	.100	.253
24 Full, Cov	.076	.282
24 Full, Sum Indiv	.068	.268
24 Full, Sum+128 Diag	.063	.196
24 Full, Cov+128 Diag	.056	.183

Table 2. EERs for GMM sample statistics.

GMM Sample Statistic Scoring The numbers presented in Table 2 indicate the performance of the various GMM sample statistic scores. Recall that the sum of the sample satistic log-likelihoods is the same as the sum of individual Gaussian log-likelihoods score. Note also that the covariance scores are best, followed by the mean scores. The weights of the Gaussian mixture models are of limited value. Perhaps given more training data, using more mixture terms, etc., would change the balance in the effectiveness of the various sample statistic log-likelihood scores. When combined with the baseline system, equal error rates improve on both same and different channel tests.

2m Train 30s Male Test	EER Same	EER Diff
128 Diag	.075	.214
128 Diag, Cep	.076	.238
128 Diag, DCep	.086	.280
128 Diag, Cep+DCep	.076	.257
24 Full, Cep Weight	.604	.566
24 Full, Cep Mean	.101	.275
24 Full, Cep Cov	.076	.274
24 Full, Cep Sum Sep	.075	.281
24 Full, DCep Weight	.514	.499
24 Full, DCep Mean	.101	.281
24 Full, DCep Cov	.063	.304
24 Full, DCep Sum Sep	.063	.305
24 Full, Sep Cep+DCep	.076	.301

Table 3. Separate Cep and DCep results.

Separate Cepstra and Difference Cepstra Models In addition to modeling cepstra and difference

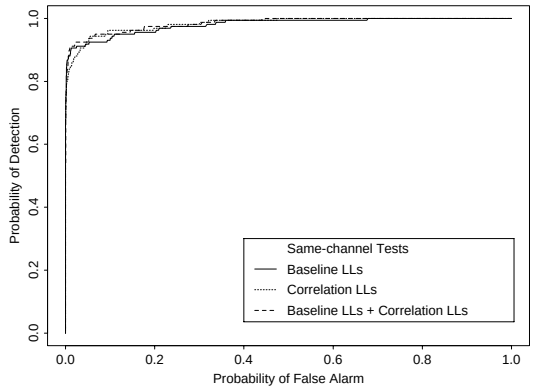

Figure 1. Comparison of baseline, correlation and combined systems on same channel tests.

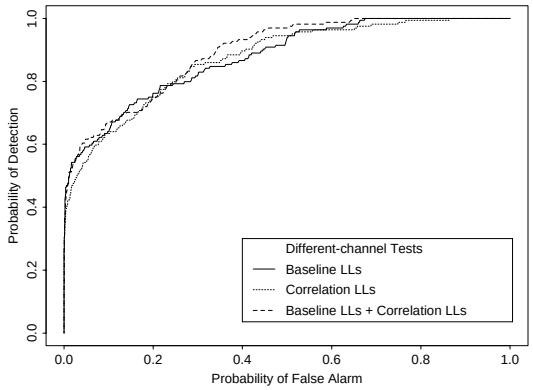

Figure 2. Comparison of baseline, correlation and combined systems on different channel tests.

cepstra jointly, they can be modeled separately. Performance on same channel tests does not degrade, however different channel results are worse when all the scores are combined. Apparently correlations between cepstra and difference cepstra are useful on different channel tests.

Correlation Log-Likelihood Results In Table 4 and Figures 1 and 2 it is demonstrated that correlation log-likelihoods results are very close to baseline results. The original motivation behind using correlations was to try to compensate for additive noise. Cepstral angles are more robust than magnitudes to additive noise [5]. For this reason correlation log-likelihoods are only applied to cepstra and not to derivative cepstra. Given that GMMs with 0 mean Gaussians perform as well as general GMMs, we looked at correlation results for 0 mean models thinking that the additive noise compensation argument might be more applicable. Results did not improve.

2m Train 30s Male Test	EER Same	EER Diff
128 Diag	.075	.214
24 Term Corr	.141	.301
12 Term Corr	.088	.252
6 Term Corr	.057	.226
6 Term Corr, 0 Mean	.083	.243
128 Diag + 6 Corr	.063	.225

Table 4. Correlation likelihood performance.

Interestingly, optimal performance is obtained with six Gaussians using only cepstra. The number of model parameters to be estimated is just 1140, nearly a tenth the number of parameters of the baseline system. Perhaps similar performance could be obtained with less training data.

7. Conclusion

A novel approach to scoring GMMs via sample statistic log-likelihoods is presented and is applied to the text-independent speaker recognition task. The method affords a greater flexibility in log-likelihood scoring. The various systems, (full and diagonal covariance systems, standard, new and sample statistic log-likelihood scoring even 0 mean models,) despite being very different all perform at roughly the same level.

8. References

[1] D. A. Reynolds "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication,* Vol 17, August 1995.

[2] H. Gish, M. Schmidt, A. Mielke "A Robust Segmental Method for Text-Independent Speaker Identification", Proc. ICASSP '94, April 1994, Adelaide, South Australia, pp. 145-148.

[3] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis,* 2nd Ed., New York, J, Wiley & Sons, 1971.

[4] M. Schmidt, A. Mielke, H. Gish "Covariance Estimation Methods for Channel Robust Text-Independent Identification", Proc. ICASSP '95, May 1995, Detroit, pp. 333-336.

[5] D. Mansour, B. H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," Proc. ICASSP '88, April 1988, New York, pp. 36-39.

GMM SAMPLE STATISTIC LOG-LIKELIHOODS FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Michael Schmidt , John Golden and Herbert Gish

BBN Systems and Technologies
70 Fawcett St., Cambridge, MA 02138 USA
mschmidt@bbn.com

A novel approach to scoring Gaussian mixture models is presented. Feature vectors are assigned to the individual Gaussians making up the model and log-likelihoods of the separate Gaussians are computed and summed. Furthermore, the log-likelihoods of the individual Gaussians can be decomposed into sample weight, mean, and covariance log-likelihoods. Correlation likelihoods can also be computed. The results of the various systems are comparable on text-independent speaker recognition experiments despite the fact that the models and scoring are all quite different. By decomposing log-likelihoods of models into various sample statistic log-likelihoods, it is possible to diagnose which part of the model has the greatest discriminative power, whether the location of the Gaussians or their shapes.