DISTANCE-BASED GAUSSIAN MIXTURE MODEL FOR SPEAKER RECOGNITION OVER THE TELEPHONE

R. D. Zilca
Research and Development Division
Amdocs Israel
8 Hapnina St. Raanana, ISRAEL
ranzilca@ieee.org

Y. Bistritz
Department of Electrical Engineering
Tel Aviv University
Tel Aviv 69978, ISRAEL
bistritz@eng.tau.ac.il

ABSTRACT

The paper considers text independent speaker identification over the telephone using short training and testing data. Gaussian Mixture Modeling (GMM) is used in the testing phase, but the parameters of the model are taken from clusters obtained for the training data by an adequate choice of feature vectors and a distance measure without optimization in the maximum likelihood (ML) sense. This distance-based GMM (DB-GMM) approach was evaluated by experiments in speaker identification from short telephone-speech data for a few feature vectors and distance measures. The selected feature vectors were Line Spectra Pairs (LSP) and Mel Frequency Cepstra (MFC). The selected distance measures were weighted Euclidean distance with IHM and BPL, respectively. DB-GMM showed consistently better performance than GMM trained by the expectation-maximization (EM) algorithm. Another notable observation is that a full covariance GMM (that is more comfortably trained by DB-GMM) always achieved significantly better performance than diagonal covariance GMM.

1. INTRODUCTION

Gaussian Mixture Modeling (GMM) provides a good approach to text-independent speaker recognition [1]. The testing phase is done by Maximum-Likelihood (ML) classification; therefore the trained parameters should be optimized in the ML sense. To meet this requirement, the Expectation-Maximization (EM) algorithm [2] was proposed to estimate the parameters of the Gaussian mixture probability density function in the training process [3]. However, in many practical situations too little available data causes the EM iterations to become ill conditioned or over fitted to the trained data. The problem is partly alleviated by limiting the model to diagonal covariance matrices rather than full covariance matrices.

The approach that we call Distance-Based GMM proposes to obtain parameters for the GMM model without optimization in the ML sense. Instead, the relative population, centers and spread of clusters of ("well chosen") feature vectors obtained from the training data by a ("well chosen") distance measure are used as the weights, average, and covariance (respectively) of the GMM. A VQ trained GMM was applied before to speaker verification from clean speech in [4] and shown to achieve almost the same speaker verification as a GMM trained by the EM algorithm. The somewhat heuristic separation of the principle of optimality used at the training and the testing processes may be justified in cases when it is not necessarily clear that the EM algorithm converges to the GMM that provide the best speaker recognition performance. For instance, when recognition is requested from a short duration of training data, the EM iterations may become ill conditioned and not converge to a numerically robust (especially a full covariance) GMM. Also, in a realistic application, when the speaker's voice is available via a communication channel, the training by EM iterations may reduce recognition rate by over-fitting the GMM to the specifics of the channel. Therefore, a DB-GMM not only guarantees a simpler training than the EM algorithm but it may also provide competing performance to EM-GMM for recognition of speakers over the telephone (and other communication channels) using short training data.

The paper reports some experiments in closed set speaker recognition task that used short duration training and testing of telephone speech in order to explore the performance of distance-based GMM (DB-GMM) with a few selected distance measures and feature vectors. The selected vectors were Mel Frequency Cepstra (MFC), the most commonly used feature vector for this model, and Line Spectra Pairs (LSP) that showed good performance in simple speaker identification schemes in [5]. The selected distance measures were the Euclidean and weighted-Euclidean with inverse harmonic measure (IHM) (for LSP) and Band Passed Liftering weighting (for MFC). The experiments revealed encouraging results in which DB-GMM consistently performed equally well or better than EM-GMM in comparable model architectures. Also, the full covariance GMM (that is more easily trained by the distance based approach) always outperformed diagonal GMM in fair comparisons that take into account the number of parameters that participate in a model.

2. DISTANCE BASED GAUSSIAN MIXTURE MODEL (DB-GMM)

A Gaussian mixture speaker model consists of a weighted sum of M Gaussian probability density functions, and is given by [1]:

6th International Conference on Spoken Language Processing, ICSLP 2000, 16-20 October 2000, Beijing, China.

$$p(x \mid \lambda) = \sum_{m=1}^{M} w_i \, p_i(x)$$

where x is a n-dimensional feature vector, w_i are the mixture weights $\sum_{m=1}^{M} w_m = 1$, and $p_m(x),\ m = 1,...,M$, are the (Gaussian) probability density functions. Each branch density is a Gaussian function of the form

$$p_m(x) = \frac{1}{(2\pi)^{n/2} \mid \Sigma_m \mid^{1/2}} e^{-\frac{1}{2}(x-\mu_m)' \Sigma_m^{-1} (x-\mu_m)}$$

where μ_m is the mean vector of length n and Σ_m is the n x n covariance matrix.

The GMM is represented by the set of parameters λ that includes the means vectors, covariance matrices and mixture weights:

$$\lambda = \{ w_m, \mu_m, \Sigma_m \quad ; \quad m = 1,...,M \ \}$$

During recognition, a sequence of N feature vectors $X = x_1, x_2, ..., x_N$ that were extracted from speech frames are presented to the GMM classifier, and the likelihood for each speaker is calculated using an assumption of statistical independence between frames as follows:

$$L(X \mid \lambda) = \log p(X \mid \lambda) = \sum_{j=1}^{N} \log p(x_j \mid \lambda)$$

At training the chosen distance measure is used to obtain from the feature vectors $T = \{v_1, v_2, \ldots, v_{n(T)}\}$ extracted from the speaker's speech M clusters $T_m = \{v_1^{(m)}, v_2^{(m)}, \ldots, v_{n(T_m)}^{(m)}\}$ with centers c_m, $m = 1,...,M$. Then, the GMM parameters are determined from the clusters as follows

$$\mu_m = c_m \quad , \quad w_m = \frac{n(T_m)}{n(T)}$$

$$\Sigma_m = \frac{1}{n(T_m) - 1} \sum_{i=1}^{n(T_m)} (v_i^{(m)} - \mu_m)(v_i^{(m)} - \mu_m)^t$$

where $n(T)$ denotes the number of elements in the set T.

3. EXPERIMENTS AND EVALUATION

3.1. Experiment Settings

The database was extracted from the NTIMIT [6]. The NTIMIT includes speakers from 8 different American dialects, male and female. For each speaker, there are 10 sentences, as follows:

- 2 "sa" sentences which include identical text for all of the speakers.
- 3 "si" sentences which include a different textual context for all of the speakers (the "phonetically diverse" sentences).
- 5 "sx" sentences which are different for some of the speakers.

We selected a subset of the NTIMIT database, that includes 32 male speakers, 4 from each dialect. This selection was intended to span the various American dialects, while restricting speaker diversity (all males, 4 speakers per dialect). The 2 "sa" sentences, 3 "si" and 2 of the "sx" sentences were used for training. This results in a training session with a length of approximately 12 to 20 seconds, depending upon the speaker's typical rate of speech. The remaining 3 "sx" sentences were used for testing, resulting in 96 test utterances (3 for each speaker), each of them approximately 1 to 3 seconds long.

The same preprocessing procedure was used for training and for testing. The speech samples were downsampled to 8KHz sampling rate, and segmented into 25ms frames with 50% overlap. A Hamming window then multiplies the speech samples of each frame, and only voiced frames are selected for further use.

3.2. Distance Measures and Features

We selected for our examination two feature vectors: Mel Frequency Cepstra (MFC), and Line Spectra Pairs (LSP). For LSP we used 10th order LPC analysis. The resulting LPC filter was then subject to a 15Hz bandwidth expansion and the LSP frequencies [7] were derived. 18th order Mel Cepstra Coefficients (MFC) were derived using the triangle filter-banks suggested by [8]. MFC was chosen because it is the most commonly used feature for speaker recognition today. LSP is the most widely used feature in speech coding. It was selected because we found it recently to be useful also for speaker recognition [5]. We used it with an Inverse Harmonic Measure (IHM) weighting that was proposed to improve low bit rate coding with LSP in [9]. For a n dimensional LSP vector, $(x_1, x_2, \ldots, x_n)$, the IHM weights

$$\omega_k = \frac{1}{x_k - x_{k-1}} + \frac{1}{x_{k+1} - x_k} \quad , \quad k = 1, \ldots, n$$

where $x_0 = 0$ and $x_{n+1} = f_{sampling} / 2$ were used in a weighted Euclidean distance. The IHM distance measure was shown to be a first order approximation of the log-spectral distortion [10]. For the MFC features of dimension n, we used weighted Euclidean distance with Band-Pass Liftering (BPL) weights

$$\omega_k = 1 + \frac{k}{2} \sin\left(\frac{k\pi}{k}\right) \quad , \quad k = 1, \ldots, n$$

The identification experiments were performed modeling the speakers by (plain) VQ, by DB-GMM and by EM-GMM. Both full covariance and diagonal covariance models were considered.

3.3. Identification Performance

The identification rates that were obtained are presented in Tables 1 and 2 for LSP and MFC respectively. The results are also illustrated in Figures 1 and 2. Note that the tables are indexed by the order of the models that were used (number of codebook vectors for VQ and number of Gaussians for GMM), while in the figures the performance is displayed with respect to the number of parameters that constitute a speaker model. (For example, an LSP based diagonal 4^{th} order GMM, the models consist of 4 means and 4 variances for each of the 10 LSP parameters plus 4 weights resulting in a total of 84 parameters.) This display reveals more directly how different architectures exploit the parameter budget.

The results indicate that for a given parameter budget, for both LSP and MFC, the best performance is achieved when full covariance GMM models are used. Also it is seen that DB-GMM consistently outperforms the EM-GMM when identical GMM architecture and features are compared. This observation is significant because distance based training of GMM is always simpler than its training by EM. The training of full covariance GMM by EM algorithm is more problematic than the diagonal model. By contrast, the training of DB-GMM is equally simple for full and diagonal covariances. We also draw attention to that for LSP, all VQ models outperform all diagonal GMM's, suggesting that the off-diagonals of the speakers' covariance matrices contribute significantly to speaker discrimination while the contribution of LSP variances seems to be negligible. No similarly unambiguous conclusions can be drawn in comparing VQ models and diagonal GMM models when they use MFC as feature vector. Also it is seen that the IHM and BPL weighted distance measure based GMM perform better for most orders than same models with plain Euclidean distance based GMM.

VQ - Euclidean	**41.7**		**50.0**		**59.4**		**65.6**		**63.5**	
	40	4	80	8	160	16	320	32	640	64
VQ - IHM	**39.6**		**53.1**		**64.6**		**69.8**		**68.8**	
	40	4	80	8	160	16	320	32	640	64
Euclidean DB-GMM/ Diagonal	**49.0**		**57.3**		**61.5**		-		-	
	84	4	168	8	336	16	672	32	1344	64
Euclidean DB-GMM/ Full	**76.0**		**75.0**		**78.1**		-		-	
	264	4	528	8	1056	16	2112	32	4224	64
EM-GMM - Diagonal	**46.9**		**55.2**		**60.4**		-		-	
	84	4	168	8	336	16	672	32	1344	64
DB-GMM (IHM) - Diagonal	**50.0**		**58.3**		**61.5**		**60.4**		-	
	84	4	168	8	336	16	672	32	1344	64
EM-GMM - Full	**74.0**		**72.9**		**76.0**		-		-	
	264	4	528	8	1056	16	2112	32	4224	64
DB-GMM (IHM) - Fulll	**72.9**		**79.2**		**74.0**		-		-	
	264	4	528	8	1056	16	2112	32	4224	64

Table 1. Identification Rates, LSP. (Identification percents are in bold figures at the upper part of each block; the participating number of parameters and number of Gaussians/codebook size in the lower part of each block.)

VQ - Euclidean	**39.6**		**47.9**		**54.2**		**67.7**		**74.0**	
	72	4	144	8	288	16	576	32	1152	64
VQ - BPL	**38.5**		**47.9**		**59.4**		**64.6**		**72.9**	
	72	4	144	8	288	16	576	32	1152	64
Euclidean DB-GMM/ Diagonal	**63.5**		**65.6**		**67.7**		**70.8**		-	
	148	4	296	8	592	16	1184	32	2368	64
Euclidean DB-GMM/ Full	**82.3**		**82.3**		-		-		-	
	760	4	1520	8	3040	16	6080	32	12160	64
EM-GMM - Diagonal	**63.5**		**65.6**		**68.8**		-		-	
	148	4	296	8	592	16	1184	32	2368	64
DB-GMM (BPL) - Diagonal	**61.5**		**68.8**		**69.8**		-		-	
	148	4	296	8	592	16	1184	32	2368	64
EM-GMM - Full	**77.1**		**77.1**		-		-		-	
	760	4	1520	8	3040	16	6080	32	12160	64
DB-GMM (BPL) - Fulll	**84.4**		**80.2**		-		-		-	
	760	4	1520	8	3040	16	6080	32	12160	64

Table 2. Identification Rates, MFC. Same format as in Table 1.

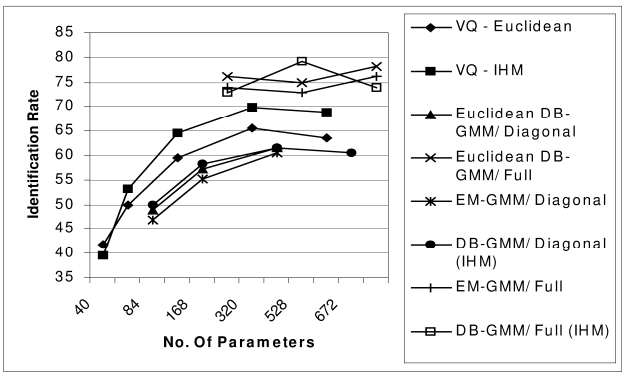

Figure 1. Identification Rates, LSP

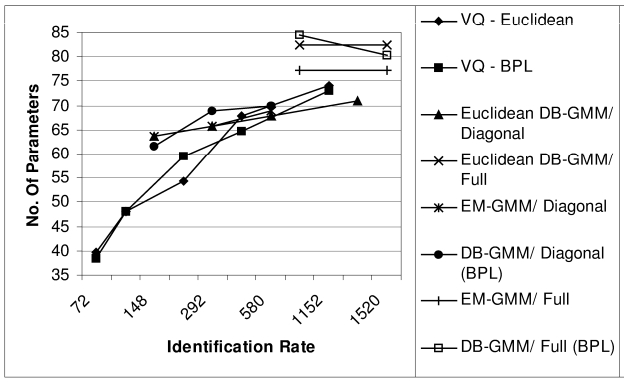

Figure 2. Identification Rates, MFC

Considering the fact that DB-GMM, that is not optimized at training in the ML sense, is used by the classifier at the testing phase, it is expected that DB-GMM would trade some decrease in performance for simpler training by comparison with EM-GMM. In this view, the observation from our experiments - that DB-GMM outperforms the EM-GMM - is surprising. Some degradation in performance for DB-GMM compared to EM-GMM has been observed in complementary experiments with clean speech (with TIMIT replacing the NTIMIT) not included in this report. It was also observed in the experiments with clean

speech in [4]. Thus, the better performance of DB-GMM with telephone speech may be attributed to that the EM iterations, more than a DB training, causes the GMM to model not just the speaker but also the phone channel that changes in the testing session.

4. CONCLUDING REMARKS

A new approach for GMM speaker models, termed DB-GMM, has been introduced and its viability assessed by some preliminary experiments in closed set speaker recognition from telephone speech with very short training and testing sessions. In this approach, ML classifier does the speaker recognition but at training, the model is not optimized in the maximum-likelihood sense. Instead parameter values for the mixture of Gaussians are obtained from clusters formed using judicious choices of feature vectors and distance measures. In our experiments with a few selected features (LSP and MFC) and distance measures, DB-GMM consistently outperformed the EM-GMM.

The good performance of DB-GMM indicates that the more significant contribution to the performance of the GMM recognizer comes from the smoothness of the probabilistic classifier during the recognition tests. The observed better performance of DB-GMM implies that for telephone quality speech the EM iterations do not converge to the minimal classification error. A possible explanation may be that the EM training causes an over-fitting of the model to not just the speaker but also the phone line characteristics that varies from session to session.

A span of approximately 20% between the poorest and best identification rates for a given number of model parameters demonstrate the crucial trade-off in selecting the model's architecture. It is demonstrated that the best deployment of a given number of parameters is to use a full covariance GMM. A full covariance model is trained by DB-GMM more easily than by EM-GMM. The best identification rate was achieved by a full covariance DB-GMM using MFC. It presents an about 16% improvement over the baseline performance of the EM trained diagonal GMM with a comparable number of parameters.

Further study, in particular a more complete search for better performing pair of feature vectors and a distance measures (newly proposed or adopted from the existing speech processing arsenal), is needed to exploit more fully the capacity of the DB-GMM concept for speaker recognition in various settings.

REFERENCES

[1] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, pp. 91-108, 1995.

[2] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM-Algorithm," *J. Royal Stat. Soc, Vol. 39, pp. 1-38, 1977.*

[3] D. A. Reynolds and R. C. Rose ``Robust Text-Independent Speaker Identification using Mixture Speaker Models", *IEEE Trans. Speech Audio Process.,* Vol . 3 pp. 72-83, 1995.

[4] G. Kolano and P. Regel-Brietzmann, "Combination of vector quantization and Gaussian mixture models for speaker verification with sparse training data" *Proc. European Conf. Speech Comm. Tech., Eurospeech'99,* pp. 1203-1206, Budapest, 1999.

[5] R. D. Zilca and Y. Bistritz "Text independent speaker identification using LSP codebook speaker models and linear discriminant analysis", *Proc. European Conf. Speech Comm. Tech., Eurospeech'99,* pp. 799-802, Budapest, 1999.

[6] C. R. Jankowski, A. Kalyanswamy, S. Basoon, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc., Albuquerque*, NM, pp. 109-112, 1990

[7] F. K. Soong and B. H. Juang, "Line Spectrum Pair and Speech Data Compression," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.,* pp. 1.10.1-1.10.4, 1984

[8] S.B. Davis and P. Marmelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, Aug. 1980.

[9] R. Laroia, N. Phamdo, and N. Farvaradin, "Robust and Efficient quantization of Speech LSP Parameters using Structured Vector Quantizers," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc*, pp. 641-644, 1991.

[10] W. R. Gardner and Bhaskar D. Rao, "Theoretical Analysis of the High Rate Vector Quantization of LPC Parameters," *IEEE Trans. Speech. Audio. Proc, Vol. 3, No. 5, Sept. 1995.*