Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis

Luo Si[1] and Rong Jin[2]

[1] School of Computer Science, Carnegie Mellon University.
5000 Forbes Ave, Pittsburgh PA, U.S.A.
lsi@cs.cmu.edu
[2] Department of Computer Science and Engineering, Michigan State University
East Lansing, MI, U.S.A.
rongjin@cse.msu.edu

Abstract. Mixture models, such as Gaussian Mixture Model, have been widely used in many applications for modeling data. Gaussian mixture model (GMM) assumes that data points are generated from a set of Gaussian models with the same set of mixture weights. A natural extension of GMM is the probabilistic latent semantic analysis (PLSA) model, which assigns different mixture weights for each data point. Thus, PLSA is more flexible than the GMM method. However, as a tradeoff, PLSA usually suffers from the overfitting problem. In this paper, we propose a regularized probabilistic latent semantic analysis model (RPLSA), which can properly adjust the amount of model flexibility so that not only the training data can be fit well but also the model is robust to avoid the overfitting problem. We conduct empirical study for the application of speaker identification to show the effectiveness of the new model. The experiment results on the NIST speaker recognition dataset indicate that the RPLSA model outperforms both the GMM and PLSA models substantially. The principle of RPLSA of appropriately adjusting model flexibility can be naturally extended to other applications and other types of mixture models.

1 Introduction

Mixture models, such as Gaussian Mixture Model, have been widely used throughout the applications of data mining and machine learning. For example, Gaussian Mixture model (GMM) has been applied for time series classification [8], image texture detection [7] and speaker identification [9]. In these tasks, the GMM model assumes that data points from a specific object or class (e.g., a speaker in speaker identification) are generated from a pool of Gaussian models with fixed mixture weights; it estimates mixture models from the training data using a maximum likelihood method; it predicts test data with the classes that generate the test data with the largest probabilities.

One general problem of modeling data with GMM is that GMM uses the same set of mixture weights for all the data points of a particular class, which limits the power of the mixture model in fitting the training data accurately. In contrast, a probabilistic latent semantic analysis (PLSA) [5][6] model allows each data point to choose its own

mixture weights. Apparently, PLSA model is more flexible than GMM model in that a different set of mixture weights is introduced for each data point. However, as a trade-off, PLSA has a substantially larger parameter space than the GMM model; the excessive freedom of assigning data point dependant mixture weights invites the PLSA model to the potential overfitting problem given the limited amount of training data.

In this paper, we propose a *regularized probabilistic latent semantic analysis* (RPLSA) model that addresses the overfitting problem in PLSA by regularizing the mixture weights. In particular, a regularization term is introduced in RPLSA, which punishes the objective function in RPLSA when different data points of the same class choose mixture weights that are far away from each other. It is an intermediate model between GMM and PLSA: different mixture weights are allowed for data points; but similar mixture weights are favored for different data points in the same class.

Empirical study for the application of speaker identification was conducted to show the effectiveness of the new RPLSA model. The NIST 1999 speaker recognition evaluation dataset with 539 speakers were used and the experiment results indicate that the RPLSA model achieves better results than both the GMM and PLSA. Furthermore, careful analysis shows that the advantage of RPLSA comes from the power of properly adjusting model flexibility.

2 Previous Research of Mixture Model

In this section, we only survey the most related research of mixture model.

2.1 Gaussian Mixture Model

GMM is one of the most widely used mixture modeling techniques [4][7][8][9]. It is a simple model and is reasonably accurate when data are generated from a set of Gaussian distributions. Let $X^i = \{x_t, 1 \le t \le T^i\}$ denote the feature vectors for data points from the ith class (e.g., a particular speaker). They are modeled by a total number of J Gaussians as follows:

$$P(X^i \mid \theta^i_{GMM}) = \prod_{t=1}^{T^i} \sum_{j=1}^{J} P(z_j) P_{z_j}(x_t \mid u_j, \Sigma_j) \tag{1}$$

where θ^i_{GMM} includes all the model parameters, i.e., $\{ P(z_j), u_j, \Sigma_j, 1 \le j \le J \}$. $P_{z_j}(x_t \mid u_j, \Sigma_j)$ is the Gaussian distribution for the j-th class, with a mean vector u_j and a covariance matrix Σ_j as:

$$P_{z_j}(x_t \mid u_j, \Sigma_j) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_j\right|^{1/2}} \exp\{-\frac{1}{2}(x_t - u_j)^T \Sigma_j^{-1} (x_t - u_j)\} \tag{2}$$

where D is the dimension of the feature vector x_t. Usually Σ_j is set to be a diagonal matrix as $diag\{\sigma^2_{jd} : 1 \le d \le D\}$ in order to reduce the size of the parameter space [4].

It can be seen from Equation (1) that the data points of a specific class are generated from multiple Gaussian models with an identical set of mixture weights (i.e., $P(z_j)$). This constraint may not be valid in the data modeling process. For example, in speaker identification, mixture weights for a vowel can be significantly different from the mixture weights for a consonant. Therefore, it is important to incorporate data point dependent mixture weights into the framework of mixture models.

2.2 Probabilistic Latent Semantic Analysis

Unlike the Gaussian Mixture Model, the probabilistic latent semantic analysis model (PLSA) allows for data point specific mixture weights. Formally, the likelihood of training data for the ith class is written as:

$$P(X^i \mid \theta^i_{PLSA}) = \prod_{t=1}^{T^i} \sum_{j=1}^{J} P(z_j \mid d_t) P_{z_j}(x_t \mid u_j, \Sigma_j) \tag{3}$$

where θ^i_{PLSA} includes $\{ u_j, \Sigma_j,\ 1 \le j \le J;\ P(z_j \mid d_t),\ 1 \le j \le J,\ 1 \le t \le T^i \}$. Note that a dummy variable d_t is introduced for every data point, and therefore the mixture weights $P(z_j \mid d_t)$ are data point dependent. The PLSA model was originally proposed for the probabilistic semantic indexing (PLSI) technique of information retrieval [5][6]. Both PLSI and PLSA allow data point specific mixture weights, but the PLSI model is based on multinomial distributions to model documents while the PLSA model is used here for modeling continuous data with Gaussian distributions. Note that the PLSA model shares the same idea with the tied-mixture model technique [1], which assumes that speech data is generated from a common pool of Gaussian models and each data point can choose its own mixture weights independently.

Because the mixture weights are data point dependent, PLSA is capable to fit training data better than GMM. However, a potential problem with PLSA is that it has a significantly larger parameter space than GMM, thus is prone to overfitting training data. To alleviate this problem, a maximum posterior (MAP) smoothing technique can be used for estimating PLSA. In particular, priors are introduced for parameters in the Gaussian models, and the parameters are estimated by maximizing the posterior of training data:

$$\begin{aligned} \log P(\theta^i_{PLSA} \mid X^i) \propto & \sum_{t=1}^{T^i} \log\Big(\sum_{j=1}^{J} P(z_j \mid d_t) P_{z_j}(x_t \mid u_j, \Sigma_j)\Big) \\ & + A \sum_{j=1}^{J} \log P(u_j \mid u_0, \Sigma_0) + B \sum_{j=1}^{J} \sum_{d=1}^{D} \log P(\sigma^2_{jd} \mid a_{0d}, \beta_{0d}) \end{aligned} \tag{4}$$

The first item on the right hand side is the likelihood of training data. The next two items are the conjugate priors for the means and variances in the Gaussian models. A and B are two constants that adjust the weights of priors. $P(u_j \mid u_0, \Sigma_0)$ is a Gaussian distribution with mean u_0 and variance Σ_0 as a diagonal matrix $diag\{\sigma^2_{0d}\}$; $P(\sigma^2_{jd} \mid a_{0d}, \beta_{0d})$ is an inverse gamma distribution with parameters a_{0d}, β_{0d} as:

$$P(\sigma_{jd}^2 \mid a_{0d}, \beta_{0d}) \propto (\sigma_{jd}^2)^{-(a_{0d}+1)} e^{\frac{-\beta_{0d}}{\sigma_{jd}^2}} \tag{5}$$

Although maximum posterior smoothing can alleviate the overfitting problem in some extent, the PLSA model still suffers from the excessive freedom of assigning totally independent data point specific mixture weights. To further address this problem, a novel method of regularizing mixture weights is proposed in this paper.

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is a generative model for collections of discrete data such as text. In LDA, each item (document) of a class (text collection) is modeled as a finite mixture over a set of topics (mixture models). LDA shares a common feature with the new research in this paper in that both of them choose moderate amount of model flexibility. LDA assumes that the mixture weights of items in a class are generated from a common Dirichlet distribution so that the weights for different data points in the same class are coupled instead of being chosen independently.

However, LDA model requires sophiscated variational methods to calculate the model parameters both in training and testing phrases, which is time consuming and thus limits the application of LDA in practical work. Furthermore, LDA model does not work well when each item contains a very small number of data points (like documents contain small number of words by average, or in speaker identification each item of a speaker utterance is a single vector of acoustic features in multi-dimensional space). Specifically consider the extreme case when each item only contains a single data point. LDA models a class X^i with single data point items as:

$$P(X^i \mid \theta_{LDA}) = \prod_{t=1}^{T^i} \int \left(P(u \mid \alpha)(\sum_{j=1}^{J} P(z_j \mid u) P_{z_j}(x_t \mid u_j, \Sigma_j)) \right) du \tag{6}$$

Where $P(u \mid \alpha)$ is the Dirichlet distribution that generates the mixture weights for all data points. By switching the order of integration and summation and integrating out the parameter u, Equation (6) becomes:

$$P(X^i \mid \theta_{LDA}) = \prod_{t=1}^{T^i} \sum_{j=1}^{J} \frac{\alpha_j}{\sum \alpha_{j'}} P_{z_j}(x_t \mid u_j, \Sigma_j) \tag{7}$$

This is essentially a GMM model if we set $\alpha_j / \sum_{j'} a_{j'}$ as the mixture weight $P(z_j)$ in the GMM model.

3 Regularized Probabilistic Latent Semantic Analysis Model

From the above research, we find that both GMM and PLSA are two extreme cases of the mixture model family: GMM uses the same set of mixture weights for all data

points of the same class, thus lacking flexibility; PLSA model allows each data point to choose its own mixture weights and therefore is prone to overfitting training data. A better idea is to develop an algorithm that can properly adjust the amount of model flexibility so that not only the training data can be fit well but also the model is robust to overfitting problems. This is the motivation of the regularized probabilistic latent semantic analysis model (RPLSA).

3.1 Model Description

Similar to the PLSA model, RPLSA allows each data point to choose its own mixture weights. At the meantime, it requires mixture weights from different data points to be similar in order to avoid overfitting. This is realized by assuming that there is a common set of mixture weights and mixture weights for different training data points should be close to the common set of mixture weights, formally as:

$$\begin{aligned}\log P(\theta^{i}_{RPLSA} \mid X^{i}) \propto & \sum_{t=1}^{T^i} \log(\sum_{j=1}^{J} P(z_j \mid d_t) P_{z_j}(x_t \mid u_j, \Sigma_j)) + A\sum_{j=1}^{J} \log P(u_j \mid u_0, \Sigma_0) \\ & + B\sum_{j=1}^{J}\sum_{d=1}^{D} \log P(\sigma^2_{jd} \mid a_{0d}, \beta_{0d}) - C\sum_{t=1}^{T^i}\sum_{j=1}^{J} P_c(z_j) \log \frac{P(z_j \mid d_t)}{P_c(z_j)}\end{aligned} \tag{8}$$

Compared to the PLSA model in Equation (4), the above equation introduces a new regularization term, i.e., $C\sum_{t=1}^{T^i}\sum_{j=1}^{J} P_c(z_j) \log \frac{P(z_j \mid d_t)}{P_c(z_j)}$, into the objective function. It is a weighted sum of the Kullback-Leibler (KL) divergence between the common mixture weights (i.e., $P_c(z_j)$) and the mixture weights that are specific to each data point (i.e., $P(z_j \mid d_t)$). C is the regularization constant that controls the amount of model flexibility.

The role of the regularization term is to enforce mixture weights for different data points to be close to each other. In general, the closer the data-dependent mixture weights are to the common set of mixture weights, the smaller the KL divergence will be. Thus, by adjusting the constant C, we are able to adjust the flexibility of the RPLSA model: A small C will lead to a large freedom in assigning different mixture weights to different data points, thus exhibiting a behavior similar to the PLSA model; A large C will strongly enforce different data points to choose similar mixture weights, thus close to the behavior of the GMM method. Therefore, the RPLSA model connects the spectrum of mixture models between GMM and PLSA.

3.2 Parameter Estimation

The Expectation-Maximization (EM) algorithm [1] is used to estimate the model parameters of the RPLSA model. In the E step, the posterior probability of which mixture model each data point belongs to is calculated as follows:

$$P'(z_j \mid d_t) = \frac{P(z_j \mid d_t)P(x_t \mid u_j, \Sigma_j = diag\{\sigma_{jd}^2\}) + CP_c(z_j)}{\sum_{j'} P(z_{j'} \mid d_t)P(x_t \mid u_{j'}, \Sigma_{j'} = diag\{\sigma_{jd}^2\}) + C} \tag{9}$$

In the M step, the $P^{new}(z_j \mid d_t)$, u_j^{new} and Σ^{new} parameters are updated using Equations (10), (11) and (12) separately.

$$P^{new}(z_j \mid d_t) = P'(z_j \mid d_t) \tag{10}$$

$$u_{jd}^{new} = \frac{\sum_{t=1}^{T^i} P'(z_j \mid d_t)\sigma_{jd}^{-2} x_{td} + A\sigma_{0d}^{-2} u_{0d}}{\sum_{t'=1}^{T^i} P'(z_j \mid d_{t'})\sigma_{jd}^{-2} + A\sigma_{0d}^{-2}} \tag{11}$$

$$\left[\sigma_{jd}^{new}\right]^2 = \frac{\sum_{t=1}^{T^i} P'(z_j \mid d_t)(x_{td} - u_{jd})^2 + 2B\beta_{0d}}{\sum_{t'=1}^{T^i} P'(z_j \mid d_{t'}) + 2B(\alpha_{0d} + 1)} \tag{12}$$

where u_{jd} and σ_{jd} are the dth element of the mean and variance respectively for the jth mixture, and x_{td} is the dth element of the feature vector x_t.

Finally, the common set of mixture weights is updated as follows:

$$P_c^{New}(z_j) \propto \exp\{1/T^i * \sum_{t=1}^{T^i} \log(P'(z_j \mid d_t))\} \tag{13}$$

which is essentially the geometric mean of the corresponding mixture weights that are attached to each data point. Note that choice of adaptively adjusting the common set of mixture weights in Equation (13) is different from the method that simply selecting a prior distribution of the mixture weights and estimating the model with maximum posterior smoothing. It can be imagined that the same set of prior of mixture weights (e.g., the Dirichlet prior distribution with uniform parameter values of the mixture weights) does not fit data with different characteristics. The adaptive estimation of the common set of mixture weights in RPLSA is a more reasonable choice.

The parameter estimation procedure for PLSA is a simplified version of that for RPLSA. In the expectation step, the posterior probability is calculated by a similar formula as Equation (9) without the factor of the regularization item. In the maximization step, the new parameters $P^{new}(z_j \mid d_t)$, u_j^{new} and Σ^{new} of PLSA are updated in a similar way as the Equations (10), (11) and (12).

3.3 Identification

The RPLSA model is different from the GMM model in that some parameters $P(z_j \mid d_t)$ need to be estimated for the test data in the identification phase. A plug-in

EM procedure is used to accomplish this. Specifically, the EM algorithm described in Section 3.2 is rerun to estimate $P(z_j \mid d_t)$ for each test data point while all the other parameters are fixed. With the estimated new mixture weights, we can identify the test item (e.g., a vector of acoustic features) for a particular class (e.g., a speaker in the training set) whose model has the largest generation probabilities of test item X^{test} as:

$$ID\_Rst(X^{test}) = \arg\max_i P(X^{test} \mid \theta^i_{RPLSA}) \tag{14}$$

The identification process of PLSA is almost the same as the procedure of RPLSA, which is not described due to space limit.

4. Experimental Results

This section shows empirical study that demonstrates the advantage of the new regularized probabilistic latent semantic model (RPLSA). Specifically, three models of GMM, PLSA and RPLSA are compared for the application of speaker identification.

4.1 Experiment Methodology

The experiments were conducted on the NIST 1999 speaker recognition evaluation dataset[1]. There are a total of 309 female speakers and 230 male speakers. The speech signal was pre-emphasized using a coefficient of 0.95. Each frame was windowed with a Hamming window and set to 30ms long with 50% frame overlap. 10 mel frequency cepstral coefficients were extracted from each speech frame. Both the training data and the test data come from the same channel. The training data consists of speech data of 7.5 seconds for each training speaker.

We present experiment results to address two issues: 1) Will the proposed RPLSA be more effective than the GMM and the PLSA models? 2) What is the power of the RPLSA model? What is the behavior of the RPLSA model with different amount of model flexibility by choosing different values for the regularization parameter C?

4.2 Experiment Results of Different Algorithms

The first set of experiments was conducted to study the effectiveness of the three mixture models. The numbers of mixture models were chosen by cross-validation for the three models. Specifically, 30 mixtures for GMM model, 50 for both PLSA and RPLSA. The smoothing prior parameters of PLSA and RPLSA were set as follows: u_0 is the mean value of the training data; Σ_0 is identity matrix; a_{0d} is 1 and β_{0d} is twice the variance of the dth value of the training data. The smoothing constants in

[1] http://www.nist.gov/speech/tests/spk/

Table 1: Speaker identification errors for the Gaussian mixture model (GMM), the probabilistic latent semantic analysis model (PLSA) and the regularized probabilistic latent semantic analysis model (RPLSA).

Test Data Length	**GMM**	**PLSA**	**RPLSA**
2 Sec	37.8%	33.9%	**31.2%**
3 Sec	31.5%	24.7%	**21.8%**
5 Sec	27.3%	22.5%	**20.1%**

Table 2: Speaker identification errors for the smoothed Gaussian mixture model (GMM), the probabilistic latent semantic analysis model (PLSA) with uniform Dirichlet prior ($\alpha = 100$) and the RPLSA model.

Test Data Length	**GMM (Smoothed)**	**PLSA (Dirichlet Prior)**	**RPLSA**
2 Sec	36.1%	33.2%	**31.2%**
3 Sec	30.2%	24.3%	**21.8%**
5 Sec	26.0%	22.3%	**20.1%**

Equations (4) and (8) were set as: A is $|T_i|/(10 * J)$ and B is $|T_i| / J$ (where $|\bullet|$ indicates the number of items within a class). The regularization constant C of RPLSA was set to be 20 by cross-validation.

To compare the algorithms in a wide range we tried various lengths of test data. The results are shown in Table1. Clearly, both PLSA and RPLSA are more effective than the GMM in all cases. This can be attributed to the fact that both PLSA and RPLSA relax the constraint on mixture weights imposed by GMM. Furthermore, the RPLSA model outperforms the PLSA model. This is consistent with the motivation of the RPLSA model as it automatically adjusts the model flexibility for better recognition accuracy.

To further confirm the hypothesis that RPLSA model has advantage than both the GMM and PLSA methods, two more sets of experiments were conducted. The first set of extended experiments was to train a GMM model with smoothed Gaussian model parameters like that used for PLSA (Two smoothed items of Gaussian model parameters like that of Equation (4) were introduced into the GMM objective function with A and B roughly tuned to be five times smaller than that of the RPLSA setting). The second set of extended experiments was to regularize the mixture weights in PLSA using a Dirichlet prior as described in Section 3.2. It is different from the regularization scheme of Equation (9) in that a Dirichlet prior uses a fixed set of common mixture weights (uniform) that is unable to adapt to the training data. The modified PLSA is trained with a new likelihood function of Equation (4) with an additional item of a Dirichlet prior with the parameter values of 100 (roughly tuned).

It can be seen from Table 2 that the new versions of GMM and PLSA give very small improvement of the original algorithms. The behavior of GMM model can be explained as that GMM has a much smaller parameter space than PLSA and RPLSA, smoothing does not give too much help. The results of the PLSA model with uniform Dirichlet prior indicates that the simple method of smoothing the mixture weights with a single prior does not successfully solve the overfitting problem.

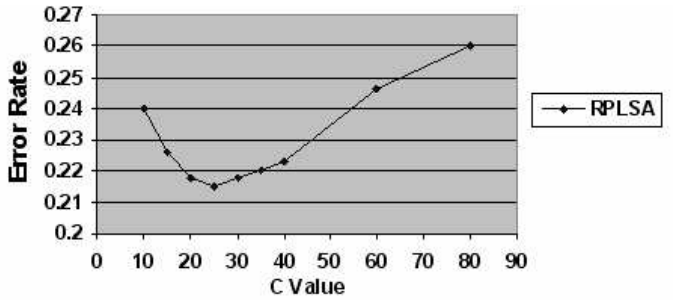

Figure 1. Behavior of RPLSA Model with Different Values of Regularization Const

4.3 Study the Behavior of the RPLSA Method

The new proposed RPLSA is an intermediate model between GMM and PLSA: different mixture weights are allowed for each data point; but similar mixture weights for different data points are encouraged. The RPLSA is the bridge to connect a spectrum of mixture models with two extreme cases of GMM and RPLSA models. Therefore, it is very interesting to investigate the behavior of the RPLSA method with different amount of model flexibly and its relationship with the GMM and RPLSA models.

Specifically, different values of parameter C in the RPLSA model of Equation (8) were investigated. 3 seconds' test data was used in this set of experiments. The detailed results are shown in Figure 1.

According to previous analysis in Section 3.1, we know that a smaller C value gives more freedom to the data points in choosing their own mixture weights, which leads to a behavior closer to that of the PLSA model. This is consistent with the observation from Figure 1. When C is as small as 10, RPLSA acquires a similar recognition accuracy with that of PLSA. On the other hand, a larger value for C makes RPLSA behave more like GMM. As indicated in Figure 1, a larger C leads to worse recognition accuracy.

For the middle part of the curve, with C ranging from 15 to 40, RPLSA acquires the best recognition accuracy; this suggests that RPLSA with reasonable amount of model flexibility reaches a better trade-off between enough model flexibility and model robustness.

The experiments in this section show the behavior of the new RPLSA model with different amount of model flexibility. It is consistent with our theoretical analysis that RPLSA has advantage than the GMM model and the RPLSA model in its better ability to adjust the appropriate amount of model flexibility.

5. Conclusion

Mixture models such as Gaussian mixture model (GMM) are very important tools for data mining and machine learning applications. However, classic mixture models like

GMM have limitations in their modeling abilities as all data points of an object are required to be generated from a pool of mixtures with the same set of mixture weights. Previous research such as the probabilistic latent semantic analysis (PLSA) model has been proposed to release this constraint. PLSA allows totally independent data point specific mixture weights. But the excessive model flexibility makes PLSA tend to suffer from the overfitting problem.

This paper proposes a new regularized PLSA (RPLSA) model: On one hand, it is similar to the original PLSA model in that a different set of mixture weights is used for different data points; on the other hand, it is similar to GMM in that mixture weights for different data points are required to be similar to each other. In particular, the new model has the ability in adjusting the model flexibility of the mixture weights through the regularization term. Experiment results for speaker identification application have shown that the new RPLSA model outperforms both the GMM and the PLSA models substantially. Choosing the appropriate amount of modeling flexibility is a general problem for all mixture modeling techniques. The new research in this paper can be naturally incorporated with other types of mixture models than the GMM model and be applied for other applications.

6. Acknowledgements

We thank Alex Waibel and Qin Jin for their helpful discussion of this work.

References

1. Bellegarda J. R., Nahamoo, D.: Tied mixture continuous parameter modeling for speech recognition, IEEE Trans. Acoustic., Speech, Signal Processing, vol. 38, (1990).
2. Blei, D., Ng, A., Jordan., M.: Latent Dirichlet allocation. Journal of Machine Learning Research. (2003) 993-1022.
3. Dempster, A. P., Laird N. M., Rubin D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B39. (1977) 1-38.
4. Furui, S.: An overview of speaker recognition technology. Automatic speech and speaker Recognition. Edited by Lee, C., Soong, F., Paliwal, K. Kluwer Academic Press. (1996).
5. Hofmann, T.: Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI (UAI). (1999).
6. Hofmann, T.: Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR) (1999).
7. Permuter, H., Francos J., Jermyn I. H.: Gaussian mixture models of texture and colour for image database retrieval. in Proc . ICASSP, vol. 1. (2003) 25-88.
8. Povinelli R. J., Johnson M. T., Lindgren A. C., Ye J. J.: Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces. IEEE Transactions on Knowledge and Data Engineering. Vol. 16. No6. (2004).
9. Reynolds, D. A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication (17) (1998) 91-108.