SPEAKER VERIFICATION USING GMM MODELLING

Svetlana SEGĂRCEANU[1], Tiberius ZAHARIA[2], Anamaria RĂDOI[3]

Authentication based on voiceprint is a simple and user-friendly biometric technology to address the overcoming security issues. We present a GMM-UBM approach to speaker verification based on one and two-factor schemes and compare their verification performance. In this framework we evaluated several score normalization and adaptation approaches. In the feature extraction stage of the speaker verification experiments we used the three classical cepstral processing methodologies, based on linear prediction, Mel and Bark scale analysis, with a second purpose of assessing their performances. The experiments are based on a speech corpus of 26 speakers, who pronounced several compulsory sentences in Romanian, and arbitrary text. Several combinations of vocabulary were tested.

Keywords: speaker verification, Gaussian mixtures modelling, threshold setting, normalization, universal background models, perceptual analysis of speech, biometric measures.

1. Introduction

The speaker verification issue is to decide on the invoked identity of a client. So it can be used as an access gate to a secured system, such as telephone banking. Two decisions are possible: client and impostor. As any pattern recognition problem, it involves two aspects: training and testing (the verification itself). In the training phase the user must pronounce a number of utterances in order to create her or his model. At verification the user's processed signal output is compared to the model of the invoked speaker. S. Furui [1] suggested that an impostor model, trained with several "impostor" users, should also be considered.

Most of the latest commercial security schemes use two-factor authentication based on combinations of passwords or personal data. Although this improves the security compared to that provided by one factor, it does not guarantee that the claimed identity is the real one, as PIN codes, personal data, can be obtained by the fake [2]. The role of biometric technologies is to confirm that the users, by their unique biological information, are what they pretend they are.

[1] PhD student, Depart. of Electronic and Telecommunication Engineering, University POLITEHNICA Bucharest, Romania, e-mail: svet_segarceanu@yahoo.com
[2] PhD student, Depart. of Electronic and Telecommunication Engineering, University POLITEHNICA Bucharest, Romania, e-mail: tezeu2000@gmail.com
[3] Ph.D. student, Department of Applied Electronics and Information Engineering, University POLITEHNICA Bucharest, Romania, email: rdi_ana@yahoo.com

Our paper investigates speaker verification in the frameworks of one factor and two-factor schemes. We used the Gaussian Mixture Modelling (GMM) of the feature space and applied the GMM-UBM methodology in the verification stage. Several techniques, such as normalization techniques, or adaptation techniques were operated to improve the performance in both schemes. As characteristic features we used cepstral coefficients derived from linear prediction (LPC), and perceptual analysis in MEL and Bark scales, and compared the three approaches. The material is organized as follows.

The next section presents the general framework based on GMM modelling used in speaker verification. Some specific aspects of this approach are explained, such as background universal models (UBM), score normalization, model adaptation techniques. The third section reviews the feature extraction approaches used throughout the experiments. The forth section presents the organization of verification experiments, grouped as one-factor and two factor verification trials. The last part of the paper is devoted to conclusions of the work.

2. Gaussian Models in Speaker Verification

Speaker verification is a speaker recognition application by which a speaker claiming a certain identity is either accepted or rejected. The components of a verification system include speech feature extraction, feature space modelling in the training stage and decision at verification.

Many current verification systems are based on statistical approaches, among which Gaussian mixtures modelling is considered very suitable. GMM may be regarded as a statistical clustering method by which each cluster is represented by a Gaussian distribution. The speaker's feature space is thus modelled by a mixture of Gaussian distributions [3]:

$$f(x)=\sum_{k=1}^{K} c_k \phi(x,/\mu_k,\Sigma_k) \text{ with } \sum_{k=1}^{K} c_k = 1 \text{ and } 0 \le c_k \le 1 \text{ for } 1 \le k \le K \quad (1)$$

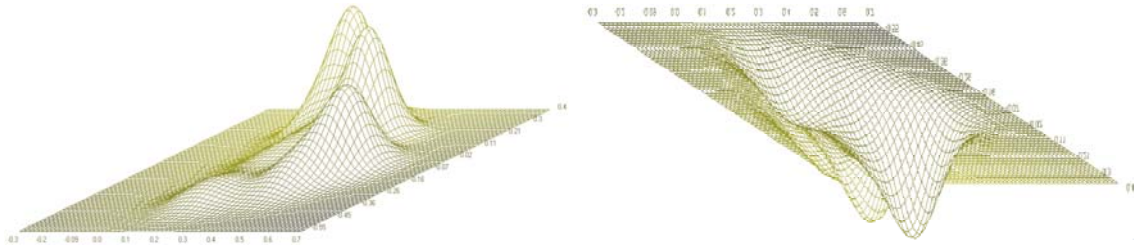

Fig. 1. The Gaussian models mixtures using diagonal matrices, of a female(left) and male(right) voices using the sentence „Meniul moliei e lîna"..

Although the general model involves a real covariance matrix, where all the elements are calculated, in applications based on speech signal processing,

diagonal matrices are used, for several reasons, such as easiness of their computing, or the statistical independence of the components of the feature vectors. The parameters of the Gaussian models mixtures in (1) are usually computed by applying the EM algorithm. Fig. 1 presents the Gaussian models mixtures using diagonal matrices, of male and female voices.

In the decision stage, GMM, as a statistical approach, is a typical example of hypotheses testing, among:

H_0 : *X* comes from the invoked speaker S

H_1: *X* does not come from the invoked speaker S

The decision is given by the test of the ratio between the likelihood rates:

$$R = p(X / H_0) / p(X / H_1) \tag{2}$$

and a decision threshold θ, where $P(X/H_0)$ and $P(X/H_1)$ are the probabilistic density functions for hypotheses H_0, H_1, evaluated on the speech segment *X*. H_0 is represented by a model λ_s, which characterizes the client speaker *S*, and is a mixture of Gaussian models. H_1, modelled by $\lambda_{\bar{s}}$**, is the alternative hypothesis.** The log-likelihood is used instead [3], [4], [5], so the evaluated expression is:

$$\Lambda(X) = \log P(X / \lambda_s) - \log P(X / \lambda_{\bar{s}}) \tag{3}$$

While λ_s is well defined and can be trained by using the client's speech, $\lambda_{\bar{s}}$ is more uncertain as it should characterize the alternative space of the client speaker. One approach to represent the alternative model uses a set of models, possibly speaker related, to cover the alternative speaker space. The second major approach stores the speech of several speakers to generate one only general model, called, universal background model"- UBM.[4], [5], [6]. In the GMM modeling, UBM is a mixture of Gaussian models to represent the alternative $\lambda_{\bar{s}}$ to the speakers participating in the verification system. In a GMM-UBM it is often represented as a sum of models of different subpopulations in the feature space.

In the decision taking stage the score obtained by evaluation of (3) is compared to the threshold θ and the speaker is accepted if the score is above it. There are two main approaches to set the threshold θ:

- Each client speaker *n* of an authentication system is associated her/his own threshold. θ_n, or
- One common value is used, as a common threshold for all the speakers.

Although the second variant is less efficient it is usually operated to represent the performance of a verification system on a DET curve [7].

In order to contain score variability and take an easier decision to establish and adapt the thresholds, score normalization was introduced. The basic idea of

the normalization techniques, introduced in [8], is to center impostor score distribution by applying, for each generated score, the normalization:

$$\tilde{L}_\lambda(X) = \frac{L_\lambda(X) - \mu_\lambda}{\sigma_\lambda} \tag{4}$$

where μ_λ, σ_λ are normalization parameters estimated for the speaker modelled by λ. *Znorm* is among the best known normalization technique, popular in the '90, where μ_λ, σ_λ are estimated individually, for each client speaker.

In the verification experiments presented in [9] the authors use a dynamic version of *znorm*, which they call "unsupervised normalization".

In practice, a speaker verification system is evaluated by appraisal of two types of errors: the false acceptance and false rejection rates, denoted FAR and FRR, measured by counting the errors of each type for a certain threshold value, on a test corpus. Another popular measure is the EER (Equal Error Rate), corresponding to the threshold value for which FAR=EER.

2.1 Adaptation of a Gaussian Mixture Model.

The basic idea of the adaptation technique introduced by de Reynolds, Quatieri, and Dunn in [10] is to update the already trained parameters of a Gaussian mixture to better outline the speaker model from the UBM. The adapting process is accomplished in two steps. First, the sufficient statistics (frequency, and order 1 and 2 moments of the mixtures of the model) are estimated based on the speakers' training data. That is, given a Gaussian mixture model (either UBM or speakers' models) and some training vectors, $X = \{x_1, x_2, \ldots, x_T\}$, the probabilistic superposition is assessed by calculating for each Gaussian component:

$$\Pr(i \mid x_t) = w_i p_i(x_t) \Big/ \sum_{j=1}^{M} w_j p_j(x_t) \tag{5}$$

Based on (5) the sufficient statistics are derived from the training data:

$$\begin{aligned} &n_i = \sum_{t=1}^{T} \Pr(i \mid x_t); \\ &E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid x_t) x_t; \quad E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid x_t) x_t^2 \end{aligned} \tag{6}$$

In the second step the sufficient statistics are used to revise the parameters of the old Gaussian mixture, by:

$$\hat{w}_i = [\alpha_i^w n_i / T + (1 - \alpha_i^w) w_i] \gamma \tag{7}$$

$$\hat{\mu}_i = \alpha_i^m E(x) + (1 - \alpha_i^m) \mu_i \tag{8}$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + (1 - \alpha_i^v)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 \tag{9}$$

Adaptation coefficients α_i^{ρ}, $\rho \in \{w, m, v\}$ for weights, means and variances, to contain the relation between old the new estimates, are defined as:

$$\alpha_i^{\rho} = n_i / (n_i + r^{\rho}) \quad (10)$$

where r^{ρ}=16, is a relevancy factor for the parameter ρ. γ is a scaling factor for weights.

$$\gamma = 1 \Big/ \sum_{k=1}^{K} \hat{w}_k \quad (11)$$

The systems that implement this method do not necessarily adapt all the parameters; they might adjust only means or means and weights, etc.

3 Feature Extraction

In the feature extraction step we used the cepstral processing based on three different approaches: the linear prediction [11] and two popular perceptual methodologies, the Mel-scale [12] and the Bark scale [13], [14] processing.

In the Mel-scale approach we used the scale representation:

$$f_{mel}(f) = 1125 \ln(1 + f/700) \quad (12)$$

and the bank of 36 filters, 1≤k≤:36:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m-1] \le k < f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m] \le k < f[m+1] \\ 0 & k \ge f[m+1] \end{cases} \quad (13)$$

In assessing the perceptual features based on the Bark scale we applied the Bark scale representation:

$$f_{bark}(f) = 7 \ln\left\{ (f/650) + \left[(f/650)^2 + 1 \right]^{0.5} \right\} \quad (14)$$

and the filter bank (15) where ϖ=2πf and f_{bark} is defined by (14) and $f_{bark}^k \approx k$.:

$$C_k'(\omega) = \begin{cases} 10^{-2.5(f_{bark} - f_{bark}^k) + 0.5} & f_{bark} \le f_{bark}^k - 0.5 \\ 1 & f_{bark}^k - 0.5 < f_{bark} < f_{bark}^k + 0.5 \\ 10^{f_{bark} - f_{bark}^k} & f_{bark} \ge f_{bark}^k + 0.5 \end{cases} \quad (15)$$

4. Experimental results

We investigated the way several factors influence a GMM-UBM framework performance; among them:

- The type of characteristic features;
- Number of Gaussian components in client speakers models and UBM;
- Type of text uttered by speakers;

- Type of normalization applied;
- Various adapting schemes on applied to speaker models or on the UBM;
- One factor (one password for all client speakers) versus the two-factor scheme (two passwords uttered by each speakers)

4.1. Speech database

In our experiments we used speech database containing speech samples of 12 male and 14 female. Each speaker recorded a number, between 4 and 11, of sessions, uttering each time some required and arbitrary text. The obligatory text included six Romanian sentences: “Eu iau nouă ouă moi”, “Meniul moliei e lâna”, “Aureola e o lumină”, ”Lamiia ia anemia unui om”, “Ei au o inimă imună” “Eu îi iau o anemonă”, pronounced once in each recording session. The arbitrary text contained, for certain speakers, besides spontaneous speech, the utterances of their names. The speech signal was sampled 11.025 KHz and represented on 8 bits.

In the verification experiments, 21 (9 male and 12 female) of the 26 speakers were considered client speakers, the other 5 sheer impostors. In addition, the speech of each client speaker was used as impostor speech for any other client speaker. The first two sessions were used for training and the rest for verification. In the one-factor experiments we tested two sentences as compulsory passwords: “Eu iau nouă ouă moi”, “Meniul moliei e lâna”. In the two-factor approach framework, in the training and verification stages, the speakers uttered two sentences: one obligatory sentence, and a second one which was either their names (for those who pronounced it in the recording sessions), or a certain one, depending on the speaker and the sentence used as the first password, of the following: “Eu iau nouă ouă moi”, “Meniul moliei e lâna”, “Aureola e o lumină”, ”Lamâia ia anemia unui om”, “Ei au o inimă imună”. We tested the combinations using as first “password” “Eu iau nouă ouă moi”, and “Meniul moliei e lâna”. Because the total number of utterances uttered was quite large, in order to assess the False Acceptance Rate we tested a limited number of combinations uttered by the “impostors”. These “combinations” contained the first compulsory text and a number of sentences as the second “password”, among which the above mentioned ones, depending on the reference client speaker.

At training and at recognition we extracted 14 cepstral coefficients on each speech frame, derived either from the PLP or from Mel-scale analysis. We used a spectrum – based criterion to remove the non-voiced frames. The first two sessions were used for training. In all these approaches we found useful to drop the first coefficient.

4.2. Speaker verification using one factor

We tested the one-factor scheme in the case of two Romanian sentences used as passwords. "Eu iau ouă moi" and "Meniul moliei e lâna". The models of the client speakers were trained from the utterances of the password recorded in the first two sessions. To calculate the log-likelihood of we used the expression:

$$p(X) = \log P(X / \lambda_s) - \log P(X / \lambda_{ubm}) \tag{16}$$

where $P(X/\lambda)$ was calculated as:

$$\log(\ p(X \mid \lambda)) = \sum_{t=1}^{T} \log(\ p(x_t \mid \lambda) \Big/ T \tag{17}$$

In the above expressions λ_s, λ_{ubm} represent the client speaker, and the UBM models, T is the length of the test feature sequence. The UBM was represented as sum of two subpopulation models, female (UBM*f*) and male (UBM*m*), trained with password utterances of the 14 female, 12 male, respectively, from the first two recording sessions; with equal contributions to the whole UBM [4], [5], [6]:

$$p(x_t / \lambda_{ubm}) = \sum_{k=1}^{K_f} \frac{1}{2} w_k^f \phi(x_t / \mu_k^f, \Sigma_k^f) + \sum_{k=1}^{K_m} \frac{1}{2} w_k^m \phi(x_t / \mu_k^m, \Sigma_k^m) \tag{18}$$

where $w_k^f, \mu_k^f, \Sigma_k^f$, $w_k^m, \mu_k^m, \Sigma_k^m$ are the parameters of UBM*f*, UBM*m*, K_f, K_m their orders. Based on (17) and (18) we evaluated (16). $p(X)$ was compared to thresholds θ varying from –2.7 and 2.7 to represent the system's performance on DET curves. Fig. 2 presents male and female Gaussian mixture models with 30 Gaussian components, derived from bi-dimensional PLP cepstral coefficients. Their superposition generates the final UBM.

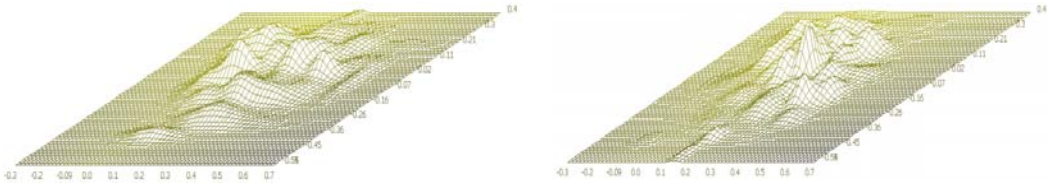

Fig.2. Male and female GMMs with 30 Gaussian components, derived from bi-dimensional PLP cepstral coefficients using the sentence „Meniul moliei e lina"

Concerning the number of Gaussian components we tested the situations:

-3 components for each client speaker, and 30 components for the female and male background models (the final UBM has 60 components).

-4 components for each client speaker, and 40 components for the female and male background models (the final UBM has 80 components).

In order to improve the performance, we tested three normalizing schemes, by applying (4) on the score of each speaker:

$-\mu_\lambda=\mu_i-1.5\cdot\sigma_i$; $\sigma_\lambda=1$ (denoted *norm1*)

-$\mu_\lambda=\mu_i-2\cdot\sigma_i$; $\sigma_\lambda=1$-(denoted *norm2*)

-$\mu_\lambda=\mu_i$; $\sigma_\lambda=\sigma_i$-(*znorm*)

In the expressions above $\mu_i,\cdot\sigma_i$ are the mean, and standard deviation of the all impostors (with regard to speaker *i*) scores, estimated from the first two recording sessions utterances. By applying *norm1* we tried to constrain the FRR to be situated around 1-0.9332, and by *norm2* to force the FRR to stay around 0.022.

We tried to improve the system's performances by adapting the universal model by removing for each tested speaker the closest components with regard to the invoked speaker. Thus we tried to simulate a speaker specific UBM. We used two ways to accomplish this:

-By removing two components of the UBM whose centers are the closest to the center of one of the components of the speaker model (*ubm1*).

-By removing two UBM components, having the highest likelihood with regard to the invoked speaker's model.

We also applied the Reynolds adaptation approach in the following ways:

-By applying (7)–(9) to speakers' models, and their training utterances

-By adapting the background model based on the same equations, from the training utterances of the five "impostor"-speakers, who have not participated as client speakers in the experiment. The UBM sub-models were adapted and the final adapted model was the combined model.

-By using both adapted speakers and universal background model.

Table 1

EERs for the one-factor scheme for two passwords, several cesptral feature sets, different sizes of λ_s, and λ_{ubm} applying normalization techniques, and adaptation methods

	3 components in the speakers'models and 60 components for UBM					
	Password1			Password2		
	LPC	MEL	PLP	LPC	MEL	PLP
Basic	11.15	15.15	14.70	11.24	12.30	13.75
Norm1	10.99	14.80	16.20	10.20	13.80	13.85
Norm2	11.47	16.88	20.10	11.75	15.90	14.10
Znorm	10.75	14.80	16.35	11.50	14.87	13.80
Ubm1	10.70	14.90	14.55	11.25	12.40	13.70
Adapt_sp	9.70	14.30	14.76	10.80	13.60	13.60
Adapt_ubm	10.99	11.60	12.50	10.90	10.50	13.40
	4 components in the speakers'models and 80 components for UBM					
Basic	12.20	14.00	11.45	10.10	10.50	12.90
Norm1	8.70	13.90	15.20	8.25	13.10	13.50
Norm2	11.00	15.70	17.34	10.40	13.80	14.40
Znorm	10.00	14.50	15.20	9.75	14.36	13.30
Ubm1	11.90	14.00	11.30	9.15	10.50	12.30
Adapt_sp	12.45	15.20	13.50	9.40	10.51	12.90
Adapt_ubm	10.35	11.40	9.75	9.20	9.25	11.69

To evaluate verification performance with *password1*, „Eu iau nouă ouă moi", we used 154 pronunciations to assess the false rejection rate (FRR) and 4026 to evaluate the false acceptance rate (FAR). For *password2*, „Meniul moliei e lîna", FRR was estimated based on 151 utterances, and FAR on 3925. Which means that while the estimation of FAR is fairly reliable, the FRR confidence is weak. As the training speech material was collected during only two recording sessions, UBM was derived from about 40 minutes of speech, while the speakers' models were based on less than one minute of speech. Table 1 presents the speaker verification EERs obtained by applying the one-factor scheme, using several feature extraction and normalization techniques, and the proposed adaptation methods. The performance of methods *ubm1* and *ubm2* were approximately equal so, we figured only those obtained using *ubm1*. Adapting background models (*adapt_ubm*) using the Reynolds methods produced the best results especially when using the Mel-approach, while adapting speakers' models (*adapt_sp*) generated poorer performance. Using adapted speakers' models and UBM behaved as if combining the two methods. Among the three normalization techniques *norm1* produced the best verification rates.

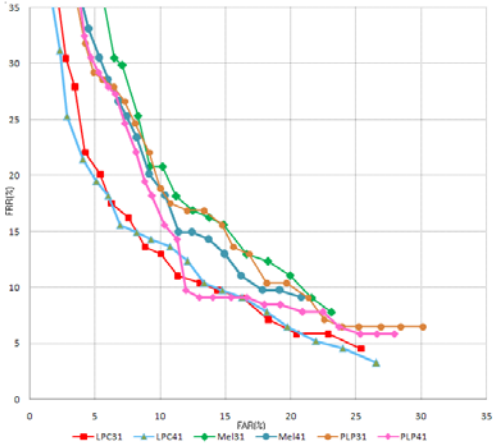

Fig. 3. DET curves using *password1*, 3 and 4-order speaker GMMs, 60 respectively 80 order UBMs, LPC, MEL and PLP cepstral features.

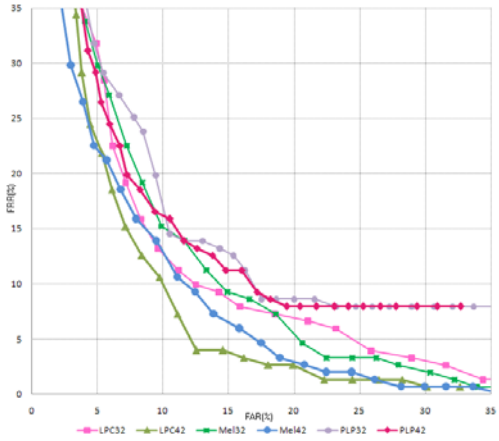

Fig. 4. DET curves using *password2*,3 and 4-order speaker GMMs, 60 respectively 80 order UBMs, LPC, MEL and PLP cepstral features .

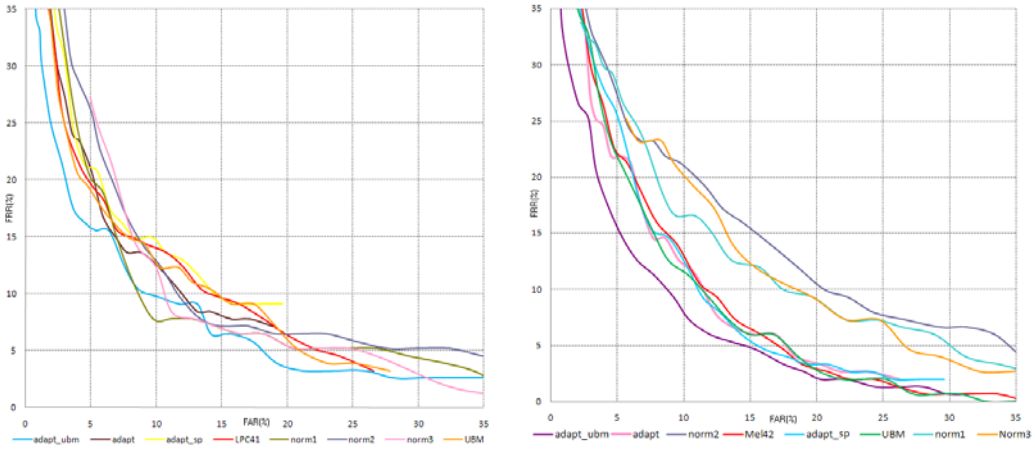

Fig. 5. Verification error rates FAR against FRR obtained using the LP cepstral features (left) and Mel-cepstral features (right), 4component speaker GMMs, and 80 order UBMs, using the unique passwords "Eu iau nouă ouă moi" (left) and "Meniul moliei e lîna"(right), applying several normalization, and adapting methods. While for the LP coefficients *norm1* is the most beneficial, for the Mel approach adapting the UBM clearly leads to the best improvement.

Fig. 3 presents the DET curves for one-factor scheme, using *password1*, 3 and 4 order speaker GMMs, 60 respectively 80 order UBMs, LPC, MEL and PLP cepstral features. The curves corresponding to the 3-order GMMs and 60-order UBMs are denoted by *LPC31*, *Mel31*, *PLP31*. The ones corresponding to the 4-order GMMs and 80-order UBMs are denoted by *LPC41*, *Mel41*, *and PLP41.*

Fig. 4 presents the DET curves for one-factor scheme using *password2*, with similar notations as in Fig. 3. The image reveals that the best performance was obtained using the higher order mixtures, and the LPC coefficients. Figs. 5 present the verification error rates FAR against FRR, obtained using the LP cepstral features (left) and Mel-cepstral features (right), 4-component speaker GMMs, and 80 components UBMs, using as unique passwords "Eu iau nouă ouă moi" (left) and "Meniul moliei e lîna"(right), several normalization, and adapting methods. While in the case of the LP coefficients *norm1* is the most beneficial, for the Mel approach, adapting the UBM clearly leads to the best improvement.

4.3. Speaker verification based on two factors

The scenarios for the two-factor authentication might be:

- The speaker utters the fist password and if accepted the process stops. If not, he utters the second password, and is accepted if the sum of the scores for the two passwords is situated conveniently with regard to the established threshold

- The speaker utters the first and the second password and is accepted based on the cumulated score.

In our experiments we tested the second approach. We used as passwords two texts, a first one common for all speakers, the other was intended to be the name of the client speaker. As not all the speakers had enough records of their names, the second text was either the speaker's name, for those who had enough such records, or another obligatory text, depending on the speaker, chosen from: "Eu iau nouă ouă moi", "Meniul moliei e lâna", "Aureola e o lumină", "Lamâia ia anemia unui om", "Ei au o inimă imună". As first password we tested two variants: "Eu iau nouă ouă moi", and "Meniul moliei e lâna". All speakers pronounced two different passwords. According to the above framework, the two-factor verification scheme is similar to the one factor scheme using longer "password" utterances, involving an invariable component and a variable one.

We modelled the UBM as sum of background models generated from the utterances of the first „password" (*UBM1*), and *UBM2* obtained from the utterances of "Aureola e o lumină", and „Eu îi iau o anemonă" to better represent the phoneme space of the Romanian language. *UBM2* was obtained as sum of models for these two sentences, *UBM21*, and *UBM22*. Each of *UBM1, UBM21,* and *UBM22* was calculated as sum of female and male subpopulations uttering the respective sentences, models. Concerning the size of the speakers and background models we tested the following situations:

- 3 components for each client speaker and each password (6 components for the final speaker's model), and 30 components for the female and male models for the three sentences based on which the model was built (180 order UBM);

- 4 components for each password (final speaker's model of order 8), and 40 components for the 6 sub-models (the final UBM has 240 components).

The overall score was established based on:

$$S=p(X_1) + p(X_2) \tag{19}$$

where X_1 and X_2 are the sequences of characteristic features resulted from the pronunciation of the two passwords by the tested speaker and:

$$p(X_i) = \log P(X / \lambda_{si}) - \log P(X / \lambda_{ubmi}) \;\; i = 1,2 \tag{20}$$

where $P(X/\lambda)$ was calculated by (17), $P(X/\lambda_{ubm1})$ by (18), λ_{s1}, λ_{s2}, are the speakers models for each of the two passwords,. $P(X/\lambda_{ubm2})$ was similarly evaluated with:

$$p(x_t / \lambda_{ubm\,2}) = \frac{1}{4}\sum_{i=1}^{2}(\sum_{k=1}^{K_f}(w_k^{f\,2i}\phi(x_t / \mu_k^{f\,2i}, \Sigma_k^{f\,2i}) + w_k^{m\,2i}\phi(x_t / \mu_k^{m\,2i}, \Sigma_k^{m\,2i})) \tag{21}$$

with $w_k^{f\,2i}, \mu_k^{f\,2i}, \Sigma_k^{f\,2i}$, $w_k^{m2i}, \mu_k^{m2i}, \Sigma_k^{m2i}$ the parameters of UBM*f2i*, UBM*m2i,* the female and male subpopulations models for UBM*2i*, i=1,2.

P(X) has been compared to thresholds θ, $-4.5 \le \theta \le 4.5$. We have studied the influence of the normalization methods *znorm* and *norm1* presented above, on the verification rates. We have studied the impact on the system's performance of the Reynolds's adaptation approach, applied to the UBM. UBM adaptation was operated using the training utterances of the sheer impostors, applying the

relations (7)-(9) to the sub-components of the background model. The final UBM was the sum of the adapted components. Table II presents speaker verification EERs obtained using several feature sets, different orders for speakers models and UBM, applying *norm1*, *znorm* and Reynolds's UBM adaptation methods. Figs. 6 present DET curves using *password2* as first password, 6 and 8 order speaker GMMs, 180 and 240 order UBM, LPC (left) and MEL(right) cepstral features.

Table II

EERs for the one-factor scheme for two passwords, several cesptral feature sets, different sizes of λ_s, and λ_{ubm} applying normalization techniques, and adaptation

	6 components in the speakers'models and 180 components for UBM					
	Password 1			Password 2		
	LPC	MEL	PLP	LPC	MEL	PLP
Basic	7.01	7.15	9.80	6.06	6.88	10.22
Norm1	5.35	7.22	9.70	4.90	7.37	9.10
Znorm	5.45	7.28	9.39	5.72	7.07	8.88
Adapt	6.30	7.75	11.80	6.01	6.58	10.22
	8 components in the speakers'models and 240 components for UBM					
Basic	6.30	5.85	8.08	5.57	5.37	9.50
Norm1	4.60	7.40	8.76	4.66	6.90	8.63
Znorm	5.10	7.40	8.49	4.72	6.75	7.85
Adapt	6.0	7.70	10.46	4.67	5.90	11.80

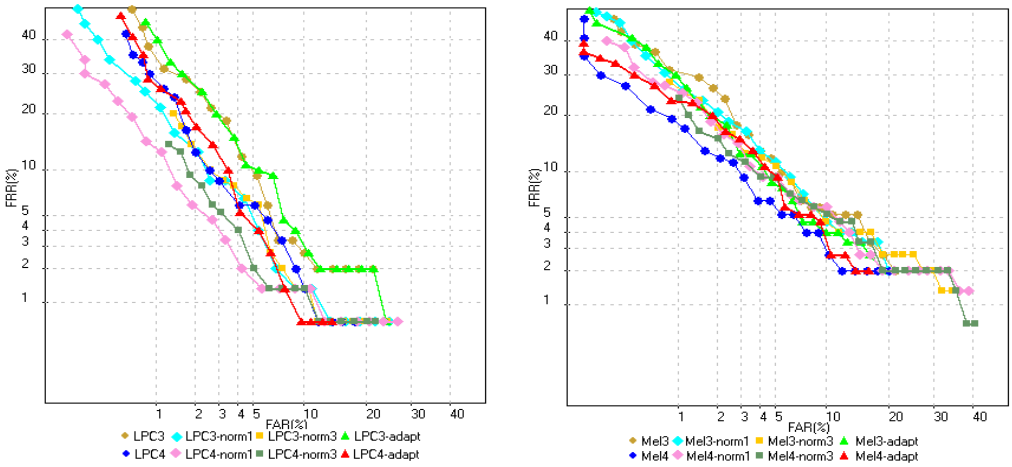

Fig. 6. DET curves for the two-factor scheme, using 6 and 8-order speaker GMMs, 180 and 240 order UBMs, respectively, *password2* as first password, LPC (left) and MEL(right) cepstral features.

5. Conclusions

Our paper presents comparative results of speaker verification experiments using one factor and two factor schemes. The two-factor scheme increased significantly the verification performance, especially when using the LP and Mel based features. This might be the consequence of the fact that two passwords involve richer speech material, at training and at testing. This involved a larger number of components in the authentic speakers and the background models.

While the results obtained using different types of features (mainly LP and Mel cepstral features in the two-factor scheme), are comparable, the normalization techniques worked better in the LP based experiments. The best results were obtained using *norm1*. On the other hand the Reynolds's adaptation techniques produced very good results in the one-factor experiments, especially in the case of Mel-cepstral coefficients. However in the two-factor experiments, the UBM adapting method has not produced any increase of the performance. Several might be the causes. One of them might be the speech material used to develop the two-factor UBM, as it was difficult to pick the right material to represent the variable part of the passwords. The UBM adaptation methods based on the removal of certain components of the universal model produced very poor increase in performance. *Norm2*, out of the three normalization techniques, worked the worse; one reason might be the initial performance of the system.

As expected, using higher order for speaker models and UBMs boosted the performance, in both types of experiments, although more visible in the two-factor scheme. The results clearly differ depending on the type of text used.

As compared with the results obtained by other researchers, the best EERs achieved in [6] are about 4.5. The experiments carried out by Daniel Neiberg, used a subset of the full Swedish SpeechDat database, applying the UBM-GMM approach and the UBM-adaptation and normalization techniques. The verification equal error rates obtained by Wu Guo, Li-Rong Dai and Ren-Hua Wang also using the GMM-UBM approach, factor analysis to reduce channel bias, and their unsupervised normalization approach, vary between and 4.28 and 9.2. The experiments were carried out on the NIST SRE 2006 corpus. The experiments presented in [15] use reflection coefficients as characteristic features, and the GMM-UBM modelling, for subsets of the TIMIT and Kiel databases. Their performance (EER) is 3.4%. The best verification EER obtained in our experiments was 4.6% using the LPC coefficients and a certain normalization scheme.

One of the drawbacks of our experiments is the scarceness of the corpus we used. For instance, in the two-factor scheme, the UBM was derived from about 1½hour of speech, while the speakers' models used less than 1 minute of utterances.

A future objective would be to resume the experiments in the context of a more consistent test corpus, more suitable choice for the passwords text, in order to produce more reliable results. Another drawback was the criterion used to eliminate the unvoiced frames, slightly modified from one feature set to another. We intend to provide a unitary framework to cope with this problem.

REFERENCES

[1] *Furui S.*, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 254–272, 1981.

[2] *Pawlewski M.and Jones J*, "Survey. Speaker verification: Part 1", Biometric Technology Today Vol. 14, pp 9-11, June 2006.

[3] *Reynolds, D. A.*, „Speaker identification and verification using Gaussian mixture speaker models", Speech Commun. 17, pp. 91–108, 1995.

[4] *Bimbot, F., et. al.*, "Tutorial on Text-independent Speaker Verification" EURASIP Journal on Applied Signal Processing, pp. 430-451, 2004.

[5] *Reynolds D. A., Quatieri T. F., and. Dunn R. B*, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.

[6] *Neiberg D.*, „Text Independent Speaker Verifcation Using Adapted Gaussian Mixture Models", Ph. D.. Thesis, Centre for Speech Technology (CTT) Department of Speech, Music and Hearing KTH, Stockholm, Sweden supervisor: Hakan Melin 2001-12-11

[7] *Doddington et. al.*, "The DET curve in assessment of detection task performance", Proc. European Conference on Speech Communication and Technology (Eurospeech '97*)*, vol. 4, pp 1895–1898, Rhodes, Greece, September 1997.

[8] *Li K. P., and Porter J. E.*, "Normalizations and selection of speech segments for speaker recognition scoring", Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '88), vol. 1, pp. 595–598, New York, NY, USA, April 1988.

[9] *Guo, W., Dai, L., Wang, R.*, "Double Gauss Based Unsupervised Score Normalization in Speaker Verification", ISCSLP 2008, pp. 165-168

[10] *Reynolds D. A., Quatieri T. F, and R. B. Dunn*, "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.

[11] *Markel, J. D., Gray Jr, A. H.*, Linear Prediction of Speech, Springer Verlag - New York, 1976.

[12] *Davis, S. B., Mermelstein P.*, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on ASSP 28, pp 357–366, 1980.

[13] *Woelfel, M. C., McDonough, J.*, "*Distant Speech Recognition*", John Wiley& Sons, 2009.

[14] *Hermansky, H.*, "Perceptual linear predictive (PLP) analysis of speech", Jour. of ASA 87(4), pp. 1738–1752, 1990.

[15] *Enzinger, E., Kasess, C. H.*, "Experiments on using Vocal Tract Estimates of Nasal Stops for Speaker Verification", Proceedings of the 7th International Conference Speech Technology and Human-Computer Dialogue, Bucharest, Oct. 2013.