PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 3, Number 3/2002.

COHORT METHOD OPTIMIZATION FOR SPEAKER RECOGNITION

Corneliu BURILEANU *, Daniel MORARU **, Alina STAN*

* University "POLITEHNICA" of Bucharest, Faculty of Electronics and Telecommunications
** "CLIPS" Laboratory, Institut National Politechnique de Grenoble, France

Corresponding author: Corneliu BURILEANU
Bd. Iuliu Maniu 1-3, Sector 6, Bucharest 77202, Romania
cburileanu@mESsnet.pub.ro

This paper presents the use of distance normalization techniques in order to improve the speaker verification systems performances. These techniques provide a dynamic threshold that compensates the trial-to-trial variations and replaces the fixed threshold used in the classical speaker verification approach. The use of cohort method normalization is described. The paper also presents a theoretical approach to the world model method normalization, which is an alternative of the cohort method. The algorithm is evaluated using the YOHO database and a proprietary database. The method is also studied from the point of view of storage space requirements and computational effort. The results showed that the use of cohort normalization increases the system performances. The algorithm also involves small computational resources, making them more suitable for a commercial application.

Key words: speaker verification, cohort normalization, world model method.

1. INTRODUCTION

The speaker verification applications are based on matching a voice sample acquired in the recognition phase with a speaker model created in the training phase [7]. This is a classification task that can yield two results: true customer and impostor. The classic approach is to measure the matching degree between the voice sample and the speaker model by computing a distance and comparing it with a fixed threshold. The value of this threshold can be set during the training phase and can be updated periodically, but this operation requires a priori knowledge of both true customers and impostors. The decision is based exclusively on the voice sample and speaker model data, without considering any information about possible impostors.

Given a sequence of vectors O obtained from a verification phrase, a test speaker identity is authenticated if and only if:

$$d(O, Y) < T(Y) \qquad (1)$$

otherwise the speaker is declared an impostor and will be rejected.

- $d(O, Y)$ is the distance between the sequence of feature vectors and the model λ_Y which was obtained in the training phase.
- $T(Y)$ is the decision threshold for the speaker Y; usually this threshold is set to the same value for all speakers.

The decision threshold is computed during the training phase by minimizing the classification error of the available voice samples. Even if the decision threshold is modified subsequently, it is not adapted in real time based on the testing conditions (noise, speaker's mood, etc).

The classic methods obtain the verification score (distance or likelihood) using only the model of the speaker who claimed the identity. This way, only information about the claimed identity is used, and no information about the potential impostors.

The variations in time of the testing conditions (noise, speaker's mood, inherent variations among utterances of the same speaker, etc.) cannot be handled by this method. Adaptation of the customer's model as well as the verification threshold for each speaker is a solution to the problem of obtaining high recognition accuracy over a long period of time. In order to compensate

the effects of these variations, three types of normalization/adaptation techniques have been attempted: in the parameter domain, in the model domain, and in the distance domain.

This paper focuses only on normalization in the distance domain and describes the cohort normalization method and the world model method.

2. THE COHORT METHOD AND THE WORLD MODEL METHOD

The cohort method [1] consists in selecting a number of speakers, different from the speaker who claimed the identity; these speakers will form the cohort. The ensemble of speakers available for the selection process is called the cohort pool. Usually, the cohort pool is identical to the ensemble of users of a given system. If m speakers are selected in a cohort, the number m is called the cohort dimension [10], [11]. The cohort method uses a threshold computed in the verification phase, instead of a fixed threshold. Thus, two distances are computed: one to the speaker model of the person whose identity is claimed, and another to the models of the cohort speakers associated with that speaker. If we agree that the acceptance of an impostor and the rejection of a true customer are of equal importance, then expression (1) becomes:

$$d(O, Y) < d(O, \overline{Y}) \qquad (2)$$

where $d(O,\overline{Y})$ is the distance between the sequence of feature vectors O and the cohort models. The term $T(Y)$ from expression (1) was replaced by $d(O,\overline{Y})$, called the normalization term.

The normalization term in expression (2) explicitly depends on the sequence of test vectors O. This way, the new obtained "threshold" is now sensitive to the variation of the sequence O from one utterance to another.

The cohort method raises two questions:

- How do we select the cohort members?
- How do we compute the distance $d(O,\overline{Y})$ when $\overline{Y}$ is in fact a composite model made of m individual models?

The differences between various implementations of the cohort method result from the different techniques being used to select the models and to compute the distance $d(O,\overline{Y})$.

There are several techniques for selecting the speakers who will form the cohort:

- Selection of the closest speakers to a given speaker; the selection criterion from the cohort pool is the matching degree with the given speaker model.
- Random selection of m speakers from the cohort pool.
- Selection based on the minimization of the error rate. This way, only one cohort model is used for all the speakers. The selection of the cohort speakers is performed by trying different models until the error rate is minimized. This technique has the inconvenience of requiring a high computation volume in the training phase.

Choosing cohort speakers close to the target speaker model increases the system performance against attack by impostors whose voices are similar to the target speaker's voice [1], [2], [12], [14].

The model $\overline{Y}$ is a composite model including m speaker models. The computation of $d(O,\overline{Y})$ requires the use of an operator over m distances, one for each cohort speaker. Examples of methods used are: the arithmetic or geometric mean method, the minimum method, and the weighted average method. Using one of these methods, the decision of acceptance/rejection for a speaker is based on the score S, which is computed as follows [1]:

- *The arithmetic mean method*
  The distance from the test phrase to the cohort models is the arithmetic mean of the distances to the cohort speaker models

$$S = \frac{1}{m}\sum_{i=1}^{m} d(X, y'_i) - d(X, y) \qquad (3)$$

  where y'_i is the model for the cohort speaker i of the cohort associated with the speaker y, and X is the test phrase.
- *The geometric mean method*
  The distance from the test phrase to the cohort models is the geometric mean of the distances to the cohort speaker models

$$S = \sqrt[m]{\prod_{i=1}^{m} d(X, y'_i)} - d(X, y) \qquad (4)$$

- *The minimum method*
  The distance between the test phrase and the impostor model is the minimum of the

distances to the cohort speaker models. For one phrase that is

$$S = \min_{y' \in \emptyset_y} [d(X, y')] - d(X, y) \quad (5)$$

where Y_y is the cohort associated with the speaker y.

- *The weighted average method [10]*
  The distance to the impostor model is a weighted average of the distances to the cohort speakers. The weights can remain fixed after the training phase or they can be modified during the test itself. The score is computed as follows

$$S = \sum_{i=1}^{m} w_i d(X, y'_i) - d(X, y) \quad (6)$$

where the weights w_i satisfy the relation

$$\sum_{i=1}^{m} w_i = 1.$$

This method is applied to a feature vector level corresponding to a window frame and not to a phrase level.

For all these above expressions, the unknown speaker is accepted as a true customer if S has a positive value (i.e., the distance to the impostor model is greater than the other term) and is rejected if the score is negative or zero.

The actions performed are:

A. In the training phase:
- build the voice model for each speaker;
- select m speakers from the cohort pool and use them to build the cohort.

B. In the verification phase:
- the unknown speaker claims an identity and utters a phrase;
- the phrase is compared to the model associated with the claimed identity and to the cohort models; for each comparison a distance is computed; using the $m+1$ distances, a score is computed using one of the methods presented above and a decision is made; the decision is either "accept" or "reject", depending on the sign of the score.

The world model method [5], [7], [9] reduces the dimension of the normalization problem from m to 1 by using a single impostor model that is synthesized. Similarly with the cohort method a speaker is accepted if the sequence of feature vectors satisfies the expression

$$d(O, Y) < d(O, \bar{Y}) \quad (7)$$

where $d(O,\bar{Y})$ is the distance between the sequences of feature vectors O and the world model.

The main question here is "how do we generate the world model?"

Two different methods are used to generate the world model [2], [11]:
- The world model is generated as the average of the feature vectors of all speakers.
- The world model is generated from a subset of training phrases. The selection criterion is the maximum average distance to the speaker models.

The operations performed are:

A. In the training phase:
- build the voice model for each speaker;
- generate the world model using one of the above mentioned methods.

B. In the verification phase:
- the unknown speaker claims an identity and utters a phrase;
- the phrase is compared to the model associated with the claimed identity and to the world model; the decision to accept or reject the speaker is made by comparing the two distances.

The performance of this method depends on the number and the specific features of the available speakers.

The world model method increases the system confidence with regard to attacks by impostors whose voices are different than the target speaker's voice.

For all the mentioned methods, depending on the type of parameters used in the feature vector, various distances can be used: log-spectrum, cepstral, probability distortions, etc.

- A typical example is the Euclidean distance

$$d(x, y) = \|x - y\|_m \quad (8)$$

the most natural measure of the match being the square error function ($m = 2$).

- A more general method is the weighted square error function:

$$d(x, y) = (x - y)^t \cdot w(x - y) \quad (9)$$

where w is a symmetric matrix non-negative defined.

- The Itakura-Saito distance computes a distance between two random input vectors, using their spectral densities:

$$d(x, y) = \left\| \frac{f_x}{f_y} - \ln\left(\frac{f_x}{f_y}\right) - 1 \right\| \qquad (10)$$

where f_x, f_y are the spectral densities of the input vectors x, respectively y.

3. THE COHORT NORMALIZATION ALGORITHM

This section describes the algorithm used to implement the cohort normalization method. Both training and verification phases are described.

Assuming that there are N available speakers – each of them having an ID and a voice model – for each speaker we create a cohort made of m speakers similar to the target speaker.

The number m is related to the system performance; a larger value yields better performance but implies more processing time.

A. The training phase:

- for each speaker, we compute the distance between the training phrases of all the other speakers (the rest of $N-1$ speakers) and this speaker model;
- we sort these training phrases ascending, based on their distances;
- we select the first m speakers thus sorted. During this process, a specific training phrase may emerge in the higher range of the sorted list more often than others. This means that the information carried by this specific phrase is not sufficiently speaker dependent.

B. The verification phase:

- we compute the distance between the test phrase and the model associated with the claimed identity;
- we compute the distances to the other m speaker models (the total number of distances computed increases from 1, when not using cohort normalization, to $m+1$ for this algorithm);
- if the lowest distance of all $m+1$ speakers is the distance to the claimed identity model, the speaker is accepted, otherwise the speaker is rejected. The minimum method is used to compute the distances to the cohort models in order to decide the impostor.

In this algorithm, the fixed threshold was replaced by the lowest distance to the m cohort models. This way, the threshold becomes explicitly dependent on the test phrase and implicitly dependent on the claimed identity.

4. TEST DESCRIPTION

The tests were performed using a text-independent speaker recognition system based on the vector quantization technique.

The tests were performed on two databases: the YOHO database and the DiSPPALL database a proprietary speech database.

The YOHO database consists of 138 speakers (106 male, 32 female) producing short combination-lock phrases consisting of three doublets (e.g., "twenty-six, fifty-one, eighty-seven"). Each speaker participated in 4 enrollment sessions consisting of 24 phrases each. In addition there are 10 verification sessions, each of which consists of 4 phrases. The YOHO database is a clear speech database.

The DiSPPALL database includes 26 speakers and the training set contains 11 phonetically balanced phrases. For verification there are 20 (4 sets of 5) phrases recorded in two sessions spaced over one month. The phrases are not restricted to number sequences. Every phrase is validated to contain at least 2 sec of speech in terms of an energy threshold. The speech sampling rate was 8kHz and the samples coding was 12-bit linear. The DiSPPALL database is a real (real noise) speech database.

Acoustic feature and codebook generation

We used 12 LPCC (Linear Prediction Cepstrum Coefficients) coefficients computed every 20 ms using 30 ms Hamming windows. The silence was removed from each phrase using an adaptive algorithm based on energy and zero crossing rate criteria. The signal was pre-emphasized using the filter with the impulse response:

$$h[n] = \begin{cases} 1 & n = 0 \\ -0.95 & n = 1 \\ 0 & n > 1 \end{cases} \qquad (11)$$

Using the cepstral vectors obtained from the training phrases, we computed a codebook for each speaker. The codebooks have 128 code vectors and

were generated using the Linde-Buzo-Gray vector quantization algorithm.

Performance criterion

In the speaker verification experiments, a test phrase is compared to the voice model of the speaker whose identity is verified, and an average distance is computed. If this distance is lower than a fixed or dynamically computed threshold, the speaker is accepted, otherwise the speaker is rejected. There are two types of errors associated with the verification process: the rejection of a true customer, called *type I error* – FRE *(false rejection)* and the acceptance of an impostor, called *type II error* – FAE *(false acceptance)*. The compromise between the two types of errors is generally balanced by using a decision threshold. These thresholds are not a priori established. Instead, the total average distance is computed, for which the type I and type II errors are equal: this value determines the *equal-error-rate* (EER).

5. EXPERIMENTAL RESULTS

This section presents the results obtained one the two databases. We tested the classical verification system and EER curves were plotted. The cohort algorithm was tested on both closed (all impostors known) and open database (unknown impostors). Finally we tried to decrease the FRE using a combined threshold-cohort approach.

The reference values for the FRE and FAE curves for the two databases, YOHO and DiSPPALL, are presented in Fig.1 and Fig.2, respectively. T_{eer} is the decision threshold for the equal-error-rate (EER).

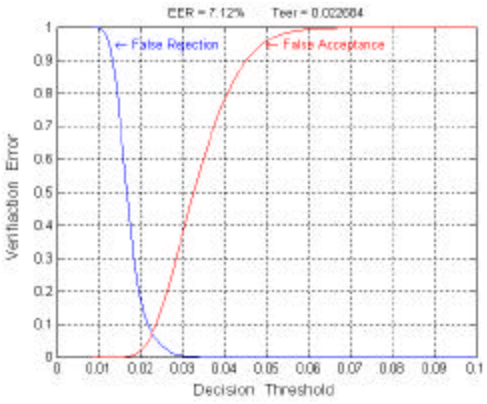

Fig.1. Verification error for the speaker recognition system using the YOHO database; EER = 7.12%

The DiSPPALL database is a real speech database giving a higher EER.

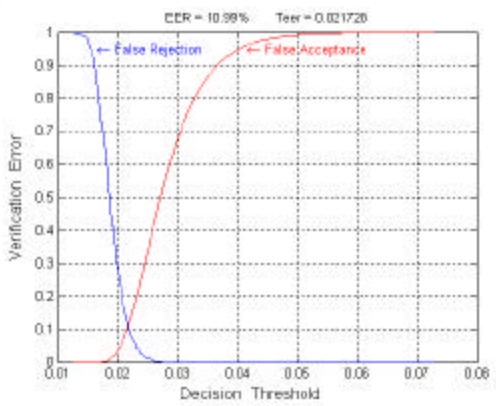

Fig.2. Verification error for the speaker recognition system using the DiSPPALL database; EER = 10.99%

We are now going to use the cohort method in order to improve the performances of our speaker verification system.

We used *the minimum method* to estimate the distance between the test phrase and the impostor model. No significant improvement occurred once the cohort dimension exceeded than a certain value (20 for YOHO, 10 for DiSPPALL).

a. The False Acceptance Error and False Rejection Error using the cohort normalization method on the closed YOHO database are presented in Fig.3. FRE is increasing from 7.11% for 5 speakers in the cohort to 11.82% for a cohort with 20 members. In the same time, FAE is decreasing from 8.94% to 1.29% for the same cohort sizes. We obtain a FAE four times smaller for a FRE less than two times higher.

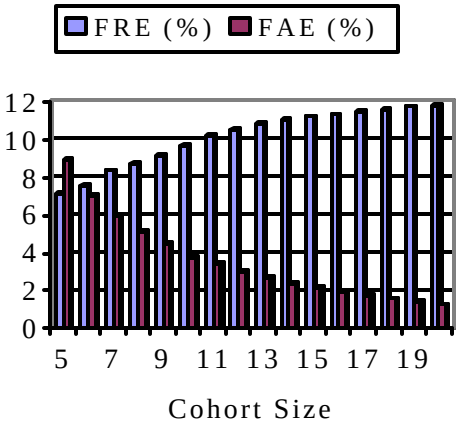

Fig.3. The False Acceptance Error (FAE) and False Rejection Error (FRE) using the cohort method on the closed YOHO database

b. The False Acceptance Error and False Rejection Error using the cohort normalization method on the closed DiSPPALL database are presented in Fig.4.

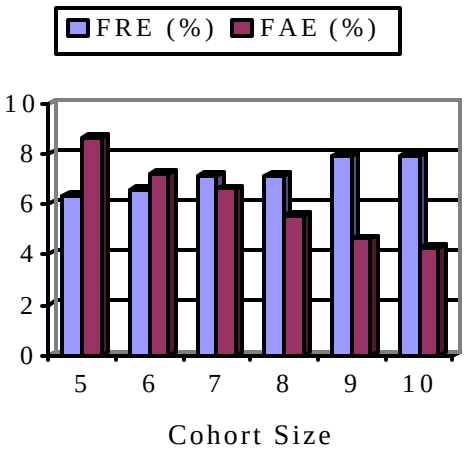

Fig.4. The False Acceptance Error (FAE) and False Rejection Error (FRE) using the cohort method on the closed DiSPPALL database.

For cohort size increasing from 5 to 10 members, FRE has values from 6.34% to 7.88% and FAE is decreasing from 8.61% to 4.23%.

c. Further experiments performed on the open YOHO database are presented in Fig.5**:**

- We removed one speaker and we applied the algorithm for the remaining of 137 speakers; in this case, the removed speaker acted as an unknown impostor.
- We split the database in two parts: the first 69 speakers were retained and the last 69 were removed. We built speaker models and cohort models for the first half of the database and we used the remaining 69 speakers only as impostors.
- We reversed the two parts of the database and repeated the above process.

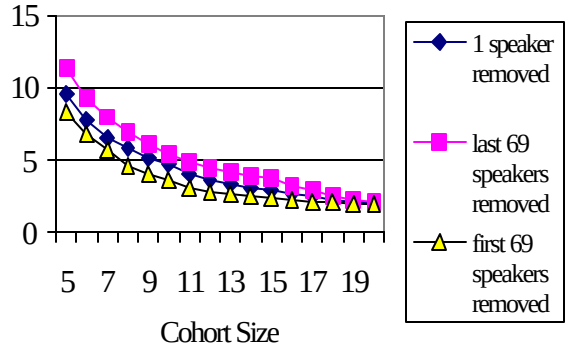

Fig.5. The False Acceptance Error (FAE %) on the open YOHO database

The values of FAE decreased from 9.59% to 2.03% in the experiments with 1 speaker removed, decreased from 11.36% to 2.06% when the last 69 speakers were removed and decreased from 8.37% to 1.90% when the last 69 speakers were retained.

d) During the tests we noticed that a large number of false acceptance errors occurred in association with high distances. Therefore, we used a fixed threshold to eliminate these errors. The threshold was set high enough (T=0.0275) so it would not influence the false rejection errors.

Test results using cohort normalization and a fixed threshold on the closed DiSPPALL database are presented in Fig.6. FRE is increasing from 6.34% for 5 speakers in the cohort to 7.88% for a cohort with 10 members. At the same time, FAE decreased from 8.61% to 4.23% for the same cohort sizes without a threshold and decreased from 6.21% to 3.11% with the threshold set to 0.0275. Thus the threshold reduces the FAE by 35% compared to the cohort FAE.

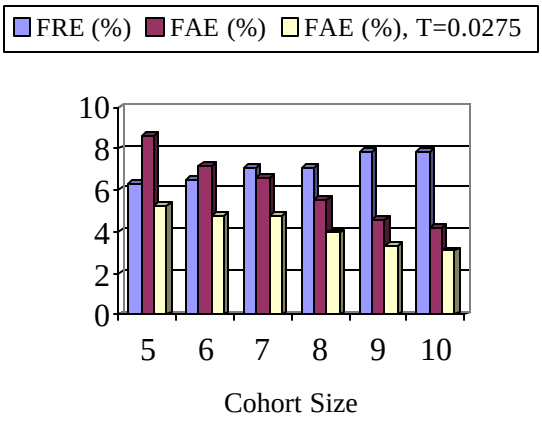

Fig.6. The False Rejection Error (FRE) and the False Acceptance Error (FAE) on the closed DiSPPALL database using cohort normalization and a fixed threshold

e) Test results using cohort normalization and fixed threshold on the closed YOHO database are presented in Fig.7.

For cohort size increasing from 5 to 20 members, FRE had values from 7.11% to 11.82% and FAE without threshold is decreasing from 8.94% to 1.29%. Using a fixed threshold, the improvements are obvious: decreases from 4.99% to 0.83%.

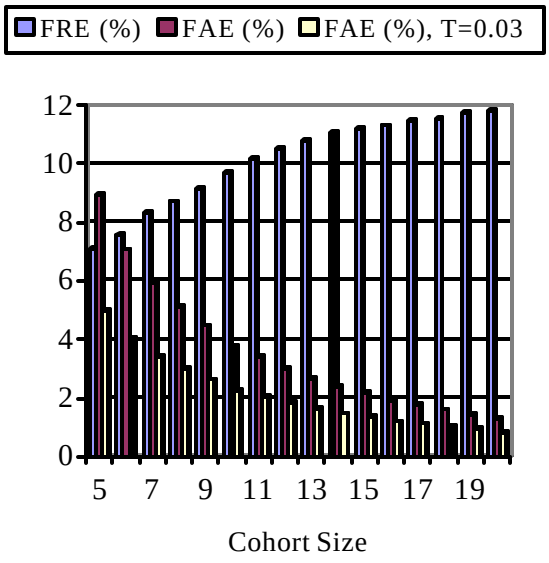

Fig.7. The False Rejection Error (FRE) and the False Acceptance Error (FAE) on the closed YOHO database using cohort normalization and a fixed threshold

f) Fig.8 presents the test results using cohort normalization on the open YOHO database, with variable number of speakers.

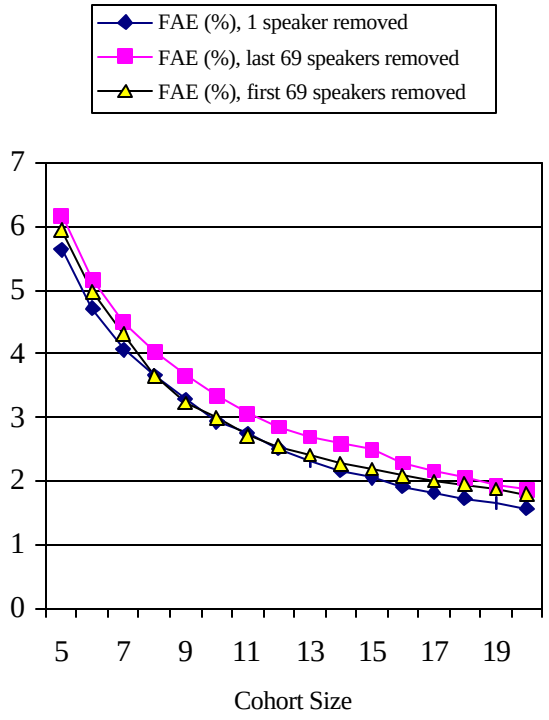

Fig.8. The False Acceptance Error (FAE) on the open YOHO database with variable number of speakers and a fixed threshold

The False Acceptance Error decreased from 5.64% to 1.57% with 1 speaker removed, from 6.17% to 1.86% for the first 69 speakers retained and from 5.95% to 1.79% for the last 69 speakers retained.

In all these experiments, the fixed threshold was set to T=0.03.

6. CONCLUSIONS

Two normalization techniques in the distance domain were presented: the cohort method and the world model method. Evaluation tests were accomplished for a speaker verification system in the following situations:

- in a "classical" manner (without cohorts)
- using cohort models
- using cohort models with a fixed threshold

YOHO is still one of the most comprehensive freely available speaker verification corpora [8]. The experimental results using the YOHO database [3], [4], [6], [8], [13] were very different depending on the complexity of the speaker models and on the number of speakers used for testing. In [8] the EER was about 0.06%, but the test conditions are not defined and the models used were HMMs with one state per phoneme and 5 Gaussians per state. In [3] the models were 32 GMMs (Gaussian Mixture Models), 10 cohorts speakers and 19 MFC (Mel Frequency Cepstrum) coefficients. The EER was 1.45% but there was a matching condition between the train and test sentences. In [13], using 32 GMMs with 26 MFC coefficients the EER were 6%, but only 10 speakers are used. In [6] the EER without cohort was 7.79% and with cohort 3%, but the models were 60 states HMMs with 3 GMMs per state and 39 acoustic parameters. Finally, in [4] the EER was less than 5%, the tests are conducted on the entire database, but with a different approach based on SVM (Support Vector Machine).

The basic approach in our paper is vector quantization, each speaker being characterized by a codebook. The main advantage of the vector quantization approach is the computational resources necessary to implement this approach in terms of processing time and memory required.

In conclusion, performance is improved using the cohorts relative to the "classical system" (without cohorts) and is even better with a fixed threshold.

Several methods were used for the computation of the distance between the sequence of feature vectors and the cohort models: the arithmetic mean method, the geometric mean method, the minimum

method and the weighted average method. In our opinion, the tendency of the two errors (error of false acceptance and error of false rejection) to modify their values in an opposite manner (one decreasing and the other increasing) could be the effect of the specific method used for the computation of the distance.

The results show that the tests on an open data base is strongly dependent on the available set of speakers for the cohort. The speaker recognition algorithms work better in the case of a closed database, where all the possible impostors are known, than in the case of an open database, where there may be unknown impostors.

On a closed database, the improvements are grater if the signal acquisition is performed under noisy conditions. For example, in the case of the DiSPPALL database, which is noisier, the decrease of the EER is almost 50% for a 6-speaker cohort, while with the YOHO database the decrease is much smaller.

The main performance improvement provided by the proposed algorithm is the decrease of the false rejection error, for a constant false acceptance error, as the cohort size increases. Also, the algorithm involves less computational resources than other algorithms, making them more suitable for a commercial application. We use only a 128 codebook-size for every speaker and 12 LPCC coefficients. The use of codebooks instead of more complicated models like GMMs or HMMs has another advantage: it requires less memory space, a digital signal processor being more appropriate to implement the algorithm.

REFERENCES

1. BEIGI H., MAES S, SORENSEN J., *A Distance Measure Between Collections of Distributions and Its Application to Speaker Recognition.* IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP'98 Proceedings. ICASSP, **Vol. 2**, pp. 753-756, 1998.
2. BESACIER L., BONASTRE J.F., *Frame Pruning for Speaker Recognition.* IEEE International Conference on Acoustics Speech and Signal Processing. ICASSP'98 Proceedings. ICASSP, pp. 211-214, 1998.
3. BRYAN M., PELLOM L., HANSEN J.H.L., *An Experimental Study of Speaker Verification Sensitivity to Computer Voice-Altered Imposters.* IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 99 Proceedings. Phoenix Arizona: ICASSP, Vol.2, pp. 837-840, 1999.
4. CAMPBELL W.M., *A Sequence Kernel and its Application to Speaker Recognition.* Neural Information Processing System. NIPS 2001 Proceedings. Vancouver, Canada, 2001.
5. CAREY M.J., PARRIS E. S., *Speaker Verification Using Connected Words.* Proc. Institute of Acoustics, **vol. 14**, Part 6, pp. 95-100, 1992.
6. CHE C.W., LIN Q., YUK D.-S., *A HMM Approach to Text-Prompted Speaker Verification.* IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'96 Proceedings. Atlanta, Georgia: ICASSP, **Vol.2**, pp. 673-676, 1996.
7. FURUI S., *An Overview of Speaker Recognition Technology. Workshop on Automatic Speaker Recognition, Identification and Verification.* ESCA'94 Proceedings. ESCA, pp. 1-9, 1994.
8. JAMES D., HUTTER H.P., BIMBOT F., *The CAVE Speaker Verification Project – Experiments on the YOHO and SESP Corpora. International Conference on Audio- and Video-Based Biometric Personal Authentication. AVBPA'97 Proceedings.* Crans-Montana, Switzerland, 1997.
9. LIU C.-S., WANG H.-C. LEE C., *Speaker Verification Using Normalized Log-Likelihood Score.* IEEE Transactions on Speech and Audio Processing. **Vol. 4**, Issue 1, pp. 56-64, 1996.
10. NAKAGAWA S., AND MARKOV K. P., *Speaker Verification Using Frame and Utterance Level Likelihood Normalization,* SPCHL'97 Proceedings. SPCHL, **Vol. 2**, pp. 1087-1091, 1997.
11. ROSENBERG A.E., DELONG LEE J.C.-H., JUANG B.-H., SOONG F.-K., *The Use of Cohort Normalized Scores for Speaker Verification.* International Conference on Spoken Language Processing. ICSLP'92 Proceedings. ICSLP, pp. 599-602, 1992.
12. ROSENBERG A.E., PARTHASARATHY, S., *Speaker Background Models for Connected Digit Password Speaker Verification.* IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'96 Proceedings. ICASSP, pp. 81-84, 1996.
13. THYES O., KUHN R., NGUYEN P., JUNQUA J.-C., *Speaker Identification And Verification Using Eigenvoices.* International Conference on Spoken Language Processing. ICSLP 2000 Proceedings. Beijing, China, 2000.
14. YU G., GISH H., *Identification of Speakers Engaged in Dialog.* IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'93 Proceedings. ICASSP, **Vol.II**, pp.383-386, 1993.

Received June 24, 2002