Design of an Automatic Speaker Recognition System Based on Adapted MFCC and GMM Methods for Arabic Speech

El Bachir TAZI[†]*, Abderrahim BENABBOU*[††] *and Mostafa. HARTI*[†]

† UFR Informatique et Nouvelles Technologies d'Information et de Communication
B.P. 1796, Dhar Mehraz Fès Maroc
†† Département d'Informatique, Faculté des Sciences et Techniques
FST Fès-Saiss, Fès Maroc

Summary
This paper presents a design of an automatic speaker-independent speech recognition system based on adapted Mel Frequency Cepstrum Coefficients (MFCC) associated to Gaussian Mixture Model (GMM) Methods. This experimental study which has performed for various learning times was conducted around MATLAB®7 language environment. Firstly our goal is to design a robust system that is able to identify any Arabic speaker with a good performance in order to implement it later as the embedded system for access control to high secure areas. Results of the experiments using 72 Arabic speakers indicate that recognition error ratio of 2.15 percent or less can be reaches if the learning and the test utterances times are superiors respectively to ten and five seconds.

Key words:
Arabic speaker recognition, embedded system, GMM modelization, maximum likelihood estimation, MFCC parameterization, time learning adaptation.

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker specific information included in speech waves to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice [15,21]. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers. Another important application of speaker recognition technology is as a forensics tool.

Speaker recognition can be classified into speaker identification and speaker verification:

Speaker identification is the process of determining from which of the registered speakers a given utterance comes. In the speaker identification task, a speech utterance from an unknown speaker is analyzed and compared with speech models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance.

Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Most of the applications in which voice is used to confirm the identity of a speaker are classified as speaker verification. In the speaker verification task, an identity is claimed by an unknown speaker, and an utterance of this unknown speaker is compared with a model for the speaker whose identity is being claimed. If the match is good enough, that is, above a threshold, the identity claim is accepted. A high threshold makes it difficult for impostors to be accepted by the system, but with the risk of falsely rejecting valid users. Conversely, a low threshold enables valid users to be accepted consistently, but with the risk of accepting impostors. To set the threshold at the desired level of customer rejection (false rejection) and impostor acceptance (false acceptance), data showing distributions of customer and impostor scores are necessary.

The fundamental difference between identification and verification is the number of decision alternatives. In identification, the number of decision alternatives is dependent to the size of the population, whereas in verification there are only two choices, acceptance or rejection, regardless of the population size.

Speaker recognition methods can also be divided into text-dependent (fixed passwords) and text-independent (no specified passwords) methods. The former require the speaker to provide utterances of key words or sentences, the same text being used for both learning and recognition, whereas the latter do not rely on a specific text being spoken. The text-dependent methods are usually based on template/model sequence matching techniques in which the time axes of an input speech sample and reference templates or reference models of

Manuscript received January 5, 2010
Manuscript revised January 20, 2010

the recorded speakers are aligned, and the similarities between them are accumulated from the beginning to the end of the utterance. Since this method can directly exploit voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent method. But there are several applications, such as forensics and surveillance applications, in which predetermined key words cannot be used.

In text-independent speaker recognition, a technique based on maximum likelihood estimation of a Gaussian mixture model (GMM) representation of speaker identity, which we have applied in this work, is one of the most popular methods frequently used [22,23]. This method corresponds to the single-state continuous ergodic HMM. Gaussian mixtures model are used for their robustness as a parametric model and for their ability to form smooth estimates of rather arbitrary underlying densities.

The rest of this paper is organized as follows: Section 2 presents a brief description of the suggested architecture for our automatic speaker recognition system. Section 3 is devoted to the state of the art and the mathematical formulation of MFCC parameterization and modelling GMM algorithms used. Experiments and performances evaluation of our system are presented in Section 4. Section 5 concludes and presents perspectives of this study. The last section lists the main references which we have used in this work.

2. Speaker recognition system description

The scheme at figure 1 represents the basic elements of our Automatic Speaker Recognition System for Arabic Speech (ASRSAS). This scheme contains four main modules:

Feature extraction module: it is responsible for the acoustic analysis of voice signal. Thus for each time signal, we extracted a matrix equivalent of features vectors.

Modelization module: it determines the models parameters from those extracted at the previous module. In the decision module following the discrimination between speakers will be made on the basis of these models.

Decision module: a decision on the identity of a speaker is taken on the basis of a similarity measure between his test model and all models of reference contained in the database.

Adaptation Module: A stage adaptation of the learning time has been inserted to optimize the system performance in terms of accuracy and speed.

This is a speaker independent system that uses the approach of global modelization GMM (Gaussian Mixture Model) for modeling and recognition associated with classical parameterization MFCC (Mel-Frequency Cepstrum Coefficients) technique for acoustical analysis.

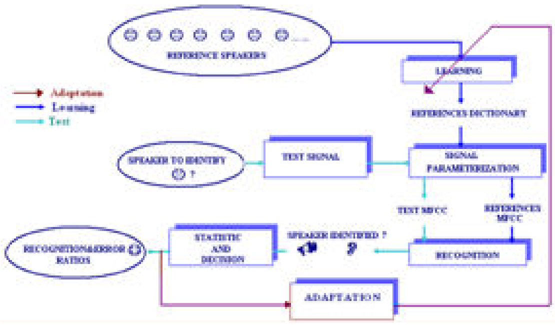

Fig. 1: Block diagram of the system suggested

3. MFCC parameterization and GMM Modelization

A. MFCC feature extraction

This step consists to obtain for each voice signal a "fingerprint", which could then be used for recognition. Several parameterization techniques of speech signals for speech and speaker recognition are cited in the literature [1,3,5,10,12]. Among these techniques we mention the most important are: MFCC and PLP (Perceptual Linear Predictive) that replaces today parameterization by Linear Prediction (LPC) which has the major weakness of estimating spectrum evenly on all frequencies in the audible band. In our system we chose the classical MFCC parameterization which is the state of the art in speaker recognition [7,10,11,15].

The technique of acoustical analysis MFCC involves calculating for each frame 12 cepstral coefficients on a Mel scale which reflects the perceived frequency of the ear [14]. The relationship between the classical frequency in Hertz and Mel frequency is given by (1).

$$m = 2595\log_{10}(1 + f/700) \qquad (1)$$

Where f in Hertz and m in Mel

After applying the Fourier transformation in the short term, the energy is calculated in the critical branches modeled by triangular filters which allowing giving the cepstral coefficients. For obtaining the cepstral coefficients Ci,k, it suffices to operate according to (2), the inverse Fourier transformation(which in practice corresponds to the inverse cosine transformation IDCT).

$$C_{i,k} = \sum_{k=0}^{K}(\log_{10}(\sum_{n=0}^{\frac{N}{2}} Y_{i,n}M_{n,k})\cos(n\left(k-\frac{1}{2}\right)\frac{\pi}{K})) \quad (2)$$

Where $Y_{i,n}$ is the n^{th} Є[1, N] coefficient of the i^{th} Є[1,I] frame Fourier transform. and $M_{n,k}$ the n^{th} Є[1, N] coefficient of the k^{th} Є[1, K] filter.

The block diagram shown in figure 2 summarizes the different steps of extraction block for the parametric calculation of these coefficients with a scale Mel, which we have implemented in MATLAB®7 language.

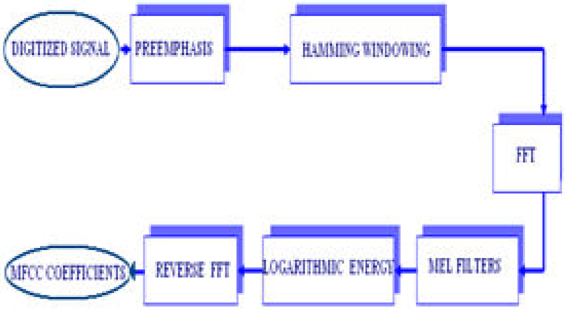

Fig. 2: MFCC features extraction block diagram.

B. GMM Modelization

1) The state of the art

In speaker recognition, there are two types of modeling methods that provide the best results [9,15,17]: The deterministic methods (Dynamic Time Warping DTW and Quantization Vector QV) and statistics methods(Gaussian Model Mixture GMM and Hidden Markov Model HMM). These last are the most used in this field.

In this study, we have chosen to use a modeling based on Gaussian mixture model GMM and we have adapted it to the identification of speakers. This technique which constitute the state of the art was selected for its flexibility at the type of signal and its good compromise between system performance in terms of accuracy and speed and complexity of algorithms [2,4,6,8,13].

The identification using GMM comprises two stages: *A learning/enrolment/training phase* on all files in the database supposed representative of all reference users and a second phase of identifying an unknown speaker called *identification/recognition/test phase.*

The Learning phase aims to estimate the parameters of Gaussian distributions that make up the models corresponding to all acoustics vectors in the database. These parameters are obtained by the K-means algorithm, and then the optimization of the values of these parameters is provided by the EM algorithm (Expectation Maximization) [20,6] whose flowchart is shown in figure 3.

The identification phase allows determining the reference model most likely from the calculation of the probability for each vector acoustic of the signal test. The likelihood of a vocal sound made of a temporal sequence of several vectors is the geometric mean of the probabilities of each of its vectors. The model of the speaker elected as one that matches the test signal is one for which the value of average likelihood is maximum [7,12,16].

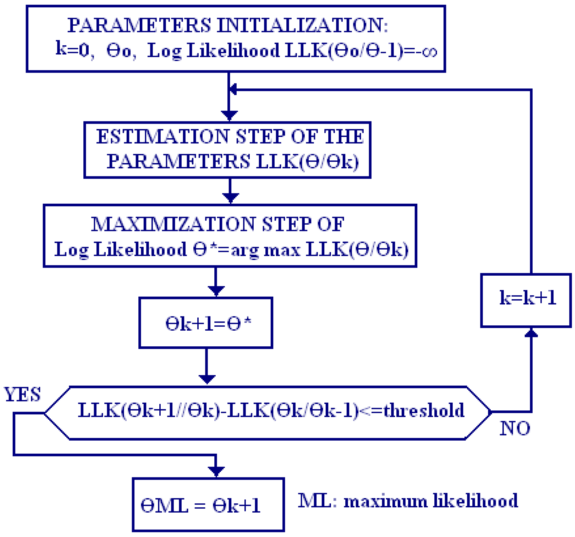
PARAMETERS INITIALIZATION:
k=0, Θo, Log Likelihood LLK(Θo/Θ-1)=-∞
ESTIMATION STEP OF THE PARAMETERS LLK(Θ/Θk)
MAXIMIZATION STEP OF Log Likelihood Θ*=arg max LLK(Θ/Θk)
Θk+1=Θ*
k=k+1
YES
LLK(Θk+1//Θk)-LLK(Θk/Θk-1)<=threshold
NO
ΘML = Θk+1
ML: maximum likelihood

Fig. 3: EM algorithm implemented for maximum likelihood estimation computation.

2) Mathematical Formulation

Let i be the number corresponding to one speaker in the database, x_i represents a signal belonging to the speaker i and Xx_i represents the model of speaker i resulting of the signal x_i.We also note £(x_i/ Xx_j) the likelihood of x_i knowing the model Xx_j.

For a y_t vector of d dimension, the multi-dimensional Gaussian distribution denoted N(μ,Σ) has a probability density function $\mathcal{F}\mu,\Sigma(y_t)$ given by (3)

$$\mathcal{F}\mu,\Sigma(yt) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} e^{(-\frac{1}{2}(yt-\mu)^{T}\Sigma^{-1}(yt-\mu))} \quad (3)$$

Where μ and Σ are respectively the average vector of d dimension and the covariance matrix of dxd dimension of the distribution. The function $\pounds(y_t/\mu,\Sigma)=\mathcal{F}\mu,\Sigma(y_t)$ is called the likelihood function of the distribution.

The Xx_i models used are the GMM (Gaussian Mixture Models). Each GMM X is a weighted sum of multivariate Gaussians (3) defined by the vector of parameters $\Theta x=(c1, \ldots. ck, \mu 1, \ldots. \mu k, \Sigma 1, \ldots, \Sigma k)$.

Where k is the number of Gaussian components and ck the weight of the mixture associated with the k^{th} component given that:

$Ck>=0$ and $\sum_{i=1}^{K} ci = 1$

The likelihood for a test vector y_t is produced by the mixture of Gaussian GMM X is expressed by (4)

$$\pounds(yt/X) = \pounds(yt/\Theta x) = \sum_{i=1}^{K} ci\,\pounds(yt/\mu i\,,\Sigma i) \quad (4)$$

For a speech signal y containing n samples y=(y1, y2, y3, ..., yn), the likelihood of this signal knowing the GMM X model is given by (5)

$$\pounds\left(y/X\right) = \prod_{i=1}^{N} \pounds\left(yi/X\right) \qquad (5)$$

Where y_i is the i^{th} sample of y signal.

4. Experiments and results

In this section, we describe the steps followed for implementing our automatic speaker recognition system.

A. Speaker database description

In this phase we have used the WAVESURFER1.8.5 free software tool [24] that allows a series of pre-treatment of the recorded signal before proceeding to parameterization phase. Thus we have established four databases, containing each of them 136 files corresponding on a population of 72 Arabic speaker of mixed sex (45 males and 27 females). Each individual had participated with 8 different recordings: 4 for learning and 4 others for testing. In this study we set the learning time to twice that of the test and that these durations vary practically of one second to half a minute. All productions sound from the speakers, were directly digitized in WAV PCM format and sampling at 16 kHz frequency with 16 bit mono quantization. Figure 4 bellow, shows a part of learning recording for the fifth speaker in our database for a period of about 30 seconds.

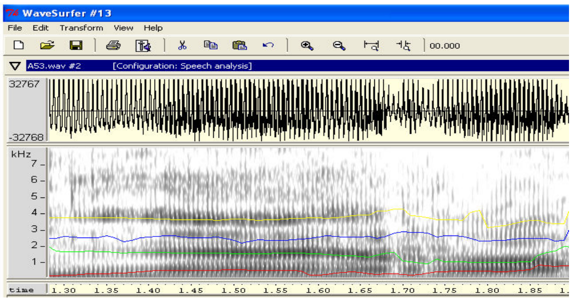

Fig. 4: Temporal spectre of the fifth speaker recorded with wavesurfer tool

B. Parametric characteristics extraction

In this phase we have developed a program to determine the matrix of vectors corresponding to acoustic MFCC parameters characterizing the speech signal equivalent to each record. Figure 5 bellow shows the 12 MFCC coefficients corresponding to this last learning record for the first speaker.

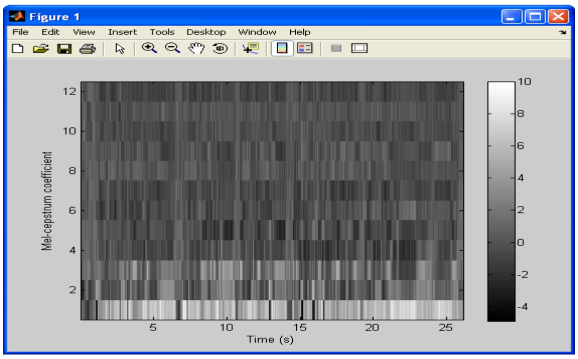

Fig. 5: MFCC Coefficients of learning recording for speaker 1.

The FFT has created these cepstral coefficients was computed for each acoustic vector on 512 points as the value that we considered the optimal resolution between time and frequency, with an overlap of 256 samples. Given the value of the sampling frequency used (16 Khz), this corresponds to an analysis window of width equal to 32 ms with an overlap of 16ms.

C. GMM Modelization

In this section we have conducted a GMM modeling of the training and test corpus for each speaker. Figure 6 shows the results of modeling parameters corresponding to the twelve parametric characteristics of reference model of the first speaker of the database. After a series of tests, the number of Gaussians was fixed at 4 as the value that gives optimal performance between speed and accuracy of the system [19].

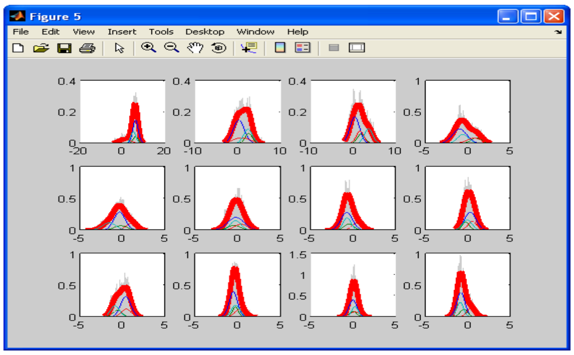

Fig. 6: *GMM reference models for the fifth speaker in the database.*

For an accuracy threshold fixed at 10^{-3}, we noted that the algorithm converges quickly (making two iterations at the estimation phase and a number of iterations less than 10 at the maximization phase). These results approve our choice of this technique for an embedded system with limited resources**.**

D. Tests of speakers recognition

By applying the classical relationship of identification ratio [15] given by (6) we have obtained the results of system performance reported in Table 1. We note that all users of the database have participated at the identification tests.

$$\textit{Identification ratio} = \frac{\textit{positive test number}}{\textit{total test}} \qquad (6)$$

Learning Time	***Mean Identification. ratio***	***Mean errors ratio***
Isolated word	95,53%	4,47%
Simple sentence	96,23%	3,77%
Speech of 10 s	97,84%	2,16%
Speech of 30 s	97,88%	2,12%

Tab.1: *Performance results of our system.*

It is clear that the best results are obtained from a learning time of the order of ten seconds and that beyond this level, the performances remains stable and practically unchanged, the identification ratio is around 97.85%, but the computing time increases very significantly. This justifies our approach of seeking a minimum level of learning that ensures both acceptable accuracy of the system without affecting its response time by unnecessary complexity of calculation. Especially since it contributes to the minimization of material resources deployed in terms of memory and CPU processing power in an embedded system [2,4,17,18].

Given the fact that the evaluation of system performance is dependent on the number of records available in database [7,12] and the comparison with other systems may not be meaningful and true only if the same database is used [15], then it will be difficult to make the comparative evaluation. Knowing that there is no standard corpus available in Arabic language, we were forced to create our own and use it to test our system. Given the obtained scores, it seems that the first results are encouraging and promising.

5. Conclusion and further work

In this paper, we presented the performance of the global modelization approach GMM and classical parameterization MFCC for an automatic independent speaker recognition system. This in four cases varied depending on the length of learning time. The experimental tests of our system implemented in the MATLAB®7 programming platform, show that these techniques of parameterization and modelization associated with proper adaptation of learning time give promising results.

In the study presented no adjustment to the number of MFCC coefficients, the accuracy threshold and the language of the speaker has been achieved. A logical extension of this work is to consider different forms of adaptation. On the other hand, we plan to implement this design on an embedded system using a DSP (Digital Signal Processor) to achieve a voice electronic lock to control access to highly secure areas. For that we are currently working on optimizing our algorithms in order to ensure the best performance in terms of response time, accuracy and robustness for the future embedded system with limited resources.

References

[1] M.F. AbuEl-Yazeed, M.A. El gamal, M.M.H. El ayadi, On the determination of optimal Model Order for GMM-Based Text-Independent Speaker Identification, EURASIP Journal on Applied Signal Processing 2004.

[2] Astrov and Sergey, Memory Space Reduction for Hidden Markov models in Low-Resource Speech Recognition systems, in 7th International Conference on Spoken Language Processing. September16-20, 2002 Denver, Colorado, USA.

[3] M. Barakat, détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabe, Thèse de doctorat, université lumière LYON2, 2000.

[4] Eloi Batlle, José A. R. Fonollosa, Computational real time cost and memory requirements for speech recognition systems, department of signal theory and communications polytechnic university Barcelona Spain, 2000.

[5] Campbell, Joseph P.,Jr. Speaker Recognition: A Tutorial, Proceedings of IEEE, vol. 85,no. 9, pp. 1437-1462, September 1997.

[6] Reynolds, Douglas A. Speaker Identification and Verification Using Gaussian Mixture Speaker Models, Speech Communication. vol. 17, pp. 91-108, 1995

[7] Reynolds, Douglas A. A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification, PhD Thesis. Georgia Institute of Technology, August 1992.

[8] Reynolds, Douglas A. Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. vol. 10, pp. 19-41, 2000.

[9] Reynolds, Douglas A. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker, Model. IEEE Transactions on Speech and Audio Processing. vol. 3, n. 1, pp. 72-83, January, 1995.

[10] Rabiner, Lawrence R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of The IEEE, vol. 77, no. 2, February 1989

[11] Jayant M. Naik. Speaker Verification: A Tutorial. IEEE Communication Magazine, pp. 42-47, January 1990.

[12] Frédéric Bimbot, jean-Francois Bonastre et all, A tutoriel on text-independent speaker verification, EURASIP Journal on Applied Signal Processing vol. 4 pp 430-451 2004

[13] Benoit G. B. Fauve, Driss Matouf, Nicolas Scheffer, Jean Froncois Bonastre et John S. D. Mason, State of the art performance in text Independent speaker verification through open source software, IEEE transactions on audio, speech and language processing, vol.15 no. 7, September 2007

[14] Othman Khalifa, S. Khan, Md. Rafiqul Islam, M. Faizal, D. Dol, Text Independent Automatic Speaker Recognition, 3rd International Conference on Electrical&Engineering ICECE 2004.

[15] Yassine Mami, Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence, Thèse de ENST, Paris France octobre 2003.

[16] Jamel Price and Ali Eydgahi, "Design of Matlab®-Based Automatic Speaker Recognition Systems" 9th International Conference on Engineering Education 2006.

[17] Mosur K. Ravishankar, Efficient Algorithms for Speech Recognition, PhD Thesis of University Pittsburgh, 15 May 1996.

[18] Tomi Kinnunen, Evgency Karpov, and Pasi Frant, Real-Time Speaker identification and verification, IEEE transactions on audio speech, and language processing, vol. 14, no 1, January 2006.

[19] E. B. Tazi, A. Benabbou, M. Harti, Conception d'un Système de Reconnaissance Automatique du Locuteur en Langue Arabe sous MATLAB. JOSTIC'08 Rabat 3-4 Novembre 2008

[20] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, B, 39, 1–38. December 1976.

[21] S. Furui, An Overview of speaker recognition technology In Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pages 1-9, Martigny, Switzerland, April 1994.

[22] Furui sadaoki and all, Research on individuality features in speech waves and automatic speaker recognition techniques; Speech Communication journal. Vol. 5, no. 2, pp. 183-197. 1986

[23] Furui sadaoki and all, Comparison of speaker recognition methods using statistical features and dynamic features; Acoustics, Speech and Signal Processing, IEEE Journal 1981

[24] http://www.speech.kth.se/wavesurfer/

Mr El Bachir TAZI graduated in Electronic Engineering from ENSET Mohammedia, in 1992. He received his postgraduate diploma DEA in Automatics and Signal Processing and University Doctorate in Electronics, Automatics and Computer Sciences from USMBA University, Fez in 1995 and 1999, respectively. Now he is working on his PhD. His thesis is on the Speaker Recognition. He is a Professor of Electronics and Computer sciences in Technical School in Fez. He is a member of the Research Unit UFR: Informatics and New Technologies of Information and Communication at USMBA Fez, Morocco.

Mr Mostafa HARTI received the PhD in Computer Sciences and Statistics from ULB University, Belgium, in 1996 and University Doctorate in Computer Sciences from University, Nancy I, France in 1986. Professor in Sidi Mohamed Ben Abdellah University, Fez, Morocco. Major Fields: G. I. S, Information System Governance, Databases, development, Language Processing Text and Speech, Statistics. Chairman of the Research Unit UFR: Informatics and New Technologies of Information and Communication at the Faculty of Sciences Dhar Mahraz, Sidi Mohamed Ben Abdellah University USMBA Fez, Morocco.

Mr Abderrahim BENABBOU received the PhD in Applied Computer Sciences from ENSIAS, Mohammed V-Souissi University, Rabat, in 2002 and University Doctorate in Computer Sciences from Mohammed V University, Rabat, in 1997. Professor in Sidi Mohamed Ben Abdellah University USMBA, FST, Department of Computer Engineering Fez, Morocco. Major Fields: Natural Language Processing, Speech processing, Human-Machine Interfaces, Embedded Systems, Artificial Intelligence and Object Paradigm.