International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 184-194

Automated Extraction of Features from Arabic Emotional Speech Corpus

Mohamed Meddeb, Hichem Karray and Adel.M.Alimi

REGIM: REsearch Groups on Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia
{*dr.mohamed.meddeb, dr.hichem.karray, Adel.alimi}@ieee.org*

Abstract---This paper presents the principal phase of extraction and recognition of the basic emotions in the Arabic speech applied to five emotional states were taken into effect; neutral, sadness, fear, anger and happiness. Emotional speech database REGIM_TES was created and evaluated to provide all practical experiences of extraction. The database described in section 3 has been recorded and processed in this vein. The selected descriptors in our study are; Pitch of voice, Energy, MFCCs, Formant, LPC and the spectrogram. Descriptors showed the importance of the Arabic language on the physiological events and the influence of culture on emotional behavior. Results performed in this work showed that pooling together features extracted at different sites indeed improved classification performance.

A comparative study between the kernel functions has enabled us to promote the RBF kernel SVMs multiclass classifier performing the classification phase.

Keywords: Arabic speech, multiclass classifier, SVMs, descriptors, emotions.

I. Introduction

Speech is the oldest form of communication that can interpret thoughts. To be intelligible to others the word is organized according to the syntactic and semantic rules of the considered language. Beyond words and semantic content of the message is the entire body which is expressed through vocalization. We cannot recognize physical ache without the latter is explicitly verbalized. Thus the voice modulation allows a transmission of a wide spectrum of emotions, feelings, desires and intentions.

Emotions have an important role in human communication [20]. However, the emotion itself is a multimodal concept. Detecting emotion requires interdisciplinary studies, including visual, textual, acoustic and physiological signal areas.

Despite great progress in speech recognition [10], we are still far from a natural interaction between man and machine, because the machine itself does not understand the emotional state of the speaker. Applications directly affected by verbal information relevant to the emotions of the mental state of the driver system. It can also be used as a diagnostic tool for therapists. It can provide assistance to elderly gents and motor disabilities [26][17]. As it takes place in the daily life of gents for better managed waves of information flow following the progress of digital information technology. Emotion recognition from speech utterance is complicated relatively to the following reasons. First, it is not obvious which voice features are more powerful to separate between emotions. Secondly, there may be more than one emotion perceived in the same statement. In this paper, we consider the huge of utterances, specifically, we will proceed to the segmentation of Arabic corpus without working out neither syntactic nor delimiters linguistics. Acoustic flow is considered continuous so that our system can estimate the emotional responses along the conversation as well as in the silent sound (energy zero, unvoiced sound, noise). According to previous studies on the emotions [5][7][8], we notice the absence of the emotional data bases sound of the Arabic language. Environment and speaking style attribute a distinct position to Arabic speech; which entails the need to create an emotional basis called REGIM_TES [1][26]. Besides, it is very difficult to determine the boundaries (Endpoints) between these parties. In other hand how emotions expressions generally depends on the speaker, culture [14][15] and environment. Emotion has no commonly accepted academic definition. However, people know their emotions when they feel them. For this reason, the researchers were able to study and define the different aspects of emotions. It is widely accepted that emotion can be characterized by two dimensions: activation and valence. Activation refers to the amount of energy required to express a certain emotion. According to some physiological studies on the mechanisms of emotions productions that were made by Williams and Stevens [9], the sympathetic nervous system is excited with the emotions of joy, anger and fear.

Problem statements

In this work, our goal is to select the efficient features for Arabic speech to develop an automatic emotion recognition system using the techniques of signal processing and pattern recognition. We focus exclusively on the paralinguistic information (prosodic and spectral) for recognition of the emotional state of the speaker. The system will be trained and tested to extract the features and identifiers of each patent emotion in the corpus data. The emotional states of the speaker will be classified using SVM about one of the basic emotions Ekman [16].

II. Related Works

A fair amount of approaches have been utilized for affect classification in the majorities of languages unless Arabic one. The most naïve of which is Keyword Spotting where emotion categorization occurs based on the detection of unambiguous affect words like "sad, happy. In [23] Affective Lexicon and [24] Affective Reasoner are two examples for this approach.

Another approach is Lexical Affinity which assigns a probabilistic affinity for particular emotions to arbitrary words. Other methods require a deep understanding of the linguistic models (fig.1) to classify the emotion. Another method suggested by [25] rests on psychological theories about human goals, desires and needs to build models of emotion.

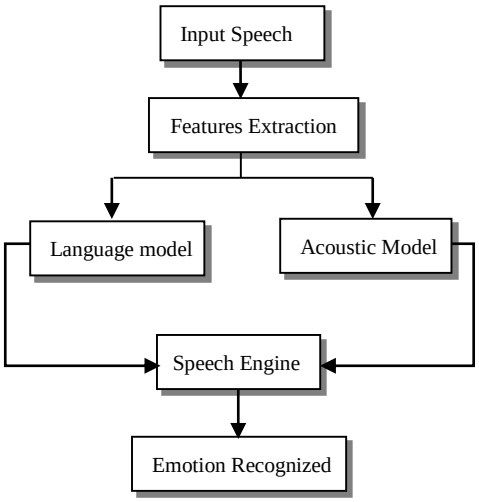

Fig.1. General Model of emotion recognition

Machine learning algorithms have also been used. Those systems can take into account not just the emotion keywords, but also probabilistic affinity. This resulted in an improved accuracy flat classification on the same corpus and using the same machine learning algorithm (SVM).

Most previous research on assessment of non-native speech has focused on restricted, predictable speech. When assessing spontaneous speech, due to relatively high word error rates of current state-of-the-art ASR systems, predominantly features related to low-level information have been used, such as features related to fluency, pronunciation or prosody [18]. Several features related to the high level aspects have been used previously, such as the content and the discourse information.

There has been less work measuring spoken responses in terms of the higher level aspects. [Zechner and Xi 2008] used a content feature together with other features related to vocabulary, pronunciation and fluency to build an automated scoring system for spontaneous high-entropy responses. This content feature was the cosine word vector product between a test response and the training responses which have the highest human score.

The experimental results showed that this feature did not provide any further contribution above a baseline of only using non-content features, and for some tasks the system performance was even slightly worse after including this feature.

In the past the main focus was on prosodic features, in particular pitch, durations and intensity [20]. Comparably small feature sets (10-100) were first utilized. In only a few studies, low level feature modeling on a frame level was pursued, usually by HMM or GMM. The higher success of static feature vectors derived by projection of the LLD such as pitch or energy by descriptive statistical functional application such as lower order moments (mean, standard deviation) or extrema is probably justified by the supra-segmental nature of the phenomena occurring with respect to emotional content in speech. In more recent research, also voice quality features such as HNR, jitter, or shimmer, and spectral and cepstral features such as formants and MFCC have become more or less the "new standard". At the same time, brute-forcing of features by analytical feature generation, has become popular. Within expert-based hand-crafted features, perceptually more adequate features have been investigated, reaching from simple log-pitch to Teager energy or more complex features such as articulatory features (e. g. centralization of vowels). Further, linguistic features are often added these days, and will certainly also be in the future.

III. Assessment Of Database

a) Text corpus of data base

Emotion can be recognized as confidential in very short phrases like " Yes, نعم " or " No,لا ". This means short sentences or even single words are appropriate to analyze the emotional characteristics in speech. But it can be interesting to analyze passages to study speech pauses and specific emotional sounds like laughter or sighs. It is best to choose statements that often appear in everyday communication. It is very difficult to know how emotions are perceived by listeners. In normal situations of communication, the listener does not really make conscious decisions on the emotional state. This should be taken into consideration during a listening test with native listeners is designed. If asked to judge the emotional content of an utterance, a conscious decision is required. Therefore, the prompt text is semantically neutral, meaning that it should not involve a specific emotional significance. To assess emotion in speech, we decided to record a sequence of isolated words and sentences that express independent emotion. The choice of words and sentences came out of the Tunisian simple well known in the social environment table.1 language.

Table.1 Text corpus of REGIM_TES

كلمة Words	نطق الكلمة Spelling	جملة Phrases	نطق الجملة Spelling
تفضّل	Tfadhal	اسمع نقلك	Ism3 nkollik
عسلامة	asslama	شوف قدامك	Shouf koddamek
شدّ	chid	فك علي	Fok aliya
خوذ	khouth	قداش عمرك	Kaddach omrik
هات	h â	صلي على النبي	Salli ala ennabi
اعطيني	a3tini	بارك اله فيك	Baraka allah fik
شبيك	chbik	لا اله الا الله	La ilaha illa allah
اقعد	ok3ad	اخرج عليّ	Okhrij alia
قوم	koum	شنو إسمك	Chnoua ismik
إجري	ijri	سّيب من يدّك	Saib min iddik
نقّص	nakkas	أقسم بالله	Oksom billah
قريب	karib	ازرب روحك	Isrib rouhik
بعيد	ba3id	السلام عليكم	Assalamou alikom
لاباس	labes	قول الحق	Koul alhak

b) Listening Test

It is very important to collect speech with a clear emotional content this is guaranteed by a listening test, in which listeners evaluate the emotional content of registration statements. There are considerable differences in the recognizability of different emotions in spoken utterance. To evaluate our database we invited ten people (not in the registration phase) who listened to each recording and gave their opinions and decision according to each record the corresponding emotion.

The operating principle is to charge randomly the record which will be listened from our emotional database, the auditor is required to listen to the recording shown on the screen of the computer, and then it should be pronounced and decided on emotion expressed in the registration by checking the box in the application correspondent made available. The controller can listen as many times as he wants before it ticks the box. Once the box is checked you can no longer change. The same action is repeated ten times by different reviewers. The decision message will be the average of the responses of all the opinions of reviewers for each type of emotion. The message viewed will confirm or not the presented emotion stored in the emotions speech database. Emotion was correctly identified in 67 % of cases, between 55% and 80%, Surprise and Happiness were often confused and Neutral and Sadness. All listeners have not followed training before the listening test, it was clear that the highest scores in the last 20 words on the first 20 items. 63% of the top 20 were set correctly perceived, but 73 % of the last 20 statements were perceived correct. There was a 10% difference between the score of the first 20 and last 20 statements showing that the auditors had been adapted to the voice in question. Was also examined whether listeners females were more responsive to listeners male. The listeners perceived the correct emotions in 69% of cases in which the male listeners perceived the correct emotions in 66 % of cases.

c) Analysis of Results

According to the public experimental results, to evaluate speech corpus. It can be shown that the judgment on the emotional state of a human is always mixed uncertainty. It was always an overlap between the basic emotion and a percentage of other emotions as shown in the following Table.2

Table.2 Listening tests of the ***REGIM_TES*** database

Actors \ Listeners	RESPONSE in %				
	Neutral	Happiness	Sadness	Anger	Fear
Neutral	60.2	06.1	27.1	03.2	02.4
Happiness	04.5	75.2	00.3	19.2	01.2
Sadness	25.5	00.2	61.7	00.4	09.4
Anger	04.5	19.8	01.3	73.3	03.3
Fear	02.2	12.5	05.7	15.5	66.5

After each listening test, subjects were asked to indicate whether they found the task very easy, easy, difficult or very difficult. They were then asked to describe the factors that facilitate the choice of different emotions. 75% of listeners found it difficult or easy or difficult to identify emotions. Except for some listeners, young actors have found it difficult to easy, whereas children find it very difficult neither easy nor difficult, except for a listener.

Table.3 Files distributions over the 720 sound recordings

		Happiness	Anger	Neutral	Fear	Sadness
720 Words	423 correct choices	70	112	68	80	93
	297 failed choices	74	32	76	64	51

The total number of successful selected records is 423 Table.3 while the failed tests are 297 sound recordings, distributed as follows: Happiness 70 elements, Anger 112 items, Neutral 68 items, Fear 80 elements and finally Sadness 93 items. About the 297 failed choices, we note Happiness 74 elements, Anger 32 elements, Neutral 76 elements, Fear 64 elements and finally Sadness 51 éléments.

Features selection About 720 sound files stored in the database sound data, 423 audio recordings have been recognized by listeners during the phase of listening done in the professional recording studio tests. Volunteers for the test

have successfully detected the correct emotion in 423 recordings while it was sometimes difficult

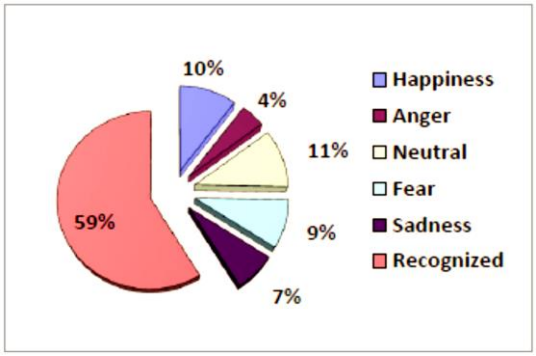

Figure.2 Emotions composition / 423 correct tests

The analysis of variance on response time and good responses show a significant effect of emotion times are shorter, and there are more good answers for joy, then to anger, fear, and finally to sadness. The first hypothesis was that the response time varies depending on the emotion shown, which is confirmed by the recognition test.

The following chart figure.3 shows the average response times depending on emotion. Joy and anger are the two best recognized by the participants emotions, and these two emotions are those represented by the schematic figures. We decided to use later for the experimentation. Presumably these two emotions are very common indeed, and they are more easily recognizable as facial expressions that correspond to them.

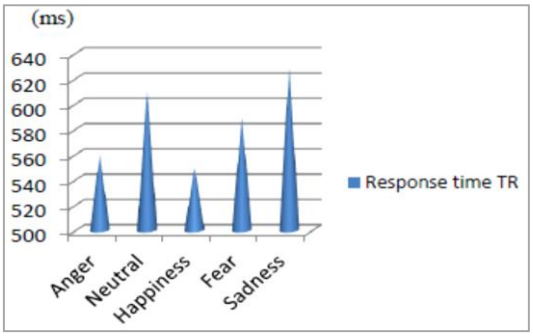

Figure.3 Response time according to the emotion

The analysis also shows a significant effect of duration: false responses, and reaction time decreases when the duration of stimulus presentation increases, with some leveling off after 150 ms. This effect could therefore mean that emotion recognition is more difficult when the presentation is too fast, especially below the threshold of 100 ms, which partly confirms the second hypothesis of this study was that the processing of stimuli becomes aware only after a certain presentation period.

The following chart figure.4 shows the average response times depending on the duration of stimulus presentation

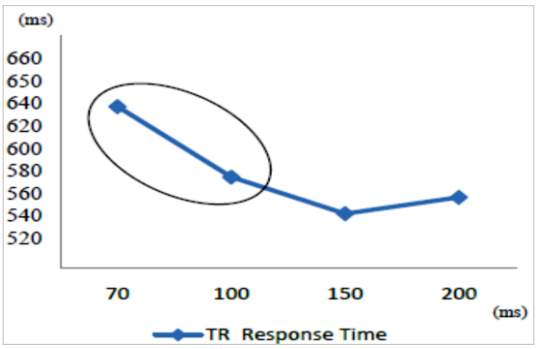

Figure.4 Response time based on the presentation time

In addition, this analysis of response time shows a significant interaction between time and emotion. Indeed, the response time to recognize emotions decreases when the duration of stimulus presentation increases, but the anger, sadness and disgust , the reaction times are longer for 200 ms to 150 ms presentation .

We can therefore ask whether these negative emotions can stop the participants' reaction, or whether different time involves different cortical processing, as some authors assume. It must still be noted that the analysis of the percentage of correct answers does not give the same interaction.

For the experiment, I will use the presentation time of 70 ms and 150 ms, in order to differentiate a very brief presentation, for which recognition of emotions is more difficult and almost "subliminal ", and a slower presentation, which is treated more aware of the subject, based on the results and feedback from participants

IV. Selections Descriptors

a) Reduction of space data representation

In theory, increasing the number of descriptors could improve system performance. However, in practice, the use of too great number of descriptors, in addition to the problem of complexity generated by a high dimension of the representation space data, can eventually put down the performance results. This step of reducing the dimension of the data space representation precedes the learning steps and decision of the classification system. To reduce the space of descriptors, we propose two options:

- The projection of the space of data representation on a smaller space (eg ACP, AD) dimension.
- The selection of the subset of the most discriminating (eg Fisher selection algorithm , genetic algorithm) descriptors , this option has the advantage of being able to extract the most relevant descriptors test step while methods directly require spraying the pre-calculation of the set of descriptors of the test sample.

b) Features Selection & Evaluation

To choose the set of descriptor and the most effective extraction method for classification task we studied a subset of our corpus of audio database REGIM_TES [17], which we

have created through our work and we tested the models based on sets of different descriptors.
Figure.5. shows the different classes of the best known descriptors of speech utterance and addressed in previous research. In the course of our work, the speech is partitioned into small intervals known as frames.

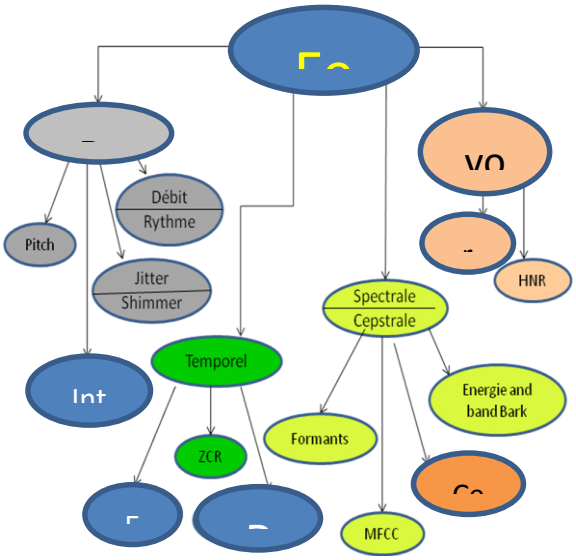

Fig.5. A sample of different classes of descriptors of human voice

The process of partitioning speech into frames based on the information they are carrying about emotion is known as feature extraction. Feature Extraction is a vital step in SER (Speech Emotion Recognition) System. Some of the features which helps to figure out emotions from the speech are-

c) Pitch and Formant

Spectral and formants: Formants (i.e. spectral maxima) are known to model spoken [2][4] content, especially lower ones. Higher ones however also represent speaker characteristics. Each one is fully represented by position [20], amplitude and bandwidth. As further spectral features band-energies, roll-off, centroid or flux are used. Long term average spectrum over a chunk averages out formant information, giving general spectral trends.

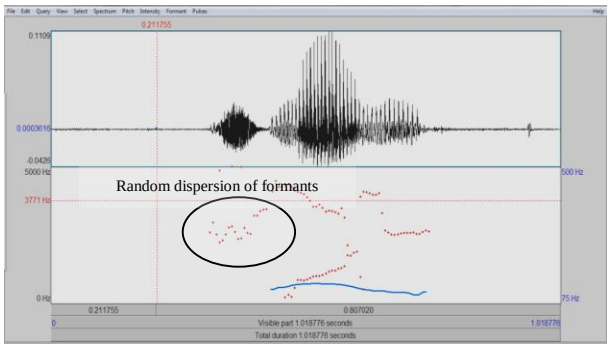

Fig.6. Graphic demonstration of pitch and formants neutral state (Praat)

Figure.6 and Figure.7: handle إسمعني Arabic word simulated by two actors Bilel and Meriam Among 12 professionals actors whom we recruited within the framework of the creation of our emotional database REGIM_TES. This graphical representation shows a group of features, pitch and formants two descriptors by monitoring the waveform of the signal. In case of neutral state the formants shape show a focuses train and irregular cloud randomly dispersed.

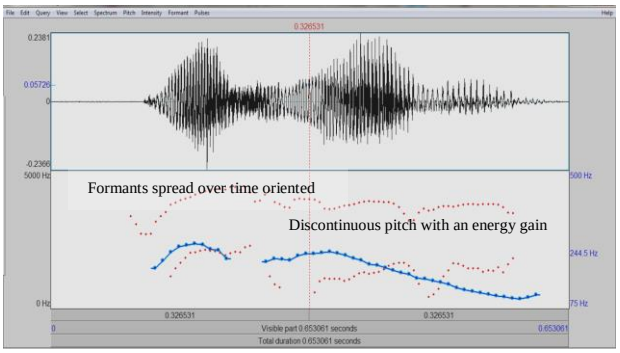

Fig.7. Graphic demonstration of pitch and formants anger state (Praat)

Under the influence of anger he seems to have held a magnetization. Indeed, the formants change the tempo by taking a specific direction with this excitement lasts energy gain almost similar to the pitch during a lapse time necessary for the emotional reaction of the person. The differences in the waveforms of figure.5 and figure.6, return essentially the effect of the frequency output of the woman who is characterized by a fundamental frequency F0 = 350Hz whereas man possesses a more serious frequency F0 = 250Hz, the blue curve shows the behavior of the fundamental frequency, the slowest rate in the sound spectrum called Pitch. To fully experience this descriptor is made to a graphical representation figure.7

d) Influence of fundamental F0 (Pitch)

F0, this is the acoustic equivalent to the perceptual unit pitch; it is measured in Hz and often made perceptually more adequate by logarithmic transformation etc [20]. Intervals, characterising points, or contours are being modelled.

To investigate the importance and influence of the fundamental frequency F0 (pitch) we proceeded as follows: set of all emotions expressed by one person using the same Arabic word قداش analysis, the blue curve shows the pitch in the neutral state, the path is nearly straight with a slight fluctuation in relation to the set of curves.

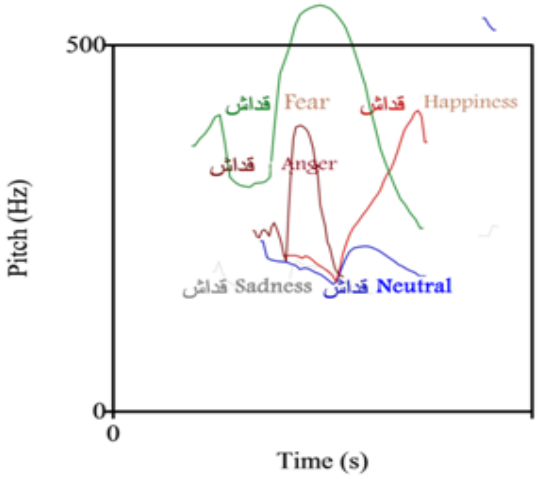

Fig.8. Pitch of word قداش by Najouwa in differents emotions

The state of the neutral emotion is shown in figure.8 by a stable curve (blue) relatively to the other representation and low energy. We note a low pitch frequency variation.
From a physiological point of view, the pitch is a characteristic measure of a person, it represents the sound footprint. Further, thanks to the pitch we can recognize the human being.
The fear and anger curves in figure.8 take the bell-shaped, with a considerable gain in frequency and mainly a decrease in the time interval. We also note that the time axis is different. Pitch explains clearly the sensitivity and strength determine and characterize the types of emotion. The brown curve expresses the anger emotion distinguished by a very sharp peak frequency and a period limited in time.

e) Specificity of Arabic speech

Examples shows the specificity of Arabic speech, the addition of a phoneme figure.9 in some affective expression as anger in particular, increases the probability of detecting an emotion over another, this phenomenon is a little contribution we announced only in the Arabic language, you can see from the chart above that the waveform of the phoneme "ت" that just stick to the front of the word "تفضل" and completely changes the shape of the curve and the values of descriptors.

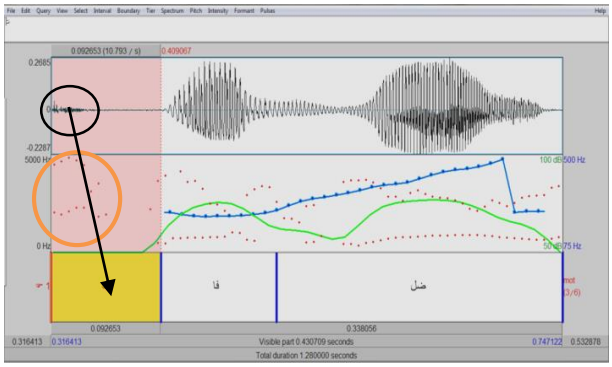

Fig.9. Annotation of the Arabic word "تفضل"

Figure.9: shows the temporal signal of word تفضل by formants, intensity and pitch.
The formants indicate the presence of phoneme "ت" which had stacked in front of the word «تفضل ». To the left of the figure.9, the phoneme is not sensitive to either the "intensity or the pitch, despite the power of these two descriptors [1]. Thanks to harmonics (formants) we could detect this variation. On the other hand the Arabic language is very rich in phonemes and syllables, so it is interesting to note that the Arabic word or phrase changes composition and order of words with random addition of a few words with respect to the composition in the neutral state. This is a common area of study with researchers working on linguistics.

f) Energy and Intensity

Energy, these feature model intensity, based on the amplitude in different intervals, with implicit or explicit normalization. They can model intervals or characterizing points.
Descriptor intensity presents an extreme importance in determining the emotional state of a person. From figure.10 the average value of the intensity is almost constant, the most sensitive factor is the time taken to express emotion.

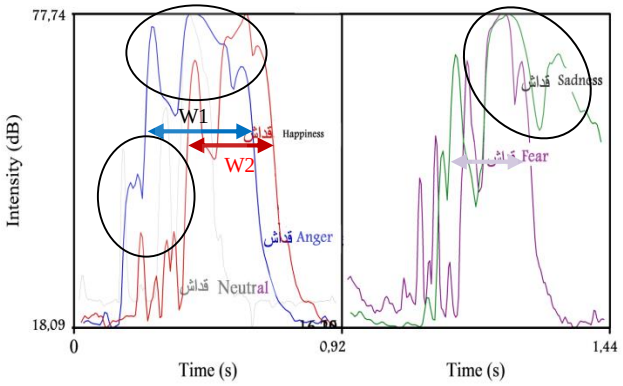

Fig.10. Intensity of word قداش by moez in different emotions

The second area of figure.10 can confirm that both anger emotions and happiness behaves in the same way through the paces of their curve. Which is distinguished, the central width W1>W2 of the globe of the Gaussian and the response time.
The third area of the same figure.10 shows divergence curve emotion of sadness and phase shift with the curve of the emotion fear. In zone one, anger takes over the starting happiness rather with a broader spectrum. Glottal sound production is influenced by the features, namely, blood circulation, heart rate, the number of pulses. In the case of the emotion of sadness, the shape of the infinite intensity curve reflects a physiological behavior of a slow and desperate person.

g) Analysis of descriptors MFCC, LPC, Formant, spectrogram

Cepstrum: MFCC features — as homomorphic transform with equidistant band-pass-filters on the Mel-scale — tend to strongly depend on the spoken content [2][3]. Yet, they

have been proven beneficial in practically any speech processing task. They emphasise changes or periodicity in the spectrum, while being relatively robust against noise.

According to figure.11, we actually discover the physiological behaviour of the human being for a moment of affection (happiness or anger). For a first analysis, we watch a shift in the spectra of the four cases to a left area of energy figure.11.c and frequency formic concentration. In the state of anger, the spectrogram shows an asymmetry. Indeed, the left side of the spectrogram became opaque figure.11.a due to the concentration of frequencies. In the case of cepstral LPC coefficients figure.11.d, the entire distribution is moved to a transitional state of anger and excitation, the same applies to MFCC descriptor study figure.11.b. It is interesting to note that the changes made to the descriptors are in the same geometrical sense.

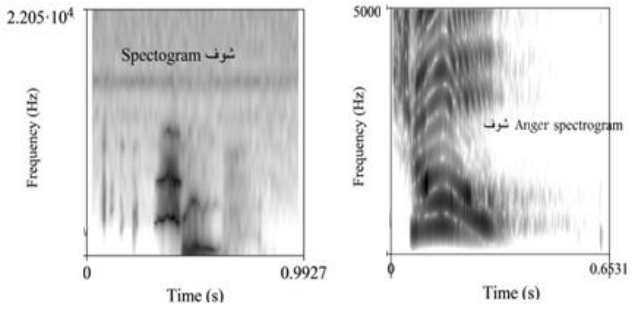

Fig.11.a Spectrogram of arabic word شوف in neutral and anger states

Fig.11.b MFCCs of arabic word شوف in neutral and anger states

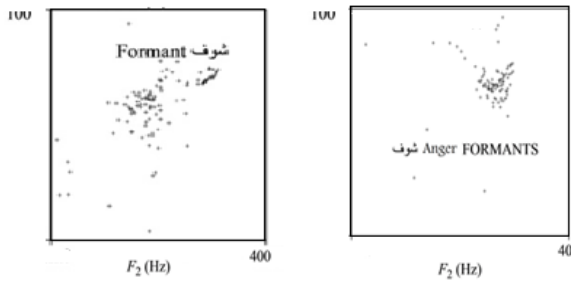

Fig.11.c Formants of arabic word شوف in neutral and anger states

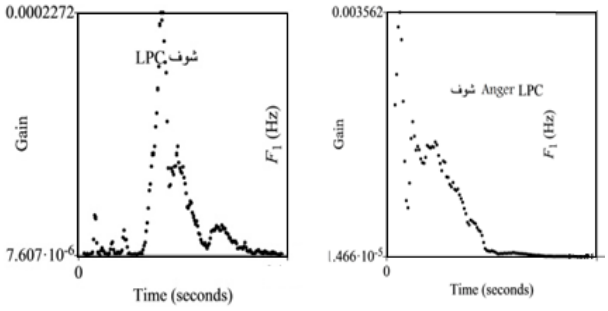

Fig.11.d LPC of arabic word شوف in neutral and anger states

The descriptors MFCC, LPC, Formant and spectrogram are represented from figure.11.a to figure.11.d to study the overall behavior of the main features. Cepstral is the spectrum with an expanded displacement of the neutral axis at the time the same observation state for LPC. The spectrogram shows an acute agitation with anger emotion, the dominant aspect of the opacity clearly explains the concentration of formants in relation to the distribution of formants in the case of emotion in the neutral state.

h) Comparative Study and Optimization

Analysis شوف Keyword (yousra 15 years) based on the variation of the descriptors by the associated numerical values. It is clear that some descriptors vary slightly compared to others and may fall some overlap or similarity between some descriptors. As one can also notice the stability or indifference in the values taken by NZR for all emotions.

TABLE.4 DESCRIPTORS IN ANGER STATE OF WORD شوف

	NZC sec	RMS Pascal	Energy Pa2 S	Intensity dB
Neutral	0.499	0.005	2.7 e-005	48.365
Fear	0.499	0.004	2.9 e-005	46.965
Happiness	0.499	0.026	6.6 e-005	62.565
Anger	0.499	0.053	0.00187110	68.550

Following the acoustic analyzes descriptors NZC, RMS and Energy Intensity previously measured and visualized by the table.4 we note that NZC is a parameter to remove because it is clear that it is useless to overwhelm the measurements.

To conclude, the most relevant to the Arab language descriptors are as follows: the MFCC (12), the Spectrogram, LPC (16) Formant (3), Autocorrelation, fundamental, RMS, energy and intensity. The number of descriptor we have advanced is experimental; it is closely related to our case. It was listed 36 the most meaningful for the Arabic corpus descriptors. Indeed, they have the highest scores on the entire corpus in Arabic.

V. Evaluations & Choice of Kernel

i) Representation and evaluation of results

We seek the best possible detection of emotions without favouring or penalizing others. A complete representation of the results is the confusion matrix, but most studies also report their findings in the form of a classification rate, which makes it easier to compare different experiences. Another interesting measure is the accuracy by emotion; the number of times an emotion is correctly identified divided by the number of times it is identified (good or bad). We chose to evaluate our results using the CL score (Class_wise: average of the diagonal of the matrix). Thus, the detection scores do not depend on the distribution of the test set and the best models for the rate of emotion recognition is about that given by the LC score.

j) Choice of SVMs classifier

A support vector machine (SVM) is a non-probabilistic linear classifier. Unlike k-nearest neighbours (KNN), SVMs take a geometric optimization approach to the classification problem, seeking to construct a hyperplane or a set of hyperplanes in a high dimensional space. The algorithm relies on the intuition that the generalization error is minimized by maximizing the geometric distance between the separating set of hyperplanes and the closest training data point. This formulation gives SVMs some neat properties such as their immunity to local minima (as compared to Neural Networks)

SVMs are also universal learners [17], that is, by using an appropriate kernel, SVMs can be extended beyond linear classification to learn polynomial classifiers. There are several reasons for experimenting with SVMs for the task at hand. Primarily, [26] lists four main points on why SVMs would work for a text categorization problem: high dimensional input space, few irrelevant features, sparseness of document vectors and the susceptibility of text categorization problems to linear separation.
Experimental results on SVMs for text categorization show significant improvements over other classifiers. The work in [1] attributes the highest classification accuracy to SVMs (compared with k-nearest neighbour KNN). In [24] also, in other works Yang and Liu, rank SVMs as the best classification algorithm compared to k-nearest neighbor, Linear Least Squares Fit, Na ïve Bayes and Neural Networks.

At first we used the method to classify k-nearest neighbours (KNN). Results are divided into three parts according to the type of distance functions. The table.5 shows a very low result of recognition about 33 %. This rate explains that our system is not linear and requires a high-dimensional space. On the other hand, the five basic emotions constitute a multi-classes system Ekman [16]. To solve this problem we apply the multi-classy SVM classifier model which proved a big stability.

TABLE.5 K-NEAREST NEIGHBOURS (KNN) EVALUATIONS

KNN	k=1	k=3	k=5	k=7
Euclidennne distance	25,83	**31,25**	27,08	28,75
Tchebychev distance	**33,33**	24,58	26,67	29,58
Hamming distance	29,17	**34,17**	27,92	26,67

k) Maximizing the margin SVM

At the beginning of the classification phase, the results were not entirely consistent; several tests were applied to identify the causes of disturbance of the decision. At the first observation, we noticed the influence of the nature of the sound recording on the extraction operation descriptors and mainly the learning phase. The second problem concerns the acoustic processing (filtering) brought the sound file. To remedy this problem we conducted maximizing the margin SVM. This criterion provides a look at the sensitivity of each variable on the margin of the classifier. The introduction of unitary variables v_i level expression given Gaussian kernel:

$$K(v.x, v.y) = \exp\left(-\frac{\sum_{i=1}^{d}\left(v_i(x_i - y_i)\right)^2}{2\sigma^2}\right) \quad (1)$$
$$x, y \in \mathrm{R}^d$$

- Or is the term by term product of two vectors.
  The margin is of the form.

$$M = 2/\|w\|,$$
$$\|w\|^2 = \sum_i(\alpha_i\, \alpha_j y_i y_j K(vx, vy)) \quad (2)$$

Finally, the derivative of $\|w\|^2$ compared to v_i is proportional to the influence of each variable on the line i. The least sensitive variables are eliminated backward selection. This criterion can be easily extended to SVM one-against-one and one-against-all considering:

$$\sum_{k=1}^{N} \frac{\partial \|w\|^2}{\partial v_i} \quad (3)$$

Where N it is the number of classifiers. The nuclei of the SVM are in competition with each other in order to select the most efficient and most appropriate nucleus to our study on the emotions of Arabic speech.

VI. SVM Kernel Selection

Recognition results using SFS are shown in figure.12 and figure.13, as a function of the number of features selected by SFS. The results have been averaged over the 10 cross-validation trials, and thus the accuracy trajectories in the

figures depict the average recognition rate calculated as the number of samples from all emotions correctly recognized divided by the total number of samples.

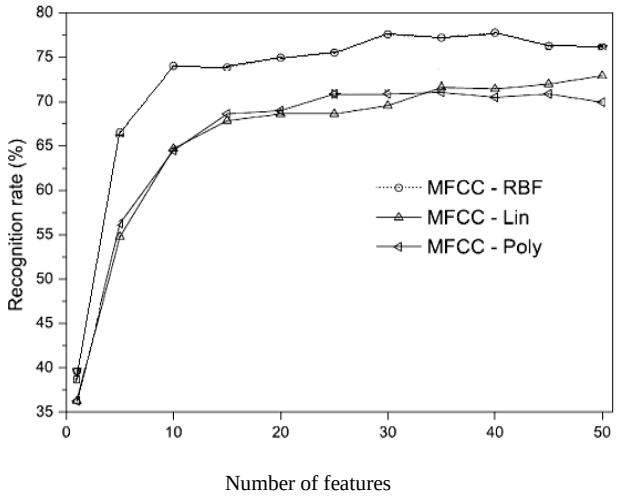

Fig.12. SVM kernel selection / MFCC

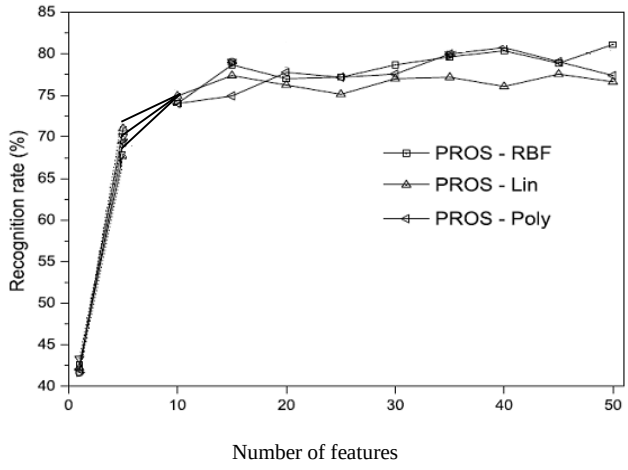

Fig.13. SVM kernel selection / PROS

Figure.12 and figure.13: Shows the RBF kernel furnishes slightly better accuracy than the other two kernels, especially for MFCC features (illustrated in figure.13), and it might also be concluded that the linear kernel often leads to the lowest accuracy. It is justified that the RBF kernel can be a good choice of kernel function in general, as it has a number of advantages over other kernels:

- It can model the nonlinear relation between attributes and target values well.
- The linear kernel is a special case of the RBF kernel.
- It has fewer parameters (only γ) than the polynomial kernel (both c0 and degree).
- It has fewer numerical difficulties compared to the polynomial and the sigmoid kernels.
- Therefore, the RBF kernel is used in all the following SVM classification experiments.

VII. Study of Results SVMs Classifier

The purpose of discussing the results of the SVM classifier is firstly to see how the prosodic, frequency, temporal and spectral / cepstral descriptors react towards classes of emotions expressed in Arabic speech, on the other hand even the opportunities to reduce the number of descriptors in use, reduce the complexity of the classification system for better recognition and reach the optimal time.
As Experimental Strategy Tests we adopted:

- The subjective evaluation of emotions, by performing the subjective listening tests with listeners candidates. The test stimuli were presented in random order to eliminate all the consequential effects in decision making.
- The objective evaluation of emotions, operated by automatic emotion recognition of speech tests.

TABLE.6 CONFUSION MATRIX OF THE GLOBAL CLASSIFIER INCLUDING ALL FEATURES (%)

	Emotion	Anger	Fear	Happiness	Neutral	Sadness
Female	Anger	**75,44**	2,66	21,71	0,46	0,12
	Fear	12,56	**41,79**	23,59	5,9	10,51
	Happiness	38,03	6,51	**50,53**	1,94	2,11
	Neutral	1,12	0,61	7,02	**60**	31,22
	Sadness	0	0,68	0,14	10,3	**86,68**
Male	Anger	**84,93**	5,33	9,33	0,27	0,13
	Fear	11,03	**55,38**	18,72	7,44	7,4
	Happiness	30,88	17,65	**46,18**	2,94	2,2
	Neutral	4,26	3,26	2,65	**58,04**	28,4
	Sadness	1,63	3,06	4,29	17,76	**73,27**

Table.6: The results are in perfect agreement with real feelings and emotions of the human being. Indeed, the feeling of anger is more predictable in men, while the happiness and sadness is a lot better detect in women.

The series of Matlab experiences applied to SVM classifier by way of its kernels (polynomial, linear, radial). Summary the measured average values. The results analysis of table.6 and table.7 confirms the instability of polynomial kernel in this case study, despite the high rate of classification it embodies an excessive overlap between emotions confirmed by the increased error rate.

TABLE.7 PROSODIC AND CEPSTRAL FEATURES

Emotion	Anger	Fear	Happiness	Neutral	Sadness	Rate
Anger	**121**	3	6	0	0	93,07%
Fear	3	**91**	1	3	0	92,85%
Happiness	7	0	**131**	0	0	94,92%
Neutral	0	0	1	**77**	3	95,06%
Sadness	0	1	0	9	**103**	91,15%
Precision	92,36%	95,78%	94,24%	86,51%	97,16%	

TABLE.8 ONLY PROSODIC FEATURES

Emotion	Anger	Fear	Happiness	Neutral	Sadness	Rate
Precision	93,65%	96,62%	90,97%	86,36%	96,19%	

Table.8 accuracy on emotion recognition results by our SVM model is dependent on the type and number of descriptors used. The relatively low accuracy rate presented in the table.5 explains the influence of the type and number of descriptors on the precision value. Indeed, precision is inversely proportional to real-time calculations. i.e. uses more descriptors further increasing the recognition accuracy, more sets of time resolution and calculation.

VIII. Conclusion and Future Works

The speech signal is unquestionably changed when psycho physiological disturbances affect a speaker in an intercultural way. Work overwhelmingly concerned emotions, especially those that psychologist's call primary they were simulated by actors or identified in real situations. Studies on well-defined emotions led to define families of acoustic parameters a sensitive to this phenomenon. A large number of parameters were tested. It appears that the parameters of prosody are the most reactive, but they are not sufficient for reliable detection, or even to define degrees in individual response. Measurements related to speech rate and spectrum also show sensitivity and are necessary complements.

Moreover, the automatic recognition, it should be remembered that it is based on an analysis of the signal, the work on the acoustic analysis leads to precise the nature and the degree of the variations. Recognition and detection of emotion from the Arabic speech is not very well known in research on signal.

In the end, this work identified a challenge of the Arabic language, the influence of culture on vocalization following a psychological excitation, Mediterranean southern region. This culture acts differently on descriptors as shown by the results of the confusion matrix of this study.

It's clear that the choice of descriptors to characterize the Arabic language requires more attention in terms of its relationship to the word of speed and energy.

For example, the emotional state of anger is shown by a transplant of a phoneme or word in Arabic syntax, with varying flow rate characterized sometimes by a long emotional silence, (low value of jitter). Arabic speech is characterized by a wide variation in the energy descriptors, especially the variation of the fundamental (pitch), the MFCC, RMS, LPC and intensity.

Emotion does not have a universal definition, especially the recognition and detection for each medium, each culture and its secret language, among other Arabic. In the future we propose to improve Arabic emotional database, the actual recordings with sequences of spontaneous conversation. The combination of several classifiers may lead to improved precision, especially the method of the fuzzy logic to the real-time recognition. In particular, we think to develop a system feedback learning to update the database [17] of emotion intelligent remote (ITV) models.

Work successor aim to introduce two new parameters, the Jitter and Shimmer representative micro variations of the fundamental frequency and energy, and to test their influence on the proposed recognition system respectively. Under the project of our laboratory, we believe in integrating the module neuronal recognition and language pack to improve the state of the neutral emotion that marks an ambiguity for most gents and automatic systems, to improve the rate recognition of emotions in the Arab word.

IX. References

[1] M. Meddeb , K. Hichem, A. Alimi: " Speech Emotion Recognition Based on Arabic Features" *in 15th international conference on Intelligent Systems design and Applications (ISDA15), Marrakesh, Morocco, IEEE conference, December 14-16, 2015.*

[2] Pooja Yadav & Gaurav Aggarwal: "Speech Emotion Classification using Machine Learning". *International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 13, May 2015.*

[3] S. Demircan and H. Kahramanlı: "Feature Extraction from Speech Data for Emotion Recognition". *Journal of Advances in Computer Networks, Vol. 2, No. 1, March 2014.*

[4] Mayank Bhargavaa,a*,Tim Polzehlb: " Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature". *In Proceedings of ICECIT-2012 Published by Elsevier.*

[5] C. Clavel,"Analysis and recognition of acoustic manifestations of emotions of fear type abnormal situations", March 15, 20.

[6] K. Alter, E. Rank, & S. A. Kotz, "Accentuation and emotions - Two different systems", *in Proc. ISCA Workshop on Speech and Emotion, Belfast, 2000.*

[7] D. C. Ambrus, "Collecting and recording of an emotional speech database technical Report", University of Maribor.

[8] N. Amir, S. Ron, & N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria", *in Proc. ISCA Workshop (ITRW) Speech and Emotion, pp. 29-33, Belfast, 2000.*

[9] C. Williams, & K. Stevens, "Emotions and speech: some acoustical correlates", J. Acoust. Soc. Am. 52 (4 Pt 2) (1972) 1238–1250.

[10] A. Yazi, "Automatic recognition of emotions part of the acoustic signal", Montreal, 17 February 2008.

[11] F. Burkhardt, & W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis", *in Proc. ISCA Workshop (ITRW) Speech and Emotion, pp. 29-33, Belfast, 2000.*

[12] P. Boersma & D. Weenink. (2009) Praat: doing phonetics by computer. Retrieved May 1, 2009.

[13] B. Bouchon-Meunier. (1995). "Fuzzy Logic and Its Applications". Artificial life. Addison-Wesley.

[14] N. Dibben, "Emotion and music". 2009, 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pages 1–3, September 2009.

[15] S.Beller, & A. Bender. (2008). "The limits of ounting: Numerical cognition between evolution and culture". Science, 319 (5860), 213.

[16] P. Ekman. (1984), "Expression and the nature of emotion". In K.Scherer & P.Ekman (eds). Approaches to emotion. Hillsdale, pp. 31.

[17] M. Meddeb , K. Hichem, & A. Alimi, "Intelligent Remote Control for TV Program based on Emotion in Arabic Speech", *International Journal of Scientific Research & Engineering Technology (IJSET), ISSN: (2277-1581) volume 1, 2014.*

[18] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication, 51(10):883–895.

[19] Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. Journal of Technology, Learning, and Assessment, Volume 4, Number 3.

[20] B. Schuller, A. Batliner, D. Seppi, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," *in Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07), vol. 2, pp. 2253–2256, Antwerp, Belgium, August 2007.*

[21] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, 2001.*

[22] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," *in Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech '09), pp. 312–315, Brighton, UK, 2009.*

[23] A. Ortony, G. L. Clore, A. Collins, The cognitive structure of emotions, New York: Cambridge University Press. 1988.

[24] C. Elliott, "The Affective Reasoner: A Process Model of Emotions in a Multi-agent System," Ph.D. thesis, Inst. Learning Sciences, Northwestern Univ.,Tech. Rep. 32, 1992.

[25] M.G. Dyer. (1987). Emotions and Their Computations: Three Computer Models. Cognition and Emotion. 1(3), 323-347.

[26] M. Meddeb, & A . Alimi,"Real Time Automatic Emotion Recognition from Speech (RTAERS*)", International Conference on Engineering and Research 2013.Marocco 2013.*

Author Biographies

MEDDEB Mohamed Masters degree in Electronic (1994), the postgraduate diploma in electrical engineering, department of automatic and signal processing from the National Engineering School of Tunis - Tunisia (ENIT), in 1999. And the PhD degree in computer science (2014) from the National Engineering School of Sfax (ENIS). Member of the REsearch Group on Intelligent Machines (REGIM). Research interests include Human Computer interaction and speech analysis. Former CEO of the public broadcasting institution, expert in new technologies at the University of Mannouba – Tunisia (ESEN).

KARRAY Hichem Received the Eng degree in computer science (1990), the Master degrees in Computer Science from the National Engineering School of Sfax - Tunisia (ENIS), in 2003. and the PhD degree in computer science (2009) from the National School of Engineers . He is member of the REsearch Group on Intelligent Machines (REGIM). His research interests include Computer Vision and Image and video .analysis. These research activities are centered around objet detection in video, video indexing.
He was the organizer of the special session Advanced Video Techniques for Indexing, Browsing, and Retrieval (in 7th International Conference on Innovations in Information Technology (Innovations 2011) in UAE (United Arab Emirates).
He is an IEEE member.

Adel M.ALIMI (IEEE Student Member'91, Member'96, Senior Member'00). He graduated in Electrical Engineering in 1990. He obtained a PhD and then an HDR both in Electrical & Computer Engineering in 1995 and 2000 respectively. He is full Professor in Electrical Engineering at the University of Sfax, ENIS, since 2006. He is founder and dire ctor of the REGIM.
Lab. on intelligent Machines. Prof. Alimi served as associate editor and member of the editorial board of several international scientific journals (IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Fuzzy Systems, Pattern Recognition Letters, NeuroComputing, Neural Processing Letters, International Journal of Image and Graphics, Neural Computing and Applications, International Journal of Robotics and Automation, International Journal of Systems Science, EURASI P Journal on Advances in Signal Processing, Journal of Universal Computer Science). He was guest editor of several special issues of international journals (e.g. Fuzzy Sets & Systems, Soft Computing, Journal of Decision Systems, Integrated Computer Aided E ngineering, Systems Analysis Modelling and Simulations).