International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 6, December 2021, pp. 5438~5449
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i6.pp5438-5449

Emotion recognition from syllabic units using k-nearest-neighbor classification and energy distribution

Abdellah Agrima[1], Ilham Mounir[2], Abdelmajid Farchi[3], Laila Elmaazouzi[4], Badia Mounir[5]
[1,3]IMMI Laboratory, Faculty of Sciences and Technics, University Hassan First, Settat, Morocco
[2,4,5]LAPSSII Laboratory, High School of Technology, University Cadi Ayyad, Safi, Morocco

Article Info

Article history:

Received Jul 21, 2020
Revised Apr 26, 2021
Accepted Jun 12, 2021

Keywords:

Emotion recognition
KNN
Speech signal
Syllables units

ABSTRACT

In this article, we present an automatic technique for recognizing emotional states from speech signals. The main focus of this paper is to present an efficient and reduced set of acoustic features that allows us to recognize the four basic human emotions (anger, sadness, joy, and neutral). The proposed features vector is composed by twenty-eight measurements corresponding to standard acoustic features such as formants, fundamental frequency (obtained by Praat software) as well as introducing new features based on the calculation of the energies in some specific frequency bands and their distributions (thanks to MATLAB codes). The extracted measurements are obtained from syllabic units' consonant/vowel (CV) derived from Moroccan Arabic dialect emotional database (MADED) corpus. Thereafter, the data which has been collected is then trained by a k-nearest-neighbor (KNN) classifier to perform the automated recognition phase. The results reach 64.65% in the multi-class classification and 94.95% for classification between positive and negative emotions.

Corresponding Author:

Agrima Abdellah
IMII Laboratory
Faculty of Sciences and Technics
University Hassan First
FST of Settat, Km 3, B.P: 577 Road of Casablanca, Settat, Morocco
Email: agrima.abdellah@gmail.com

1. INTRODUCTION

Interpreting emotional information is imperative for the social interactions that we have every day [1], [2]. This involves many different components such as body language, posture, facial and vocal expressions. Any information that relates to these components is usually obtained from physiological sensors, sound, or image [3]-[8]. The data is manipulated of a very low level (sound samples or even pixels of images). Between this low-level data and the interpretation that humans make of it, the gap is enormous. Indeed, the manifestation of emotions is an especially intricate area of human communication lying at the intersection of multidisciplinary sciences such as psychology, psychiatry, audiology, and computer science [9]-[11]. The analysis of conversations 'speech analytics', is one of the recent challenges in many applications, for example, health monitoring [12], video games [13], and computer science [14]. A typical domain of learning emotional state from analysis of conversations is call centers. Indeed, a better understanding of the needs of a customer means for the enterprise a better management and greater benefit [15]-[17].

The science of learning lays on data which come usually from different signals. One of the most frequently used signals is the speech signal. Speech is indeed one of the fundamental modalities that men use to communicate. Automatic speech recognition systems give the machine the ability to transform the sound signal into a series of words. The field of automatic language processing provides access to the meaning of this series of words. Starting from these tools (relatively effective), it is necessary to go further. The question is no longer merely about knowing what is said, but also to know the context of the pronunciation of the sentence. It is at this level that the emotional dimension intervenes. Emotion recognition is a difficult errand since its expression and perception vary extraordinarily across cultures, spoken languages, and sentences [18]-[21]. Like any emerging field, it hasn't been completely well-established yet, but to qualify emotional speech, researchers have attempted to represent it by various descriptors related to spoken content (If we do not take into account the variation of the sentence, it is difficult to differentiate between a linguistic and emotional variation), articulatory position (position of the jaw according to the pronounced emotion) [22], and acoustic properties [23], [24].

The contribution we are presenting here wish to gain an understanding of the subject and exposing an exploratory aspect of it. First, unlike the majority of the existing work, we will consider Arabic phonemes CV (plosive consonant/vowel) instead of whole sentences. Second, we will look at data acoustic features in terms of energy and its distribution in some specific bands. Finally, to learn from the constructed data sets and predict emotional state of speaker, we use the instance-based classification algorithms k-nearest-neighbor (KNN).

Over the past fifteen years, increasing number of researchers have been interested in the study of emotions in speech [25]. From data labelled emotionally, a feature set based on spectro-temporal features (for instance pitch, intensity, and energy of the speech that are extracted using algorithms) is developed for processing the voice signal. Supra-segmental representations (also called low-level descriptors (LLD)) are derived from linguistic units such as sentences, words, or phonemes [26], [27]. High-level descriptors (HLD) such as in [28]-[30] are commonly extracted by computing various statistics from the LLDs over the defined linguistic units. The size of the feature set exceeds thousands (HLD and LLD) contingent upon the number of statistics extracted. The discriminating classifiers are then trained on these high-level characteristics for a multitude of tasks such as binary (valence, activation) or categorical (happy, sad, joy, sad, etc.) classifications as well as regression on continuous emotional attributes. Many classification approaches have been applied to emotion recognition systems (SER). Indeed, the hybrid Gaussian mixture model (GMM), neural network with (RBF, PNN, SVM, ELM) [31], [32], and hidden Markov model (HMM) [33] were used to generate a model of recognition by emotion and by gender (man/woman) of the speaker. Also, a support vector machine (SVM) combined with a polynomial nucleus was applied for multi-class and multi-corpus recognition systems [34], [35]. Recently, deep neural networks (DNNs) [36] have been designed to address SER problems. One of the advantages of this approach is transfer learning. Transfer learning consists of learning the first layers of the network performing joint operations from one corpus to another, or from one emotion to another-on more data than the deeper layers assigned to a specific task [37]. Kaya and Karpov [38] used kernel extreme learning machines (ELM), adapted to a base of few samples for many characteristics. They also proposed a new method of normalizing the characteristics as an alternative to the standard normalization of centering-reduction. To summarize the state-of-the-art of the field, we refer the reader to Table 1.

In our study, we investigate the utilization of speech signals for detecting emotion. We propose an emotion recognition method from vocal data based on a signal processing algorithm as shown in Figure 1. Our method consists of three steps: (1) data preparation (2) speech features extraction as shown in Table 2, and (3) classification utilizing the KNN algorithm.

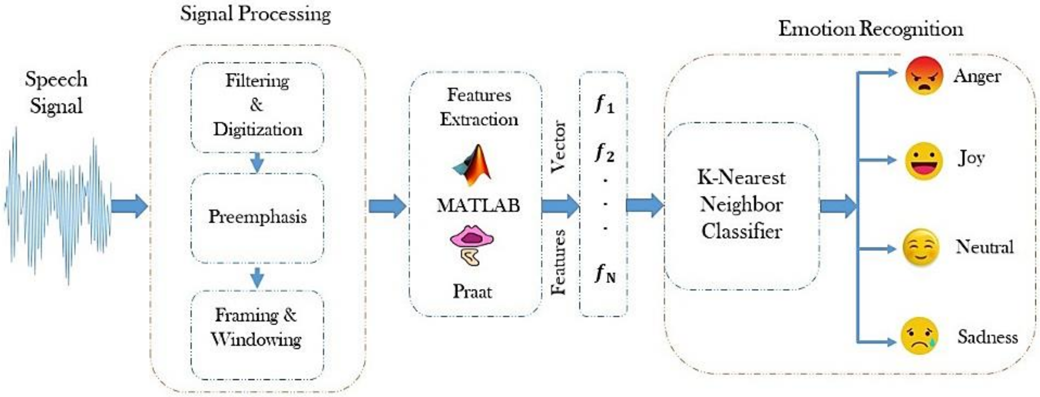


Figure 1. Global speech emotion recognition system

Table 1. A brief summary of SER with different databases, different features and classification algorithms

References	Dataset	Features	Models/Classifiers	Best result
Han *et al.* [39]	IEMOCAP-DB	Pitch-based features, and their delta feature across time frames. MFCC Coefficients,	DNN and Extreme Learning Machine	• 54.30% average recognition rate.
Deng *et al.* [40]	ABC DB, AIBO DB, SUSAS DB	Low-Level Descriptors	Denoising auto-encoders and SVM	• 64.18% recognition rate for ABC DB. • 62.74% average recognition rate for SUSAS DB.
Mirsamadi *et al.* [41]	IEMOCAP corpus	Automatically learned by DRNN, as well as hand-crafted LLDs consisting of F0, voicing probability, frame energy, ZCR, and MFCC	Deep RNN	• Proposed system with raw spectral features has 61.8% recognition rate. • Proposed system with LLD features has 63.5% recognition rate.
Mao *et al.* [42]	SAVEE DB, Berlin EMO DB, DES DB, MES DB	Automatically learned by CNN	CNN	• 73.6% accuracy for SAVEE DB. • 79.9% for DES DB. • 78.3% for MES DB. • 85.2% for EMODB.
Issa, Dias, *et al.* [43]	RAVDESS, Berlin EMO DB, IEMOCAP	Mel-frequency cepstral coefficients (MFCCs), Chroma-gram, Mel-scaled spectrogram, Spectral contrast feature, Tonnetz representation.	Deep CNN	• 71.61% for RAVDESS DB with 8 classes. • 86.1% for EMO-DB 7 classes. • 95.71% for EMO-DB with 7 classes. • 64.3% for IEMOCAP with 4 classes.
Sinith *et al.* [44]	Berlin-Emo SAVEE	Pitch, intensity, MFCC	Support Vector Regression	• Males: 67.5 %. • Females: 70%. • Both: 75%, 61.25%

2. PROPOSED METHOD

In this article, a new characteristic extraction scheme is proposed which is based on a pseudo-phonetic approach [45], [46]. The key point of our work is to extract the characteristics according to different segments such as the syllabic units to remedy the linguistic variation constraint. These segments are identified by manual segmentation of the speech signal. Our developed method is based on extracting clues utilizing signal processing methods. Low-level descriptors (LLDs) and high-level descriptors (HLDs) as shown in Table 2 are obtained from a voice signal labelled by four emotions (anger, sadness, joy, and neutral). Each chosen audio sequence must firstly satisfy the audibility criterion and thereafter passes through the following process:

- Modulating the signal according to a set of contextual, cultural, and linguistic variables whose purpose is to allow communication (emotional or not),
- Capturing the signal produced by the speaker,
- Annotating the signal [47],
- Segmenting the signal in syllabic format,
- Extracting acoustic features by using signal processing techniques,
- Classifying using K-NN algorithm.

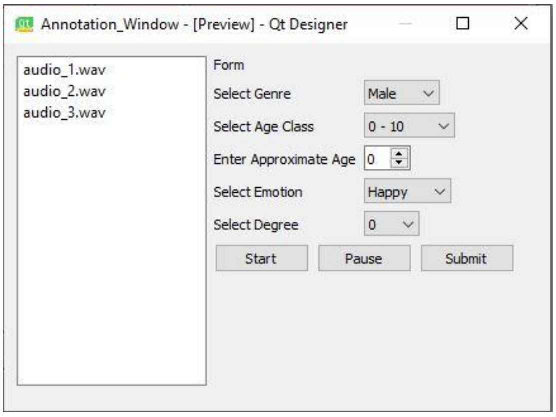
Table 2. High and low-level settings for the emotion recognition system

Low-level descriptors	High-level descriptors
The four first formants	Standard deviation
Intensity	Median pitch
Pitch	Mean pitch
Jitter (Local, Absolute)	Maximum pitch
Shimmer (Local, absolute)	Minimum pitch
Pulses	
The energy in the six bands (E_1, E_2, E_3, E_3, E_4, E_5, E_6)	
Percentage of energy in the six bands (E_{1n}, E_{2n}, E_{3n}, E_{3n}, E_{4n}, E_{5n}, E_{6n})	

2.1. Data preparation

Any scientific study in machine learning is extremely dependent on the data that is used to describe the phenomenon to be modelled. Therefore, the collection of information adapted to the task that we want to model becomes a major issue for obtaining a detection model with a sufficiently strong generalization power. Nowadays there has been some genuine work in the zone of emotion recognition in general and emotion from sound specifically; however, an enormous portion of this work has been assessed on acted speech [48], [49], and very little work has been done on real and spontaneous speech [50].

The study presented in this article is located in the context of emotion detection during interactions between Moroccan citizens, aged 16–60 years expressing four basic emotions: happiness, anger, neutral state, and sadness. Our corpus Moroccan Arabic dialect emotional database (MADED) is obtained from uncontrolled recordings, collected from real situations that can be extremely diverse. The selected subset includes situations taking place in different contexts (indoor, outdoor scenes, monologue, and dialogue). The emotions are validated and labelled by the interface shown in Figure 2 built by our team. Firstly, the data set is changed over to .wav format and cut into a syllabic structure as being the basic unit similar to the essential unit of our handling. Then it goes through another classification step of the syllabic type stored in the same folder.



Annotation_Window - [Preview] - Qt Designer

audio_1.wav
audio_2.wav
audio_3.wav

Form

Select Genre: Male

Select Age Class: 0 - 10

Enter Approximate Age: 0

Select Emotion: Happy

Select Degree: 0

Figure 2. Desktop annotation tool used for emotion evaluation

In the literature, there is a great deal of discussion on the length of the audio for which the emotion can be extracted dependably [51]-[54]. In our study, we propose a basic methodology for segmenting the audio based on a pseudo-phonetic approach. The main idea is to extract some features like (formants, pitch, and energies in six bands) from the speech signal according to the syllabic segment (CV), where C indicates Consonant and V Vowel. For all our experiments, we considered only four emotions, in particular, joy, anger, sadness, and neutral because the rule-based emotion extraction system that we used catered to only these four emotions. There were 979 sound records relating to these 4 emotions, specifically 250 syllables for /ba/, 240 for /Du/, 270 for /Ki/, and 219 for /Ta/ as shown in Table 3.

Table 3. Number of audio files corresponding to 4 emotions

Variable	N	J	S	A
/ba/	51	63	66	70
/Du/	62	62	50	66
/Ki/	70	60	54	86
/Ta/	54	50	54	61

Note: N: neutral, J: joy, S: sadness, A: anger

2.2. Speech features extraction

At present, there is no agreement on the best arrangement of important descriptors for an automatic emotion recognition system. The most well-known practice is to select a large number to have a richer classification. However, increasing the number of descriptors too high for a corpus of reduced size can possibly lead to performance degradation and accordingly be counterproductive. To solve this problem, it is necessary to adopt a new strategy that will make it possible to reduce the number of descriptors while keeping an acceptable recognition rate.

The system we propose is composed of standard acoustic features as shown in Table 2 that served as the challenge baseline set since. The interspeech emotion challenge 2009 [55]. The novelty comes from the way we computed the energy of the CV segments [56]. The main tools we used are MATLAB codes and the toolbox Praat [57], [58].

2.3. Computation of logarithmic energy characteristics based on DFT

In the pre-processing phase, we divided the speech signal, sampled at 22050 Hz, into time segments of 11.6 ms with an overlap of 9.6 ms. Thereafter we applied a hamming windowing to each segment followed by zero-padding. Finally, we calculated a 512-point discrete fourier transform. Now, to compute the energy and its distribution for each syllable as shown in Figure 3, we chose six specific bands of frequency as in [59], [60]: $(B_1: 0–400$ Hz); $B_2: 400–800$ Hz; $B_3: 800–1200$ Hz; $B_4: 1200–2000$ Hz; $B_5: 2000–3500$ Hz and $B_6: 3500–5000$ Hz).

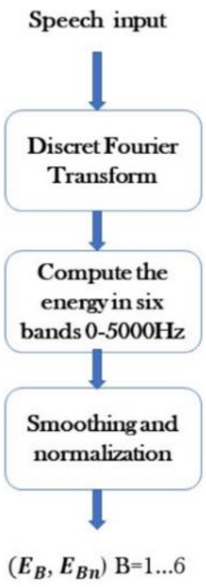

Speech input
Discret Fourier Transform
Compute the energy in six bands 0-5000Hz
Smoothing and normalization
(E_B, E_{Bn}) B=1...6

Figure 3. Block diagram of the compute of energy in six bands

These frequency bands correspond in fact to each part of the vocal tract. From the smoothed spectrum X(n,s) (the scale spectrum for each frame was smoothed by a 20 point moving average taken along the time index n), we calculated the energy in each band by (1):

$$E_B(n) = \sum_s 10\log_{10} |X(n,s)|^2 \tag{1}$$

where the band record B goes from 1 to 6. The frequency index s ranges from the DFT (Discrete Fourier Transform) indices representing boundaries (the lower and upper) for each frequency band.After that, the normalization is applied to each frame, according to (2):

$$E_{Bn}(n) = \frac{E_B(n)}{E_T(n)} \tag{2}$$

where $E_{Bn}(n)$ is the standardized band energy B in the frame n, E_T(n) is the general energy in the frame n and $E_B(n)$is the band energy B in the frame n.

3. THE RECOGNITION MODEL

The recognition model consists of two steps: (1) extracting the acoustic features to have a data set (2) applying the classification model. The data set is obtained by using Praat and MATLAB codes. For the task of classification, we utilized the KNN [61] to classify an instance of a data set into an emotion class.

The KNN algorithm is a supervised learning technique; it can be used for both classification and regression. To make a prediction, the KNN algorithm will be founded on the entire dataset. Indeed, for an observation, which is not part of the dataset, that we want to predict, the algorithm will search for the K instances of the dataset closest to our observation. Then for these K neighbors, the algorithm is based on their output variables Y to calculate the value of the variable Y of the observation that we want to predict. Also if K-NN is used for the classification, it is the mode of the variables Y of the K closest observations which will be used for the prediction.

3.1. Algorithmic composition

We can represent the operations of KNN by writing the following pseudo-code:

```
Algorithm 1: Start Algorithm
Input data:
  a) A data set D.
  b) A function for defining the distance d.
  c) An integer K
For a new observation X for which we want to predict its output variable Y do:
    Step 1: Calculate all the distances of this observation X with the other observations of the data set D.
    Step 2: Retain the K observations from the data set D closest to X using the function of calculating the
            distance d.
    Step 3: Take the values of Y from the K observations retained:
     a)  If we perform a regression, calculate the mean (or median) of Y retained.
     b)  If we perform a classification, calculate the mode of Y retained (this is our case).
     c)  Return the value calculated in step 3 as the value that was predicted by K-NN for observation X.
End
```

K-NN needs a distance calculation function between two observations. In our case, we have continuous data; hence the Euclidean distance is a good candidate.

3.2. Euclidean distance

It is a distance that computes the square root of the sum of the square differences between the coordinates of two points:

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{j=1}^{n} \left(x_j - y_j\right)^2} \tag{3}$$

where x= (x_j) and y= (y_j); *j=1...n*

For all our experiments, we used the free software STATISTICA [62] which is a set of data mining tools allowing the processing and selection of the parameters and proposing different learning algorithms. This software is currently increasingly used in the pattern recognition community. It includes many known algorithms such as KNN, SVM, decision trees (J48), as well as Meta algorithms.

STATISTICA KNN is a memory-based model characterized by a bunch of examples (objects) for which the results are known (i.e., the examples are labeled). In KNN the independent and dependent variables can be either categorical or continuous. The problem is the regression for continuous dependent variables, otherwise, the problem is the classification. Hence, KNN in STATISTICA can handle both classification and regression problems. In the event that we have another model (object), we would like to approximate the outcome dependent on the KNN examples. To settle on the choice KNN should discover K models that are nearest in distance to our new model (object). KNN predictions depend on averaging the results of the K-Nearest-Neighbor for the regression problems. For the classification problems, KNN utilizes the vote dominant part rule. The value of K strongly impacts the prediction accuracy. To find the optimal value for K we can utilize the cross-validation algorithm in STATISTICA.

4. EXPERIMENTAL RESULTS

We have run the KNN algorithm several times, each time aiming for a different goal. The classifications we made were accordingly binary or multiple. All experiments were performed on data sets collected from syllabic units: /ba/, /du/, /ki/, /ta/. There are four classification tasks presented in this work:

a. We tested the ability of the proposed model to detect each of the studied emotions. The classification in that case is binary. The targeted emotion (N: neutral, H: joy, S: sadness, A: anger) was labelled by its name and the others by O (others).
b. Still from the binary classification perspective, we tried to see to what extent our model is able to separate among positive and negative emotions.
c. Within each group of emotion (positive and negative), we tested if the proposed feature vector is a satisfactory tool to distinguish between them (according to [63], positive emotions include (joy and neutral) and negative emotions include (sadness and anger)).
d. At last, a multiple classification is performed to evaluate the whole system.

From these experiments, results are presented below. As shown in Figure 4, the analysis of the syllabic unit /ba/ using the proposed set of features gives high accuracy percentage of detection. Indeed, we obtained for example 60.20% for all emotions and 73.63% to distinguish between positive and negative emotions. In the same way, the rates vary respectively from 86.21% to 78.95% for positive emotions (neutral vs joy) and negative emotions (anger vs sadness). The KNN algorithm recognized a neutral state with more than 89.55%, sadness with 85.07%, anger with 79.10%, and joy with 76.12%.

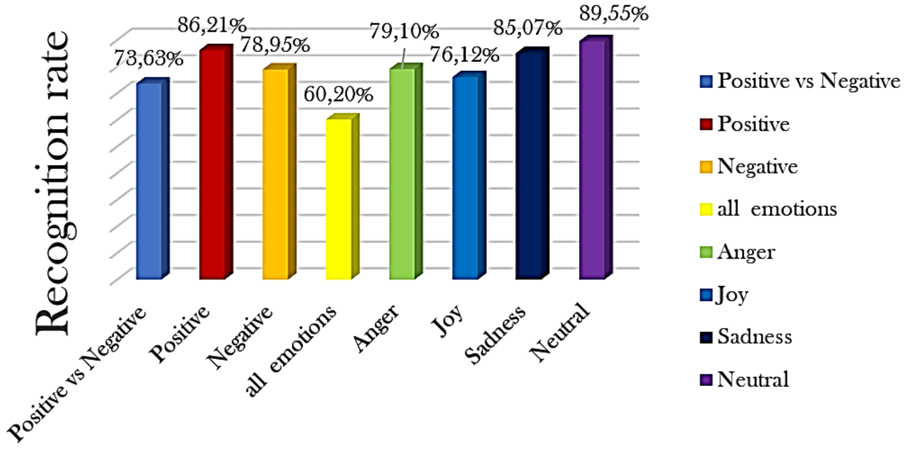


Figure 4. Analysis of the syllable /ba/

In the same way, an analysis performed on the CV /du/ shows results with rates that can go up to 94.95% to detect neutral emotion. We obtained 91.41% to recognize joy, 83.84% for anger, and 79.80% for sadness. For analysis between positive and negative emotions, the rate reached 84.34%, and 91.57%, 70.18% for positive emotions (neutral vs joy), and negative emotions (anger vs sadness). The global classification including all emotions reached a rate of 64.65% as shown in Figure 5.

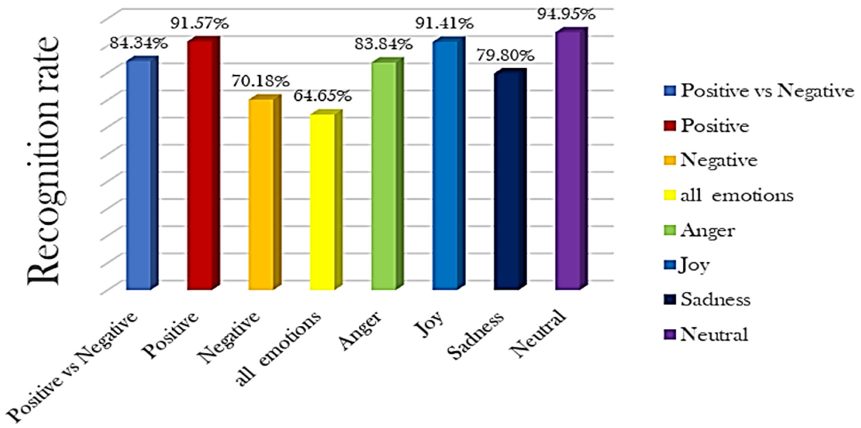


Figure 5. Analysis of the syllable /du/

Similarly, Figure 6 shows the accuracies that we obtained for the CV /ki/. The KNN algorithm recognized joy with more than 93.20%, anger with 84.47%, 81.55% for the neutral state, and 78.15% for sadness. Also, the binary classification between positive and negative emotions achieved a recognition rate of 72.33%. Furthermore, by considering each group of emotions (positive and negative), the results reached 87.62 % for the positive emotions and 75% for the negative ones. Concerning the classification of all emotions, the performed rate reached its minimum value of 52.43%.

Finally, the study of the CV /ta/ shows that: neutral state presents a high recognition rate of 90.13%, followed by anger with 84.57%, sadness with 78.15%, and lastly joy with 77.16%. For negative vs positive emotions, the accuracy is 75.92%. Within each category of emotion (positive and negative), the accuracy is 89.74% and 77.38% respectively. The rate obtained for the whole emotions is 61.73% which can be considered as an acceptable rate as shown in Figure 7.

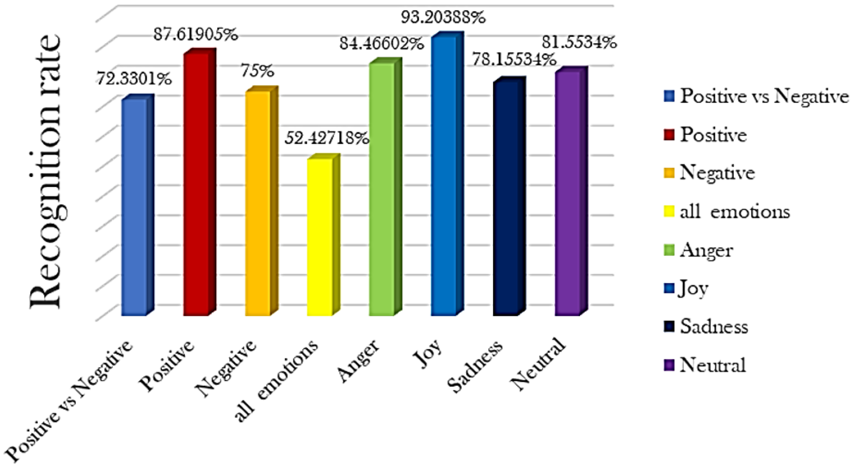


Figure 6. Analysis of the syllable /ki/

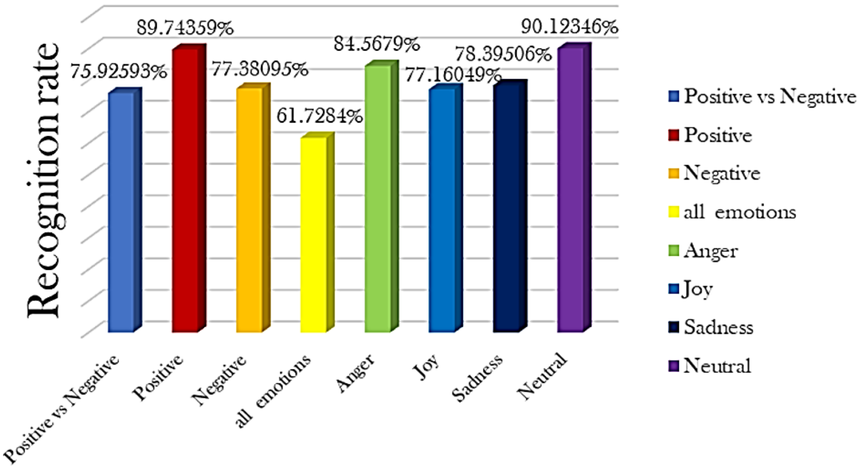


Figure 7. Analysis of the syllable /ta/

5. DISCUSSION

Our study aims to provide an automatic emotion recognition model with a reduced set of acoustic features. Measurements were carried out using the MADED database from which we extracted four plosive consonant /b/, /d/, /k/, /t/ associated with the three vowels /a/, /u/ and /i/. These choices are based on previous works [64], where the authors proved that the energy and its distribution in specific bands play an important role in characterising arabic plosive consonants. The results we obtained are quite satisfactory comparing to previous works as shown in Table 1.

It should be pointed out that our study raised many questions. Indeed, it is seen from the results that consonant /d/ achieves the best rates in almost all cases (especially in the neutral case 94.95%) which may lead us to think that place of articulation of the consonant has a role to play in determining the emotion under consideration. Moreover, syllables associated with the same vowel (/ba/ and /ta/) seem to present almost the same results. But as we come to investigate more carefully the recognition rates, we can see that:

a. Negative emotions present the best rates (78.95% for /ba/ and 77.38% for /ta/ while for /du/ 70.18% and for /ki/ 75%)
b. When the vowel /a/ is associated with the plosive /b/, sadness is more recognized than anger. The opposite occurs when /a/ is associated with /t/.

A slight comparison between syllables /ki/ and /du/ shows that for both the joy presents the best recognition rates (93.20% resp. 91.41%). But differences occur in the neutral and multiple classification cases. These rates establish in fact how far objects from the emotional representation we propose are close to each other. Indeed, the exploratory nature of our study has dictated the choice of KNN algorithm rather than SVM or artificial neuronal network (ANN). in the classification task. Our main concern is to establish to how extent the features vector succeeds to evaluate similarities between the same emotions.

6. CONCLUSION

This work gives a good grounding in modeling emotion with acoustic features. The method given here uses energy and its distribution in six bands as a principal tool for distinguishing between the four basic emotions: neutral, sadness, joy, and anger. The classical KNN algorithm is used to perform the classification task. In some cases, the results were conclusive but not exhaustive. This study can be extended in future works to richer corpora with different utterance representations in different languages and with different algorithms like neural networks and support vector machine algorithms.

REFERENCES

[1] L. A. P. Gaspar, S. O. C. Morales, and F. T. Romero, "Multimodal Emotion Recognition with Evolutionary Computation for Human-Robot Interaction," Expert Systems with Applications, vol. 66, pp. 42-61, 2016, doi: 10.1016/j.eswa.2016.08.047.

[2] L. F. Chen, Z. T. Liu, M. Wu, M. Ding, F. Y. Dong, and K. Hirota, "Emotion-Age-Gender-Nationality Based Intention Understanding in Human-Robot Interaction Using Two-Layer Fuzzy Support Vector Regression," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 709-729, 2015, doi: 10.1007/s12369-015-0290-2

[3] A. F. Caballero *et al*., "Smart Environment Architecture for Emotion Detection and Regulation," *Journal of Biomedical Informatics*, vol. 64, pp. 55-73, 2016, doi: 10.1016/j.jbi.2016.09.015.

[4] M. Egger, M. Ley, and S. Hanke, "Emotion Recognition from Physiological Signal Analysis: A Review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35-55, 2019, doi: 10.1016/j.entcs.2019.04.009.

[5] H. Boubenna and D. Lee, "Image-Based Emotion Recognition Using Evolutionary Algorithms," *Biologically Inspired Cognitive Architectures*, vol. 24, pp. 70-76, 2018, doi: 10.1016/j.bica.2018.04.008.

[6] A. Raheel, M. Majid, M. Alnowami, and S. M. Anwar, "Physiological Sensors Based Emotion Recognition While Experiencing Tactile Enhanced Multimedia," *Sensors*, vol. 20, no. 14, 2020, Art. no. 4037, doi: 10.3390/s20144037

[7] W. Mellouk and W. Handouzi, "Facial Emotion Recognition Using Deep Learning: Review and Insights," *Procedia Computer Science*, vol. 175, pp. 689-694, 2020, doi: 10.1016/j.procs.2020.07.101.

[8] J. Wu, S. Guo, H. Huang, W. Liu, and Y. Xiang, "Information and Communications Technologies for Sustainable Development Goals: State-of-the-Art, Needs and Perspectives," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 3, pp. 2389-2406, 2018, doi: 10.1109/COMST.2018.2812301.

[9] M. J. West, A. J. Angwin, D. A. Copland, W. L. Arnott, and N. L. Nelson, "Cross-Modal Emotion Recognition and Autism-like Traits in Typically Developing Children," *Journal of Experimental Child Psychology*, vol. 191, 2020, Art. no. 104737, doi: 10.1016/j.jecp.2019.104737.

[10] M. Janssens *et al*., "Emotion Recognition in Psychosis: No Evidence for an Association with Real World Social Functioning," *Schizophrenia Research*, vol. 142, no. 1-3, pp. 116-121, 2012, doi: 10.1016/j.schres.2012.10.003.

[11] R. Nakatsu, A. Solomides, and N. Tosa, "Emotion Recognition and Its Application to Computer Agents with Spontaneous Interactive Capabilities," *Knowledge-Based Systems*, vol. 13, no. 7-8, pp. 497-504, 2000, doi: 10.1145/319463.319641.

[12] S. Tivatansakul, M. Ohkura, S. Puangpontip, and T. Achalakul, "Emotional Healthcare System: Emotion Detection by Facial Expressions Using Japanese Database," *2014 6th Computer Science and Electronic Engineering Conference (CEEC)*, 2014, pp. 41-46, doi: 10.1109/CEEC.2014.6958552.

[13] R. L. Mandryk and M. S. Atkins, "A Fuzzy Physiological Approach for Continuously Modeling Emotion during Interaction with Play Technologies," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 329-347, 2007, doi: 10.1016/j.ijhcs.2006.11.011.

[14] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, 2004, pp. 351-358, doi: 10.3115/1218955.1219000.

[15] M. Bojanić, V. Delic, and A. Kapov, “Call Redistribution for a Call Center Based on Speech Emotion Recognition,” *Applied Sciences*, vol. 10, no. 13, 2020, Art. no. 4653, doi: 10.3390/app10134653.

[16] L. Vidrascu and L. Devillers, “Real-Life Emotion Representation and Detection in Call Centers Data,” *Affective Computing and Intelligent Interaction*, 2005, pp. 739-746, doi: 10.1007/11573548_95.

[17] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble Methods for Spoken Emotion Recognition in Call-Centres,” *Speech Communication*, vol. 49, no. 2, pp. 98-112, 2007, doi: 10.1016/j.specom.2006.11.004.

[18] C. M. Whiting, S. A. Kotz, J. Gross, B. L. Giordano, and P. Belin, “The Perception of Caricatured Emotion in Voice,” *Cognition*, vol. 200, 2020, Art. no. 104249, doi: 10.1016/j.cognition.2020.104249.

[19] N. Kamaruddin, A. Wahab, and C. Quek, “Cultural Dependency Analysis for Understanding Speech Emotion,” *Expert Systems with Applications*, vol. 39, no. 5, pp. 5115-5133, 2012, doi: 10.1016/j.eswa.2011.11.028.

[20] G. David, J. M. Molina, and Z. Callejas, “Combining Speech-Based and Linguistic Classifiers to Recognize Emotion in User Spoken Utterances,” *Neurocomputing*, vol. 326-327, pp. 132-140, 2019, doi: 10.1016/j.neucom.2017.01.120.

[21] S. L Castro and C. F Lima, “Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody,” *Behavior Research Methods,* vol. 42, no. 1, pp. 74-81, 2010, doi: 10.3758/BRM.42.1.74.

[22] S. Mohit, M. Tu, V. Berisha, C. Chakrabarti, and A. Spanias, “Articulation Constrained Learning with Application to Speech Emotion Recognition,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, 2019, Art. no. 14, doi: 10.1186/s13636-019-0157-9.

[23] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech Emotion Recognition: Features and Classification Models,” *Digital Signal Processing*, vol. 22, no. 6, pp. 1154-1160, 2012, doi: 10.1016/j.dsp.2012.05.007.

[24] Ö. Turgut, “A Novel Feature Selection Method for Speech Emotion Recognition,” *Applied Acoustics*, vol. 146, pp. 320-326, 2019, doi: 10.1016/j.apacoust.2018.11.028.

[25] S. K. Davis, M. Morningstar, M. A. Dirks, and P. Qualter, “Ability Emotional Intelligence: What about Recognition of Emotion in Voices?,” *Personality and Individual Differences*, vol. 160, 2020, Art. no. 109938, doi: 10.1016/j.paid.2020.109938.

[26] C. S. Luís and C. F. Lima, “Recognizing Emotions in Spoken Language: A Validated Set of Portuguese Sentences and Pseudo-sentences for Research on Emotional Prosody,” *Behavior Research Methods*, vol. 42, no. 1, pp. 74-81, 2010, doi: 10.3758/BRM.42.1.74.

[27] K. H. Hyun, E. H. Kim, and Y. K Kwak, “Emotional Feature Extraction Method Based on the Concentration of Phoneme Influence for Human-Robot Interaction,” *Advanced Robotics*, vol. 24, no. 1-2, pp. 47-67, 2010, doi: 10.1163/016918609X12585530487822.

[28] D. Kaminska, T. Sapinski, and G. Anbarjafari, “Efficiency of Chosen Speech Descriptors in Relation to Emotion Recognition,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, 2017, doi: 10.1186/s13636-017-0100-x.

[29] A. M. Berkehan and K. Oğuz, “Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers,” *Speech Communication*, vol. 116, pp. 56-76, 2020, doi: 10.1016/j.specom.2019.12.001.

[30] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011, doi: 10.1016/j.patcog.2010.09.020.

[31] Y. D. Chavhan, B. S. Yelure, and K. N. Tayade, “Speech Emotion Recognition Using RBF Kernel of LIBSVM,” *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 2015, pp. 1132-1135, doi: 10.1109/ECS.2015.7124760.

[32] F. Albu, D. Hagiescu, L. Vladutu, and M. A. Puica, “Neural Network approaches for children’s emotion recognition in intelligent learning applications,” *7th International Conference on Education and New Learning Technologies*, 2015, pp. 3229-3239.

[33] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech Emotion Recognition Using Hidden Markov Models,” *Speech Communication*, vol. 41, no. 4, pp. 603-623, 2003, doi: 10.1016/S0167-6393(03)00099-2.

[34] Z. Wang, R. Jiao, and H. Jiang, “Emotion Recognition Using WT-SVM in Human-Computer Interaction,” *Journal of New Media*, vol. 2, no. 3, pp. 121-130, 2020, doi: 10.32604/jnm.2020.010674.

[35] W. Zhang, X. Meng, Z. Li, Q. Lu, and S. Tan, “Emotion Recognition in Speech Using Multi-Classification SVM,” *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom),* 2015, pp. 1181-1186, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.215.

[36] D. Balaji, “Speech Emotion Recognition Using Deep Neural Networks,” *International Journal for Research in Applied Science and Engineering Technology (IJRASET),* vol. 8, no. 6, pp. 2460-2465, 2020, doi: 10.22214/ijraset.2020.5359.

[37] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, “Two-Layer Fuzzy Multiple Random Forest for Speech Emotion Recognition in Human-Robot Interaction,” *Information Sciences,* vol. 509, pp. 150-163, 2020, doi: 10.1016/j.ins.2019.09.005.

[38] H. Kaya and A. A. Karpov, “Efficient and effective strategies for cross-corpus acoustic emotion recognition,” *Neurocomputing*, vol. 275, pp. 1028-1034, 2018, doi: 10.1016/j.neucom.2017.09.049.

[39] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” *Fifteenth annual conference of the international speech communication association*, 2014.

[40] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, “Autoencoder-based unsupervised domain adaptation for speech emotion recognition,” *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068-1072, 2014, doi: 10.1109/LSP.2014.2324759.

[41] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.

[42] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014, doi: 10.1109/TMM.2014.2360798.

[43] D. Issa, M. F. Demirci, and A. Yazici, “Speech Emotion Recognition with Deep Convolutional Neural Networks,” *Biomedical Signal Processing and Control*, vol. 59, 2020, Art. no. 101894, doi: 10.1016/j.bspc.2020.101894.

[44] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, “Emotion recognition from audio signals using Support Vector Machine,” *Proceedings of the IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015*, 2016, pp. 139-144, doi: 10.1109/RAICS.2015.7488403.

[45] S. G. Koolagudi and S. R. Krothapalli, “Emotion Recognition from Speech Using Sub-Syllabic and Pitch Synchronous Spectral Features,” *International Journal of Speech Technology*, vol. 15, no. 4, pp. 495-511, 2012, doi: 10.1007/s10772-012-9150-8.

[46] A. Origlia, F. Cutugno, and V. Galatà, “Continuous Emotion Recognition with Phonetic Syllables,” *Speech Communication*, vol. 57, pp. 155-169, 2014, doi: 10.1016/j.specom.2013.09.012.

[47] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection,” *Neural Networks*, vol. 18, no. 4, pp. 407-422, 2005, doi: 10.1016/j.neunet.2005.03.007.

[48] Berlin database of emotional speech, 2005. [Online]. Available: http://emodb.bilderbar.info/index-1280.html.

[49] CASIA Chinese emotion corpus, 2008. [Online]. Available: http://www.chineseldc.org/resourceinfo.php?rid=76.

[50] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2009, doi: 10.1109/TPAMI.2008.52.

[51] M. Bayer, W. Sommer, and A. Schacht, “Reading Emotional Words within Sentences: The Impact of Arousal and Valence on Event-Related Potentials,” *International Journal of Psychophysiology*, vol. 78, no. 3, pp. 299-307, 2010, doi: 10.1016/j.ijpsycho.2010.09.004.

[52] E. A. Kensinger and S. Corkin, “Memory Enhancement for Emotional Words: Are Emotional Words More Vividly Remembered than Neutral Words?,” *Memory and Cognition*, vol. 31, no. 8, pp. 1169-1180, 2003, doi: 10.3758/BF03195800.

[53] J. Zhao, S. Chen, J. Liang, and Q. Jin, “Speech Emotion Recognition in Dyadic Dialogues with Attentive Interaction Modeling,” *Interspeech 2019*, Graz, Austria, 2019, doi: 10.21437/Interspeech.2019-2103.

[54] L. E. Ling, E. Grabe, and F. Nolan, “QUantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English,” *Language and Speech*, vol. 43, no. 4, pp. 377-401, 2000, doi: 10.1177/00238309000430040301.

[55] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge,” *Speech Communication*, vol. 53, no. 9-10, pp. 1062-1087, 2011, doi: 10.1016/j.specom.2011.01.011.

[56] S. B. Alex, L. Mary, and B. P. Babu, “Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features,” *Circuits, Systems, and Signal Processing*, vol. 39, no. 3, pp. 5681-5709, 2020, doi: 10.1007/s00034-020-01429-3.

[57] MATLAB, “The MathWorks,” *Inc., Natick, Massachusetts*, United States, 2018.

[58] M. J. Owren, “GSU Praat Tools: Scripts for Modifying and Analyzing Sounds Using Praat Acoustics Software,” *Behavior Research Methods*, vol. 40, no. 3, pp. 822-829, 2008, doi: 10.3758/BRM.40.3.822.

[59] K. Tahiry, B. Mounir, I. Mounir, and A. Farchi, “Energy Bands and Spectral Cues for Arabic Vowels Recognition,” *International Journal of Speech Technology*, vol. 19, no. 4, pp. 707-716, 2016, doi: 10.1007/s10772-016-9363-3.

[60] M. Farchi, K. Tahiry, S. Mounir, B. Mounir, A. Mouhsen, “Energy distribution in formant bands for arabic vowels,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 1163-1167, 2019, doi: ijece.v9i2.pp1163-1167.

[61] A. Meftah, M. Qamhan, Y. A. Alotaibi, and M. Zakariah, “Arabic Speech Emotion Recognition Using KNN and KSU Emotions Corpus,” *International Journal of Simulation Systems Science and Technology*, pp. 21.1-21.5, 2020, doi: 10.5013/IJSSST.a.21.02.21.

[62] C. H. Weiß, “StatSoft Inc., Tulsa., OK.: STATISTICA., Version 8,” *ASt A. Adv Statist Anal*, vol. 91, no. 3, pp. 339-341, 2007, doi: 10.1007/s10182-007-0038-x.

[63] Z. Hao and K. Chen, “Transferable Positive/Negative Speech Emotion Recognition via Class-Wise Adversarial Domain Adaptation,” *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)*, 2019, pp. 3732-3736, doi: 10.1109/ICASSP.2019.8683299.

[64] K. Tahiry, B. Mounir, I. Mounir, and A. Farchi, “Arabic Stop Consonants Characterisation and Classification Using the Normalized Energy in Frequency Bands,” *International Journal of Speech Technology*, vol. 20, no. 4, pp. 869-880, 2017, doi: 10.1007/s10772-017-9454-9.

BIOGRAPHIES OF AUTHORS

Abdellah Agrima was born in 1992. He received the engineering degree in network and telecommunications from the Faculty of Sciences and Technics, Hassan First University, Morocco in 2016. He is currently a Ph.D. student in Engineering, mechanical, Industrial Management, and Innovation (IMMII) Laboratory research Laboratory, Faculty of Sciences & Technics, Hassan First University with a thesis on emotion recognition using speech.

Ilham Mounir Ph.D. In Applied Mathematics Professeur Habilité à Diriger les Recherches at High School of Technology- Cadi Ayyad University Member of the LAPSSII Laboratory (Laboratory of Process, Signals, Industrial Systems, informatic) Research interests: Applied Mathematics, signal processing, emotion recognition, speech recognition and energy: optimization and modeling.

Laila Elmazouzi Ing Ph. D. In Telecommunication and Networks Professeur Habilité à Diriger les Recherches at High School of Technology- Cadi Ayyad University Member of the LAPSSII Laboratory (Laboratory of Process, Signals, Industrial Systems, informatic) Research interests: Telecommunication, signal processing, emotion recognition, machine learning.

Badia Mounir was born in Casablanca, Morocco, in 1968. Engineer degree (1992) in “Automatic and Industrial computing”, The Mohammadia School of engineering, Rabat, Morocco. Assistant Professor at Graduate School of Technology, University Cadi Ayyad since 1992. Habilitaded to supervise research (HDR) since 2007 and professor of higher education (PES) since 2017. Member of Laboratory of Process, Signals, Industrial Systems, informatic (LAPSSII) Laboratory. Her research interests include speech recognition, signal processing, energy optimization and modeling.

Abdelmajid Farchi Ing Ph.D. In Electric engineering and Telecommunications Chief of research team « Signals and Systems » in Laboratory of Engineering, Industrial Management and Innovation. Educational person responsible for the cycle engineer Electrical systems and Embedded Systems of the faculty of the sciences and technology of Settat; Morocco.