International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 6, December 2022, pp. 6159~6171
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i6.pp6159-6171

Emotion recognition based on the energy distribution of plosive syllables

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

Article Info

Article history:

Received Sep 15, 2021
Revised Aug 16, 2022
Accepted Aug 26, 2022

Keywords:

Consonant-vowel
Neural architecture search
Speech emotion recognition
energy distribution

ABSTRACT

We usually encounter two problems during speech emotion recognition (SER): expression and perception problems, which vary considerably between speakers, languages, and sentence pronunciation. In fact, finding an optimal system that characterizes the emotions overcoming all these differences is a promising prospect. In this perspective, we considered two emotional databases: Moroccan Arabic dialect emotional database (MADED), and Ryerson audio-visual database on emotional speech and song (RAVDESS) which present notable differences in terms of type (natural/acted), and language (Arabic/English). We proposed a detection process based on 27 acoustic features extracted from consonant-vowel (CV) syllabic units: *\ba*, *\du*, *\ki*, *\ta* common to both databases. We tested two classification strategies: multiclass (all emotions combined: joy, sadness, neutral, anger) and binary (neutral vs. others, positive emotions (joy) vs. negative emotions (sadness, anger), sadness vs. anger). These strategies were tested three times: i) on MADED, ii) on RAVDESS, iii) on MADED and RAVDESS. The proposed method gave better recognition accuracy in the case of binary classification. The rates reach an average of 78% for the multi-class classification, 100% for neutral vs. other cases, 100% for the negative emotions (i.e. anger vs. sadness), and 96% for the positive vs. negative emotions.

Corresponding Author:

Abdellah Agrima
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

Emotion conveys very important information about the psychological state of each individual. It thus plays a crucial role in shaping human social interactions. In this regard, several researchers have turned to the field of automatic emotion recognition to understand and retrieve the emotions of the interlocutor. This type of study is based on a wide variety of fields such as neuroscience [1], psychology [2], psychiatry [3], audiology [4], and computer science [5]. On the other hand, new communication technologies, such as call center applications, assistance to the elderly or disabled [6]–[9], and have largely contributed to the expansion and development of this field.

In the literature, there are several ways to recognize an emotional state. Indeed, researchers have used different methods based on facial expressions, speech, and physiological signals [10]–[13]. Each of these methods has its strengths and weaknesses. Those based on speech take advantage of the ease and low cost of speech acquisition. These facts have motivated researchers to exploit the features of speech in order to

build an emotion recognition model from different audios. This is not only to ensure human-machine interaction but also to analyze the interaction capabilities between humans [14], [15].

To build a speech emotion recognition system (SER), emotional speech databases are needed. Since the late 1950's, many emotional databases have been available. Interspeech emotion challenges started the process of standardizing these databases (2009, 2011, 2013...). These series of conferences addressed the problem of lack of standardized corpora and test conditions to compare SER performance under identical conditions [16]–[19].

Despite this, building an optimal emotion detection system from audio signals is still a challenge. Indeed, everyone knows that people of different ages, languages, cultures, genders express their emotions in different ways, so finding the feature vector that carries the optimal information capable of distinguishing between emotions regardless of who expresses them is an open problem. Our work attempts to address this issue by developing an automatic detection system that relies on two emotional databases expressed in English. The Ryerson audio and visual database of emotional speech and song (RAVDES) and Arabic Moroccan Arabic dialect emotional database (MADED). The proposed feature vector is obtained from the consonant-vowel (CV) syllabic units (C for a consonant and V for a vowel) belonging to both databases and which have undergone several acoustic treatments. The classification phase is performed using the artificial neural network algorithm.

The remainder of this work is arranged in the following manner. The second section of the paper gives an overview of relevant studies. The proposed method, specifically our new feature extraction method based on phonetic syllables and focusing on the spectrum and content, is discussed in section 3. Section 4 discusses the simulation results, while section 5 concludes the paper.

2. RELATED WORKS

The approach followed by the majority of works on the construction of automatic emotion recognition systems consists of three essential steps. The first one is crucial and has an impact on the whole process, it is the construction of the emotional database. In the literature, several types of corpora differ according to the way the emotion has been induced, the language, considered emotions and iterated sentences [20]–[25].

The second step involves the extraction of acoustic parameters that are robust to interfering factors such as speaker variation and environmental distortion. These features include energy, fundamental frequency, duration, loudness, formants, and voice quality parameters. They can also be correlated with data from other modalities (visual modalities) to improve the efficiency and robustness of the detection system [26], [27].

The task of classification constitutes the last step in a SER construction. Many classification approaches have been applied to the selected features, such as the gaussian mixture model (GMM) [28], support vector machines (SVM) [29], hidden Markov model (HMM) [30], random forest algorithm [31], artificial neural network (ANN) [32] and extreme learning machine (ELM) [33]. Table 1 (see in appendix) provides a summary of the research on SER. It presents the databases, the emotions, the classifiers, the acoustic features as well as the accuracy rates.

3. PROPOSED APPROACH

Our proposed method has four main components: The first one is building a consistent emotional speech database. After this, extracting effective features and applied normalization. Finally, designing reliable classifiers using machine learning algorithms. Figure 1 describes all mentioned steps of the proposed method.

3.1. Construction and preparation of databases

3.1.1. Moroccan Arabic dialect emotional database

The Arabic language is one of the ten most widely spoken languages around the world, nevertheless, there is a lack of Arabic speech emotions datasets. Wherefore, the first crucial task of our study was the construction of a natural appropriate database. Natural databases (usually obtained from direct and spontaneous interactions between interlocutors: interviews, and call centers, are the most time and resource-consuming, as they require extremely hard preliminary denoising and labeling tasks. But they offer a richness in the recorded dialogues and the emotions they contain.

In this work, the natural MADED is constructed from programs broadcast on the YouTube channel. These were interviews where the speaker is involved in an interaction that shows his emotions. Four primary emotions. Were considered: anger, joy, neutral state, and sadness.

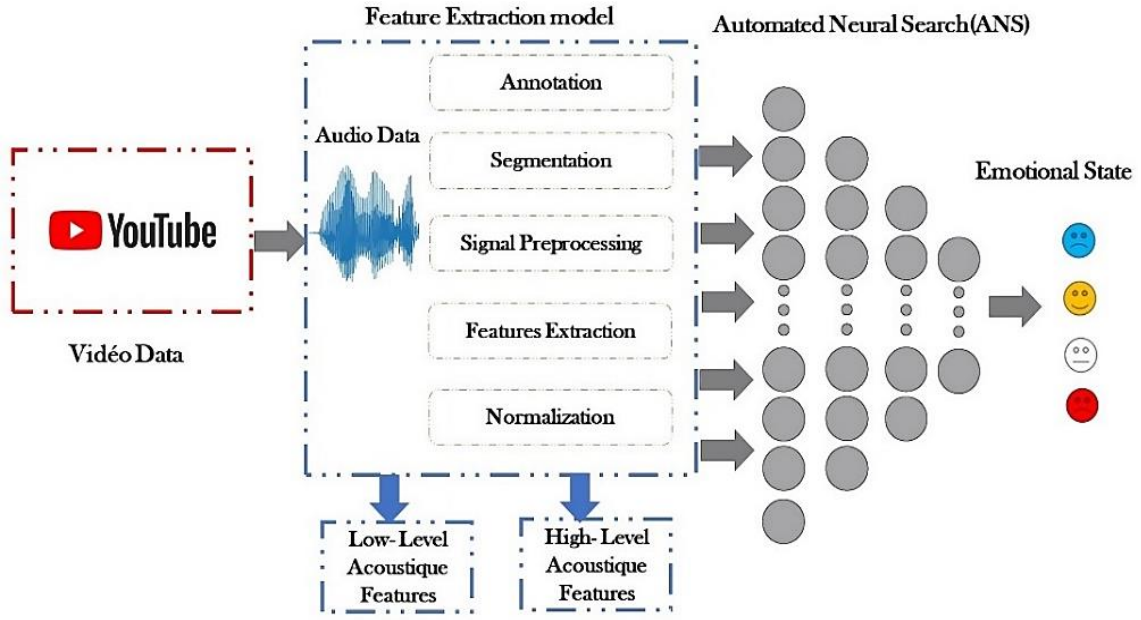

Figure 1. The overall steps of the proposed method

a. Data collection

We were only allowed to use audio support and verbal expression in our corpus. Hundreds of clips extracted from a variety of Moroccan dialect movies were utilized. Focused on speakers between the ages of 16 and 60 Figure 2 who were expressing primary emotions such as anger, happiness, neutrality, and sadness Table 2.

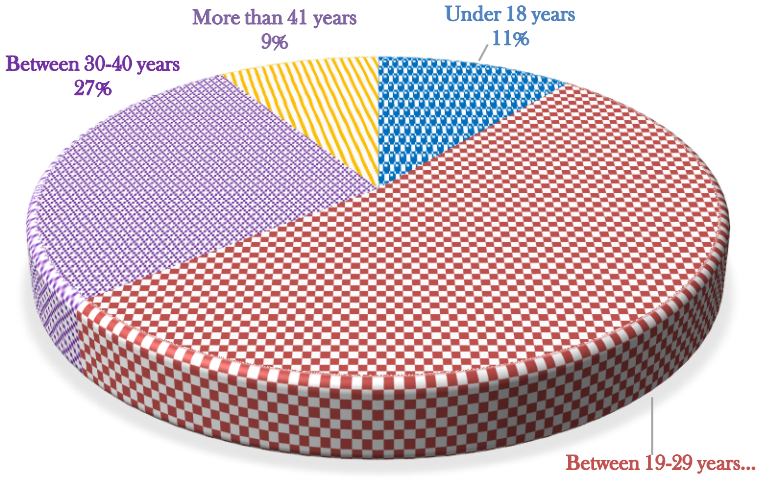

Figure 2. The age range of the speakers

Table 2. The number of audio files that correspond to each of the four emotions

Emotion	Anger	Happiness	Neutral	Sadness
Number of records (MADED)	281	171	302	271
Number of records (RAVDESS)	327	361	523	295

Our corpus is distinguished by the diversity of speakers (gender, social environment), as well as the contexts in which emotions occur (depending on the types of videos and thanks to the diversity of existing scenarios). In the emotional expressions, the degree of realistic data received is great. More than 100

audio-visual sequences in Moroccan dialect (in this study, we only focus on men) are therefore chosen, ranging in length from 5 seconds to 25 minutes, for a total of 20 hours of recording. The voice data were digitized with a sampling frequency of 22.050 kHz, the background noise is then eliminated using the software "AUDACITY" then the signal is cut in the form of syllabic units and recorded in .wav format. Table 2 contains the number of samples used in this paper.

b. Database validation task

Due to the spontaneous nature of our corpus, the different emotional states of the speaker were only identified by observing his face and listening to his voice [34]. Despite this, the identification of emotions realized in different ways requires objective validation using tools such as ELAN [35] and ANVIL [36]. Basically, listeners undergo a perceptual test in which they assess the emotion of the stimuli in terms of emotional intensity.

We took the perception test to dozens of volunteers from Cadi Ayyad University students, Ph.D. students, and professors. Individual perception tests were conducted in a soundproof room by the listeners. Each listener estimates the experienced emotion as well as its intensity after hearing a stimulus and validates his judgment using the following interface, which was created specifically for this purpose as shown in Figure 3.

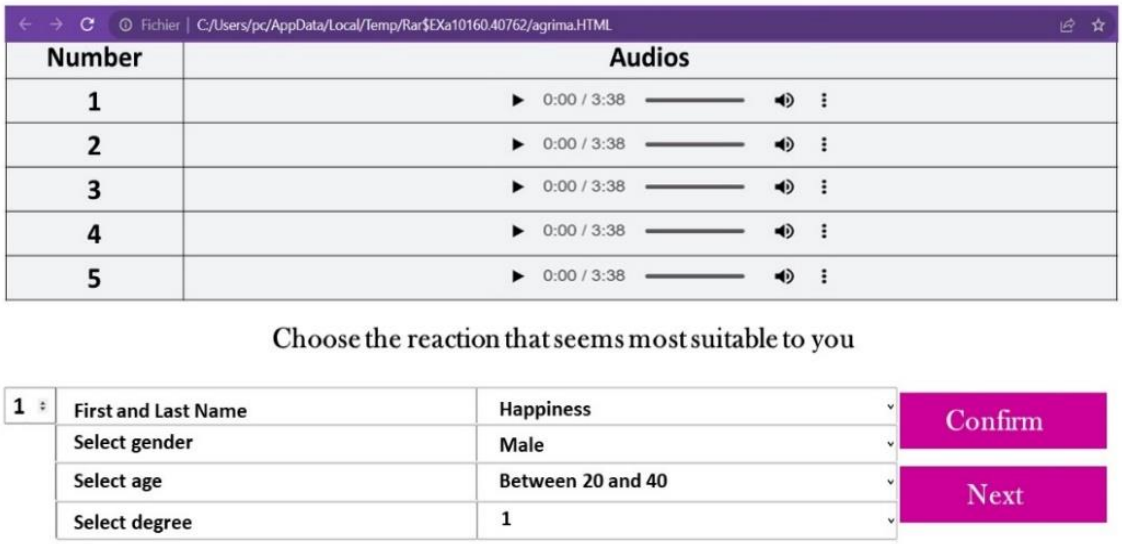

Figure 3. Web annotation tool used for emotion evaluation

3.1.2. The Ryerson audio-visual database of emotional speech and song

We use an open-access dataset called the RAVDESS that incorporates both video and audio (song and speech) data [37]. We will use only the audio (speech) files of 12 actors (12 males) vocalizing two matching lexical utterances with a North American accent. The speech includes eight expressed emotions (angry, sad, happy, fearful, surprise disgust, neutral, and calm) expressed with strong and normal intensity. We restrict just on four emotions namely (neutral, joy, anger, and sadness).

a. Syllable choice

In this work, We suggest a feature extraction method based on a phonetic approach, this method has been studied by [38]. The main goal of our research is to extract features from various segments, such as vowels and consonants. These segments are identified by manual segmentation of the speech signal in consonant-vowel format. The segments obtained during this phase are then called phonetic units. This technique was developed for automatic language recognition [39]. It allows eliminating the phoneme effect, produced during the pronunciation of words, which comes from the variation of the pronounced words, and which makes the distinction between the effect of the linguistic variation and the emotional variation complex. It also allowed us to work on the same units extracted in different languages (Arabic and English in our case). Thus, in line with the aforementioned idea of finding an optimal acoustic features vector able to distinguish between emotions whatever the speaker (language, age, culture...), the syllabic approach is

perfectly adapted [40]. Table 3 presents common syllables extracted from MADED and RAVDESS databases.

Table 3. Syllables choice

Phonetic Units	\Ba /بَ	\Du /دُ	\Ki /كِ	\Ta /تَ	Total
MADED	270	265	276	218	1029
RAVDESS	350	590	361	209	1510
MADED+ RAVDESS	620	855	637	427	2539

3.2. Speech features extraction

The first step of our process of speech emotion recognition is the extraction of speech emotional features using development tools MATLAB and Praat software [41], [42]. We propose an automatic SER based on a reduced number of acoustic indices as shown in Table 4, namely: intensity, the fundamental frequency (F0), the first four Formants, and voice quality parameters (Jitter, Shimmer) including acoustic descriptors extracted from the energy distribution in six bands. With Praat, we extracted for each utterance the fundamental frequency F0, the first four formants (F1, F2, F3, and F4), the intensity, the number of pulses, Jitter, and Shimmer. To compute energy and its distribution in the six bands for each CV, we used MATLAB codes as in [43]. The voice signal sampled at 22.050 kHz was subdivided into time segments of 11.6 ms with an overlap of 9.6 ms. Each segment was Hamming windowed and followed by zero-padding, then a 512-point fast Fourier transform was computed. The magnitude spectrum for each frame was smoothed by a 20-point moving average taken along the time index n. From the smoothed spectrum X(n, k), six different frequency bands mentioned in Table 4 were selected. Each band characterizes a part of the vocal tract. Formula (1) was used to compute the energy in each band:

$$E_{bd}(n) = \sum_n 10\log_{10}\left(|X(n,k)|^2\right) \quad (1)$$

where the band index bd is a number between 1 and 6. The frequency index k ranges from the discrete Fourier transform (DFT) indices representing the upper and lower boundaries for each band. Then, for each frame, the energy distribution was calculated by (2):

$$E_{bdpct}(n) = \frac{E_{bd}(n)}{E_T(n)} \quad (2)$$

where E_{bdpct} is the normalized band energy b in the frame n, $E_T(n)$ is the overall energy in the frame n and $E_{bd}(n)$ is the band energy bd in the frame n.

Table 4. 27 low and high-level descriptors

The energies and the energy distribution in the six bands	B_1= [100 Hz – 400 Hz]; B_2= [400 Hz – 800 Hz]; B_3 = [800 Hz – 1500 Hz]; B_4= [1200 Hz – 2000 Hz]; B_5= [2000 Hz – 3500 Hz]; B_6= [5000 Hz – 11025 Hz]
Voice quality parameters	Jitter (local), Shimmer (local), Jitter (local, absolute) Shimmer (local, dB)
Prosodic cues	The fundamental frequency F0(Maximum, Minimum, Standard deviation, median, Mean), Number of pulses (NI), and Intensity(I)
Spectral cues	4 Formants (F1, F2, F3, F4)

3.3. Normalization

Feature normalization is a crucial process for reducing speaker and recording variability while maintaining the discriminative strength of the features. By using feature normalization, the generalization ability of features is increased. Many works adopted a strategy of normalization at different levels [44]: speaker, value, and instance, to improve their classification methods. The most common approach in the literature is z-normalization. For a given speech feature from the speakers: f_S, its mean value: X, and standard deviation: Y are calculated. Then, the normalized feature Z is estimated as described in (3).

$$Z = \frac{(f_S - \mathrm{X})}{Y} \quad (3)$$

Using this technique, we noticed improvements in recognition rates in the majority of cases.

3.4. Classification

For the classification of emotions, we used machine learning algorithms. In the literature, many algorithms learn from training samples, then use the established model to classify the new observations. In our study, we chose two different classifiers, namely: ANN and SVM.

4. EXPERIMENTAL VALIDATION

This section presents the results of various experiments performed with the speech dataset for emotion recognition. It also highlights the major strengths and weaknesses observed during the evaluation process. The automatic system of emotion detection that we propose is based on 27 acoustic features extracted from 2,539 CV syllables *(/Ba, /Du, /Ki, /Ta*). These syllables are drawn from two databases MADED and RAVDESS. Some of the used indices are already presented in literature such as intensity, pitch, voice quality features, spectrum, and cepstrum coefficients. From the research done on automatic Arabic language recognition [43], [45], it has been found that the energy distribution contributes significantly to the recognition of plosive consonants. In the same vein, we have studied the variation of this energy distribution and its role in automatic emotion detection.

A first work [46], published in 2021, highlighted this finding. Indeed, by working on plosive consonants extracted from the MADED natural database, and by proposing a set of features based essentially on the energy and its variations, quite satisfactory recognition rates were obtained using the K-nearest neighbors (K-NN) algorithm. The results also raised questions about the place of articulation of the consonant, the vowel associated with the syllable, and the type of emotion. In this work, we first expanded the MADED database by balancing the number of occurrences for each considered emotion type. Then, to test our proposed system, we have chosen another database which is completely different from MADED in terms of language (English) and type of database (acted) and which contains the same considered syllables. This is the RAVDESS database.

The proposed method resolves the mutual interference between the emotions so that the recognition rate of different emotions is significantly improved in Figure 4. In the first case, a multi-class classification was run using all the emotions (step 2), the results found by applying the neural network on this dataset show that the rates obtained are relatively low compared to the final method that was adopted. Subsequently, the training set is divided into two subclasses (layer 3) that present a binary classification between the Neutral emotion and the other emotions (we assign the emotions: joy, sadness, and anger the label other). Then, each category of emotions (positive or negative) is divided into different subclasses (layer 4). Finally, in layer 5, a binary classification between negative emotions was carried out.

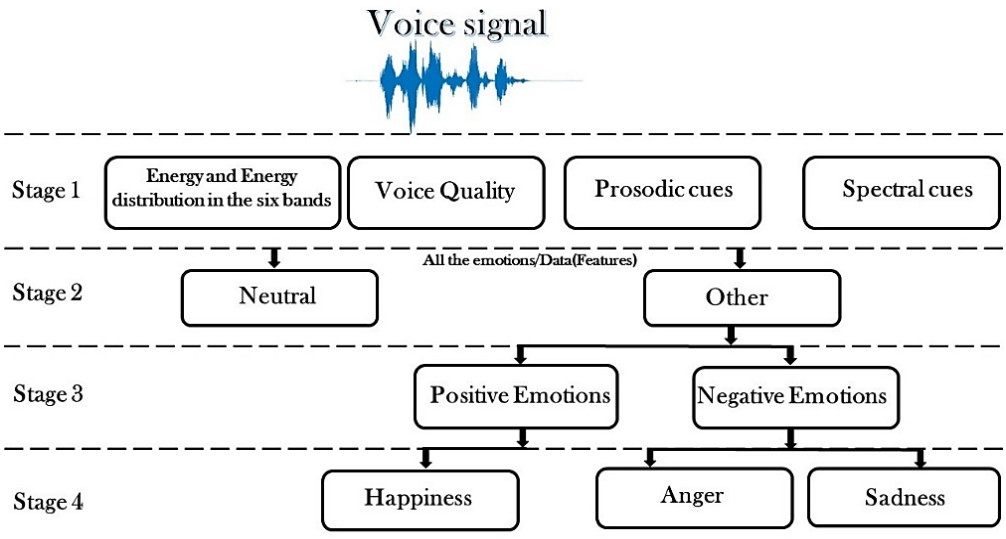

Figure 4. Speech emotion recognition stages

To verify the effectiveness of our SER, different classification models were used: SVM, decision tree, and automated neural networks algorithms [47]. In this paper, the results of the automated neural networks will be presented where the activations functions are (identity, logistic, exponential, SoftMax, and

logistic), the number of nodes in the output, and input is respectively 4 and 27. the number of hidden layers is between 7 and 21 units. 70% of the data are used for training, 15% for testing, and 15% for validation. Then the experiments are performed with different data each time. The results of the system of detection of the emotions obtained with consonant-vowel syllables drawn from RAVDESS and MADED corpora are shown in the following figures.

4.1. The MADED analysis

According to the multilayer perceptron neural network algorithm, performed with fewer parameters, we observe in Figure 5 for syllable /Ba/ that the average recognition rates reach 70% for all categories of emotions, for the neutral/other, he frequents 93%. The percentages of negative emotions and positive emotions and negative are around 88%, the negative emotions reach 82%. Likewise in the Figure 6, we notice that the results of classification of the syllable /Do/ did not change. For all categories of emotions, the value reaches 77%, the neutral/other attain 100%, the negative emotions reach 82%, and the positive and negative emotions reach 96%.

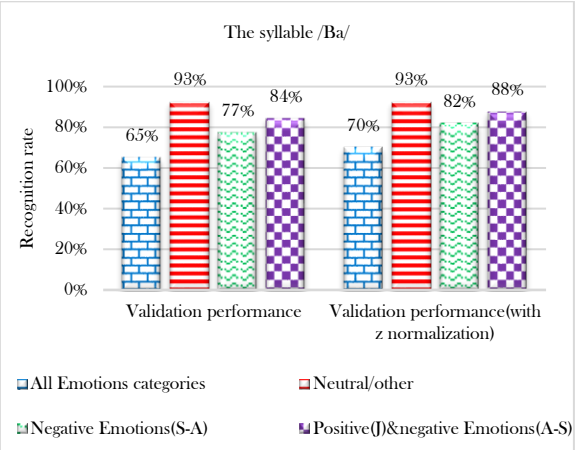

Figure 5. Classification percentage for syllable /Ba/

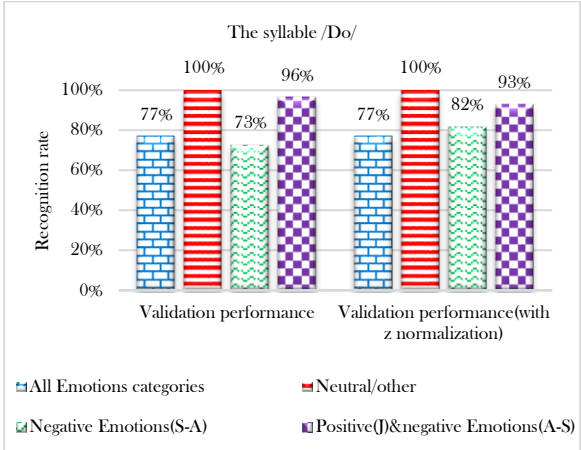

Figure 6. Classification percentage for syllable /Do/

Concerning the classification percentages for syllable */Ki/* in Figure 7, all the results stilled constant. All emotion categories attain 78%, for the neutral/other its 98%, the negative emotions reach 85% and the positive and negative ones its 91%. According to Figure 8, the classification results keep the same value (for syllable */Ta/*). For all emotions categories, the value is still around 75%. The value neutral/other is 100% and negative as the negative emotions. The percentage of 88% is the value of the classification between positive and negative emotions.

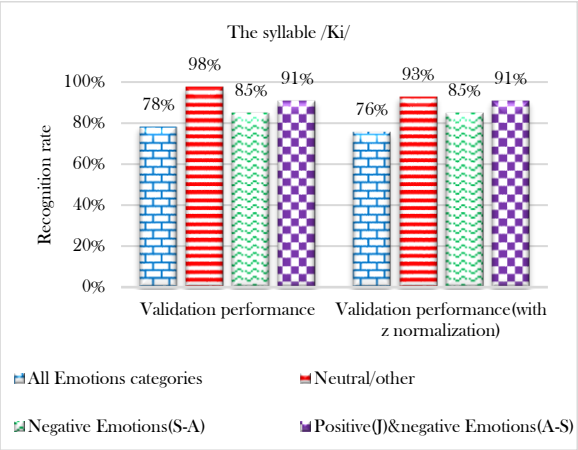

Figure 7. Classification percentage for syllable /Ki/

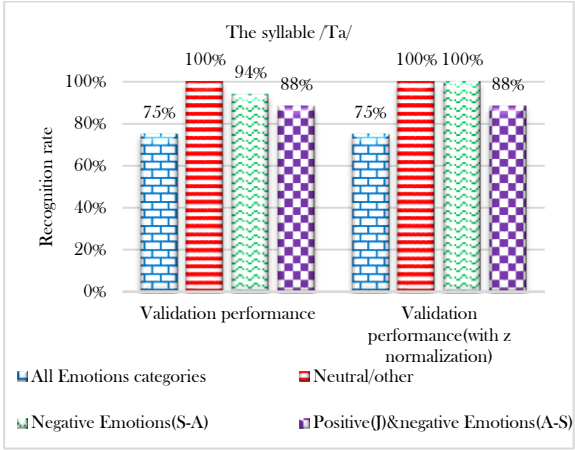

Figure 8. Classification percentage for syllable /Ta/

4.2. The RAVDESS analysis

Similarly, we will apply our model to the RAVDESS database. Figure 9 represents the recognition rate for syllable/Ba/, for all emotion categories the results of classification reach 77%, 92%for neutral/other, 95% for negative emotions, and 91% positive and negative emotions. According to Figure 10, The syllable classification results for syllable */Do/* are as follows: the rates of all emotions categories, neutral/other, negative emotions and positive & negative emotions are respectively 73%, 88%, 94% and 86%.

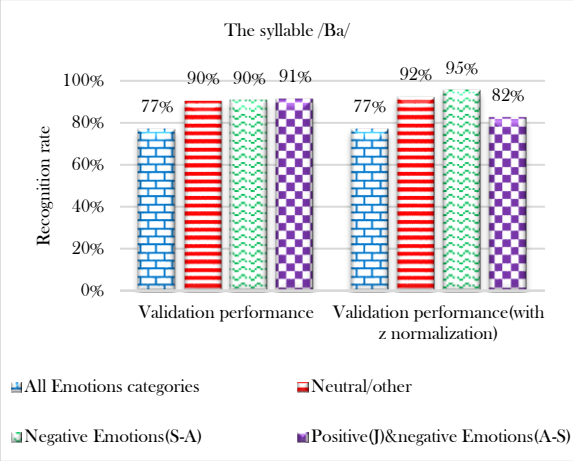

Figure 9. Classification percentage for syllable *\Ba*

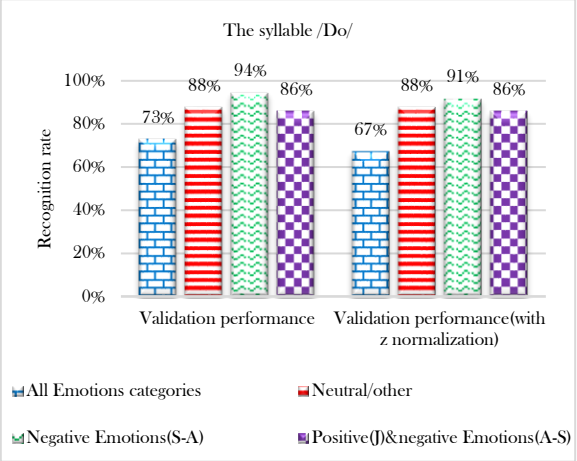

Figure 10. Classification percentage for syllable *\Do*

In Figure 11, we can see that the highest-ranking score for all sentiment categories is 72%. For neutrals/others, it frequents 89%. The percentage of negative emotions is 100% and the positive and negative emotions is about 83%. In the graph of Figure 12, we notice that the classification results for the syllable *\Ta* have not changed. Scores reached 100% for all emotions categories, 92% for neutral/other, 97% for negative emotions, and 99% for both positive and negative emotions.

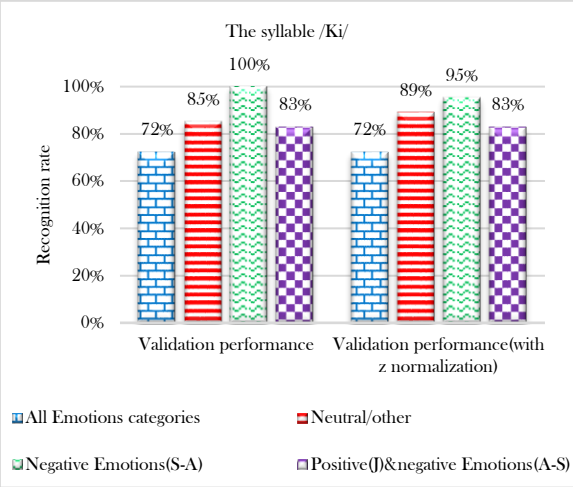

Figure 11. Classification percentage for syllable *\Ki*

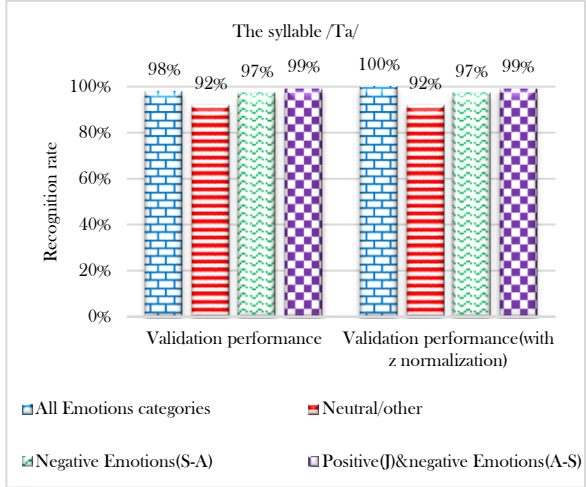

Figure 12. Classification percentage for syllable *\Ta*

4.3. Mixed RAVDESS and MADED analysis

To check the stability of our system of recognition of emotions, we mixed the two databases and tested them the same way as in the preceding parts. In the graph of Figure 13, the recognition rate for all emotions, reaches 61%. Scores reached 91% for neutral/other, 89% for negative emotions and positive and negative emotions. Which is an acceptable rates taking into consideration that the data come from two

databases that involve two different domains. Similarly, the average recognition rate for syllable *Do*\ in Figure 14 scored 65% for all emotions categories, 88% for neutral/other and negative sentiment, and 82% for both positive and negative sentiment. When applying Z-normalization, improvements in the recognition rate can go as far as +8%.

According to Figure 15, the classification results remain unchanged. Across all emotions categories, the value is still around 69%. Neutral/Other values are 87%, the value of negative emotions is 88% and 79% for positive and negative emotions. The Figure 16 represents detection rates for all sentiment categories, neutral/other, negative sentiment, and positive and negative sentiment. Classification results varied between 70% and 93%.

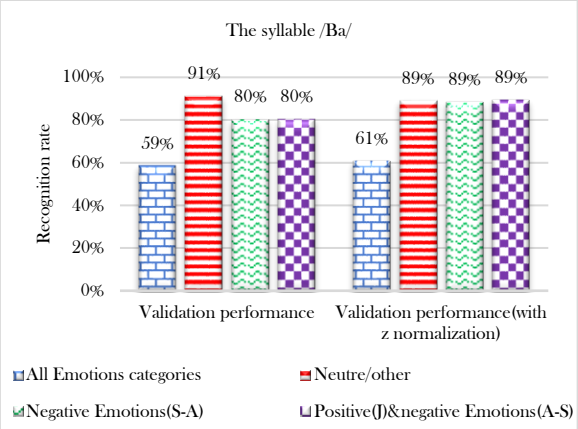

Figure 13. Classification percentage for syllable *Ba*

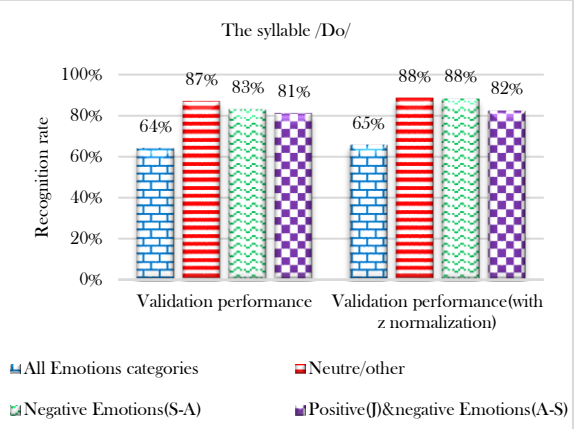

Figure 14. Classification percentage for syllable *Do*

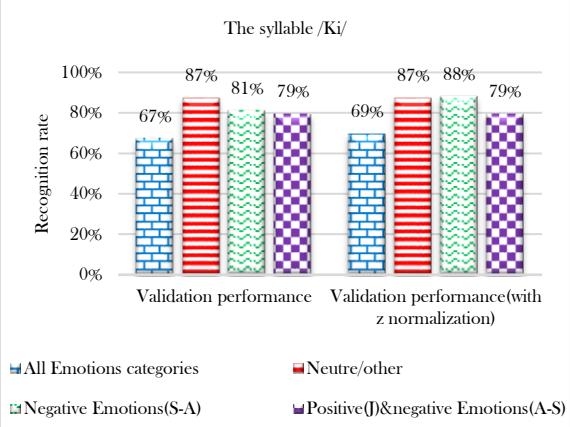

Figure 15. Classification percentage for syllable *Ki*

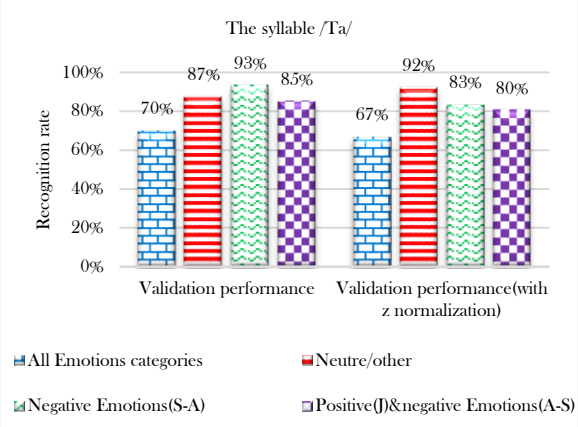

Figure 16. Classification percentage for syllable *Ta*

5. CONCLUSION

In this work, we used CV syllabic units associated with plosive consonants (*ba*, *du*, *ki*, *ta*) to analyze four emotions: joy, sadness, anger, and neutral state. For this purpose, we considered two emotional databases MADED (natural database in Moroccan Arabic) and RAVDESS (acted database in English). The proposed acoustic features vector consists of 27 measures. These measures correspond to standard indices such as the first four formants, fundamental frequency, voice quality parameters, and energy. They were, however, enriched by the specific calculation of the energy distribution in six specific bands, each of them covering a part of the vocal tract (Band 1: 0 to 400 Hz; Band 2: 400 to 800 Hz; Band 3: 800 to 1,200 Hz; Band 4: 1,200 to 2,000 Hz; Band 5: 2,000 to 3,500 Hz and Band 6: 3,500 to 5,000 Hz). The classification phase was performed with the neural network algorithm.

The obtained results show that, although the multi-class classification gives acceptable rates, they are still lower than those obtained with the proposed binary classifications (Neutral/other, positive emotions

/negative emotions, negative emotions (anger/sadness)). On the other hand, differences regarding the emotional databases were noted. For MADED, differentiating between the neutral state and the other emotions always gives the highest rates (they sometimes reach 100%). For RAVDESS, it is the distinction between negative emotions (anger/sadness) that reaches the highest scores (up to 100%).

To test the proposed system, we mixed the two bases and applied the same procedure. The obtained rates have slightly decreased but are still very satisfactory. The analysis of the results obtained with the z-normalization shows that, with respect to each type of classification, the values are very close for all the syllables, with a slight increase for the syllable /ta in the case of the discrimination between neutral/other. These results encourage us to say that the strategy we propose in this work is very conclusive and encourages us to explore other aspects of automatic emotion recognition from a speech by considering other syllabic units, other languages, and other emotions

APPENDIX

Table 1. Comparative analysis of existing techniques

Ref/year	Dataset	Emotion	Classifiers	Acoustic Features	Best result
2018/[48]	Spanish database and Berlin Emo-DB	Disgust, anger, joy, fear, sadness, surprise, and neutral	Recursive neural networks (RNN) and multivariate linear regression (MLR)	Mel-frequency-cepstrum-coefficients (MFCC), and Modulation spectral (MS).	– The recognition rate reaches 90% by applying the recurrent neural network on the Spanish database – 82.41% for the Berlin dataset.
2018/[49]	FAU-Aibo (German speech), CASIA (Chinese Speech), SAVEE (English)	Sadness, Happy, anger, surprise, fear, and neutral.	Brain emotional learning (BEL) and Genetic algorithm (GA)	INTERSPEECH 2009 standard feature set. MFCC and their first-order delta coefficients.	– Speaker-dependent: 71.05% (FAU Aibo) and 90.28% (CASIA), 76.40% (SAVEE). – Speaker-independent: 64.60% (FAU Aibo) and 38.55% (CASIA), 44.18% (SAVEE).
2019/[50]	BUEMODB (Turkish speech), EMODB (German speech), and RUSLANA (Russian speech).	Neutral, Anger, happy, Sadness, Disgust, Fear, Boredom	Principal component analysis (PCA) + Logistic regression + Long short-term memory (LSTM)	OpenSMILE statistical features	– EMODB (with 7 classes): 56.1%. – BUEMO (with 4 classes): 58.8. – RUSLANA (with 6 classes): 49.5%.
2020/[51]	Egyptian Arabic speech emotion (EYASE)	Sad, happy, neutral, and angry.	SVM and k-NN	Prosodic, spectral, and wavelet features	The SVM classifier outperformed the other classifiers in the multi-emotion classification (AHNS), with recognition rates of 69.9%, 66.3 percent, and 66.8% for males, females, and both.
2020/[52]	EmoDB and Spanish Database.	Joy, sadness, anger, disgust, neutral, fear, and sadness.	RNN, MLR, and SVM	MS+ MFCC	– The best accuracy (94%) for the Spanish database. – For the Berlin database, the rate reached 83%.
2020/[53]	IEMOCAP	Happy, sad, angry, and neutral	Deep neural network (DNN)+ HMM	MFCC + Epoch based features	MFCC has a recognition rate of 60.86 percent and MFCC + Epoch-based features has a recognition rate of 65.93 percent.
2020/[54]	RAVDESS, SAVEE, TESS, THAI and CREMA-D	Sadness, anger, disgust, and fear	deep convolutional neural network (DCNN)	MFCC	– RAVDESS: 75.83% – SAVEE: 65.83% – TESS: 55.71% – THAI: 96.60% – CREMA-D: 65.77%

REFERENCES

[1] S. N. Zisad, M. S. Hossain, and K. Andersson, "Speech emotion recognition in neurological disorders using convolutional neural network," in *Brain Informatics*, 2020, pp. 287–296.

[2] M. Chatterjee *et al.*, "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers," *Hearing Research*, vol. 322, pp. 151–162, Apr. 2015, doi: 10.1016/j.heares.2014.10.003.

[3] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, May 2017, pp. 371–375, doi: 10.1109/MDM.2017.64.

[4] E. M. Picou *et al.*, "Hearing, emotion, amplification, research, and training workshop: current understanding of hearing loss and emotion perception and priorities for future research," *Trends in Hearing*, vol. 22, Jan. 2018, doi: 10.1177/2331216518803215.

[5] K. Hartmann, I. Siegert, D. Philippou-Hübner, and A. Wendemuth, "Emotion detection in HCI: From speech features to emotion space," *IFAC Proceedings Volumes*, vol. 46, no. 15, pp. 288–295, 2013, doi: 10.3182/20130811-5-US-2037.00049.

[6] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied*

Sciences, vol. 10, no. 13, Jul. 2020, doi: 10.3390/app10134653.
[7] F.-M. Lee, L.-H. Li, and R.-Y. Huang, "Recognizing low/high anger in speech for call centers," in *Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation*, 2008, pp. 171–176.
[8] 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, Feb. 2007, doi: 10.1016/j.specom.2006.11.004.
[9] K. Wang, Z. Zhu, J. Zhang, and L. Chen, "Speech emotion recognition of Chinese elderly people," *Web Intelligence*, vol. 16, no. 3, pp. 149–157, Sep. 2018, doi: 10.3233/WEB-180382.
[10] 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.
[11] 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, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.
[12] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.
[13] 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, May 2019, doi: 10.1016/j.entcs.2019.04.009.
[14] E. Merdivan, D. Singh, S. Hanke, and A. Holzinger, "Dialogue systems for intelligent human computer interactions," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 57–71, May 2019, doi: 10.1016/j.entcs.2019.04.010.
[15] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013, doi: 10.1007/s11235-011-9624-z.
[16] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009emotion challenge," in *Interspeech 2009*, 2009, pp. 312–315.
[17] B. Schuller *et al.*, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH 2013*, 2013, pp. 148–152.
[18] B. W. Schuller *et al.*, "The Interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Interspeech 2020*, Oct. 2020, pp. 2042–2046, doi: 10.21437/Interspeech.2020-32.
[19] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Interspeech 2011*, 2011, pp. 3201–3204.
[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech 2005*, Sep. 2005, pp. 1517–1520, doi: 10.21437/Interspeech.2005-446.
[21] L. Aijun *et al.*, "CASS: A phonetically transcribed corpus of mandarin spontaneous speech," in *CASS: A phonetically transcribed corpus of mandarin spontaneous speech*, 2000, pp. 1–4.
[22] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: a Chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 913–924, Nov. 2017, doi: 10.1007/s12652-016-0406-z.
[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE & #146;05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, p. 8, doi: 10.1109/ICDEW.2006.145.
[24] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan. 2012, doi: 10.1109/T-AFFC.2011.20.
[25] A. Batliner, S. Steidl, and E. Noth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus," in *Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect Marrakesh*, 2008, pp. 1–4.
[26] A. Hassouneh, A. M. Mutawa, and M. Murugappan, "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods," *Informatics in Medicine Unlocked*, vol. 20, 2020, doi: 10.1016/j.imu.2020.100372.
[27] G. Xianhai, "Study of emotion recognition based on electrocardiogram and RBF neural network," *Procedia Engineering*, vol. 15, pp. 2408–2412, 2011, doi: 10.1016/j.proeng.2011.08.452.
[28] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using gaussian mixture model (GMM) and K- nearest neighbor (K-NN) techniques," *Procedia Computer Science*, vol. 49, pp. 50–57, 2015, doi: 10.1016/j.procs.2015.04.226.
[29] C. Yu, Q. Tian, F. Cheng, and S. Zhang, "Speech emotion recognition using support vector machines," in *Advanced Research on Computer Science and Information Engineering*, 2011, pp. 215–220.
[30] 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, Nov. 2003, doi: 10.1016/S0167-6393(03)00099-2.
[31] Z. Zhang, "Speech feature selection and emotion recognition based on weighted binary cuckoo search," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1499–1507, Feb. 2021, doi: 10.1016/j.aej.2020.11.004.
[32] M. Iqbal, S. Ali, M. Abid, F. Majeed, and A. Ali, "Artificial neural network based emotion classification and recognition from speech," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020, doi: 10.14569/IJACSA.2020.0111253.
[33] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, Jan. 2018, doi: 10.1016/j.neucom.2017.09.049.
[34] S. Rigoulot and M. D. Pell, "Emotion in the voice influences the way we scan emotional faces," *Speech Communication*, vol. 65, pp. 36–49, Nov. 2014, doi: 10.1016/j.specom.2014.05.006.
[35] H. Lausberg and H. Sloetjes, "Coding gestural behavior with the NEUROGES-ELAN system," *Behavior Research Methods*, vol. 41, no. 3, pp. 841–849, Aug. 2009, doi: 10.3758/BRM.41.3.841.
[36] M. Kipp, "ANVIL-A generic annotation tool for multimodal dialogue," in *Eurospeech 2001 - Scandinavia*, 2001, pp. 1367–1370.
[37] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0196391.
[38] A. Origlia, F. Cutugno, and V. Galatà, "Continuous emotion recognition with phonetic syllables," *Speech Communication*, vol. 57, pp. 155–169, Feb. 2014, doi: 10.1016/j.specom.2013.09.012.
[39] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, no. 4, pp. 436–456, Dec. 2005, doi: 10.1016/j.specom.2005.04.012.
[40] 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, Dec. 2012, doi: 10.1007/s10772-012-9150-8.
[41] P. Boersma and V. Van Heuven, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.
[42] MathWorks, "MATLAB and statistics toolbox release 2016." The MathWorks, Inc, Natick, Massachusetts, United States, 20016.

[43] K. Tahiry, B. Mounir, I. Mounir, L. Elmazouzi, 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, Dec. 2017, doi: 10.1007/s10772-017-9454-9.
[44] T. J. Sefara, "The effects of normalisation methods on speech emotion recognition," in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, Nov. 2019, pp. 1–8, doi: 10.1109/IMITEC45504.2019.9015895.
[45] 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, Dec. 2016, doi: 10.1007/s10772-016-9363-3.
[46] A. Agrima, I. Mounir, A. Farchi, L. Elmaazouzi, and B. Mounir, "Emotion recognition from syllabic units using k-nearest-neighbor classification and energy distribution," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5438–5449, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5438-5449.
[47] G. S. Bhat, N. Shankar, and I. M. S. Panahi, "Automated machine learning based speech classification for hearing aid applications and its real-time implementation on smartphone," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 956–959, doi: 10.1109/EMBC44109.2020.9175693.
[48] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech emotion recognition: methods and cases study," in *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2018, pp. 175–182, doi: 10.5220/0006611601750182.
[49] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018, doi: 10.1016/j.neucom.2018.05.005.
[50] O. Verkholyak, H. Kaya, and A. Karpov, "Modeling short-term and long-term dependencies of the speech signal for paralinguistic emotion classification," *SPIIRAS Proceedings*, vol. 18, no. 1, pp. 30–56, Feb. 2019, doi: 10.15622/sp.18.1.30-56.
[51] L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19–30, Sep. 2020, doi: 10.1016/j.specom.2020.04.005.
[52] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning*, IntechOpen, 2020.
[53] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features," *Circuits, Systems, and Signal Processing*, vol. 40, no. 1, pp. 466–489, Jan. 2021, doi: 10.1007/s00034-020-01486-8.
[54] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, Feb. 2021, doi: 10.3390/s21041249.

BIOGRAPHIES OF AUTHORS

Abdellah Agrima was born in 1992. In 2016 He received an engineering degree in network and telecommunications. He is a Ph.D. student in the Laboratory of Engineering, Industrial Management, and Innovation (EMIMI), at the Faculty of Sciences and Technics, Hassan First of Settat with a thesis on Emotion Recognition using Speech. He can be contacted at email: agrima.abdellah@gmail.com.

Ilham Mounir Professor of higher education (PES) at Graduate School of Technology of SAFI(ESTS)/ Cadi Ayyad University. Member of Laboratory of Process, Signals, Industrial Systems, informatic (LAPSSII) Laboratory. Applied mathematics, signal processing, emotion detection, voice recognition, and energy: modeling and optimization are among her research interests. She can be contacted at email: ilhamounir@gmail.com.

Abdelmajid Farchi Ing Ph.D. In Electric engineering and Telecommunications Chief of the research team Signals and Systems in the Laboratory of Engineering, Industrial Management, and Innovation. He is an educational person responsible for the cycle engineer Electrical Systems and Embedded Systems (ESES) of the Faculty Sciences and Technics, Hassan First of Settat, Morocco. He can be contacted at email: abdelmajid.farchi1@gmail.com.

Laila Elmazouzi Degree engineering in Telecommunication and Networks, and Professor of higher education (PES) at Graduate School of Technology of SAFI(ESTS)/ Cadi Ayyad University. Member of Laboratory of Process, Signals, Industrial Systems, informatic (LAPSSII) Laboratory. Telecommunications, Signal Processing, emotion detection, and Machine Learning are among her research interests She can be contacted at email: elmazouzi2001@yahoo.fr.

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 supervising research (HDR) since 2007 and professor of higher education (PES) since 2017. Member of the LAPSSII (Laboratory of Process, Signals, Industrial Systems, and Informatics). Speech recognition, signal processing, energy optimization, and modeling are some of her research interests. She can be contacted at email: mounirbadia2014@gmail.com.