

Convolutional neural network for speech emotion recognition in the Moroccan Arabic dialect language

Soufiyan Ouali, Said El Garouani

Department of Computer Science, Faculty of Science, Sidi Mohamed Ben Abdellah University, Fez, Morocco

Article Info

Article history:

Received Mar 28, 2024

Revised Oct 20, 2024

Accepted Oct 28, 2024

Keywords:

Arabic SER

Convolutional neural networks

Feature extraction

Signal processing

Speech emotion recognition

ABSTRACT

Extracting the speaker's emotional state has become an active research topic lately due to the demand for more human interactive applications. This field of research has noted significant advancement, especially in the English language, owing to the availability of massive speech-labeled corpora. However, the progress of analogous methodologies in the Arabic language is still in its infancy stages. This paper presents a new massive natural speech emotion dataset and a speech recognition model for the Moroccan Arabic language. Four primary emotion labels were selected: happy, sad, angry, and neutral. Various spectral features, such as the mel-frequency cepstral coefficient (MFCC), were extracted and tested to determine the optimal feature combination. A convolutional neural networks (CNNs) model was built and trained on our dataset. The results were compared between spectral features individually and combined with the CNN model resulting in the selection of MFCC, root-mean-square (RMS), mel-scaled spectrogram, and spectral, as optimal spectral features for our dataset. This selection yielded significant results, with an accuracy of 99.55% for emotion recognition, outperforming the existing research.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Soufiyan Ouali

Department of Computer Science, Faculty of Science, Sidi Mohamed Ben Abdellah University

Fez, Morocco

Email: soufiyane.ouali@usmba.ac.ma

1. INTRODUCTION

From the early decades of mechanization, humans were interested in building machines to facilitate their lives. Therefore, the advancement and innovation of many automatic machines were built to replace the daily routine jobs of humans [1]. Hence, they built smart cities, homes, cars, and phones. Nowadays, researchers are interested in building more advanced machines that mimic humankind's thinking and understanding, which revolutionized the human-computer interaction (HCI) [2] field. HCI has emerged towards innovation, design, and construction of new kinds of information and interaction technology with human-like intelligence [3] by using advanced techniques such as speech recognition (SR). SR is a sub-field of HCI interested in building smart entities that can understand and interpret voices humanly using phonetic and paralinguistic voice data, verbal and non-verbal elements such as voice-ton, pitch, and speed [4], [5]. Researchers in ref [6] highlighted that conversation is highly affected by the understanding of paralinguistic features. Therefore, a new field of study has emerged, speech-emotion recognition (SER) which is a sub-field of SR interested in extracting the speaker's emotional state based on speech signal properties [7]. One of the main questions that drew the attention of researchers in the SER field is what are the important voice features that identify a speaker's emotional state. Moreover, what makes SER important to the HCI field is that employing SER will enable computers to take a suitable action according to the speaker's emotional state which is a human-like characteristic.

The primary framework for constructing a SER system is as follows: (a) dataset building. SER datasets are categorized into three kinds [8]; natural or spontaneous speech datasets, which contain natural and authentic emotions; acted datasets, composed of recordings by a professional voice actor; and elicited dataset, which involves inducing specific emotions in a person to record their speech. (b) feature extraction. Speech is a continuous signal containing global and local information [9]. Therefore, feature extraction is the phase of extracting numerical data from speech using signal properties which are categorized into 4 classes: prosodic features, spectral features, voice quality features, and teager energy operator (TEO) based features [8]. (c) classification algorithm, after collecting and labeling the dataset it is fed into a classifier to learn the patterns and connections between each utterance to predict new or unlabeled data. Both machine learning (ML), e.g., RF, DT, KNN, and deep learning, e.g., RNN, LSTM, models are used for this task, the choice between ML and DL models depends on various factors such as the size, type, and length of the dataset.

The Arabic language can be divided into three categories. Classic Arabic, an official language of all Arabs and the language of the Quran. Modern standard Arabic (MSA), is a formal communication language commonly used in news and academic papers. Colloquial or dialectal Arabic language is used in daily or familiar conversations and it depends on regional variations. Despite the advancements in the field of SER, particularly in English, Arabic research is still in its early stages. This can be attributed to various factors, including the scarcity of well-structured datasets. In the state-of-the-art literature, we found that only three SER researches were conducted on dialectal Arabic in section 2.

Recently, more interest has been directed towards the Arabic language. Hifny and Ali [10] authors proposed a new deep neural network attention-based architecture consisting of CNN-BLSTM-DNN and trained it on the KSUEmotions dataset [11] from which they extracted 13 MFCC spectral features to build an Arabic SER able to identify five emotion labels (i.e., neutral, happy, sad, surprised, and questioning). Employing this architect enabled the researchers to achieve an accuracy of 87.2 %. Ahmed [12] built an Arabic SER based on the BAVED dataset [13] to identify three emotion states, low emotion (tired or exhausted), neutral emotion, and high emotion (positive or negative emotions, happiness, joy, sadness, or anger). The proposed model tested both Wav2vec2.0 and HuBERT for the feature extraction phase and then the output was fed to MLP classifiers to predict the emotion label. The result showed that Wav2vec2.0 achieved 89% accuracy, surpassing HuBERT, which achieved an accuracy of 84%. In a similar study by Klaylat *et al.* [14] researchers extracted only the prosodic feature from an Arabic dataset they built manually which consists of 3 emotion classes (i.e., happy, angry, surprised), and tested it with a thirty-five classifications model to find out that the best result obtained using the sequential minimal optimization with an accuracy of 95.52%. In Arabic dialects, three researches were conducted In [12] researchers built an SER system for the Saudi Arabic dialect to identify four emotion labels (i.e., happy, sad, angry, neutral), they created a dataset manually from which they extracted and tested three spectral features (i.e., MFCC, mel spectrogram, and spectral contrast) individually and combined, with three classification models (i.e., SVM, KNN, MLP), the result showed that the KNN performed better with the combination of MFCC and mel spectrogram with an accuracy of 68.57%, and the SVM and MLP model performed better with accuracies of 77.14% and 71.43%, respectively. Abdel-Hamid [15] researchers built a semi-neutral Egyptian Arabic dataset consisting of four emotions (i.e., happy, sad, angry, neutral) called 'EYASE'. Prosodic, spectral, and wavelet features were extracted and compared with different combinations with SVM and KNN classification models. The experimental results showed that SVM outperformed KNN, achieving an accuracy of 95% for detecting emotion from male voices, 84.2% for female voices, and 90.7% for both genders. Anger emotion was found to be the easiest class to detect. Mustafa *et al.* [16] research was conducted on the Algerian Arabic dialect. The researchers used the MFCC spectral feature and tested it with twelve classification models on their manually built dataset to detect four emotion labels (i.e., happy, sad, angry, neutral), the result showed that the best accuracy obtained with LSTM-CNN model with an accuracy of 93.34%.

Despite the numerous advancements in SER systems for the Arabic language, a significant challenge remains unaddressed. The vast diversity within Arabic dialects, which varies greatly due to regional characteristics, poses a substantial obstacle. These variations hinder the generalization of existing SER models, rendering them ineffective across different dialects. Consequently, the creation of specialized SER systems for each dialectal Arabic is not just important but it is imperative. Therefore, a thorough review of current research reveals a glaring gap: no researcher has yet tackled the unique complexities of the Moroccan Arabic dialect. This oversight leaves a critical void in the field and highlights an urgent need for focused research. The main contribution of our paper in the SER field is summarized as follows: i) present the first speech emotion dataset for the Moroccan Arabic language (MADES), ii) building an SER system able to extract the emotional state from the speech of the Moroccan speakers, and iii) constructing an efficient deep-learning model that outperforms the existing research in the SER field. The rest of the paper is organized as follows: In section 2, we present the Methodology of our work. The experiment, discussion, and results are presented in section 3. Finally, section 4 concludes the paper and discusses future work.

2. METHOD

2.1. Dataset building

In this work, we created a natural emotion speech dataset for the MADES. To keep the emotional naturalness of our corpus, the data were collected from popular Moroccan reality TV shows, radio programs, podcasts, and interviews. A group of audio and videos were conducted to select only meaningful ones that include clear emotional references. Subsequently, all the data are labeled and converted into audio using the WAV format. The final dataset consists of 1,505 records varied from three to ten seconds, based on the sentence length. Based on research in [16] Commonly, each speech can express one of the four emotions: happiness, anger, sadness, or neutrality. Therefore, we labeled each record with one of these emotion labels. The record labeling was conducted separately by two native Moroccan speakers. Only records that received the same label from both assessors, indicating a clear representation of the emotion, were included in the database.

To ensure the representativeness of the dataset we included audio samples from both genders across various age groups, comprising 785 records of males and 720 records of females. Moreover, to reduce the risk of overfitting the sample size for each category is approximately equal. Furthermore, each record includes three pieces of information: the emotional state, age, and gender of the speaker. Thus, it can also be utilized for gender speech recognition or age identification. To our knowledge, this is the first Dataset for emotion recognition in the Moroccan Arabic dialect. The MADES dataset is available for research purposes upon request in [17]. Table 1 summarizes the dataset distribution.

Table 1. MADES dataset distribution

Emotion/gender	Male			Female			Total
	Age: [10-20]	Age: [20-40]	Age: [40-60]	Age: [10-20]	Age: [20-40]	Age: [40-60]	
Happy	12	146	19	21	142	39	379
Sad	99	62	11	15	179	15	381
Angry	63	125	90	42	22	65	407
Neutral	68	34	56	56	102	22	338
Total	242	367	176	134	445	141	1505

2.2. Feature extraction

Building an efficient model is highly dependent on the quality of the training dataset. However, there is no standard feature extraction method in the state-of-the-art. In this work, we choose to extract the spectral features using the Librosa library [18], as they have proven promising results [12], [15], [16]. To choose only the most significant spectral features that are suitable for our dataset, we have extracted and tested five spectral features with our classifier:

- Mel-frequency cepstral coefficients (MFCC), constitute a set of coefficients capturing the shape of the power spectrum of a sound signal [19]. MFCC is widely utilized in various applications, particularly in voice signal processing, such as speaker recognition, voice recognition, and gender identification [20].
- Root-mean-square (RMS) which computes the value RMS for each frame, either from the audio samples y or from a spectrogram.
- Mel spectrogram is utilized to compute mel-scaled spectrograms, and focusing on the low-frequency part of speech.
- Spectral feature variant: spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, and spectral roll-off are all extracted.
- Zero crossing rate (ZCR) of an audio time series.

2.3. Classification model

In this work, a thorough investigation has been conducted to select the most pertinent classifier model for our dataset. Thereby, we choose to build and train a conventional neural network (CNN) model which is a deep-learning neural network known for its ability to learn complex patterns and can recognize local and global characteristics of input data, making it suitable for SR applications [21]-[23]. Moreover, they have shown good results in the literature. For example, training CNN mode for the SER system enabled the authors in [24] to achieve an accuracy of 86.06%, in [25] researchers achieved an accuracy of 79%, and in [26] authors achieved an F1-score of 86.65%.

The CNN architecture model developed in this work comprises (n=19) nineteen layers, with (n=15) fifteen of them being CNN layers, as outlined in Table 2. To stabilize and accelerate the training of the neural network, a batch normalization layer was added after each 1D convolutional layer. To enhance the model's

generalization capability, a dropout rate of 0.2 was applied between the 1D convolutional layer, promoting resilience and preventing overfitting. Moreover, we added a max-pooling layer between the 1D convolutional layer to reduce the spatial dimensions of the input data.

Table 2. The CNN architecture proposed

Layers	Type	Details
1	Conv	512 filter + kernel size = 5 + strides = 1 + padding = 'same'+ activation = relu
2	Batch Normalization	default
3	MaxPool1D	Pool size =5 +strides = 2+ padding = same
4	Conv	512 filter + kernel size = 5 + strides = 1 + padding = 'same'+ activation = relu
5	Batch Normalization	default
6	MaxPool1D	Pool size =5 +strides = 2+ padding = same
7	Conv	256 filter + kernel size = 5 + strides = 1 + padding = same + activation = relu
8	Batch Normalization	default
9	MaxPool1D	Pool size =5 +strides = 2+ padding = same
10	Conv	256 filter + kernel size = 3 + strides = 1 + padding = same + activation = relu
11	Batch Normalization	default
12	MaxPool1D	Pool size =5 +strides = 2+ padding = same
13	Conv	128 filter + kernel size = 3 + strides = 1 + padding = same + activation = relu
14	Batch Normalization	default
15	MaxPool1D	Pool size =3 +strides = 2+ padding = same
16	Flatten	default
17	Dense	512 neurons + relu activation
18	Batch Normalization	default
19	Dense	4 neurons + SoftMax activation

3. RESULTS AND DISCUSSION

3.1. Dataset preprocessing

After building the MADES dataset the total number of records is 1,505. Recognizing the significance of extensive data in training an efficient model, one key preprocessing technique employed is data augmentation [27]. This process involves creating new synthetic data samples by introducing small perturbations to the initial training set through the injection of various effects. The effects used in our dataset include: noise injection because in realistic scenarios sound audio signals frequently experience environmental noise, distortions, or interference. The model develops the ability to navigate such scenarios through training on data containing noise, leading to more accurate predictions in real-world conditions. Speed change: In practical environments, speaking speeds vary. Therefore, two versions of the original recording were created; one with speed multiplied by 1.25 and another with speed multiplied by 0.85. These values were chosen carefully to augment the data while preserving the original sense of the recording. Shifting time: A process that enhances the diversity of temporal aspects in the training data, thereby promoting greater robustness and adaptability in the model. Pitch change: generate records by changing the pitch of the audio signal. To maintain the sense of the original recording, the pitch is adjusted by a factor of 0.6.

The application of data augmentation, in which we implemented five effects, has generated 7,699 records, contributing to the model achieving notable results, as illustrated in Table 3. Beyond creating a sufficient dataset, data augmentation plays a significant role in reducing training overfit. As represented in Table 3, the difference between training and validation accuracy when training the model on the original data is 6.3%, and the validation loss is 52.54%, signifying a substantial degree of overfitting. In contrast, when training the model with augmented data, the difference between training and validation accuracy is 0.11%, and the validation loss is 2.51% highlighting the absence of overfitting and good training.

Table 3. The impact of data augmentation on model performance

Learning rate/dataset	Without augmented data %	With augmented data %
Training accuracy	87.81	99.33
Validation accuracy	81.51	99.22
Training loss	33.00	02.04
Validation loss	52.54	02.51

Another data preprocessing step involved standardizing the dataset, a crucial procedure in data analysis and machine learning [28], [29]. Given the sensitivity of the chosen model to outliers, we employed

the Standard-Scaler to standardize our dataset, and this had a positive impact on training the model, as illustrated in Table 4. (note: in this experience, we trained the model only on original data without augmented data).

Table 4. The impact of dataset standardization on model performance

Learning rate/dataset	Without standardization %	With standardization %
Training accuracy	78.92	82.00
Validation accuracy	70.19	81.51
Training loss	55.74	51.97
Validation loss	67.35	57.26

3.2. Feature selection

Five features were evaluated to select the most significant features for training our model, including MFCC, mel spectrogram, spectral with its 5 variants, RMS, and ZCR. Starting with MFCC, a crucial feature, we wanted to investigate how many coefficients to include. While the first 13 coefficients are often seen as the most relevant, our tests as shown in Table 5, revealed that opting for 40 coefficients led to a better learning rate. Hence, we decided to include 40 coefficients from the MFCC feature. (note: in this experience, we trained the model only on original data without augmented data).

Table 5. The impact of MFCC dimension number on model performance

Learning rate/dataset	With 13 MFCC features %	With 40 MFCC features %
Training accuracy	81.00	87.81
Validation accuracy	81.51	82.51
Training loss	33.00	02.04
Validation loss	52.54	02.51

After selecting the number of coefficients for the MFCC feature, we assessed the influence of other features. Table 6 presents the outcomes of various experiments conducted using individual features and combinations. Therefore, the best results were obtained by combining the following features: MFCC, RMS, Mel-spectrogram, and spectral with its 5 variants, resulting in a training and validation accuracy of 99.55% and 99.42% respectively. Therefore, the combination of features enhanced the classifiers' performance, leading to higher accuracy compared to individual features.

Table 6. Recognition rate for different feature extraction combinations

Features	Validation loss %	Training accuracy %	Validation accuracy %
MFCC 40	02.04	99.33	99.22
MFCC 40 + RMS	01.75	99.48	99.09
MFCC 40 + RMS + Melspec	01.95	99.53	99.03
MFCC 40 + RMS+ Melspec + spectral	01.35	99.55	99.42
MFCC 40 + RMS + Melspec + spectral + ZCR	01.39	99.56	98.51

The experiment was conducted using the programming language Python 3.10.12, Keras 3.0.1, and Google Colab, with computations performed on a CPU. The duration of each experiment ranged from approximately 32 to 45 minutes. Considering the classifier hyperparameter, the optimal results were achieved after 20 epochs, utilizing a batch size of 32, and implementing a learning rate reduction to 0.0005, we used the "Adam" optimizer and "categorical cross-entropy" as a loss function. This configuration played a crucial role in enhancing the model's performance.

For a deeper analysis of the results, we constructed confusion matrices illustrating the classifiers' performance in predicting each emotion as shown in Figure 1. On these matrices, the x-axis signifies the predicted labels, while the y-axis signifies the true labels. Notably, the angry emotion class exhibited robust predictions, achieving the highest accuracy rate of 99.73%. This result can be explained by the fact that the angry emotions class contains high frequency and pitch as shown in Figures 2 and 3 which are easily captured by the CNN classifier. As illustrated in Table 7, the results are compared to the state of the art, demonstrating that our model outperformed existing research.

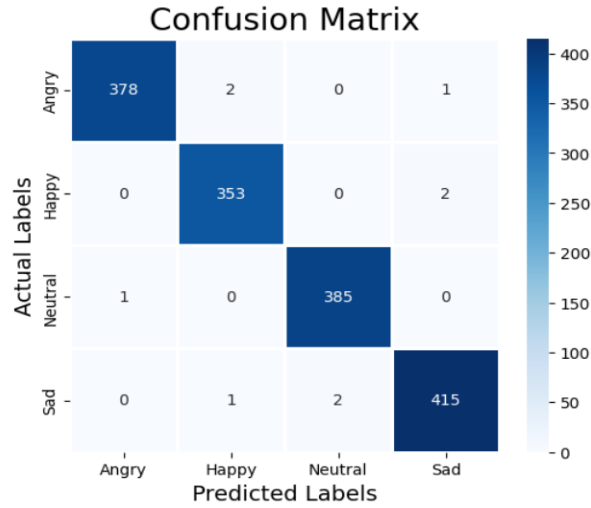


Figure 1. Confusion matrix of our Arabic SER model

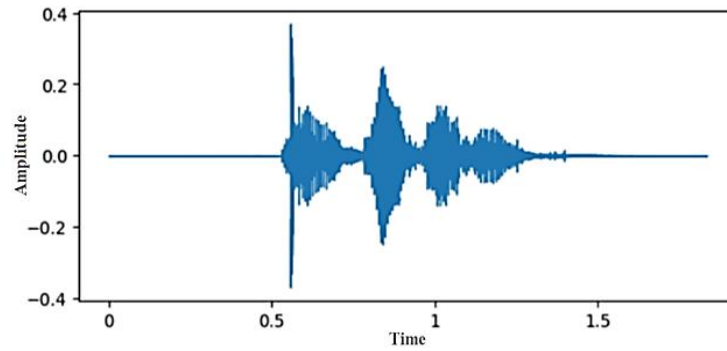


Figure 2. Wave plot for audio with happy emotion

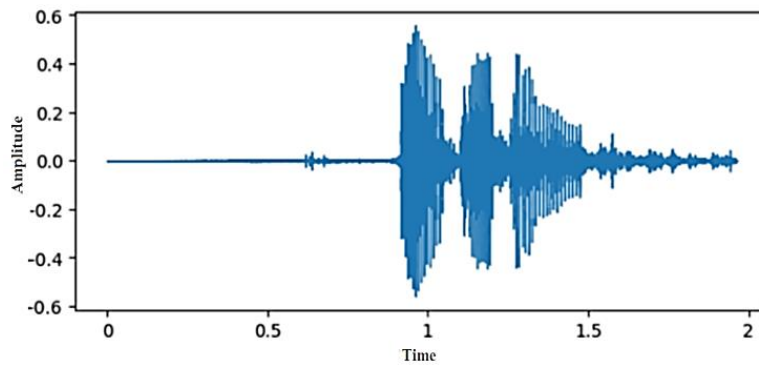


Figure 3. Wave plot for audio with angry emotion

Table 7. Comparing our model with state-of-the-art models

Model	Accuracy
[12] Arabic SER with BAVED dataset	89 %
[15] Arabic (Egyptian dialect) SER	88.3%
[30] Arabic (Saudi dialect) SER	77.14 %
[31] Arabic (Algerian dialect) SER	93.34
Our model, Arabic (Moroccan dialect) SER	99.55%

4. CONCLUSION

With the advancement of AI and the automation of various sectors, automatic speech recognition has emerged as a prominent research field, offering alternatives to routine human tasks in areas like call centers, healthcare, virtual assistants, and domains such as smart cities, smart homes, and smart devices. While extensive research has been conducted in English, the exploration of this field in the Arabic language is still in its early stages. This paper contributes to the field of SR by developing an efficient model capable of recognizing the emotional state of the speaker in the Moroccan Arabic dialect language. Through a series of experiments involving dataset creation, feature extraction, and classifier model selection, we identified optimal combinations to build an effective model; data augmentation exhibited a noteworthy enhancement in model learning, resulting in an 11.52% increase in training accuracy. Standardizing the dataset further enhanced model learning, resulting in a 2.08% increase in training accuracy, while selecting the appropriate number of MFCC coefficients boosted training accuracy with an increase of 6.81% in training accuracy. Combining MFCC, mel-spectrogram, spectral features, and RMS with a CNN model further improved the model performance, raising the accuracy from 99.33% to 99.55%. The model achieved promising results, with a 99.55% accuracy rate for the speaker's emotional state. The challenge of limited data prompted our decision to build a model predicting only four emotional states. In future work, we aim to create a larger dataset encompassing at least eight emotional states and explore additional prosodic and spectral features to further enhance the model's accuracy. Furthermore, building a SER model for the Moroccan Arabic dialect using Transformers as they have shown notable results in this field.




REFERENCES

- [1] M. G. Helander, "Emerging office automation systems," *Human Factors*, vol. 27, no. 1, pp. 3–20, Feb. 1985, doi: 10.1177/001872088502700102.
- [2] B. A. Myers, "A brief history of human-computer interaction technology," *Interactions*, vol. 5, no. 2, pp. 44–54, Mar. 1998, doi: 10.1145/274430.274436.
- [3] A. M. Ghanbari, S. Ghanbari, and Y. Norouzi, "A new approach to architecture of human-computer interaction," in *2017 IEEE International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2017*, Nov. 2017, vol. 2017-November, pp. 1–4, doi: 10.1109/ICSIMA.2017.8311991.
- [4] H. Traunmüller, "Paralinguistic Phenomena," in *Sociolinguistics*, Walter de Gruyter, 2017, pp. 653–665.
- [5] J. Hook, F. Noroozi, O. Toygar, and G. Anbarjafari, "Automatic speech based emotion recognition using paralinguistics features," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 67, no. 3, pp. 479–488, 2019, doi: 10.24425/bpasts.2019.129647.
- [6] B. Mishra, "Role of paralanguage in effective english communication," *ICFAI Journal of Soft Skills*, vol. 3, no. 2, 2009.
- [7] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018, doi: 10.1145/3129340.
- [8] 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.
- [9] Y. Gao, B. Li, N. Wang, and T. Zhu, "Speech emotion recognition using local and global features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10654 LNAI, Springer International Publishing, 2017, pp. 3–13.
- [10] Y. Hifny and A. Ali, "Efficient Arabic emotion recognition using deep neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2019, vol. 2019-May, pp. 6710–6714, doi: 10.1109/ICASSP.2019.8683632.
- [11] A. Meftah, Y. Alotaibi, and S. A. Selouani, "Designing, building, and analyzing an arabic speech emotional corpus," in *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, vol. 22, 2014.
- [12] S. Ahmed, "ASER: Arabic speech emotion recognition employing Wav2vec2.0 and HuBERT based on BAVED dataset," *Transactions on Machine Learning and Artificial Intelligence*, vol. 9, no. 6, pp. 1–8, Nov. 2021, doi: 10.14738/tmlai.96.11039.
- [13] A. Aouf, "Basic arabic vocal emotions dataset," doi: 10.34740/KAGGLE/DS/345828.
- [14] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Emotion recognition in Arabic speech," *Analog Integrated Circuits and Signal Processing*, 96, pp.337-351, 2018
- [15] 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.
- [16] M. B. Mustafa, M. A. M. Yusooif, Z. M. Don, and M. Malekzadeh, "Speech emotion recognition research: an analysis of research focus," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 137–156, Jan. 2018, doi: 10.1007/s10772-018-9493-x.
- [17] S. Ouali, "Morrocan Arabic speech emotion recognition dataset," *Github*, 2024. <https://github.com/SoufiyaneOuail/Moroccan-Arabic-Speech-Emotion-Recognition-Dataset>.
- [18] B. McFee *et al.*, "librosa: audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–24, doi: 10.25080/majora-7b98e3ed-003.
- [19] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: a review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.
- [20] L. E. Boucheron and P. L. De Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," in *ICSES'08 - ICSES 2008 International Conference on Signals and Electronic Systems, Proceedings*, 2008, pp. 485–488, doi: 10.1109/ICSES.2008.4673475.
- [21] A. Alsobhani, H. M. A. Alabboodi, and H. Mahdi, "Speech recognition using convolution deep neural networks," *Journal of Physics: Conference Series*, vol. 1973, no. 1, p. 12166, Aug. 2021, doi: 10.1088/1742-6596/1973/1/012166.
- [22] E. Bisong, "Convolutional neural networks (CNN)," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. 423–441.
- [23] S. Skansi, "Convolutional neural networks," in *Nature Methods*, vol. 20, no. 9, Springer International Publishing, 2018, pp. 121–133.




- [24] G. Liu, W. He, and B. Jin, "Feature fusion of speech emotion recognition based on deep learning," in *Proceedings of 2018 6th IEEE International Conference on Network Infrastructure and Digital Content, IC-NIDC 2018*, Aug. 2018, pp. 193–197, doi: 10.1109/ICNIDC.2018.8525706.
- [25] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks," Sep. 2020, doi: 10.1109/ICWT50448.2020.9243622.
- [26] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, Dec. 2017, pp. 1–4, doi: 10.1109/APSIPA.2016.7820699.
- [27] L. Ferreira-Paiva, E. Alfaro-Espinoza, V. M. Almeida, L. B. Felix, and R. V. A. Neves, "A survey of data augmentation for audio classification," Oct. 2022, doi: 10.20906/cba2022/3469.
- [28] K. Jajuga and M. Walesiak, "Standardisation of data set under different measurement scales," in *Classification and Information Processing at the Turn of the Millennium*, Springer Berlin Heidelberg, 2000, pp. 105–112.
- [29] M. S. Shanker, M. Y. Hu, and M. S. Hung, "Effect of data standardization on neural network training," *Omega*, vol. 24, no. 4, pp. 385–397, Aug. 1996, doi: 10.1016/0305-0483(96)00010-2.
- [30] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic speech emotion recognition from saudi dialect corpus," *IEEE Access*, vol. 9, pp. 127081–127085, 2021, doi: 10.1109/ACCESS.2021.3110992.
- [31] R. Yahia Cherif, A. Moussaoui, N. Frahta, and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," in *2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021*, Mar. 2021, pp. 1–6, doi: 10.1109/WIDSTaif52235.2021.9430224.

BIOGRAPHIES OF AUTHORS



Soufiyan Ouali    is a Ph.D. student at Sidi Mohamed Ben Abdellah Fez, Morocco. Achieved his Master's degree in Computer Science specializing in data science, Artificial intelligence, and natural language processing from the Department of Computer Science Faculty of Science, Sidi Mohamed Ben Abdellah Fez, Morocco. He can be contacted at email: soufiyane.ouali@usmba.ac.ma.



Said El Garouani    is an Assistant Professor at Abdelmalek Essaadi University, Faculty of Sciences Tetouan since 2012 specializing in the field of Technical Communication and Information and mobile systems. He works also as a consultant for several companies. He can be contacted at email: said.elgarouani@usmba.ac.ma.