[1]Soufiyan Ouali*
[2]Said El Garouani

Deep Learning for Arabic Speech Recognition Using Convolutional Neural Networks

Journal of Electrical Systems

Abstract: - Extracting the speaker's emotional state from their speech 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. In this paper, we present a Speech Recognition model for the Arabic language, proficient in discerning both the emotional state and gender of the speaker through voice analysis. Three primary emotion labels were selected: low, standard, and high levels of emotion. Various spectral features, such as the mel-frequency cepstral coefficient (MFCC), were extracted and tested to determine the optimal features. Furthermore, various Machine Learning models (SVM, KNN, and HMM) and Deep Learning models (LSTM and CNN) were evaluated for training. The results were compared between the five models using different extracted features, ultimately culminating in the selection of MFCC, root-mean-square (RMS), mel-scaled spectrogram, spectral, and zero-crossing rate as spectral features, and the CNN as a classification model. This selection yielded significant results, with an accuracy of 93% for emotion recognition and 99% for gender recognition.

Keywords: Speech Emotion Recognition, Speech Gender Recognition Arabic SER, Speech Recognition.

I. INTRODUCTION

Speaking, hearing, or understanding voice is relatively simple for humans and is considered routine, including the ability to identify various characteristics of the speaker, such as their emotional state, gender, or age based on the voice tone. However, this task is far more complex for machines, as they comprehend only binary code (0 and 1). Therefore, researchers have endeavored for many decades to identify and capture important features that characterize voice or audio data [3] [16]. They employ techniques such as window-based algorithms [4], Mel-Frequency Cepstral Coefficient (MFCC), and subsequently, these extracted features are utilized for training Machine Learning (ML) or Deep Learning (DL) models, Figure 1 represents the Pipeline of SER.

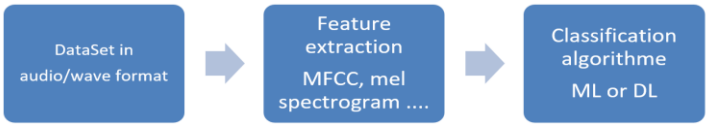

Fig. 1. Speech emotion recognition pipeline

Despite the notable advancements in the field of speech recognition in English [18], Arabic language research remains in its early stages due to the scarcity of available high-quality labeled Arabic datasets, especially in audio format. Additionally, the pronunciation of Arabic words presents complexities with challenging letters such as kha-/خ/, ha-/ح/, and aa-/ع/. Furthermore, similar pronunciations of many letters contribute to the increased complexity of SR and SER in Arabic. Many researches have been done to bridge this gap, i.e., researchers in [19] built an Arabic SER using Wav2vec2.0 and Hubert Based on the BAVED Dataset, they achieved an accuracy of 89%. Another study utilized MFCC, Mel spectrogram, and spectral contrast as spectral features with the SVM classifier for SER, reaching an accuracy of 77.14% [20]. In another study in [21] on Arabic (Egyptian dialect) SER, by using prosodic, spectral, and wavelet features researchers achieved an accuracy of 88.3%.

This paper aims to contribute to the ongoing research by constructing an efficient model capable of extracting the emotional state of the speaker and detecting the speaker's gender. different tests and experiments were conducted

[1] Department of Computer Science, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Morocco

[2] Department of Computer Science, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Morocco

*Corresponding author: Soufiyan Ouali, Email: Soufiyan.Ouali@usmba.ac.ma

to determine the optimal combination of data features and classification models for achieving good results. The rest of the article is organized as follows.

Section 2 outlines the dataset-building process. Section 3 details the feature extraction process. Section 4 introduces the classification models. The experiments result and evaluations are demonstrated in section 5. Finally, the conclusion and future work are presented in Section 6.

II. DATASET BUILDING

To train our model, we used the BAVED dataset [1], a compilation of seven Arabic words recorded in audio/wav format and spoken with various expressed emotions. Each word in the BAVED dataset is articulated on three levels of emotion, a low level of emotion (tired or exhausted), a middle level of emotion for neutral emotion, and a high level of emotion representing positive or negative emotions (happiness, joy, sadness, anger). The dataset comprises 1935 recordings made by 61 speakers, consisting of 45 males and 16 females aged between 18 and 23. Table 1 represents the distribution of recorders based on each emotion level and speakers' gender. The dataset distribution is balanced, which will have a beneficial impact on training the model [2].

using the BAVED dataset, we built two datasets, one for SER and another for speech gender recognition (SGR).

Table 1. Distribution of dataset's recorders

Learning rate / Dataset	Number of records	Number of Male	Number of Female
Low level	592	347	244
Mid-level	670	377	293
High level	674	385	289

III. FEATURE EXTRACTION

Building an efficient model is highly dependent on the quality of the training dataset. Therefore, various spectral features were extracted using the Librosa library [5] and tested to identify only the important ones. The extracted features are as follows:

- Mel-Frequency Cepstral Coefficients (MFCC), which constitute a set of coefficients capturing the shape of the power spectrum of a sound signal [6]. MFCC is widely utilized in various applications, particularly in voice signal processing, such as speaker recognition, voice recognition, and gender identification [7].
- Mel spectrogram is utilized to compute Mel-scaled spectrograms, and focusing on the low-frequency part of speech
- Spectral feature has five variants: spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, and spectral roll-off are all extracted.
- Chroma-soft, Compute chromogram from a waveform or power spectrogram.
- Root-mean-square (RMS) which computes the value RMS for each frame, either from the audio samples y or from a spectrogram
- Zero crossing rate (ZCR) of an audio time series.

IV. CLASSIFICATION MODELS

In order to select the most suitable model for our dataset, we conducted a thorough analysis of models that demonstrated high performance in SR. Therefore, five classification methods, Support Vector Machines (SVM), k-nearest Neighbors (KNN), Hidden Markov Model (HMM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) were tested. SVM, a widely used supervised learning model in SER, exhibits notable potential in feature classification and multiple regression problems [8]. KNN, another supervised learning model, identifies the k nearest neighbors to a given data point and classifies it based on the majority vote from those neighbors, as highlighted in [9] KNN shows significant potential in SR, especially for small vocabularies. HMMs, a powerful algorithm for modeling sequential data like speech signals, researchers in [10] and [11] demonstrated its ability to extract formant structure information even in noisy environments. LSTM, a type of recurrent neural network (RNN) is a pertinent algorithm for SR due to its capability to capture long-term dependencies in sequential data [12]. CNN a Deep Learning neural network architecture is known for its ability to learn complex patterns which makes it suitable for SR applications [13].

V. EXPERIMENTS AND RESULTS

1. DATASET PREPROCESSING

After downloading the BAVED dataset, two datasets were created. The first is for emotion recognition, we classified each record into one of the three emotion labels. The second dataset is for speech gender recognition, we categorized each record into male or female labels. Recognizing the significance of extensive data in training an efficient model, one key preprocessing technique employed is data augmentation [17]. 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. Through training on data containing noise, the model develops the ability to navigate such scenarios, 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 is 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 11,610 records, contributing to the model achieving notable results, as illustrated in Table 2. Beyond creating a sufficient dataset, data augmentation plays a significant role in reducing training overfit. In Table 2, it is evident that the difference between training accuracy and validation accuracy when training the model on the original data is 14%, and the validation loss is 90%, signifying a substantial degree of overfitting.

Table 2. Optimizing Model Performance: The Impact of Data Augmentation on Training

Learning rate / Dataset	Without augmented data	With augmented data
Training accuracy	0.7390	0.8271
Validation accuracy	0.5979	0.7934
Training loss	0.6093	0.4209

Validation loss	0.9032	0.7934

Another data preprocessing step involved standardizing the dataset, a crucial procedure in data analysis and machine learning [14] [15]. 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 3.

Table 3. Optimizing Model Performance: The Impact of Dataset Standardization.

Learning rate / Dataset	With OUT standardization	With standardization
Training accuracy	0.5039	0.7390
Validation accuracy	0.5258	0.5979
Training loss	0.9620	0.6093
Validation loss	0.9683	0.9032

2. FEATURE SELECTION

In order to select the most significant features for training our model (n=9), nine features were evaluated, including MFCC, Mel spectrogram, Spectral with its 5 variants, RMS, Chroma-soft, 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 4, revealed that opting for 40 coefficients led to a better learning rate. Hence, we decided to include 40 coefficients from the MFCC feature.

Table 4. The Impact of MFCC coefficient number in model training.

Learning rate / Dataset	With 10 MFCC Coefficients + (other feature)	With 40 MFCC Coefficients + (other feature)
Training accuracy	0.8271	0.9648
Validation accuracy	0.7934	0.9324
Training loss	0.4209	0.0998
Validation loss	0.7934	0.1811

After selecting the number of coefficients for the MFCC feature, we assessed the influence of other features. Figure 2 illustrates the outcomes of various experiments conducted using individual features and combinations thereof. The best results were obtained by combining the following features: MFCC, Mel-spectrogram, Spectral with its 5 variants, RMS, and ZCR, resulting in a validation accuracy of 93%. Therefore, the combination of features enhanced the classifiers' performance, leading to higher accuracy compared to individual features.

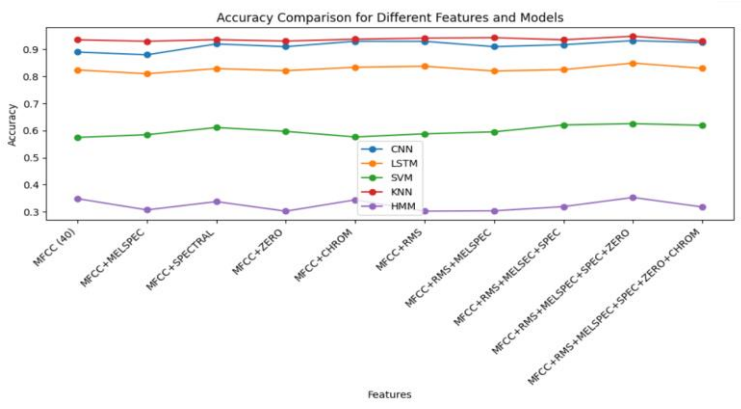

Fig. 2. Learning rate for different feature combinations.

3. CLASSIFIER ALGORITHM SELECTION

After preprocessing the dataset and selecting relevant features to choose the optimal classifier model we evaluated five classifier models (i.e., CNN, LSTM, SVM, KNN, HMM) on the SER dataset. To ensure a fair comparison and to minimize selection bias, we used the validation accuracy as a metric, which indicates the model's ability to predict unknown data. Table 5 summarizes the results obtained by each classifier, highlighting the superior performance of KNN and CNN. While the validation accuracy of the KNN model exceeds that of the CNN, relying solely on training and validation accuracy is not sufficient for a comprehensive evaluation. Therefore, we examined the cross-validation score of the KNN model, which was 0.9366. While the metrics exhibit similarity, a more profound analysis was essential to discern the distinguishing characteristics between these two models. Consequently, we selected CNN for our tasks because, in contrast to KNN, CNN has an excellent ability to learn complex patterns and hierarchical features in sequential data. Furthermore, the computational efficiency of CNNs in prediction is well suited to the real-time processing requirements of SR applications. Furthermore, KNN requires to calculate distances for all data points during predictions, and with our large dataset, this can be computationally expensive, potentially leading to predictions that are less accurate and natural.

Table 5. Model performance comparison

Learning rate / Dataset	CNN	LSTM	SVM	KNN	HMM
Training accuracy	0.9648	0.8966	0.6223	0.9736	0.3178
Validation accuracy	0.9324	0.8493	0.6256	0.9484	0.3528

The architecture of the CNN model we built consists of multiple layers for feature extraction and classification. The network begins with a series of Conv1D layers with a relu activation function, each followed by Batch Normalization and MaxPooling1D for down-sampling ((pool_size=5, strides=2, padding='same'). Dropout (20%) layers are strategically placed to prevent overfitting. The network progressively reduces the spatial dimensions of the input data, capturing hierarchical features. The Flatten layer converts the output into a one-dimensional array, which is fed into Dense layers for further processing. The final Dense layer produces predictions with a SoftMax activation function for three output classes, Low level, middle level, and high level. The model is compiled with the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. Moreover, a learning rate reduction strategy is implemented with ReduceLROnPlateau, monitoring validation accuracy, reducing the learning rate by a factor of 0.5 after 3 epochs of stagnation, with a minimum learning rate of 0.00001. The optimal results were achieved after 20 epochs (as shown in Figure 3), utilizing a batch size of 32, and implementing a learning rate reduction to 0.0005 This configuration played a crucial role in enhancing the model's performance for both tasks. The architecture explained above was adopted to train both models, Arabic Speech emotion recognition, and speech gender recognition (with modifying the final Dense layer to have only 2 neurons as it predicts two values male or female).

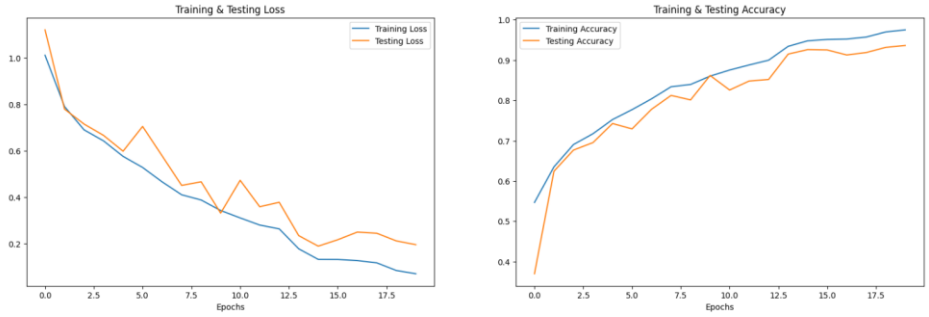

Fig. 3. Training curve over epochs for SER task with CNN

For a deeper analysis of the results, we constructed confusion matrices illustrating the classifiers' performance in predicting each emotion and the gender of the speaker (Figure 5). On these matrices, the x-axis signifies the predicted labels, while the y-axis signifies the true labels. Notably, the high emotion level and male voice categories

exhibited robust predictions, achieving the highest accuracy rates of 93% and 99%, respectively. These results can be explained by the fact that the high emotions level class and male voice class are well represented by the actor and contain high frequency and pitch (as shown in Figure 4) which are easily captured by the CNN classifier. As illustrated in Table 6, the results are compared to the state of the art, demonstrating that our model outperformed existing research.

Table 6. Comparing Our Model with State-of-the-Art Models

Model	Accuracy
[19] Arabic SER with BAVED dataset	89 %
[20] Arabic (Saudi dialect) SER	77.14 %
[21] on Arabic (Egyptian dialect) SER	88.3%
Our model, SER for Arabic	93%

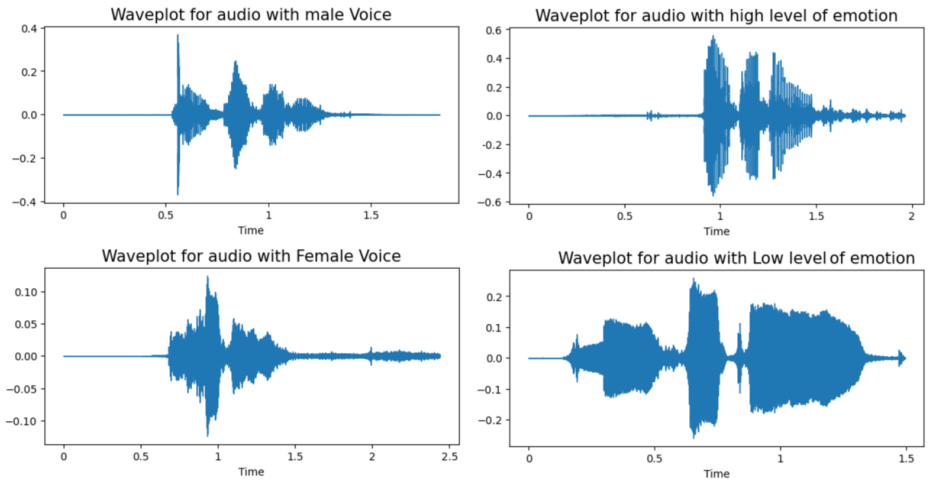

Fig. 4. Difference between high, low, male, and female voice tone using Wave plot

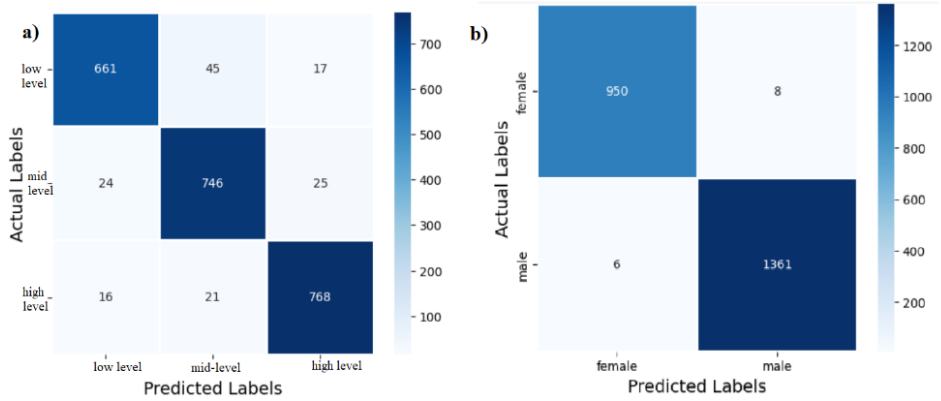

Fig. 5. a) Confusion matrix for SER. b) Confusion matrix for SGR

VI. CONCLUSION AND FUTUR WORK

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 performing two tasks: speech emotion recognition and speech gender recognition in the Arabic 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 demonstrated a significant improvement in model learning, increasing from 59% to 79%. Standardizing the dataset further enhanced model learning from 52% to 59%, while selecting the appropriate number of MFCC coefficients boosted training from 79% to 93%. Combining MFCC, Mel-spectrogram, Spectral features, RMS, and ZCR with a CNN model resulted in a remarkable improvement, raising the learning accuracy from 89% to 93%.

The model achieved promising results in both tasks, with a 95% accuracy rate for gender identification by voice and 93% for extracting the speaker's emotional state. Furthermore, adjusting the training dataset, enhanced the model's accuracy and realism, simulating natural voice identification by humans which are affected by factors such as noise, pitch changes, and speed variations. The challenge of limited data prompted our decision to build a model predicting three emotional states. In future work, we aim to create a larger dataset encompassing at least eight emotion states and explore additional spectral features to further enhance the model's accuracy.

ACKNOWLEDGMENT

Competing interests: The authors declare that they have no financial or personal relationship that may have inappropriately influenced them in writing this article.

Funding information: The authors received no financial support for the research, authorship, and publication of this article.

REFERENCES

[1] A. Aouf, “Basic Arabic vocal emotions dataset (baved) - github,” https://github.com/40uf411/Basic-Arabic-VocalEmotions-Dataset, 21 September, 2019.

[2] Olson, David L.. “Data Set Balancing.” Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management (2004).

[3] “A Survey on Different Algorithms for Automatic Speaker Recognition Systems.” (2016).

[4] Sapijaszko, Genevieve M. and Wasfy B. Mikhael. “An overview of recent window-based feature extraction algorithms for speaker recognition.” *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)* (2012): 880-883.

[5] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, ‘‘Librosa: Audio and music signal analysis in Python,’’ in Proc. 14th Python Sci. Conf., 2015, pp. 18–25

[6] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," in IEEE Access, vol. 10, pp. 122136-122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[7] Boucheron, Laura E. and P.L. de Leon. “On the inversion of Mel-frequency cepstral coefficients for speech enhancement applications.” *2008 International Conference on Signals and Electronic Systems* (2008): 485-488.

[8] Sonkamble, Balwant A. and Dharmpal D. Doye. “An overview of speech recognition system based on the support vector machines.” *2008 International Conference on Computer and Communication Engineering* (2008): 768-771.

[9] Lippmann, Richard. “Review of Neural Networks for Speech Recognition.” *Neural Computation* 1 (1989): 1-38.

[10] Weber, Katrin. “HMM Mixtures (HMM2) for Robust Speech Recognition.” (2003).

[11] Weber, Katrin, Samy Bengio and Hervé Bourlard. “HMM2- extraction of formant structures and their use for robust ASR.” *Interspeech* (2001).

[12] Zeyer, Albert, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter and Hermann Ney. “A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition.” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2016): 2462-2466.

[13] Alsobhani, Ayad, Hanaa M A ALabboodi and Haider Salih Mahdi. “Speech Recognition using Convolution Deep Neural Networks.” *Journal of Physics: Conference Series* 1973 (2021): n. pag.

[14] Jajuga, Krzysztof and Marek Walesiak. “Standardisation of Data Set under Different Measurement Scales.” (2000).

[15] McCaffrey, James P. “Standardizing Data for Neural Networks.” (2014).

[16] Furui S. "History and Development of Speech Recognition." Speech Technology, pp. 1--18, 2010.

[17] Ferreira-Paiva, Lucas, Elizabeth Alfaro-Espinoza, Vinicius M. Almeida, Leonardo B. Felix, and Rodolpho VA Neves. "A survey of data augmentation for audio classification." In *Congresso Brasileiro de Automática-CBA*, vol. 3, no. 1. 2022.

[18] Wang, Dong, Xiaodong Wang, and Shaohe Lv. "An overview of end-to-end automatic speech recognition." *Symmetry* 11, no. 8 (2019): 1018.

[19] Mohamed, Omar, and Salah A. Aly. "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset." *arXiv preprint arXiv:2110.04425* (2021).

[20] Aljuhani, Reem Hamed, Areej Alshutayri, and Shahd Alahdal. "Arabic speech emotion recognition from saudi dialect corpus." *IEEE Access* 9 (2021): 127081-127085.

[21] Abdel-Hamid, Lamiaa. "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features." *Speech Communication* 122 (2020): 19-30.