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

Arabic Speech Emotion Recognition using Convolutional Neural Networks

Journal of Electrical Systems

Abstract: - Emotions are considered an essential and fundamental aspect of human conversations. It serves as a means for opinion expression and for enlightening others about their psychological and physical well-being. Therefore, 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. In this paper, we present an effective Arabic Speech Recognition model, proficient in discerning both the emotional state and gender of the speaker through voice analysis. The model is trained to recognize six primary emotions: tiredness, sadness, anger, neutrality, happiness, and joy. For dataset preprocessing and feature extraction, various spectral features, such as the Mel-frequency Cepstral coefficient (MFCC), were extracted and tested to determine the optimal feature combination. For Classifier selection, Three Machine Learning models (SVM, KNN, and HMM) and two Deep Learning models (LSTM and CNN) were evaluated for training. The experimental results were analyzed and compared across the five models using various performance measures. This evaluation aimed to select the optimal model capable of performing well under different conditions, including noisy environments. The best results are achieved by the Combination of MFCC, root-mean-square (RMS), mel-scaled spectrogram, Spectral feature, and zero-crossing rate as spectral features, and the CNN as a classification model. This selection yielded significant results outperforming the models in the State of the Art, with a Recognition accuracy of 93% for emotion recognition and 99% for gender recognition.

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

I. INTRODUCTION

Recently, more attention has been directed towards building intelligent Generative chatbots that mimic human conversation [1]. Despite the significant advancements in text-based chatbots, the development of intelligent speech chatbots remains in its early stages [2]. Furthermore, current research highlights that creating an effective speech chatbot requires more than just answering users' questions; it also necessitates understanding the user's utterances based on their emotional state.

Humans are known for the impact their emotions have on their conversations [3]. For example, a person might say, "I'm so glad you finally showed up," with a tone that conveys frustration or disappointment rather than happiness. This highlights the need for chatbots to interpret the emotional context accurately to respond appropriately. Consequently, researchers in the chatbot domain are increasingly focusing on developing models capable of extracting the emotional state of speakers to provide appropriate responses [4].

The 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 [5] [6].

The main idea behind voice analysis is that, since sound is an electric wave that propagates through space, the resulting signal can be captured and analyzed [7]. Given that muscle tension may increase in certain emotional conditions, segmental features of speech may also be influenced by the speaker's emotional state. Therefore, researchers interested in extracting the speaker's emotional state analyze various features such as pitch and amplitude. In reference [8], X. M. Cheng et.al., analyzed features such as time, amplitude, pitch, and formant construction for emotions like happiness, anger, surprise, and sorrow. By comparing these with neutral speech, they identified nine emotional features for emotion recognition. They introduced two recognition methods based on principal component analysis, demonstrating that these methods effectively recognize emotions.

[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

Copyright © JES 2024 on-line: journal.esrgroups.org

In reference [9], D. Ververidis et.al., introduced a detailed and robust set of spectral features, including statistics of Mel-Frequency Cepstral Coefficients (MFCCs) computed over three phoneme types: stressed vowels, unstressed vowels, and consonants. They evaluated these features in speaker-independent emotion recognition using two public datasets. Their results showed that this richer set of spectral features, along with phoneme type differentiation, significantly improves classification accuracy compared to prosodic or utterance-level spectral features. Furthermore, they found that the accuracy of emotion recognition using class-level spectral features increases with utterance length, unlike utterance-level prosodic features.

Various techniques are used for feature extraction such as window-based algorithms [10], and Mel-Frequency Cepstral Coefficient (MFCC) [11]. As illustrated in ***Figure 1***, The extracted features are utilized for training Machine Learning (ML) or Deep Learning (DL) models.

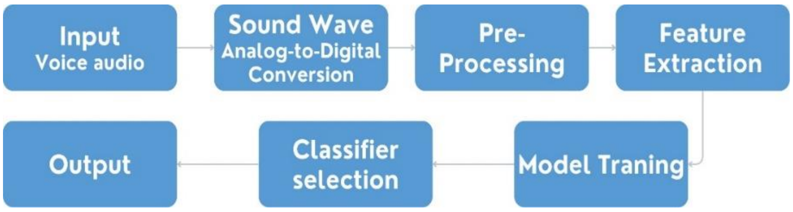

Fig. 1. Speech Emotion Recognition pipeline

Despite the notable advancements in the field of speech recognition in English [12], Arabic language research remains in its early stages due to many reasons, including the scarcity of available high-quality labeled Speech Arabic datasets. 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 increase the complexity of Speech Recognition (SR) and Speech Emotion Recognition (SER) in Arabic.

Numerous research endeavors have been conducted to address this disparity. In [13] researchers built an Arabic SER model Based on the BAVED Dataset [14] to recognize four emotional expressions (Happy, Sad, Angry, Neutral). They used Wav2vec2.0 and Hubert for feature extraction and Bidirectional-LSTM for the classification, they achieved an accuracy of 89%. In reference, [15] researchers built an emotion recognition model that recognizes four emotional expressions (Happy, Sad, Angry, Neutral) in the Saudi Arabian dialect. They constructed a new Elicited speech dataset collected from TV shows and sitcoms. For feature extraction and classification, they utilized MFCC, Mel spectrogram, spectral contrast, and the SVM classifier. They achieved a recognition accuracy of 77.14%. Similarly, in [16] researchers constructed a new dataset to train a SER model on the Egyptian Arabic dialect. By using prosodic, spectral, and wavelet features and the SVM classifier researchers achieved a recognition 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 to determine the optimal combination of data features and classification models. 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

To train our model, we used the BAVED dataset [14], a compilation of seven Arabic words (اعجبني-like, لم يعجبني-unlike, هدا-this, الفيلم-film, رائع-good, مقبول-neutral, and سيئ-bad) recorded in audio.wav format and spoken with various expressed emotions. Each word in the BAVED dataset is pronounced 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 samples recorded by 61 speakers, consisting of 45 males and 16 females aged between 18 and 23. ***Figure 2*** 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 [17]. Through using the BAVED dataset, we have built two datasets, one for SER and another for speech gender recognition (SGR).

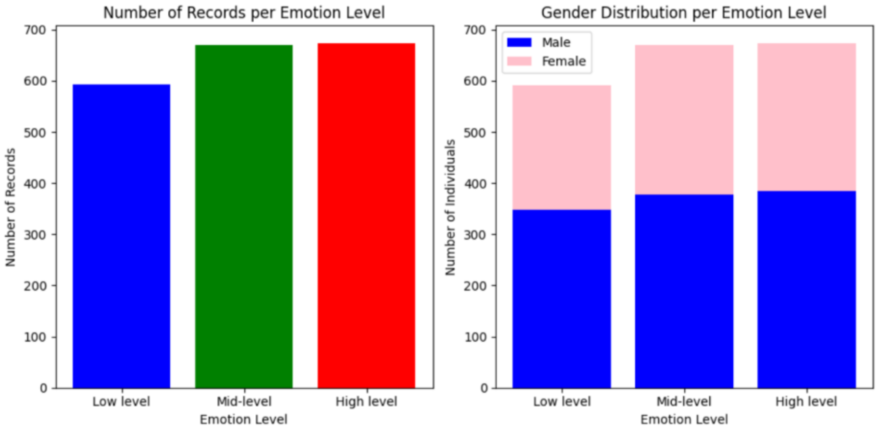

Fig. 2.Distribution of dataset's recorders

The pre-processed and combined data is then divided into two subsets: a training set and a testing set. Specifically, 80% of the data is allocated for training and 20% for testing. The split is performed in a stratified manner to ensure generalization across various classes. The training set is used to train the SER model, while the testing set is used to evaluate the model's performance. This approach ensures that the model is tested on unseen data, helping to prevent overfitting and improving its generalizability.

III. FEATURE EXTRACTION

Feature extraction is the process of transforming raw data into a set of measurable characteristics or features that can be used for analysis. Building an efficient model is highly dependent on the quality of the training dataset. Therefore, feature extraction is a crucial step.

Numerous feature extraction techniques exist in the state of the art. These techniques can be categorized into continuous features (e.g., pitch, energy, formants), qualitative features (e.g., voice quality, harshness, tenseness, breathiness), and spectral features (e.g., MFCC, LPC, LFPC) [18]. Spectral features, in particular, have shown promising results [16, 19], as they capture both local and global voice information while using a lower-dimensional space, thereby simplifying models and improving computational efficiency. Therefore, in this paper, we focus on extracting and training our model using spectral features.

The spectral features were extracted using the Librosa library [20] and tested to select the most important ones. The extracted features are:

- ***Mel-Frequency Cepstral Coefficients (MFCC)***, which constitute a set of coefficients capturing the shape of the power spectrum of a sound signal [11]. MFCC is widely utilized in various applications, particularly in voice signal processing, such as speaker recognition, voice recognition, and gender identification [21].
- ***Mel spectrogram*** is utilized to compute Mel-scaled spectrograms, and focusing on the low-frequency part of speech
- ***Spectral 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

To select the most suitable model for our dataset, we conducted a thorough analysis of models that demonstrated high performance in Speech Recognition (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 [22]. 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 [23] KNN shows significant potential in SR, especially for small vocabularies. HMMs, a powerful algorithm for modeling sequential data like speech signals, researchers in [24] and [25] 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 [26]. CNN a Deep Learning neural network architecture is known for its ability to learn complex patterns which makes it suitable for SR applications [27].

V. EXPERIMENTS AND RESULTS

1. DATASET PREPROCESSING

After downloading the BAVED dataset, two datasets were created. The first one is for the SER task, in which we classified each record into one of the three emotion labels. The second dataset is for the SGR task, in which we categorized each record into male or female labels.

The first 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 1***.

Table 1. Optimizing Model Performance: The Impact of Dataset Standardization on Recognition Rate.

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

The results demonstrate that standardizing the data leads to significant improvements in both training and validation accuracies, with training accuracy increasing from 0.5039 to 0.7390 and validation accuracy from 0.5258 to 0.5979. This corresponds to enhancements of approximately 47% and 14%, respectively. Additionally, validation loss decreases from 0.9683 to 0.9032 with standardization, indicating better model performance. Standardization aids the model in learning and generalizing more effectively. However, while the results have been enhanced, they are not considered very good for consideration.

The second dataset preprocessing is **data augmentation**. Recognizing the significance of extensive data in training an efficient model, one key preprocessing technique employed is data augmentation [30]. This process involves creating new synthetic data samples by introducing small perturbations to the initial training set by injecting various effects. The effects used in our dataset include:

- ***noise injection***, 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.
- ***Time Shifting,*** 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 as shown in ***Table 2***, before data augmentation, the model achieves a relatively high training accuracy of 0.7390, indicating that it performs well on the training data. However, the validation accuracy is significantly lower at 0.5979, suggesting that the model's performance on unseen data is not as strong. On the contrary, after data augmentation. The difference between training accuracy and validation accuracy is ≈ 0.03, which indicates that the model can generalize well on unseen data.

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

Learning rate/ Dataset	Without augmented data	With augmenteddata
Training accuracy	0.7390	0.8271
Validation accuracy	0.5979	0.7934
Training loss	0.6093	0.4209
Validation loss	0.9032	0.7934

The model trained with augmented data shows significant improvements compared to the model trained without augmented data. Specifically, training accuracy increases by 8.81% (from 0.7390 to 0.8271), and validation accuracy improves by 19.55% (from 0.5979 to 0.7934). Additionally, the validation loss decreases by 11.98% (from 0.9032 to 0.7934) with augmented data. These results highlight the effectiveness of data augmentation in enhancing both training and validation performance, as well as improving model convergence during training.

2. FEATURE SELECTION

To select the most significant features for training our model 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 3***, revealed that opting for 40 coefficients led to a better learning rate. Hence, we decided to include 40 coefficients from the MFCC feature.

Table 3. The Impact of MFCC Coefficient Number in Recognition Rates.

Learning rate/ Dataset	With 10 MFCCCoefficients	With 40 MFCCCoefficients
	+ (other feature)	+ (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 3*** illustrates the outcomes of various experiments conducted using individual features and combinations. 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.

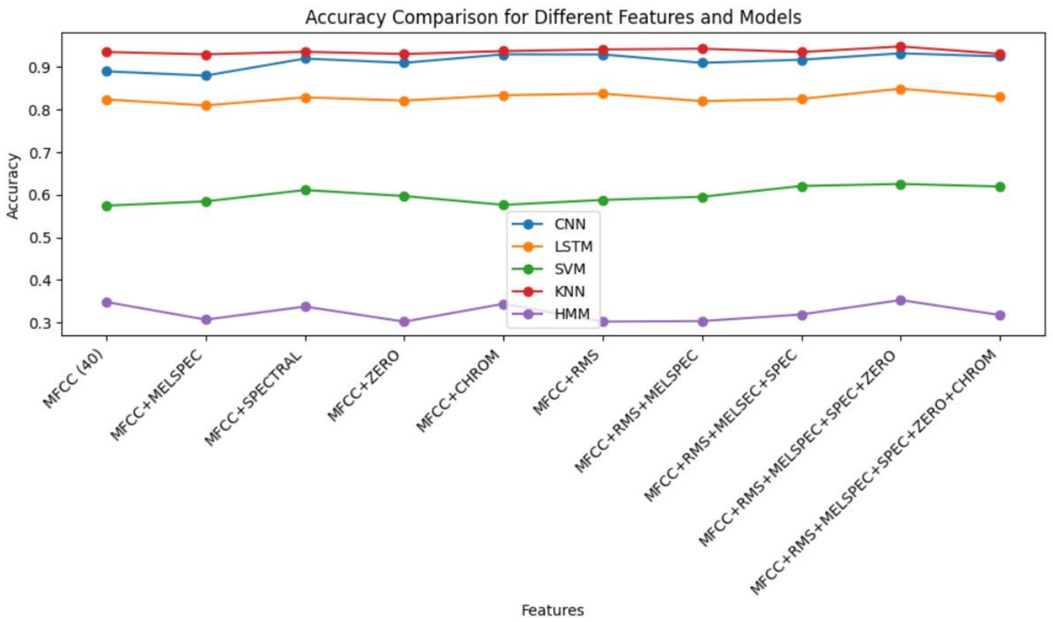

Fig. 3. Recognition accuracy for different feature combinations (measure: Validation accuracy).

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 4*** 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 insufficient 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 4. 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 4***), 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.

For the Speech Gender Recognition task, we used the same architecture explained above while modifying the final Dense layer to have only 2 neurons as it predicts two values male or female.

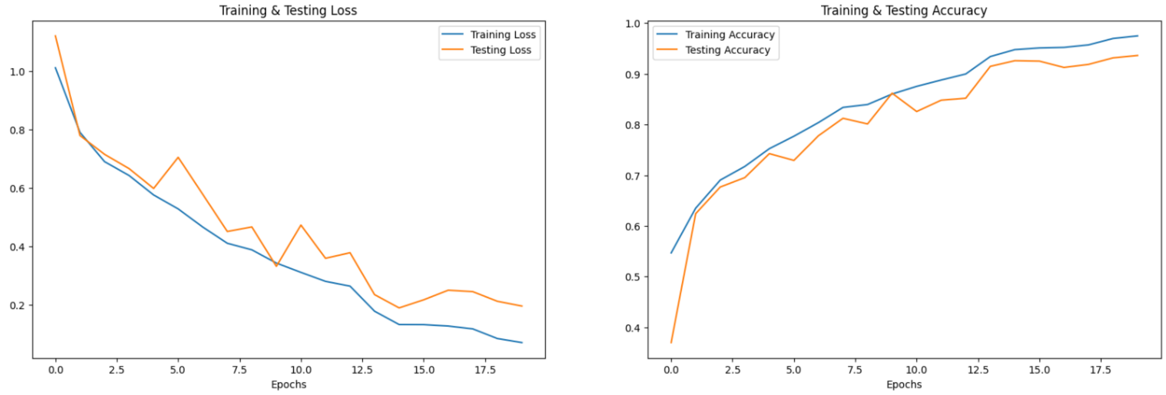

Fig. 4. Training curve over epochs for SER task with CNN

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 6***. 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 5***) which are easily captured by the CNN classifier.

As illustrated in ***Table 5***, the results are compared to the state of the art, demonstrating that our model outperformed existing research.

Table 5. 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, Arabic SER	93%

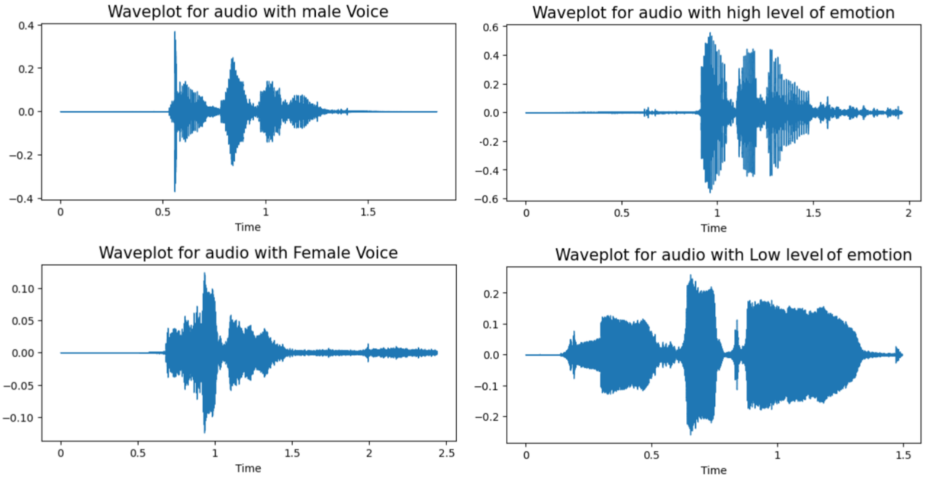

Fig. 5. Difference between high, low, male, and female voice tone using Wave plot

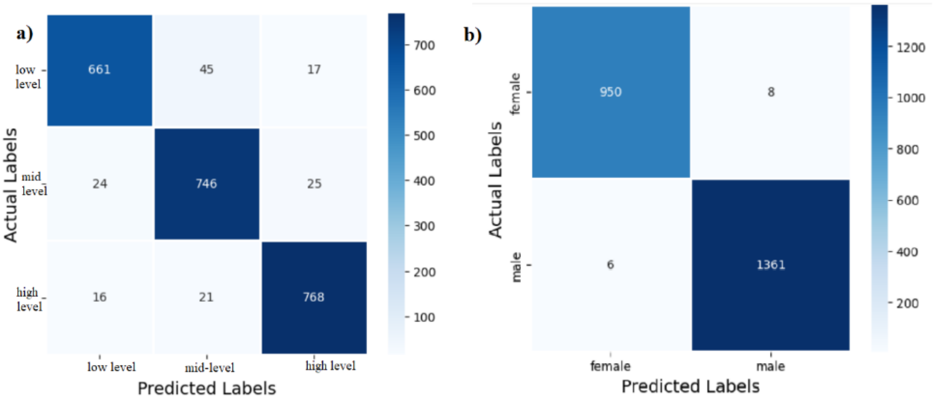

Fig. 6. a) Confusion matrix for SER. b) Confusion matrix for SGR

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 construction and selection, we identified optimal combinations to build an effective model. Implementing standardization techniques resulted in a significant enhancement in model Recognition Rate, with a 7% increase. Similarly, Data augmentation demonstrated a significant improvement in model learning, with a 20% increase in Recognition rates. Furthermore, selecting the appropriate number of MFCC coefficients enhanced training from with a 10% increase. Finally, Combining MFCC, Mel-spectrogram, Spectral features, RMS, and ZCR with a CNN model resulted in a remarkable improvement, with a 4% to 10% increase in Recognition rate across all classifiers.

The model achieved promising results in both tasks, with a 99% accuracy rate for speech gender identification 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] J. Grudin and R. Jacques, "Chatbots, humbots, and the quest for artificial general intelligence," in Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, May 2019. doi: 10.1145/3290605.3300439.

[2] S. A. and Dr. John, "Survey on Chatbot Design Techniques in Speech Conversation Systems," in International Journal of Advanced Computer Science and Applications, The Science and Information Organization, 2015. doi: 10.14569/IJACSA.2015.060712.

[3] K. Baskaran, W. Cui, and A. Kankanhalli, "A Review of Emotions in Human-Conversational Agent Interaction," in Proceedings of the AAAI Symposium Series, AAAI Press, Oct. 2023, pp. 61–67. doi: 10.1609/AAAISS.V1I1.27477.

[4] S Ramakrishnan, "Recognition of emotion from speech: A review," in Speech Enhancement, Modeling and Recognition-Algorithms and Applications., vol. 7, InTech, 2012, pp. 121–137.

[5] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," in Communications in Computer and Information Science, 2010, pp. 134–145. doi: 10.1007/978-3-642-17641-8_18.

[6] X. Xiong, "A Summary of the Development of Speech Recognition Technology," in ACM International Conference Proceeding Series, Association for Computing Machinery, Dec. 2022, pp. 768–773. doi: 10.1145/3584376.3584513.

[7] Duncan. Templeton, "Perception of sound," in Acoustic design., Van Nostrand Reinhold, 1987, pp. 13–48.

[8] X. M. Cheng, P. Y. Cheng, and L. Zhao, "A study on emotional feature analysis and recognition in speech signal," in 2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009, 2009, pp. 418–420. doi: 10.1109/ICMTMA.2009.89.

[9] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Commun, vol. 48, no. 9, pp. 1162–1181, Sep. 2006, doi: 10.1016/J.SPECOM.2006.04.003.

[10] G. I. Sapijaszko and W. B. Mikhael, "An overview of recent window based feature extraction algorithms for speaker recognition," in Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest, 2012, pp. 880–883. doi: 10.1109/MWSCAS.2012.6292161.

[11] 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.

[12] D. Wang, X. Wang, and S. Lv, "An Overview of End-to-End Automatic Speech Recognition," Symmetry 2019, Vol. 11, Page 1018, vol. 11, no. 8, p. 1018, Aug. 2019, doi: 10.3390/SYM11081018.

[13] O. Mohamed and S. A. Aly, “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.

[14] Ali Aouf., “Basic Arabic Vocal Emotions Dataset.” Kaggle, 2020. doi: 10.34740/kaggle/ds/345828.

[15] 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.

[16] L. Abdel-Hamid, “Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features,” Speech Commun, vol. 122, pp. 19–30, Sep. 2020, doi: 10.1016/J.SPECOM.2020.04.005.

[17] D. L. Olson, “Data Set Balancing,” in Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 2004, pp. 71–80. doi: 10.1007/978-3-540-30537-8_8.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognit, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/J.PATCOG.2010.09.020.

[19] 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.

[20] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” in Proceedings of the 14th Python in Science Conference, SciPy, 2015, pp. 18–24. doi: 10.25080/MAJORA-7B98E3ED-003.

[21] 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.

[22] B. A. Sonkamble and D. D. Doye, “An overview of speech recognition system based on the support vector machines,” in Proceedings of the International Conference on Computer and Communication Engineering 2008, ICCCE08: Global Links for Human Development, 2008, pp. 768–771. doi: 10.1109/ICCCE.2008.4580709.

[23] R. P. Lippmann, “Review of Neural Networks for Speech Recognition,” Neural Comput, vol. 1, no. 1, pp. 1–38, Mar. 1989, doi: 10.1162/NECO.1989.1.1.1.

[24] K. Weber, “HMM mixtures (HMM2) for robust speech recognition,” Docteur ès Sciences thesis, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland., 2003, doi: 10.5075/epfl-thesis-2790.

[25] K. Weber, S. Bengio, and H. Bourlard, “HMM2- Extraction of formant structures and their use for robust ASR,” in EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology, International Speech Communication Association, 2001, pp. 607–610. doi: 10.21437/EUROSPEECH.2001-161.

[26] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney, “A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., Jun. 2016, pp. 2462–2466. doi: 10.1109/ICASSP.2017.7952599.

[27] A. Alsobhani, H. M. A. Alabboodi, and H. Mahdi, “Speech Recognition using Convolution Deep Neural Networks,” J Phys Conf Ser, vol. 1973, no. 1, p. 012166, Aug. 2021, doi: 10.1088/1742-6596/1973/1/012166.

[28] K. Jajuga and M. Walesiak, “Standardisation of Data Set under Different Measurement Scales,” in Decker, R., Gaul, W. (eds) Classification and Information Processing at the Turn of the Millennium. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg., Springer, Berlin, Heidelberg, 2000, pp. 105–112. doi: 10.1007/978-3-642-57280-7_11.

[29] J. McCaffrey, “Standardizing Data for Neural Networks,” 2014.

[30] 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,” Congresso Brasileiro de Automática - CBA, vol. 3, no. 1, Oct. 2022, doi: 10.20906/CBA2022/3469.