NABEA AL-AJYAL FOUNDATION PRESS
Online ISSN: 3078-5413
CyberSystem Journal (CSJ)
csj.nabea.pub
CyberSystem Journal Research Article

A Hybrid Deep Learning Approach for Arabic Speech Emotion Recognition Using Acoustic and Spectrogram Features

Dalal Djeridi[1*], Bachir Said[2] and Mourad Belhadj[3]

[1,2,3] Laboratory of Artificial Intelligence and Information Technologies, University of Kasdi Merbah Ouargla, Algeria

* Corresponding Author: **Dalal Djeridi**, Email: hadjeridi.dalal@univ-ouargla.dz.

***Abstract*:** Arabic Speech Emotion Recognition (ASER) presents unique challenges due to linguistic diversity, phonetic complexity, and the limited availability of labeled datasets. This work presents a hybrid deep learning model that integrates a Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN) to effectively classify emotions from Arabic speech data, using the KEDAS dataset. The model is designed to process two complementary sets of features: acoustic features, such as low-level descriptors, as well as Mel-spectrogram features that capture time-frequency information. Experimental results demonstrate that the hybrid architecture effectively leverages the strengths of both feature types, achieving a high accuracy of 98% in emotion recognition.

Access this article online

***Keywords*:** Arabic Speech Emotion Recognition, low-level descriptors, Mel-Spectrogram, KEDAS dataset

1. Introduction

In recent years, the demand for emotionally intelligent systems has driven significant research in Speech Emotion Recognition (SER). These systems aim to identify a speaker's emotional state based solely on vocal cues, enabling more natural and adaptive human-computer interaction. While considerable progress has been made in SER for languages like English, Arabic remains underexplored, despite its widespread use and linguistic richness. Arabic poses specific challenges due to its diverse dialects, phonetic variations, and prosodic structures, which complicate the extraction and interpretation of emotional cues. To address these challenges, this research presents a hybrid deep learning model that leverages both acoustic features and deep spectral representations of Arabic speech. The model incorporates a Multi-Layer Perceptron (MLP) that processes low-level descriptors widely recognized in speech processing. Simultaneously, a Convolutional Neural Network (CNN) branch is applied to Mel-spectrogram inputs, enabling the model to detect local patterns in time and frequency. The fusion of these branches captures both global and localized emotional characteristics of speech. This dual-branch design, trained on the KEDAS dataset, provides a promising framework for overcoming the limitations of single-feature or single-architecture approaches in Arabic emotion recognition.

[1] developed an Algerian dialect dataset using MFCC features, achieving a highest accuracy of 93.34% with their CNN-LSTM model. Mohammad et al. used a combination of Linear Predictive Coding (LPC) and Periodogram Power Spectral Density (PPSD) features. The extracted features

Received February 1, 2025; Revised March 7, 2025; Accepted May 11, 2025; Published June 30, 2025
https://doi.org/10.57238/csj.2025.1008

were fed into a classifier (such as ANN, SVM, or logistic regression). The experiments demonstrated that logistic regression achieved the best result with 91.7% (Mohammad & Elhadef, 2021). Ouali & Garouani combined MFCC, mel-spectrogram, spectral features, RMS, and ZCR with a CNN model. Which was achieving a 99% accuracy rate for speech gender identification and a 93% accuracy rate for speech emotion recognition [3]. Kaloub & Elgabar used different acoustic feature sets (including MFCC, Mel spectrogram, spectral contrast, zero-crossing rate, and intensity) and applied four ML classifiers (SMO, RF, KNN, and SL). The best performance was achieved by SMO and SL classifiers, at 83.82% [4]. We organize the rest of the paper as follows: In Section 2, we present the methodology of our work, which describes the dataset used, the extracted features, and the proposed method. In Section 3, we present the results. In Section 4, we discuss the results obtained. Finally, we conclude in Section 5.

2. Materials and Methods

2.1 Dataset Description

In this work, we utilized the Kasdi Merbah Emotional Dataset in Arabic Speech (KEDAS), developed by the LINATI laboratory at Kasdi Merbah Ouargla University in partnership with the Unit of Linguistic Research and Arabic Language Issues in Algeria. The dataset was constructed using ten carefully selected sentences representing five emotional states: fear, anger, sadness, happiness, and neutrality. KEDAS includes 5,000 audio samples recorded in Modern Standard Arabic by 500 speakers (254 female and 246 male), comprising 417 young and 83 middle-aged or elderly individuals. Each emotional category is represented by 1,000 recordings [5]. The dataset is publicly accessible [6].

2.2 Features Extraction

We use two types of feature extraction methods: low-level descriptors (LLDs) and Mel-spectrogram features.

Low-Level Descriptors (LLDs). The openSMILE toolkit was used to extract 26 LLDs from the 'emobase' feature set for each time frame. The extracted acoustic features included:

- Zero-crossing rate (ZCR)
- Intensity
- Loudness
- Probability of voicing
- Fundamental frequency (F0) and its envelope
- 12 Mel-frequency cepstral coefficients (MFCCs)
- 8 line spectral pairs (LSPs)
- The following functionals were applied to the LLDs:
- Overall range
- Arithmetic mean
- Maximum and minimum values (with their relative positions within the input sequence)
- Standard deviation
- Distribution's kurtosis and skewness
- Two coefficients from linear regression
- Errors from both linear and quadratic fits
- First to third quartiles
- Three interquartile ranges

This process resulted in a feature vector size of 988 [7].

Mel-Spectrogram. The generation of a Mel-spectrogram involves several key steps:

1. Pre-emphasis is applied to the audio signal to enhance clarity and suppress low-frequency components.
2. The signal is segmented into overlapping frames to preserve temporal continuity.
3. A windowing function is applied to each frame to minimize spectral leakage and reduce discontinuities caused by digital sampling.
4. The Fast Fourier Transform (FFT) is applied to convert the framed signal from the time domain to the frequency domain.
5. The resulting spectrum is mapped onto the Mel scale using triangular bandpass filters, producing the Mel spectrogram [8].

2.3 Training and Testing Process

For training and evaluation, the data was divided into two subsets: 80% for training, and 20% for testing. Additionally, we shuffled the data using a random seed of 42 to ensure reproducibility.

2.4 Building the Hybrid Model

The proposed hybrid model combines two complementary deep learning architectures—a Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN)—to enhance the performance of Arabic Speech Emotion Recognition (ASER). The model is designed to

process two distinct types of audio features extracted from speech signals. The first input branch is an MLP that takes acoustic features such as LLD, capturing global audio characteristics. Architecture includes: fully connected layers with ReLU activation, batch normalization, and dropout for regularization, enabling it to learn complex patterns in the tabular data. The second input branch is a CNN that processes Mel-spectrogram features as sequential 1D data, allowing the model to learn local temporal patterns and frequency variations in the speech. Architecture includes Conv1D and MaxPooling1D layers followed by flattening. The outputs of both branches are concatenated and passed through additional dense layers to learn fused representations before reaching the softmax output layer, which classifies the input into one of five emotion categories. This hybrid approach leverages both global statistical features and localized spectral information, resulting in a more robust and accurate emotion recognition system. As shown in Figure 1 illustrates the proposed approach.

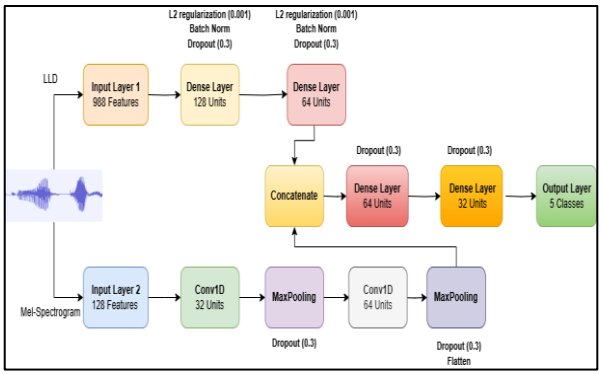

Figure 1. Proposed Hybrid MLP-CNN Architecture for Arabic SER

3. Results

The experiment evaluated the hybrid model's performance on the KEDAS dataset. The model was trained for 100 epochs with a batch size of 32, using validation data to monitor performance. To optimize training and prevent overfitting, we implemented two callbacks:

- EarlyStopping: This monitored validation loss and halted training after five epochs without improvement, while restoring the best weights.
- ReduceLROnPlateau: The reduced the learning rate by a factor of 0.5 after three epochs without improvement, enabling finer optimization steps.

This combination of callbacks enhanced the model's generalization capability while ensuring efficient training. We evaluated performance using four metrics: accuracy, precision, recall, and F1-score, supplemented by a confusion matrix and ROC curve analysis.

We conducted comparative experiments on the dataset and analyzed each result. The key findings are summarized in

Table 1.

shows the training curve across epochs for the speech emotion recognition task. Figure 2Figure 2 presents the ROC curve for the KEDAS dataset.

Table 2 details the model's performance across five emotions (Angry, Happy, Fear, Neutral, and Sad) using precision, recall, F1-score, and accuracy metrics.

Table 3 displays the confusion matrix for the KEDAS dataset.

Table 1. Accuracy of our Experiments

Model	Accuracy %
CNN	96.40
MLP	96.90
MLP-CNN	98

Table 2. Precision, recall, and F1 score on KEDAS

Measure/Emotion	Angry	Sad	Fear	Happy	Neutral
Precision	0.99	0.99	0.98	0.97	0.98
Recall	0.99	0.99	0.99	0.97	0.98
F1-score	0.99	0.99	0.99	0.97	0.98

Table 3. Confusion matrix of the KEDAS dataset

	Angry	Happy	Fear	Neutral	Sad
Angry	0.98	0.01	0.00	0.00	0.00
Happy	0.00	0.96	0.03	0.01	0.00
Fear	0.00	0.00	0.99	0.00	0.01
Neutral	0.01	0.00	0.01	0.98	0.01

Sad	0.01	0.01	0.00	0.01	0.98

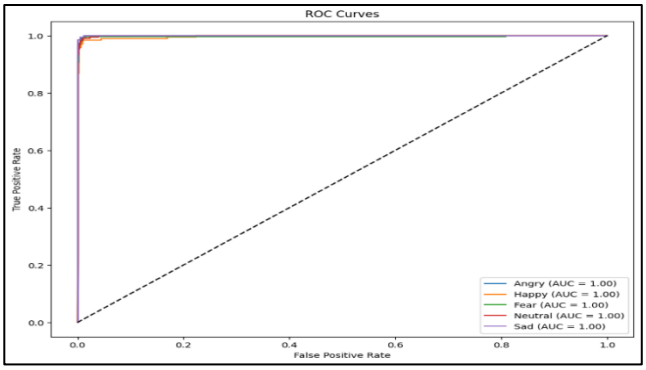

Figure 2. Roc curve of the KEDAS dataset

4. Discussion

The experimental results demonstrate outstanding performance of our emotion classification model on the KEDAS dataset, as evidenced by multiple evaluation metrics:

4.1 Training Dynamics

The training dynamics reveal exceptionally strong model performance, with near-perfect initial accuracy (≈0.9975) that stabilized at 0.98-0.99 by the fifth epoch. The close tracking of validation accuracy to training accuracy indicates minimal overfitting, while the early convergence suggests efficient learning requiring only minor optimization in later stages.

4.2 Loss Characteristics

Both training and validation loss curves demonstrate effective error minimization, converging to low values between 0.12-0.22 by epoch 5 while maintaining stable, parallel trajectories.

4.3 Classification Metrics

Classification metrics confirm outstanding performance across all emotion categories, achieving 98% overall accuracy with consistently high scores (0.97-0.99) for most metrics. The "Happy" class showed slightly lower but still excellent performance (0.97 precision/recall/F1-score). Error analysis of the confusion matrix reveals dominant diagonal values (0.96-0.99 true positives) across all classes, with primary confusion cases including: "Happy" being misclassified as "Fear" (3%) and "Neutral" (1%); "Sad" showing minimal confusion with "Angry" (1%), "Neutral" (1%), and "Happy" (1%); and "Neutral" displaying slight overlap with "Angry" (1%), "Fear" (1%), and "Sad" (1%). These minor classification errors likely stem from acoustic similarities between low-intensity "Happy" and "Neutral" expressions, as well as potential overlaps in arousal levels between certain emotion pairs.

The model demonstrates excellent generalization capability through balanced precision/recall across classes, robust learning evidenced by high accuracy with minimal overfitting, and strong potential for real-world affective computing applications.

Future improvements could focus on targeted fine-tuning for the "Happy" class, detailed analysis of edge-case misclassifications, and investigation of feature-space overlaps between frequently confused emotions.

5. Conclusion

This work introduces a hybrid deep learning framework for recognizing emotions in Arabic speech by combining acoustic features and Mel-Spectrogram representations. By integrating MLP and CNN architectures, the model successfully captures both high-level abstract patterns and fine-grained temporal information relevant to emotional expression. The approach proves particularly effective in addressing the linguistic and acoustic variability inherent in Arabic, resulting in improved classification performance across five emotion classes. Furthermore, the use of adaptive training strategies, such as early stopping and learning rate reduction, enhances model generalization and stability. Overall, the results demonstrate the viability of hybrid models for Arabic SER and open avenues for future work, including the integration of dialect-specific features, data augmentation strategies, and multimodal fusion with visual or textual cues to further advance emotion recognition in Arabic-speaking context.

References

[1] R. Yahia Cherif, A. Moussaoui, N. Frahta, and M. Berrimi, “Effective speech emotion recognition using deep learning approaches for Algerian dialect,” in Proc. Int. Conf. Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, Mar. 2021. doi: 10.1109/WIDSTAIF52235.2021.9430224.

[2] O. A. Mohammad and M. Elhadef, “Arabic speech emotion recognition method based on LPC and PPSD,” in Proc. 2nd Int. Conf. Computation, Automation and Knowledge Management (ICCAKM), Dubai, UAE, Jan.

2021, pp. 31–36. doi:10.1109/ICCAKM50778.2021.9357769.

[3] S. Ouali and S. El Garouani, "Arabic speech emotion recognition using convolutional neural networks," J. Electr. Syst., vol. 20, no. 7, pp. 2649–2657, 2024.

[4] A. Kaloub and E. A. Elgabar, "Speech-based techniques for emotion detection in natural Arabic audio files," Int. Arab J. Inf. Technol., vol. 22, no. 1, pp. 139–157, Jan. 2025. doi: 10.34028/iajit/22/1/11.

[5] M. Belhadj, I. Bendellali, and E. Lakhdari, "KEDAS: A validated Arabic speech emotion dataset," in Proc. Int. Symp. Innovative Informatics of Biskra (ISNIB), Biskra, Algeria, 2022. doi: 10.1109/ISNIB57382.2022.10075694.

[6] B. Mourad, E. Lakhdari, and I. Bendellali, "Kasdi-Merbah (University) emotional database in Arabic speech," Dec. 2023. doi: 10.35111/qqer-qz15.

[7] "openSMILE," Accessed: Apr. 4, 2025. [Online]. Available: https://audeering.github.io/opensmile/get-started.html.

[8] M. E. Elalami, S. M. K. Tobar, S. M. Khater, and E. A. Esmaeil, "Texture feature and Mel-spectrogram analysis for music sound classification," Int. J. Adv. Comput. Sci. Appl., 2024. [Online]. Available: https://www.ijacsa.thesai.org.

How to cite this article

D. Djeridi, B. Said, and M. Belhadj, "A hybrid deep learning approach for Arabic speech emotion recognition using acoustic and spectrogram features," *CyberSystem J.*, vol. 2, no. 1, pp. 71-75, 2025. doi: 10.57238/csj.2025.1008

Access this article online