University of Zawia Journal of Engineering Sciences and Technology (UZJEST)
Volume 3, Issue 2, (2025), pp 205-215, ISSN 3006-2764

Speech Emotion Feature Extraction and Classification

Hamed M Suliman [1]

[1] Computer Technology Department, High Institute of Science and Technology, Misrata, Libya
*Corresponding author email: hamedmmsuliman@gmail.com

Received: 20-06-2025 | Accepted: 06-11-2025 | Available online: 15-12-2025 | DOI:10.26629/uzjest.2025.17

ABSTRACT

Emotion speech recognition has become a prominent research area in recent years due to its extensive applications across various industries. A key challenge in this field is extracting emotional states from the audio signal waveforms. Emotions serve as a way for individuals to communicate their moods or mental states to others. People experience a range of emotions, including sadness, joy, neutrality, disgust, anger, surprise, fear, and calmness. This paper focuses on extracting emotional features from audio waveforms. We have restructured an existing state-of-the-art multilayer perceptron (MLP) design to achieve this objective. The design features a hidden layer with 300 units, utilizes a batch size of 256, and restricts training to 500 epochs. The accuracy of the MLP classifier for speech emotion recognition reaches 95.65% when tested with a limited dataset composed of RAVDESS songs, which is significantly higher than the results reported in other studies in this field.

Keywords: emotion recognition, Multilayer perceptron (MLP), feature extraction.

استخراج وتصنيف السمات الصوتية للعواطف

حمد م سليمان[1]

[1]قسم تقنيات الحاسوب، المعهد العالي للعلوم والتقنية، مصراتة، ليبيا

ملخـــــص البحـــــث

أصبح التعرف على الكلام العاطفي مجالًا بحثيًا بارزًا في السنوات الأخيرة نظرًا لتطبيقاته الواسعة في مختلف الصناعات. يتمثل أحد التحديات الرئيسية في هذا المجال في استخراج الحالات العاطفية من أشكال موجات الإشارة الصوتية. تعمل المشاعر كوسيلة للأفراد للتعبير عن مزاجهم أو حالاتهم العقلية للآخرين. يختبر الناس مجموعة من المشاعر، بما في ذلك الحزن والفرح والحياد والاشمئزاز والغضب والمفاجأة والخوف والهدوء. تركز هذه الورقة على استخراج السمات العاطفية من أشكال الموجات الصوتية. لقد أعدنا هيكلة تصميم متطور متعدد الطبقات (MLP) لتحقيق هذا الهدف. يتميز التصميم بطبقة مخفية تحتوي على 300 وحدة، ويستخدم حجم دفعة يبلغ 256، ويقيد التدريب بـ 500 حقبة. تصل دقة مصنف MLP للتعرف على المشاعر الكلامية إلى 95.65٪ عند اختباره بمجموعة بيانات محدودة تتكون من أغاني RAVDESS، وهي أعلى بكثير من النتائج المبلغ عنها في دراسات أخرى في هذا المجال.

الكلمات الدالة: التعرف على العاطفة، مدرك متعدد الطبقات، استخراج السمات.

1-Introduction

Speech is the primary means of human communication and consists of two types of information. The first type is linguistic information, while the second type is paralinguistic information. Linguistic information relates to the meaning of speech, whereas paralinguistic information is associated with the emotion conveyed in speech [1].

Human speech consists of statements built from words. These words are composed of syllables, which carry the emotional content of speech. By analyzing these speech emotions, we can classify a speaker's gender and emotional state [2].

As applied science has advanced significantly, there is a growing demand for more realistic Human-Computer Interaction (HCI). HCI systems must be able to accurately perceive and interpret speech. Therefore, HCI should analyze these emotional cues to extract meaningful information, making the interaction smoother and more contextually aware of the speaker's feelings. Consequently, Speech Emotion Recognition (SER)—the technology capable of determining emotional states from spoken words—has attracted significant attention from experts [2].

Speech recognition emerged between 1930 and 1950, which is known as the first generation period. During this time, speech recognition research focused on studying the connection between the linguistic identity of speech (such as the pronunciation of a syllable or phoneme) and the spectral characterization of that speech, which is known as the acoustic-phonetic representation of sound. Furthermore, during this period, the linkage between speech classes and the signal spectrum was discovered. By the end of this era, many forms of short-time spectral analysis devices were invented, such as filter bank analyses, cepstral analyses, and the mechanism of linear predictive coding (LPC) [3].

In the second generation of speech recognition research (1951-1959), the focus shifted to developing algorithms for quantifying the phonetic identity of speech based on weighted spectral characteristics of sound as a function of time. The main breakthrough during this period was that Denes and Fry from University College London, in 1959, invented a speech recognition device that could recognize a sequence of one of four vowels followed by one of nine consonants. The device included a spectral analyser and a spectral pattern matching algorithm. The accuracy of the system was 44% [3].

The third generation of speech recognition research, which extended from 1960 to 1980, saw significant technological progress in speech processing systems that had a substantial impact on speech recognition. The major advance was achieved by Velichko and Zagoruyko from Russia in the late 1960s. These two Russian scientists designed a speech recognizer based on a pattern recognition system and utilized dynamic programming time alignment approaches [3].

Furthermore, in 1976, the HARPY system developed by Bruce Lowerre of Carnegie Mellon University was introduced. This system utilized a conventional segmentation and labeling method. The main development of this system was the integration of acoustic models with the lexical representation of words[3].
The fourth generation of speech recognition research is characterized by the use of rigorous statistical models that rely on mathematics. The first statistical model was the Hidden Markov Model, while the major achievement of this generation is the artificial neural network (ANN) approach, which assesses class-conditional likelihood densities. The main technological advancement was the invention of the computer architecture known as the parallel distributed processing (PDP) model. This model consists of a dense interconnection of simple neural computational elements and employs a training approach known as error backpropagation. The most famous PDP application for speech recognition is the multilayer perceptron (MLP), as depicted in Figure 1[3].

Over the last decade, many technologies have been used to recognize the emotional state of a speaker from their words. Common approaches depend on extracting acoustic features from spoken words, such as Mel-frequency cepstral coefficients (MFCC), pitch, zero-crossing rate, and shimmer, then using machine learning classifiers such as HMM, GMM, and SVM to classify them. Therefore, Machine learning technologies play a vital role in enabling the advancement of algorithms that detect speech emotion by analysing patterns and features

associated with various emotional states. As a result, human-computer interaction has led to the development of applications such as mental health evaluation, emotion analysis, and improved client service through speech emotion recognition, making it an area of growing interest for experts and researchers [4].

Recently, deep learning algorithms have achieved high accuracies in emotion recognition. Recurrent neural networks (RNNs) and attention mechanisms have achieved strong performance on this task. An attention network is utilized to detect the pairing between speech and text, together with Bi-LSTM network to model the sequence in emotion recognition. Furthermore, to perform classification, it is required to have multi-hop attention to determine specific parts of the textual information and then attend to the audio feature. However, these suggested approaches are not only sophisticated in network structure but also costly according to computations [5].

A key challenge in this field is extracting emotional states from audio signal waveforms. Emotions serve as a means for individuals to communicate their moods or mental states to others. People experience a range of emotions, including sadness, joy, neutrality, disgust, anger, surprise, fear, and calmness. Due to recent advances in powerful deep learning techniques, many algorithms are now capable of accurately extracting speech emotion waveforms and features. After feature extraction, the machine algorithm must analyse these features and discover statistical relationships between specific features and emotional states. This mechanism is known as classification [6].

The most commonly used features in speech emotion recognition are the linear magnitude spectrum, logarithmic magnitude spectrum, Fast Fourier Transform, short-term energy, zero-crossing rate, and the Mel-frequency cepstral coefficient (MFCC). Deep learning algorithms classify these eight speech emotions with varying degrees of accuracy. Speech emotion recognition plays a vital role in advancing human-machine interfaces by enabling the extraction and classification of speech emotion features. These emotional features vary over time and therefore involve vocal variables such as amplitude, frequency, and spectral characteristics. Analyzing these emotion features can potentially lead to better interactions between humans and machines, resulting in more precise identification of speech emotion features [7].

1.1 **Literature Review**

Previous papers in Speech Emotion Recognition (SER) have identified many automated methods and datasets using various CNNs and classification algorithms. Damodar et al. [8] developed a hybrid system consisting of CNNs and a decision tree for audio feature extraction. The performance of the CNNs for emotion recognition was 72%, while the performance of the decision tree was 63%.

The most recent research was conducted by Abeer Alnuaim [9], titled “Human-Computer Interaction for Recognition of Speech Emotions Using a Multilayer Perceptron (MLP) Classifier.” This approach featured more layers than previous methods and used an adaptive learning rate instead of a constant one. Due to these enhancements, the approach achieved a confidence score of 81%.

Previous papers in Speech Emotion Recognition (SER) have identified many automated methods and datasets using various CNNs and classification algorithms. An article by Jeery Joy [10], titled “Speech Emotion Recognition Using Neural Network and MLP Classifier,” utilized CNNs to predict vocal mood, achieving a confidence score of 70.28%.

The gap in Speech Emotion Recognition (SER) is that all previous papers in this field reported low accuracies, with the best recorded accuracy being 81%. However, my MLP classifier achieved an accuracy of 95.65% using the RAVDESS songs dataset, after restructuring an existing state-of-the-art multilayer perceptron (MLP) design to achieve this objective.

2- Proposed Methodology

The hidden emotions in our speech are represented in tone and pitch. In this article, we extracted emotional states such as sadness, joy, neutrality, disgust, anger, surprise, fear, and calmness, and then classified them. Furthermore, convolutional neural networks are used to predict the emotions in speech, and these emotions are subsequently classified using a multilayer perceptron (MLP). The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is used in this paper.

2.1 Dataset

The RAVDESS songs dataset is integral to this study, serving as an audio speech dataset for emotion recognition [11]. It comprises 1440 audio files from 24 actors, each contributing 60 trials. The dataset features an equal distribution of male and female actors and encompasses eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised.
The dataset consists of two statements in North American accents: "Kids are talking by the door" and "Dogs are sitting by the door." In addition, each file name contains a seven-part numerical identifier to clarify the stimulus characteristics.

2.2 The Model Architecture

MLP stands for multi-layer perceptron, which is an essential neural network architecture commonly used for classification tasks. It consists of three main layers: the input layer, the hidden layer, and the output layer. The main task of the input layer is to accept the incoming audio waveforms, while the role of the output layer is to calculate the accuracy of the MLP classifier.
In the designed system, at least three extracted features of speech emotions—Mel-frequency cepstral coefficients (MFCC), features of the 12 pitch classes' chroma, and Mel spectrogram frequencies—are selected by the input layer. The major task of the hidden layer is to perform activation functions on the incoming information. Lastly, the output layer generates the output by executing classification to assign the predicted emotion. The composition of this multilayer perceptron is depicted in Figure 1.

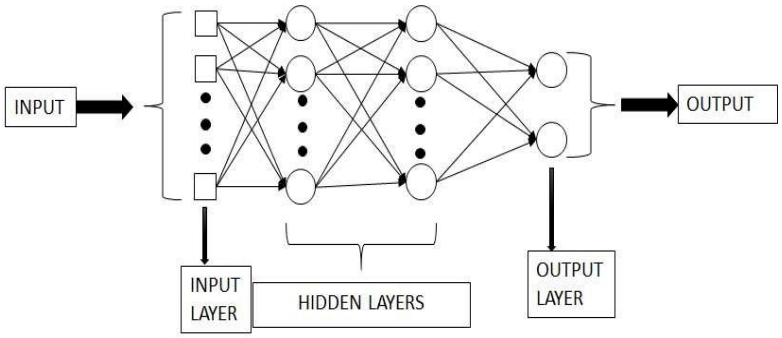


Figure1. Multi layered Perceptron [10]

This paper classified eight speech emotions—neutral, calm, happy, sad, angry, fearful, disgusted, and surprised—with an accuracy of 95.65% on the RAVDESS songs dataset. We restructured an existing state-of-the-art multilayer perceptron (MLP) design to achieve this objective. The design features a hidden layer with 300 units, utilizes a batch size of 256, and restricts training to 500 epochs. As a result, we were able to determine the speech emotion state and classify it as belonging to one of the following emotion states: sadness, joy, neutrality, disgust, anger, surprise, fear, or calmness. In addition, we plotted the loss, accuracy, and confusion matrix graphs for the MLP classifier.
The main accomplishments of this article are:

1. The utilized convolutional neural networks replace constant learning rates with an adaptive learning rate; therefore, the model achieves higher accuracy.
2. At least three features are computed by the MLP classifier: the Mel-frequency cepstral coefficient (MFCC), Mel spectrogram frequencies, and features about the 12 pitch classes (chroma).
3. The MLP classifier successfully determined eight speech emotions—neutral, calm, happy, sad, angry, fearful, disgusted, and surprised—with an accuracy of 95.65% on test datasets.
4. The MLP classifier used a limited dataset composed of RAVDESS songs; hence, data augmentation mechanisms were not needed.
5. The MLP classifier is a simple model in design and therefore requires little training time.

3-Results and Discussion

We utilized Python programming to efficiently and precisely extract speech features. By employing libraries like NumPy, SciPy, and Matplotlib, we were able to accurately plot acoustic speech signal graphics with clarity and visually appealing results, alleviating the burden of excessive manual work. In this process, we plotted, analyzed, and computed several speech features: the linear magnitude spectrum, logarithm magnitude spectrum, Fast Fourier Transform, short-term energy, zero-crossing rate, and the Mel-frequency cepstral coefficient (MFCC) for many speech emotions. Figures 2, 3, 4, 5, 6, 7, 8 and 9 illustrate these extracted features.

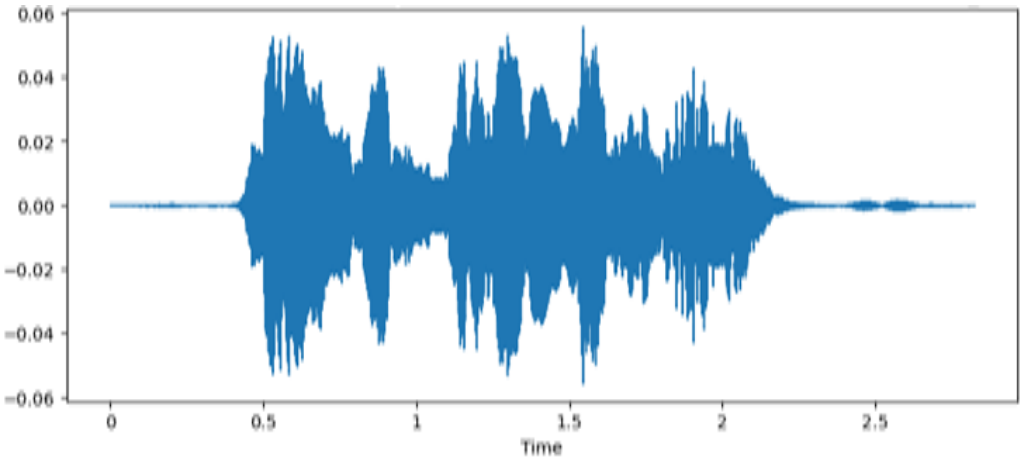


Figure 2. Audio sample of angry emotion

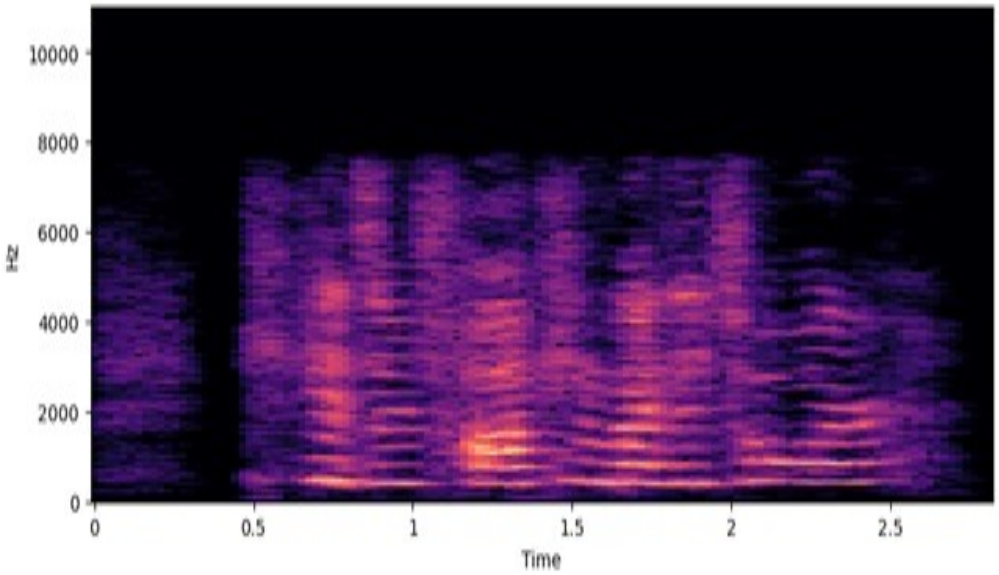


Figure 3. Spectrogram for an audio with angry emotion

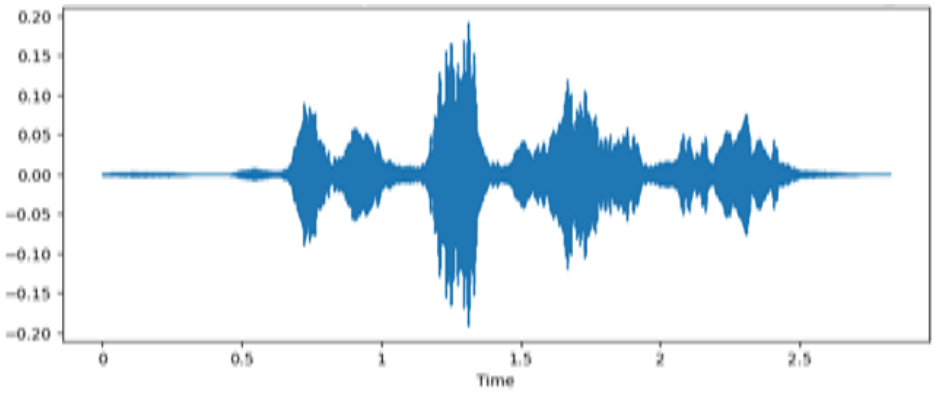


Figure 4. Audio sample of fear emotion

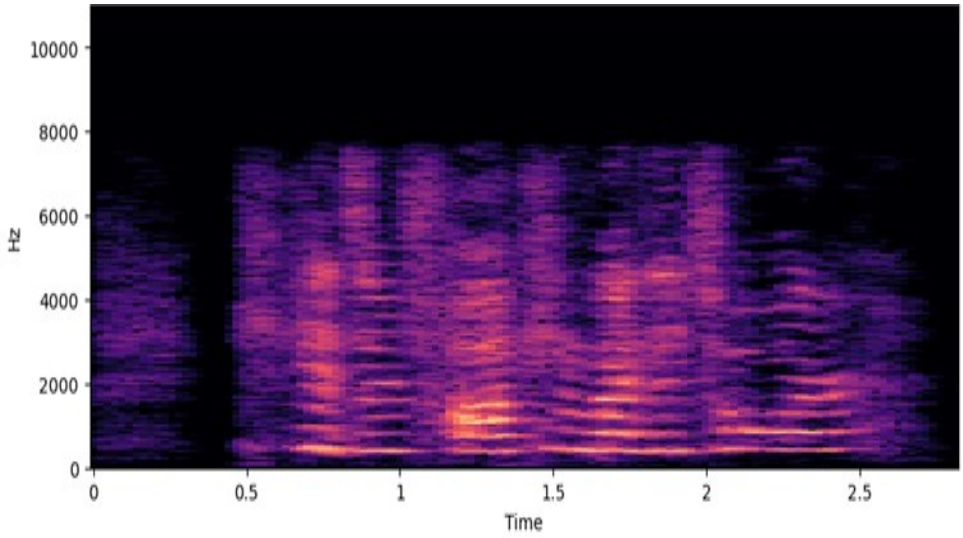


Figure 5. Spectrogram for an audio with fear emotion

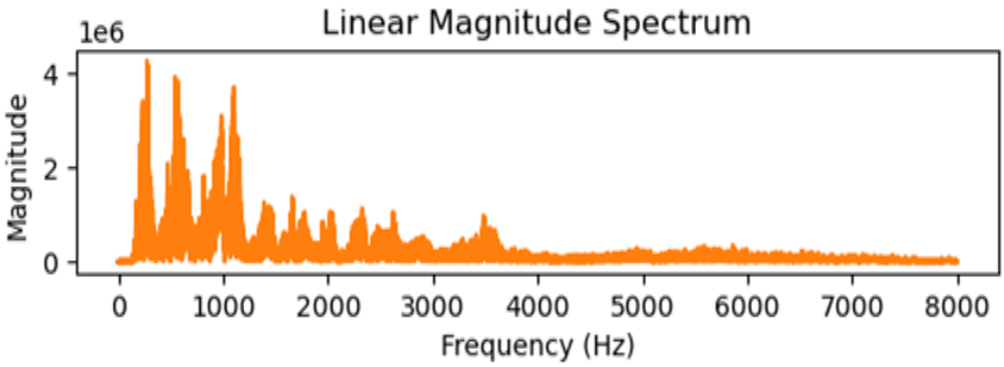


Figure 6. Linear magnitude spectrum for angry emotion

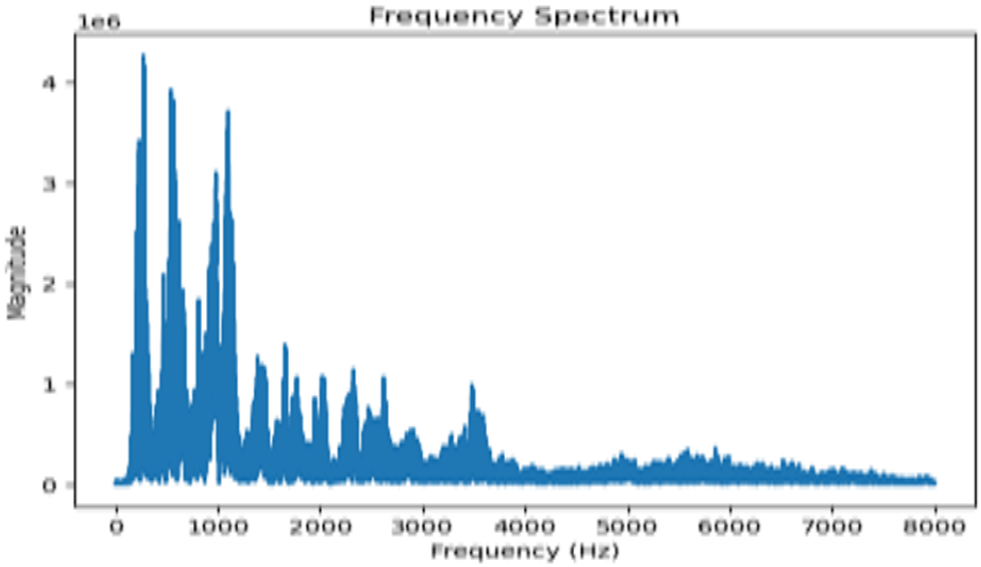


Figure 7. FFT frequency spectrum for angry emotion

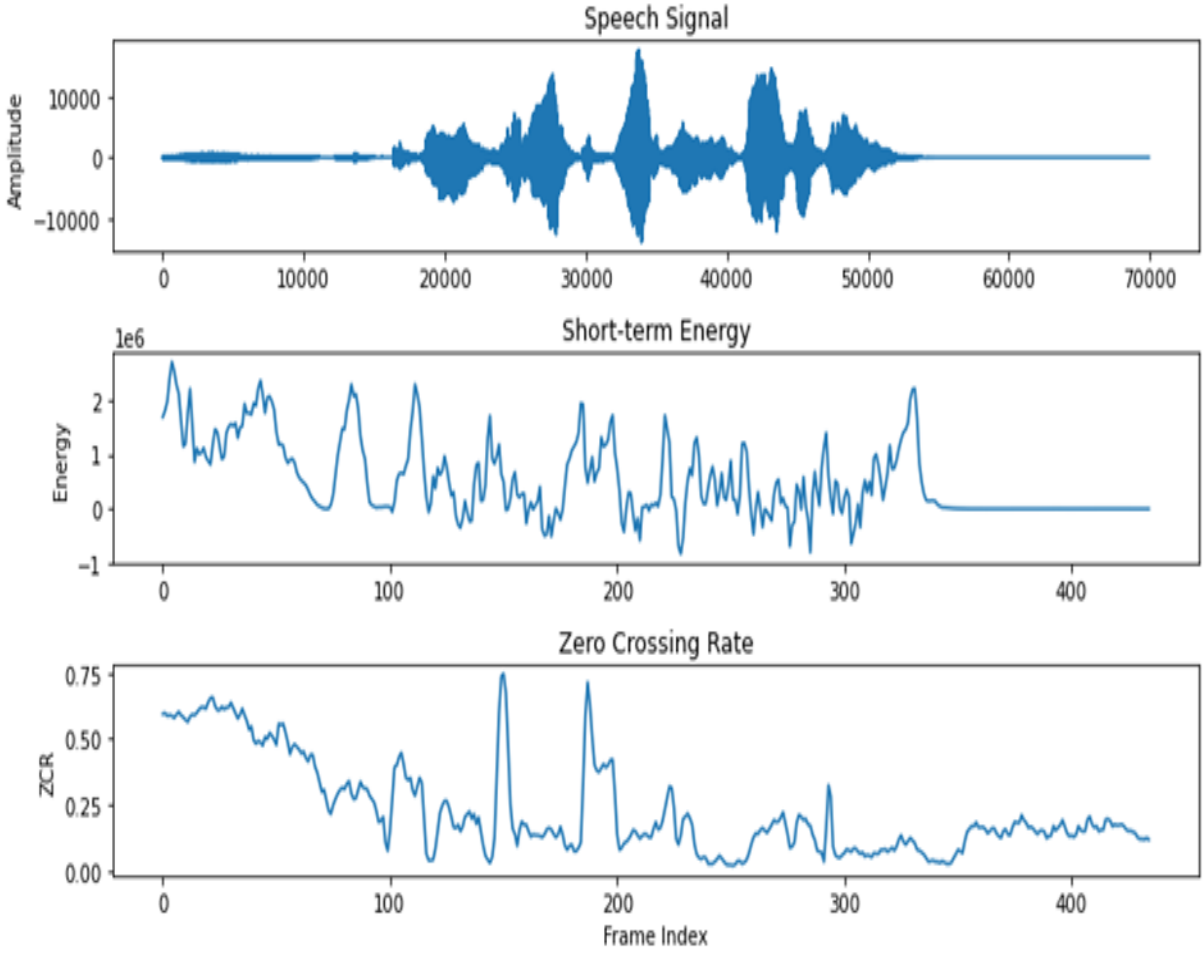


Figure 8. Short-term Energy and Zero Crossing Rate for angry emotion

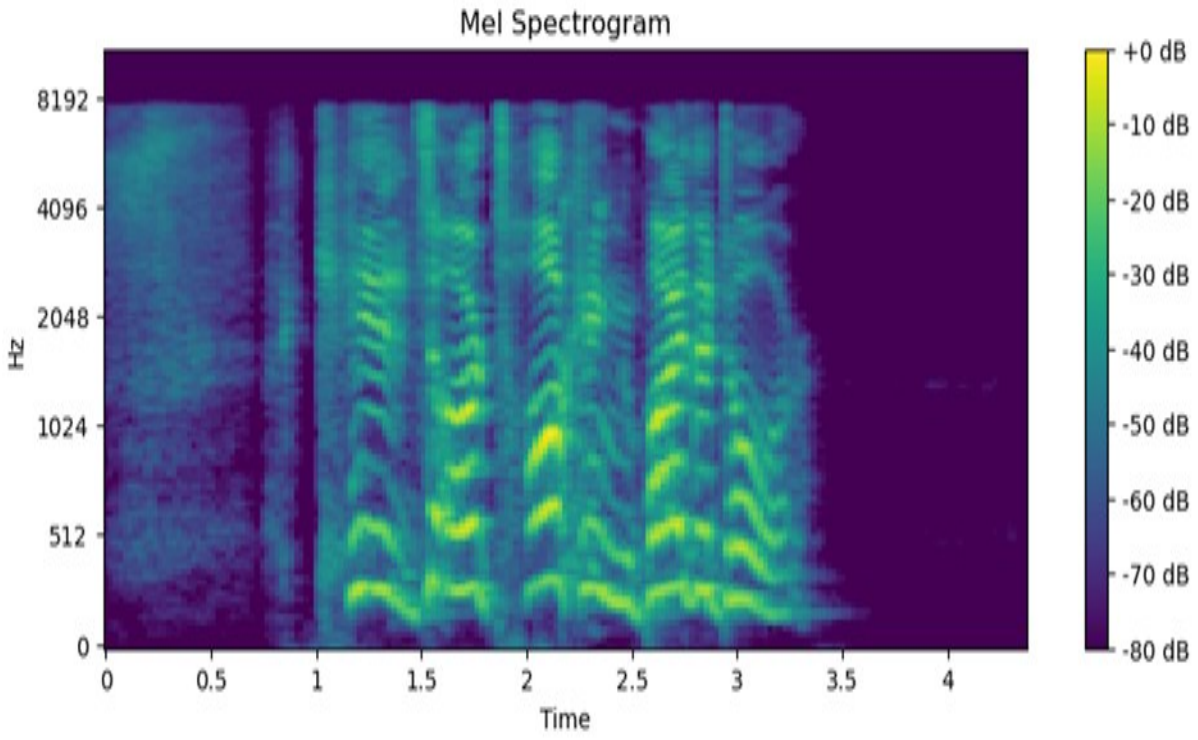


Figure 9. MFCCs for angry emotion

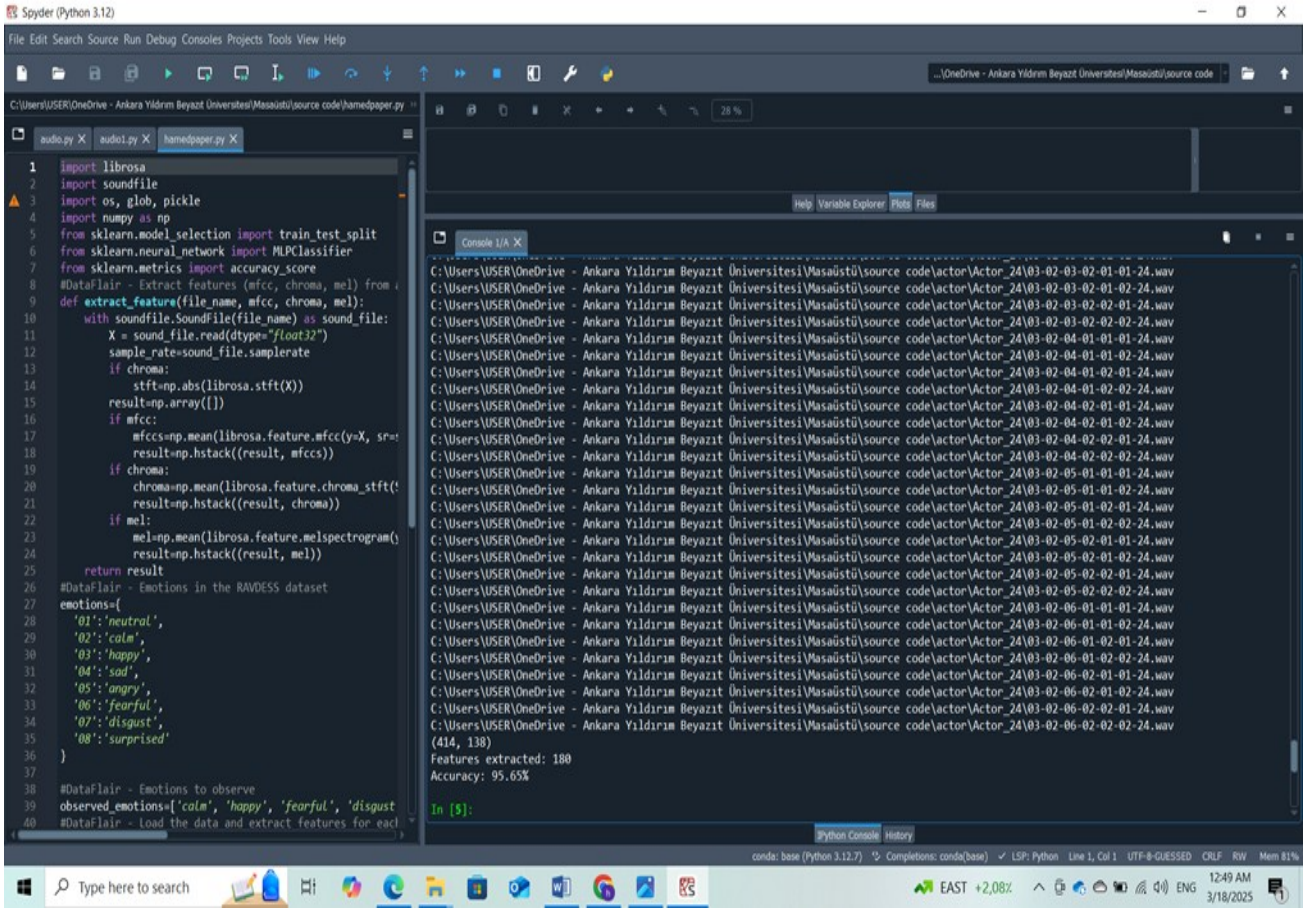
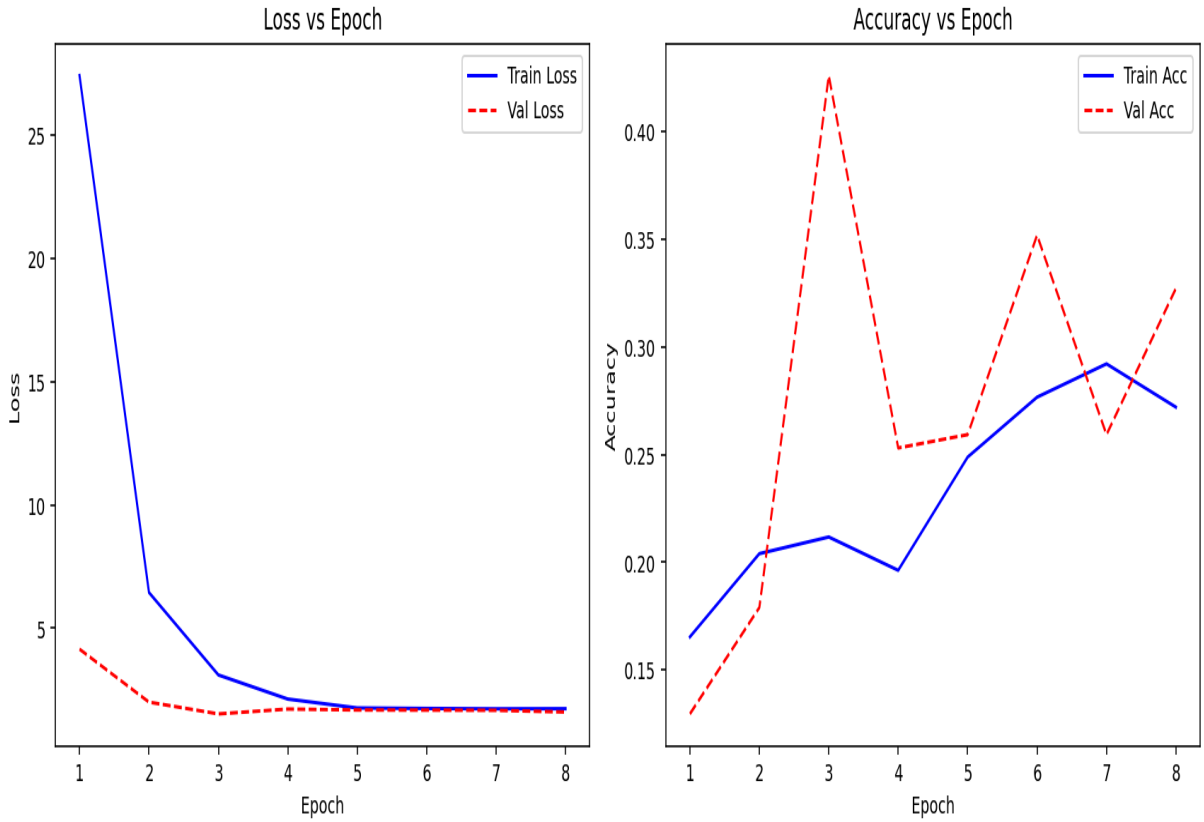


Figure 10. MLP classifier accuracy

According to Figure 10, the MLP classifier achieved a significant confidence level of 95.65% using a restricted RAVDESS songs dataset. At least three extracted features of speech emotions—Mel-frequency cepstral coefficients (MFCC), features of the 12 pitch classes' chroma, and Mel spectrogram frequencies—are used by the MLP classifier. Together, these minimum features enabled the MLP classifier to make accurate speech emotion prediction.

The RAVDESS songs dataset, used for training the MLP classifier, includes vocal patterns from twenty-four experienced actors with North American accents, representing eight distinct emotions. Furthermore, the RAVDESS dataset was partitioned into a training set and a test set. The training set comprised three-fourths of the entire dataset, while the test set consisted of the remaining one-fourth. The design features a hidden layer with 300 units, utilizes a batch size of 256, and limits training to 500 epochs. Additionally, using an adaptive learning rate instead of a constant one resulted in a better learning mechanism. The MLP classifier was able to extract 180 features from a dataset of shape (414, 138). Using larger datasets than the RAVDESS songs dataset could potentially lead to even greater confidence in the MLP classifier's predictions. Furthermore, the MLP classifier achieved very low training and validation loss but obtained low accuracy because it was difficult to differentiate between the eight speech emotions. The confusion matrix also revealed high confidence for the angry speech emotion (97%) and considerable confidence for the sad speech emotion (68%), while the model achieved low confidence for the remaining six speech emotions. Figure 11 shows loss, accuracy and confusion matrix of MLP classifier.

The main challenge for the MLP classifier is classifying eight different speech emotions. Nevertheless, the MLP classifier achieved a confidence level of 95.65% in classifying eight diverse speech emotions, which is among the highest reported in this field.



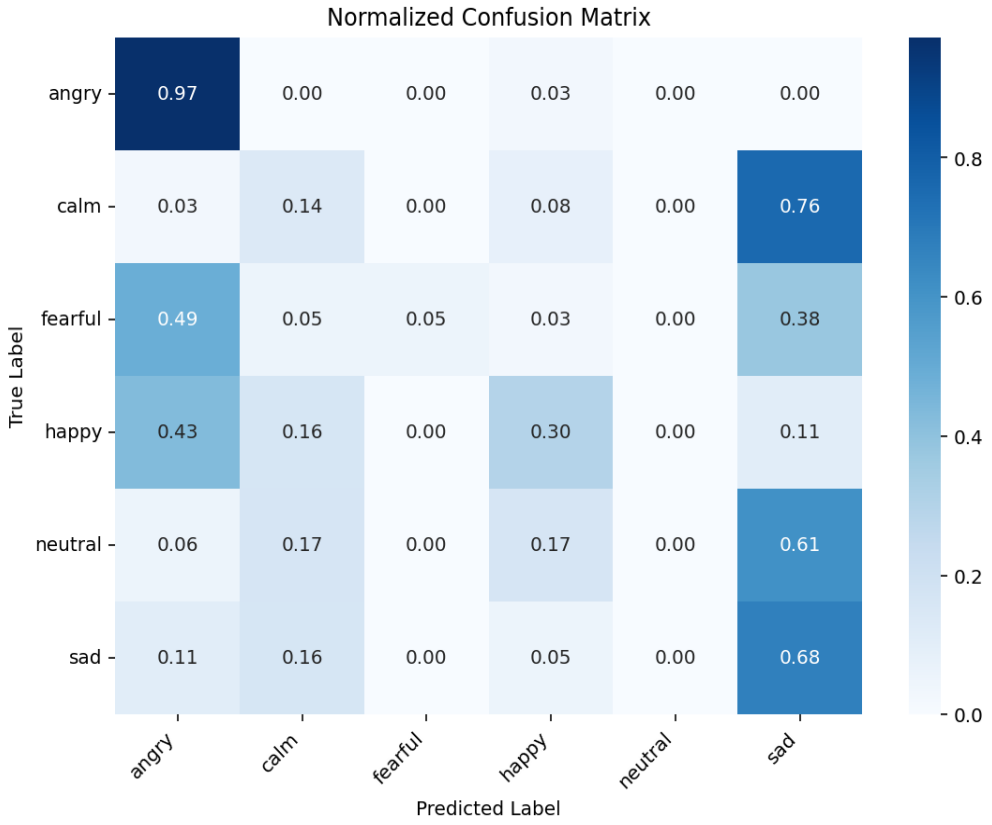


Figure 11. Loss, accuracy and confusion matrix for MLP classifier

4- Conclusions

The literature on Speech Emotion Recognition (SER) classification faces many challenges due to the limited availability of datasets. As a result, previous SER studies reported low accuracies, with the best recorded accuracy being 81%.

This paper classified eight speech emotions—neutral, calm, happy, sad, angry, fearful, disgusted, and surprised—with an accuracy of 95.65% on the RAVDESS songs dataset. We restructured an existing state-of-the-art multilayer perceptron (MLP) design to achieve this objective. The design features a hidden layer with 300 units, utilizes a batch size of 256, and restricts training to 500 epochs.

The findings of this paper show that the MLP classifier achieved a confidence level of 95.65% in classifying eight distinct speech emotions, which is among the highest reported in this field. Furthermore, at least three extracted features of speech emotions—Mel-frequency cepstral coefficients (MFCC), chroma features of the 12 pitch classes, and Mel spectrogram frequencies—are used by the MLP classifier. Together, these features enabled the MLP classifier to make highly accurate predictions of speech emotions.

In this study, we utilized a limited dataset because the RAVDESS songs dataset is currently the only available resource for speech emotion analysis. Therefore, future research should incorporate larger datasets and explore various classifiers, such as Support Vector Machines, to potentially enhance the effectiveness and accuracy of emotion classification.

REFERENCES

[1] Zhaoa, J. (2018). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035

[2] Pandey, S. (2019). *Deep Learning Techniques for Speech Emotion Recognition: A Review*. https://doi.org/DOI:%252010.1109/RADIOELEK.2019.8733432

[3] Benesty, J. (2008). *Handbook of Speech Processing*. Springer-Verlag Berlin Heidelberg.

[4] ONG, K. L. (2024). MaxMViT-MLP: Multiaxis and Multiscale Vision Transformers Fusion Network for Speech Emotion Recognition. *IEEE ACCESS*, *12, 2024*.

https://doi.org/0.1109/ACCESS.2024.3360483
[5] Peng, Z. (2021). Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. *IEEE International Conference on Acoustics*. https://ieeexplore.ieee.org/abstract/document/9414286
[6] Lugović, S. (2016). Techniques and Applications of Emotion Recognition in Speech. *39th International Convention on Information*. https://ieeexplore.ieee.org/abstract/document/7522336
[7] Palo, H. (2017). Emotion recognition using MLP and GMM for Oriya language. *Computational Vision and Robotics*, *7*(No. 4), 426–442.
[8] Damodar, N. (2019). Voice Emotion Recognition using CNN and Decision Tree. *International Journal of Innovative Technology and Exploring Engineering* (*IJITEE*).
[9] Alnuaim, A. (2022). Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. *Hindawi*, *2022* (ID 6005446), 12 pages. https://doi.org/://doi.org/10.1155/2022/6005446
[10] Jerry, J. (2020). Speech Emotion Recognition using Neural Network and MLP Classifier. *IJESC*, 10 (No. 4), 25170–25172.
[11] Poojary, N. (2021). Speech Emotion Recognition Using MLP Classifier. *IJSRCSEIT*, *7*(4), 218–222. https://doi.org/10.32628/CSEIT217446