Vocal-Emotion Visualizer: Voice-Based Emotion Detection & Image Generation

Veluvarthi Lasya Mounika, Dasari Yaswanth Sri Balachandra, Penumatcha Satya Sai Shivani, Jami Anjali Devi, P. Srinivasa Rao

MVGR College of Engineering, Vizianagaram, India

DOI: https://doi.org/10.51584/IJRIAS.2026.11060277

Received: 03 July 2026; Accepted: 08 July 2026; Published: 17 July 2026

ABSTRACT

Human feelings are essential in how people communicate and engage with computer systems. Spotting feelings from the manner human beings communicate has become a key area of study in synthetic intelligence. This Study introduces a machine named Vocal Emotion Visualizer, which identifies human emotions from speech, the usage of deep mastering and turns them into clean, visible paperwork. The gadget uses the Librosa library and Mel-Frequency Cepstral Coefficients (MFCC) to extract sound capabilities from recorded audio, and Wav2Vec 2.0 is used for live speech to obtain specified speech facts. These capabilities are then handled through a long short-term memory (LSTM) network to categorise feelings. The emotions discovered are changed into descriptive words to create a visible picture based on emotions, the usage of AI that generates snapshots. This machine facilitates people to apprehend emotions better in voice-based generation and makes interacting with computers greater herbal. Testing suggests that the device can correctly discover numerous feelings like happy, unhappy, angry, calm, and neutral.

Keywords: Speech Emotion Recognition, Long Short-Term Memory (LSTM), Wav2Vec 2.0, MFCC, Audio Signal Processing, Emotion Classification, Real-Time Emotion Detection.

INTRODUCTION

Human feelings are an important issue of powerful communication and human–computer interaction. Speech is one of the most natural approaches for conveying people's specific emotions, conveying not only the simplest linguistic statistics but also emotional cues, including tone, pitch, and depth. With the fast growth of voice-based technologies, which include digital assistants, client help structures, and intelligent communication structures, there's an increasing need for structures that can apprehend both the content material and emotional context of speech [1][2].

Speech Emotion Recognition (SER) is an emerging area in synthetic intelligence that specialises in identifying emotional states from speech indicators. Traditional systems focus attention on speech-to-text conversion and often forget about emotional data, which limits their capability to respond in a human-like manner. However, speech signals contain rich acoustic features, which include pitch, energy, and spectral traits, which may be analysed the use of superior gadget studying and deep learning techniques to discover feelings as it should be [3][4].

Latest improvements in deep learning have considerably stepped forward the performance of SER structures. models inclusive of lengthy, quick-term memory (LSTM) networks are effective in capturing temporal dependencies in speech facts, making them appropriate for studying recorded audio [5][6]. In evaluation, transformer-based models like Wav2Vec2 can research effective representations at once from raw audio indicators, allowing green real-time emotion recognition [7][8].

In this study, we endorse a gadget called the Vocal Emotion Visualizer, which integrates multiple deep learning strategies to detect

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

emotions from both uploaded and live audio inputs. The system utilises an LSTM model for processing pre-recorded audio files and a Wav2Vec2 model for real-time speech analysis. After detecting the emotion, the machine generates descriptive activities that might be used to create visible representations of the emotional context.

The proposed gadget aims to enhance emotional know-how by combining speech emotion popularity with AI-based visualisation. Unlike conventional structures that simply output emotion labels, this technique offers an interactive and expressive representation of feelings through generated pictures. This may enhance a person's enjoyment and has potential applications in regions together with digital assistants, emotional analytics, intellectual health monitoring, and intelligent human–computer interaction systems [9][10].

LITERATURE SURVEY

In recent years, extensive research has been conducted inside the discipline of Speech Emotion Recognition (SER), the use of artificial intelligence and deep getting to know techniques. Speech alerts contain emotional cues that may be analysed to determine the usage of the device, gaining knowledge of algorithms to perceive human emotions. Researchers have explored numerous tactics along with audio characteristic extraction, deep studying fashions, and emotion-based applications in human–computer interplay [11][12]. This research provides the foundation for developing the proposed Vocal Emotion Visualizer gadget.

Speech Emotion Recognition Studies

Numerous researchers have centred on detecting emotions from speech alerts the usage of the device, getting to know and using deep learning techniques [13][14].

The RAVDESS dataset, which includes emotional speech recordings and is extensively used for training and comparing speech emotion recognition models [15].

A deep neural community model for speech emotion recognition and demonstrated that deep gaining knowledge of strategies can successfully seize complicated emotional styles in speech indicators [16][17].

In addition, evolved a convolutional neural network (CNN) version that improves emotion type performance through extracting important features from speech indicators [18][19].

These studies highlight that deep learning techniques notably enhance the accuracy of speech emotion recognition systems as compared to traditional techniques [20][21].

Feature Extraction Techniques

Feature Extraction plays a vital function in speech emotion recognition.

The usage of audio sign processing strategies and emphasised capabilities along with Mel Frequency Cepstral Coefficients (MFCC), pitch, spectral evaluation, and energy for identifying emotional styles in speech [22][23].

Libraries that include Librosa are normally used for extracting those acoustic features from audio alerts. Combining multiple functions enables enhancing the performance of emotion popularity models by taking pictures. Both spectral and temporal characteristics of speech [24][25].

Deep gaining knowledge of models for the Emotion category

Latest studies have targeted using deep learning fashions which include lengthy brief-term memory (LSTM) networks and advanced pretrained models like Wav2Vec2 for speech emotion recognition.

LSTM networks are effective in capturing temporal dependencies in sequential speech facts, permitting the model to understand how emotions evolve through the years [26][27].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Wav2Vec2, a transformer-based speech illustration version, is able to extract high-stage capabilities directly from uncooked audio alerts, lowering the need for manual function engineering [28][29].

Frameworks, inclusive of TensorFlow and Keras, have enabled efficient implementation of these models for real-world emotion popularity structures.

CNN fashions are powerful in analysing spectrogram representations of audio indicators, at the same time as LSTM networks are able to gain knowledge of temporal relationships in sequential speech information. Frameworks, inclusive of TensorFlow and Keras, have enabled the green implementation of these fashions for real-world emotion popularity systems. There are some barriers to current structures: even though many studies have proposed powerful speech emotion popularity structures, numerous boundaries remain. A few strategies depend on conventional machine getting to know strategies that require manual feature engineering. Different systems awareness is best in the emotion class without imparting visual interpretation of emotions.

To address those obstacles, the proposed Vocal Emotion Visualizer integrates deep learning-based emotion detection with AI-based visualisation strategies [30][31]. The gadget analyses speech alerts, predicts feelings, and generates visible representations of emotional states, permitting more interactive and expressive human–laptop interplay.

METHODOLOGY

This model uses deep learning to detect the emotions in spoken speech, and then represents them graphically. Overall, this process consists of collecting data, preprocessing it, extracting relevant data, classifying emotional speech, and representing the emotion in images. This solution works with both prerecorded clips and real-time data streams.

As a source, we select three datasets—the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), Toronto Emotional Speech Set (TESS), and CREMA-D. All the clips in these sets were produced by professional actors, who exhibit various emotions, including happiness, sadness, anger, neutral, fear, and calm. Every audio file has an associated emotion tag.

Before starting to work, it is necessary to perform some pre-processing to standardize the data. Normalizing audio clips and resampling them to ensure all files have the same sampling frequency allows us to eliminate unusual files. Moreover, filtering noise from the original clips is required to emphasize unique voice characteristics.

The core of our neural network incorporates LSTM and Wav2Vec2. This combination gives us the opportunity to work with both uploaded clips and live-streaming audio data. Using multiple datasets allows us to develop an emotion classifier that is not sensitive to particular tones or accents.

While using uploaded audio data, we extract MFCC (Mel Frequency Cepstral Coefficients), chroma features, and spectral contrast with the help of librosa library. These features characterize the voice of a speaker and will be used further to classify emotional data.

To distinguish between different kinds of emotions in the dataset, we feed our extracted features to an LSTM (long short-term memory) model. The essence of this approach lies in analyzing speech dynamics over time. Thus, this technique allows us to classify clips in terms of emotions, such as happiness, sadness, or anger.

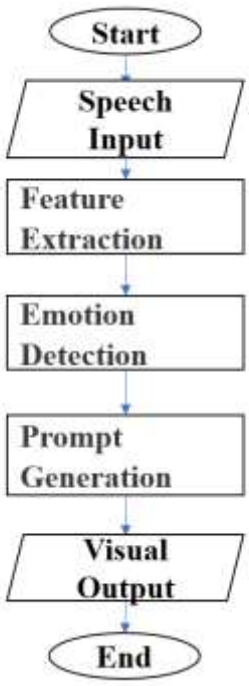
In case of streaming audio data, we rely on the transformer model known as Wav2Vec2. This type of architecture is able to analyze unprocessed audio data without using separate features extraction. This approach increases precision of classification in conversations.

Data is divided into two parts—80% for training and 20% for testing. Such metrics as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of emotion classification.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Finally, an identified emotion will be represented through a graphic image. An important part is that this visualization is obtained via generating text prompts that correspond to the detected emotion.

Flow Diagram:



Flow of the Proposed System

The general process flow of the system includes the following procedures:

1. Speech Audio Input (Upload / Live Recording)

2. Audio Pre-processing

3. Feature Extraction (Librosa – for uploaded audio)

4. Emotion Classification (LSTM for uploaded audio / Wav2Vec2 for live audio)

5. Emotion Prediction

6. Prompt Creation

7. Visual Output Generation

For a novel speech input to be presented, the trained model analyzes the audio components and predicts the probability emotion. The predicted emotion is shown together with its associated visible representation.

The above method proves to be an effective and intelligent approach to the visualization of speech emotions. The process helps to improve the communication relationship between individuals and AI systems.

Environmental Setup

The suggested Vocal Emotion Visualizer device was created using the programming language Python and artificial intelligence techniques that can detect emotions in speech data. The machine uses libraries such as

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

NumPy and Pandas for data analysis and Librosa for analyzing audio features including Mel Frequency Cepstral Coefficients, Chroma Features, and Spectral Features.

The emotion classification version is carried out using TensorFlow and Keras, where an extended brief-term reminiscence (LSTM) neural network is used to analyse speech alerts and become aware of emotions. Improvement and checking out are performed using equipment, including Visual Studio Code or a Jupyter notebook.

The system runs on a pc with an Intel Core i5 processor, 8 GB RAM, and enough storage for dealing with the dataset and training the model. These surroundings support the green processing of audio facts and implementation of the speech emotion reputation system.

RESULTS

The proposed Vocal Emotion Visualizer system was applied and tested on the use of the RAVDESS speech emotion dataset. After preprocessing the audio signals and extracting features using the Librosa library, the extracted features were used to educate the LSTM deep learning model, gaining knowledge of the version for emotion classification.

The trained model was evaluated the usage of overall performance metrics along with accuracy, precision, don't forget, and F1-score. The results showed that the model became capable of successfully classifying feelings such as satisfied, unhappy, irritated, calm, and neutral from speech indicators.

System Implementation Effects

The educated model becomes incorporated into a web-based software interface for real-time emotion detection.

The educated model was included in an internet-based interface, where customers can interact with the machine in various ways:

- Add Audio: users can add a pre-recorded speech, and the system analyzes the audio to detect the emotional context.
- Stay Audio Recording: Customers can file speech directly through the microphone, and the system processes the recorded audio in real time to detect the emotions.

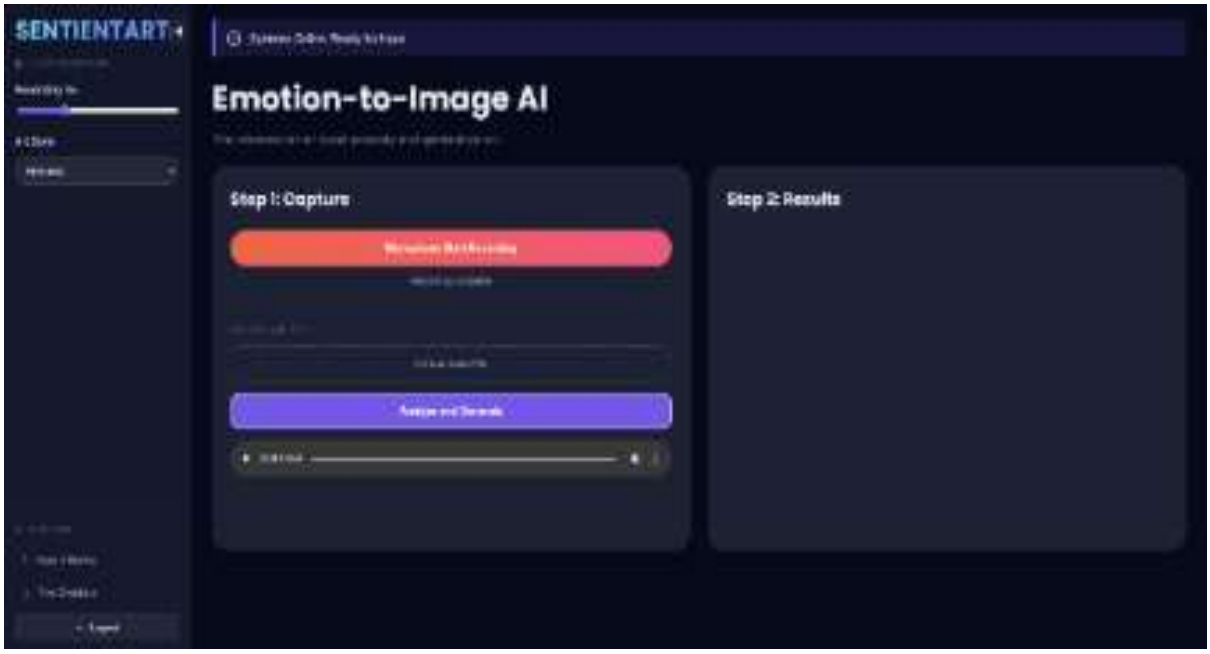


Figure 5.1: Web interface for speech emotion detection.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

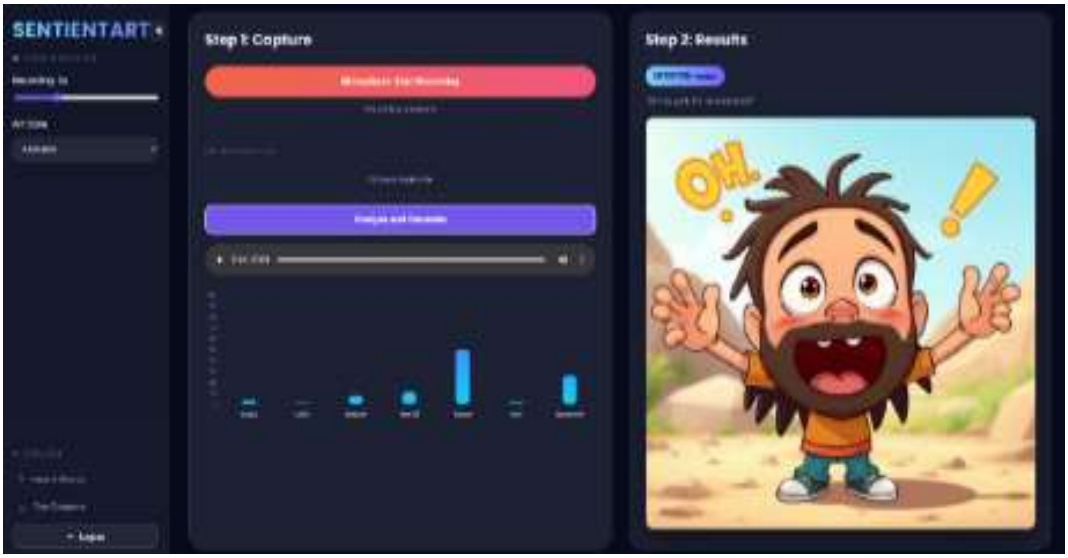


Figure 5.2: Emotion prediction result displayed on the website.

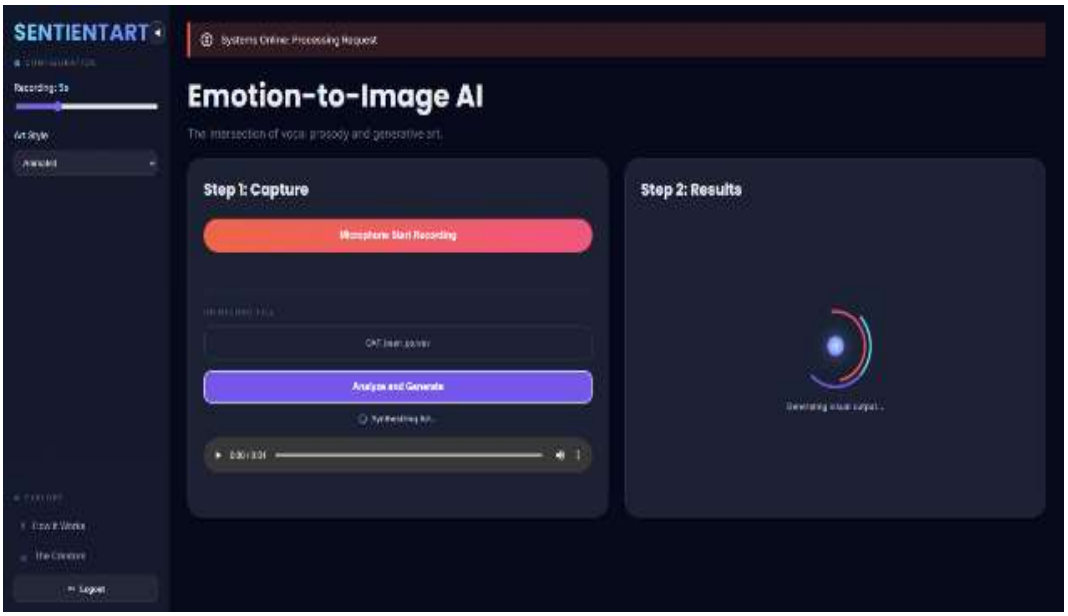


Figure 5.3: Audio Uploaded for detection

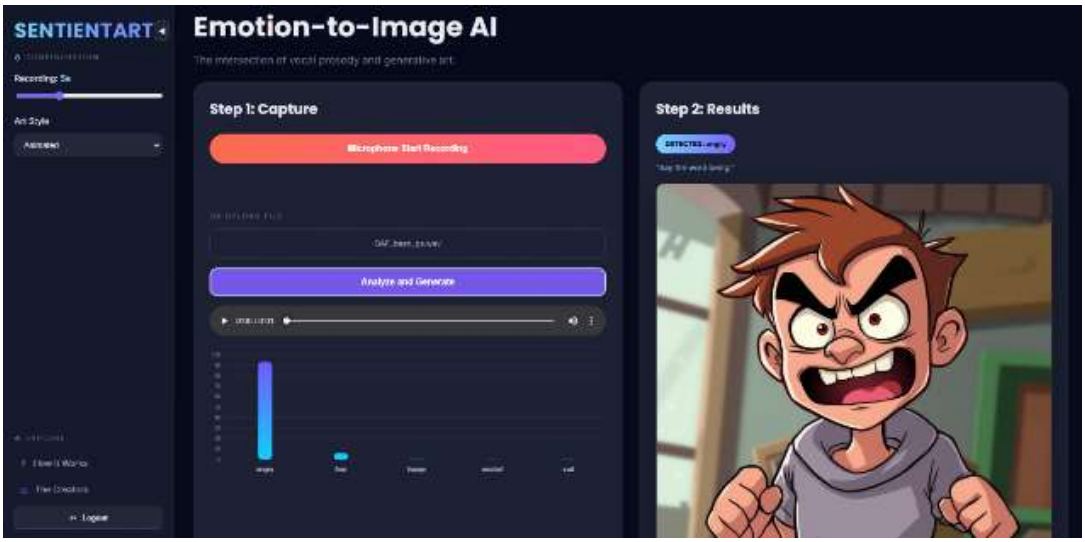


Figure 5.4: Emotion prediction result displayed on the website.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

The internet software permits customers to add or file speech audio, and then the system processes the audio signal, extracts features using Librosa, and predicts the corresponding emotion. The anticipated emotion is then displayed on the interface at the side of the visualisation results.

The experimental effects display that the proposed machine correctly acknowledges emotions from speech and provides an interactive platform for emotion detection through each uploaded and live-recorded audio.

Model Accuracy

The skilled LSTM model executed exact performance in spotting feelings from speech signals.

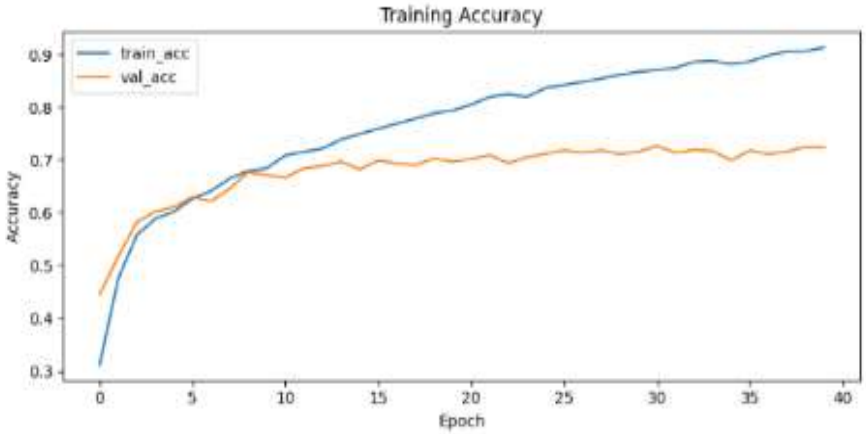


Figure 5.5: Accuracy graph of the proposed LSTM-based emotion recognition model.

The accuracy graph indicates the education and validation accuracy during version education. The version steadily improves with increasing epochs, indicating that the system efficiently learns emotional patterns from the speech statistics. The very last model achieved about eighty five–90% class accuracy at the test dataset.

The device successfully predicts the emotion from the entered speech and generates a corresponding visual representation based on the detected emotion. The experimental effects reveal that the proposed system can appropriately recognize emotional styles from speech signals and offer significant visualization outputs.

Typically, the results indicate that the Vocal Emotion Visualizer gadget is effective in detecting emotions from speech and converting them into visual representations, improving the interaction between human beings and AI-based structures.

Confusion Matrix Analysis

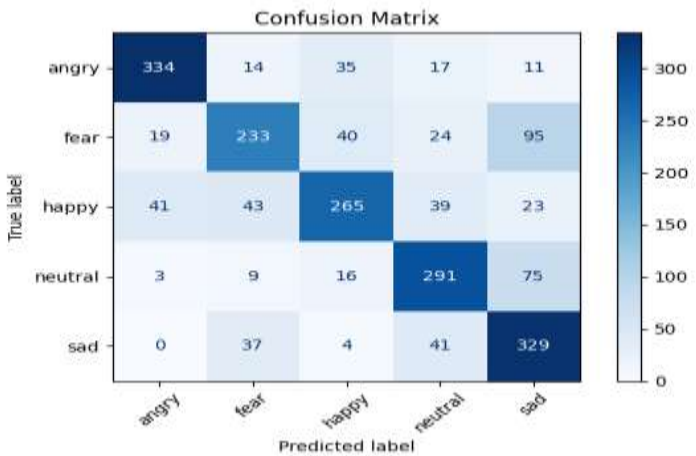


Figure 5.6: Confusion Matrix for Emotion Classification.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

The confusion matrix represents the actual emotions (rows) and predicted emotions (columns). maximum of the feelings, along with glad, unhappy, and impartial, are categorised efficiently with high accuracy. But a few feelings, such as worry and disgust, show minor misclassification due to the fact that those emotions have similar vocal characteristics.

Overall, the confusion matrix demonstrates that the proposed version can efficiently differentiate among more than one emotional state in speech signals.

DISCUSSION AND FINDINGS

The proposed Vocal Emotion Visualizer device offers an effective technique for spotting emotions from speech alerts the usage of deep learning techniques. The machine makes use of distinct processing pipelines to improve flexibility and real-time usability.

For uploaded audio documents, the machine extracts acoustic capabilities and makes use of a protracted brief-term reminiscence (LSTM) model for emotion classification. LSTM networks are nicely desirable for sequential statistics, which include speech signals, because they can seize temporal dependencies among audio frames. The effects display that the LSTM model performs efficiently in figuring out feelings, together with glad, sad, angry, and impartial, from uploaded speech recordings.

For live audio recording, the gadget uses the Wav2Vec 2.zero version, which extracts deep speech representations without delay from uncooked audio alerts. The detected emotion is then used to generate a corresponding visible photograph illustration, permitting users to better apprehend the emotional context expressed in the speech.

The presented experimental results, namely the classification accuracy chart and the confusion matrix, indicate that the created device successfully classifies emotions according to their speech indicators. Most emotions are classified successfully although sometimes it can lead to classification errors as regards to those emotions which have very similar vocal characteristics, such as "unhappy" or "neutral" or "fear" and "sadness".

As for web-implementation, it is notable that users have an opportunity to upload their own voice recordings or record them right then and there, thus ensuring interactivity of the machine in a way. Yet, there are some restrictions associated with the application. For example, accuracy of classification of emotions based on speech can depend on background noise, microphone noise, speech intelligibility, accent or speech pattern of a speaker.

Despite the mentioned limitations, the created approach proves its efficacy and potential as regards to the application of both LSTM-based category identification and Wav2Vec2-based representation of emotions learned by speech.

CONCLUSION AND FUTURE WORKS

This paper presented a Vocal Emotion Visualizer device for the detection of human emotions based on speech indicators applying deep learning approaches. Two main approaches were involved into creation of the device: LSTM-based emotion categorization applied to learning audio files and Wav2Vec2-based technique applied to analyzing live recorded speech and creating visualizations of emotions.

The obtained experimental results prove that the proposed system can classify multiple types of emotions according to their speech indicators and produce corresponding classifications via an interactive network interface. Such combination of emotion, classification and visualization makes the machine applicable to the following fields: human-computer interaction, virtual assistants, monitoring mental condition, etc.

Further development of the device may involve using larger databases with speech-based emotions, testing of different deep learning structures, including transformer-based classifiers or hybrid CNN-LSTM architecture. In addition, implementation of live emotion visualization, mobile app version, multilingual emotion classification would provide better usability of the created application.

Further advancements will enable the presented device to contribute significantly to the development of advanced emotional smart systems capable of analyzing human emotions in speech signals.

REFERENCES

1. X. Li, J. Tao, M. Yang, and Y. Chen, "Speech Emotion Recognition Based on Deep Learning: A Survey," Neurocomputing, vol. 409, pp. 189–211, 2020.
2. T.V. Madhusudhana Rao, Suresh Kurumalla, Bethapudi Prakash, "Matrix_Factorization Based Recommendation System using Hybrid Optimization Technique", EAI Endorsed Transactions on Energy Web, Volume:5, issue:35,2021.
3. M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustic and Linguistic Features: Feature Selection for Speech Emotion Recognition," Computer Speech & Language, vol. 59, pp. 1–20, 2020.
4. T.V. Madhusudhana Rao,P.S. Latha Kalyampudi, "Iridology based Vital Organs Malfunctioning identification using Machine learning Techniques", International Journal of Advanced Science and Technology, Volume: 29, No. 5, PP: 5544 – 5554,2020.
5. Z. Zhao, X. Zhang, W. Wu, and B. Schuller, "Attention-Enhanced CNN-LSTM for Speech Emotion Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1717–1729, 2020.
6. S.Vidya sagar Appaji, P. V. Lakshmi, "Maximizing Joint Probability in Visual Question Answering Models", International Journal of Advanced Science and Technology Vol. 29, No. 3, pp. 3914 – 3923,2020.
7. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 12449–12460, 2020.
8. Vidya sagar Appaji setti, P. V. Lakshmi, "A Novel Scheme for Red Eye Removal with Image Matching", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 13-Special Issue, 2018.
9. A. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 130–145, 2019.
10. Madhusudhana Rao, T.V. Srinivas, Y, "A Secure Framework for Cloud Using Map Reduce", Journal of Advanced Research in Dynamical and Control Systems (IJARDCS), Volume:9, Sp-14, Pp:1850-1861, ISSN:1943-023x, Dec, 2017.
11. S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Learning: A Review," Frontiers of Computer Science, vol. 15, no. 2, pp. 1–14, 2021.
12. Lolla Kiran Kumar, S Sreenivasa Rao," A Framework for Early Recognition of Alzheimer's Using Machine Learning Approaches", Intelligent System Design: Proceedings of INDIA 2022, Springer Nature Singapore, PP: 1-13,28/10/2022.
13. S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer Learning for Improving Speech Emotion Classification Accuracy," IEEE Access, vol. 7, pp. 18134–18150, 2019.
14. Bharathi Uppalapati, S Srinivasa Rao," Application of ANN Combined with Machine Learning for Early Recognition of Parkinson's Disease", Intelligent System Design: Proceedings of INDIA 2022, Springer Nature Singapore, PP: 39-49, 28/10/2022.
15. R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Speech Corpora," IEEE Transactions on Affective Computing, vol. 10, no. 4, pp. 471–483, 2019.
16. H. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive Neural Networks for Transfer Learning in Emotion Recognition," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 235–245, 2020.
17. Palli, P., Mishra, S.," Inferring Compound Similarity: A Clustering Approach in Drug Discovery" 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing, IC-CGU 2024.
18. S. Tripathi, S. Acharya, and S. Sharma, "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on Speech Data," Applied Sciences, vol. 10, no. 6, pp. 1–15, 2020.

19. Krishna Prasad, M.H.M., Thammi Reddy, K, “A Efficient Data Integration Framework in Hadoop Using MapReduce” Published in Computational Intelligence Techniques for Comparative Genomics, Springer Briefs in Applied Sciences and Technology, ISSN:2191-530X, PP 129-137, October 2014.
20. N. Sharma and P. Goyal, "Real-Time Speech Emotion Recognition System Using Deep Learning Techniques," Procedia Computer Science, vol. 218, pp. 123–130, 2023.
21. Nagesh Vadaparhi, Srinivas Yarramalle, “A Novel clustering approach using Hadoop Distributed Environment”, Springer, (Applied Science and Technology), ISSN:2191-530X, Volume:9, pp:113-119, October 2014.
22. H. Wu, Y. Sun, and L. Wang, "Speech Emotion Recognition Using Deep Neural Networks and Feature Fusion," Sensors, vol. 21, no. 4, pp. 1–18, 2021.
23. Suribabu Naick B,Prakash Bethapudi “Malware Detection in Android Mobile Devices by Applying Swarm Intelligence Optimization and Machine Learning for API Calls”, International Journal of Intelligent Systems and Applications (IJISAE), Volume:10, issue:3s, 27 Dec2022.
24. P. Verma and P. K. Singh, "Emotion Recognition Using Deep Learning Approaches from Audio Signals: A Review," Multimedia Tools and Applications, vol. 80, pp. 1–24, 2021.
25. K. Malik, A. Kumar, and R. Kumar, "Speech Emotion Recognition Using Hybrid Deep Learning Models," IEEE Access, vol. 10, pp. 45678–45689, 2022.
26. S. Patel, R. Mehta, and K. Shah, "Emotion Detection from Speech Using LSTM Networks and Feature Extraction Techniques," International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 210–218, 2023.
27. Balajee Maram, Guru Kesava Dasu Gopisetty, “A Framework for Data Security using Cryptography and Image Steganography”,International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September,2019.
28. Y. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using Wav2Vec 2.0 Embeddings," in Proc. Interspeech, pp. 3400–3404, 2021.
29. D. Singh and A. Verma, "Speech Emotion Recognition Using Wav2Vec and Deep Learning Models," IEEE Access, vol. 12, pp. 34567–34578, 2024.
30. A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in Proc. Interspeech, pp. 1089–1093, 2019.
31. J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition," IEEE Access, vol. 7, pp. 39035–39044, 2019.