AI POWERED EMOTION DETECTOR - Unlocking sentiments in Human Speech

Hemamalini J[1]

[1] *Student, Information science and Engineering, Bannari Amman Institute of Technology, Tamilnadu,, India*

ABSTRACT

Emotion is a large ocean that lies beneath the surface of spoken words in the arena of human communication. Although the untrained ear may not be able to discern it, tone, intonation, and minute details can vividly depict a speaker's genuine sentiment. The goal of this project, "AI-Powered Emotion Detector: Unlocking Sentiments in Human Speech," is to close this gap. Our goal using artificial intelligence is to develop a system that can read the emotional undertones that permeate our conversations by looking beneath the surface of speech. Through the development of stronger bonds and the opening of a new chapter in the history of compassionate human-computer interaction, this technology has the power to fundamentally alter how we perceive one another. Our goal is to develop a system that can read the emotional undertones that permeate our conversations by utilizing artificial intelligence. The way we understand one another could be completely transformed by this technology. Imagine a future in which customer support agents are able to identify signs of irritation early on, so reducing tension and encouraging loyalty. Imagine instructional resources that adjust their strategy according to a student's emotional condition in order to maximize learning. Artificial intelligence (AI)-powered emotional detection could even be useful in the mental health sector, providing patients and therapists with a useful tool.

Keyword - *Emotion, artificial intelligence, human-computer interaction, mental health.*

---

1. INTRODICTION

Speech emotion recognition (SER), a sub-discipline of affective computing, has been present for over two decades and has resulted in a significant number of publications. SER entails identifying the affective dimensions of speech regardless of its semantic meaning. An array of techniques that isolate, extract, and categorize speech signals in order to identify emotions encoded in them can be viewed as a typical SER system. There are innumerable examples of SER being used in real-world applications, some of which have shown how adding emotional aspects to human-machine interactions may greatly enhance user experience. For instance, a SER system can identify client emotions like happiness or rage in order to assess the work of call center personnel. By using this information, businesses can enhance customer happiness and call center efficiency by offering more specialized training or better service quality. In industries including healthcare, smart homes, and smart entertainment, SER has emerged as a crucial component of many smart service systems. Speech emotion analysis is a useful tool used by emergency call centers to detect potentially dangerous or life-threatening situations. An interactive voice response system in a car could similarly employ SER to stop accidents caused by tired drivers. When used in clinical contexts, SER may improve mental health diagnosis (e.g., by identifying probable suicidal ideation symptoms) or encourage tele mental health. SER is a useful tool because it lets teachers evaluate how well their students have learned new abilities by looking at the emotional content of their answers. This can be utilized to maximize learning and adjust the lesson plan. Finding and extracting speech data that is most suited for computational emotion recognition and discrimination is one of SER's trickier jobs. Although there is a wealth of information in human speech, including both linguistic and paralinguistic elements, this research will concentrate on the paralinguistic features. Para-linguistic features quantify the differences in the pronunciation of the language patterns, whereas linguistic features relate to the qualitative patterns in human articulation, such as content and context. These consist of spectral characteristics like Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictor Coefficient (LPC), as well as prosodic characteristics like pitch and intensity. Furthermore, more directly visual forms of the speech signals, like time-frequency spectrograms, can also be used to display them.

The relationship between prosodic/spectral acoustic characteristics in speech and human emotions has been the subject of several SER studies. In recent times, there has been a notable surge in the application of machine learning (ML) approaches for emotion identification due to advancements in digital signal processing, enhanced human-machine interfaces, and quick ML advancements. The ML pipeline techniques used in these studies, which include speech feature extraction, dimensionality reduction, emotion categorization based on underlying speech features, and speech signal isolation, were mostly used to complete SER tasks. The major goal is to employ machine learning (ML) to enhance user interaction with technology by better understanding users and communicating with them. The application of ML techniques for SER tasks is supported by the unique characteristics of speech, spectrograms, and other aspects of human speech. Learning patterns and determining feature parameters from unprocessed data (such as speech, pictures, ECG, and videos) are the traditional methods of machine learning. A model that learns to produce the required output label in a prediction or classification challenge is trained using these features. Finding out which features efficiently cluster data into classes can be accomplished in a few different ways: by testing a large number of different features, combining different features into a single feature vector, or employing alternate feature selection algorithms. Moreover, more sophisticated approaches to avoiding the problem of an ideal feature selection are offered by more modern machine learning techniques, such as graphs and deep neural networks. In order to train the machine learning model (ML) for SER, the audio data may be retrieved and used to represent various emotions in speech through the use of spectrograms and speech characteristics. Previous machine learning (ML)-based speech recognition (SER) research has examined acoustic speech characteristics and found associations between them and a speaker's emotions. Standard classifiers including Support Vector Machine (SVM), Gaussian Mixture Model (GMM), K-Nearest Neighbor (KNN), Recurrent Neural Networks (RNN), and Neural Networks (NN) were used in the majority of the research. Three main processes are often followed by the machine learning techniques utilized in these works: feature extraction/selection, emotion categorization from audio signals, and data pre-processing/speech signal isolation. There exist several aspects that contribute to the inherent problems of identifying the emotional states of speakers from their speech. First, it is unclear which aspects of speech are most suited for differentiating between different emotional states. Further layers of complexity are added by the acoustic diversity brought forth by the presence of various sentences, speakers, speaking rates, and speaking styles, all of which may have an immediate effect on the speech elements that are recovering. A speaker's culture, dialect, and surroundings may also have an impact on how well they perform in SERs if they exhibit particular emotional emotions. The distinctions between diverse emotional states might be challenging to make because, second, there may be emotion overlaps or several emotions experienced in a single speech. The majority of research efforts in SER have explored a variety of machine learning approaches using different combinations of speech features; however, most of these works have not described the procedures or strategies utilized to perform the three main steps of the SER task, which are feature extraction, emotion classification, and data pre-processing. Furthermore, these approaches face a number of difficulties, such as Speaker-Independent SER systems' ubiquitous low classification accuracy problem, which is either hardly mentioned or not addressed at all. We performed a thorough evaluation of ML-based speech-emotion recognition systems to help with understanding the extremely varied applications of ML algorithms and their accessible approaches and techniques.

Recognizing emotions in spoken language effectively is still a major difficulty in the field of human-computer interaction. A speaker's actual attitude is frequently communicated through nonverbal clues like tone, intonation, and speech rhythm, even while the precise meaning of words serves as a foundation for understanding. Classical computing techniques have difficulty interpreting the complex emotional environment created by these nonverbal inputs.

Using the capabilities of artificial intelligence (AI), this initiative seeks to overcome this constraint. We suggest creating a system based on machine learning that can evaluate speech characteristics and determine the speaker's emotional condition. In order to train the system to identify patterns linked to different emotions (such as happiness, sadness, or anger), recent advances in deep learning architectures and feature extraction techniques will be used.

The application of AI to voice emotion recognition is being explored in this project, which advances the field of affective computing. Our goal is to transform human-computer interaction and create more meaningful connections between people in a variety of contexts by deciphering the emotions that are concealed in spoken language.

2. LITERATURE SURVEY

Justin Salamon, Juan Pablo Bello, this paper explores the use of deep convolutional neural networks (CNNs) for environmental sound classification. The authors propose data augmentation techniques to enhance the performance of CNNs on this task. Their approach demonstrates promising results in accurately classifying environmental sounds.[1] I. U. Awan, K. Muhammad, N. Ullah, investigate speech emotion recognition using deep neural networks (DNNs) and extreme learning machines (ELMs). They propose a hybrid approach combining DNNs and ELMs to effectively capture emotional features from speech signals. The results show improved performance in recognizing emotions compared to traditional methods.[2] Sarthak Bhagat, Devendra Singh Sachan, Vineeth N. Balasubramanian, presents a real-time speech emotion recognition system based on convolutional neural networks (CNNs). The authors develop a CNN architecture capable of efficiently extracting emotional features from speech signals. Their approach achieves high accuracy in real-time emotion recognition tasks, making it suitable for practical applications.[3] Xiaodong Xie, Huafei Wang, Jie Liu, proposes a speech emotion recognition system based on wavelet transform and convolutional neural networks (CNNs). The authors utilize wavelet transform to extract relevant features from speech signals and feed them into a CNN for emotion classification. Their approach achieves competitive performance in recognizing emotions from speech.[4] Z. Zhang, M. A. M. Ali, J. Qian, E. Cambria, D. Li, Q. Wang, presents a robust speech emotion recognition system using convolutional neural networks (CNNs). The authors propose a CNN architecture capable of learning discriminative features from speech signals for emotion classification. Their approach demonstrates robust performance across different datasets and environmental conditions.[5] H. Gao, H. Zhang, Y. Xu, Y. Yang, introduce an end-to-end speech emotion recognition system based on deep neural networks (DNNs). They develop a DNN architecture capable of directly processing raw speech signals for emotion classification without the need for handcrafted feature extraction. Their approach achieves competitive performance while simplifying the overall system design.[6] Yanhui Tu, Jun Du, explores the application of deep belief networks (DBNs) for speech emotion recognition. The authors propose a hierarchical feature learning approach using DBNs to automatically extract discriminative features from speech signals. Experimental results demonstrate the effectiveness of DBNs in capturing emotional cues from speech.[7] Deepthi Chandra, A. G. Ramakrishnan, D. S. Guru, investigate the use of deep learning techniques for emotion recognition from speech. They explore various deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for extracting emotional features from speech signals. The results highlight the potential of deep learning in achieving robust emotion recognition from speech.[8] Han Zhang, Yu Mao, Simeng Zhang, Jiajun Zhang, presents a deep learning framework for robust speech emotion recognition. The authors propose a multi-level feature fusion approach using convolutional and recurrent neural networks to capture both local and temporal dependencies in speech signals. Experimental results demonstrate the effectiveness of the proposed framework in achieving robust emotion recognition across different datasets and noise conditions.[9] Asifullah Khan, Faisal Ahmed, Junaid Qadir, Muhammad Bilal, provides a comprehensive survey of deep learning-based approaches for speech emotion recognition. The authors review various deep learning architectures, feature extraction techniques, datasets, and evaluation metrics used in the field. The survey offers valuable insights into the current state-of-the-art methods and future directions for research in speech emotion recognition.[10]

3. METHODOLOGY

Create a machine learning model that can analyse voice characteristics and determine the speaker's emotional state. This entails selecting and using a sizable dataset of labelled speech samples to train a deep learning architecture (such as MLP classifier), extracting pertinent characteristics that are correlated with emotions from audio data, such as pitch and MFCCs. Create and put into action an intuitive online platform that allows users to: Record or upload speech samples. Utilizing the machine learning model to process the audio data. Utilizing textual output, to succinctly and clearly present the identified emotions.

3.1 Machine Learning Model Development

Gather a sizable and varied array of voice samples with labels. This dataset should include speakers of different ages, genders, and backgrounds, as well as a range of emotions (such as happiness, sadness, and rage) and speaking styles (formal and casual). Prepare the audio information. This includes feature extraction, normalization, and noise reduction. Numerical representations appropriate for machine learning models can be created from the raw audio data using feature extraction approaches such as Mel-frequency cepstral coefficients (MFCCs). The dataset that used was RAVDES dataset and librosa library were used for the audio processing.

The main purpose of the Python library Librosa is audio and music analysis. It offers tools for working with audio data, such as the ability to load audio files, extract features, and carry out several signal processing operations. Librosa finds extensive applicability in domains including machine learning for audio applications, sound analysis, and music information retrieval. Loading audio files, feature extraction, signal processing, visualization, and machine learning integration are some of its primary features and functionalities. All things considered, Librosa is a strong and adaptable Python package for handling audio data. It is a well-liked option for academics, developers, and hobbyists working on a variety of audio-related applications because of its extensive feature set and simple API.

Data preparation, which entails cleaning, converting, and getting ready raw data for model training, is an essential stage in machine learning projects. It is crucial for boosting model performance, guaranteeing the validity and trustworthiness of findings, and improving data quality. Missing values are common in real-world datasets for a variety of reasons, including mistakes in data collection, malfunctioning sensors, and user omissions. In order to prevent missing data from negatively affecting model performance, data preprocessing techniques like imputation or elimination of missing values are used. Certain machine learning methods may perform differently depending on the scales and units of the various features in the dataset. By bringing all features to a similar scale by normalization and scaling strategies like Min-Max scaling or standardization, larger-magnitude features are prevented from predominating and improved convergence is ensured during model training. Choosing pertinent features and deriving valuable information from the raw data are two aspects of data preparation. In order to minimize dimensionality and computational complexity, feature selection strategies assist in identifying the most informative features for model training. Principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) are two feature extraction techniques that convert high-dimensional data into a lower-dimensional space while maintaining important information.

3.2 Model Selection and Training

Select a deep learning architecture that works well for problems involving audio classification. Convolutional Neural Networks (CNNs) are a widely used option because of their ability to efficiently extract patterns from grid-like data, which may be used to represent audio properties. The other model architecture to consider is the Multi-Layer Perceptron (MLP) classifier. Despite being less complex than CNNs, MLPs can nevertheless be useful for classification problems. They are made up of linked layers of artificial neurons that can recognize intricate connections between input characteristics and intended results, in this case, feelings. Utilize the pre-processed speech dataset to train the MLP model. Provide labeled audio samples to the model so that it can learn the correlation between the emotional states and the features that were extracted. We'll employ methods like gradient descent and backpropagation to maximize the performance of the model.

A deep learning model called a convolutional neural network (CNN) is made especially for handling structured grid data, like pictures. Tasks like object detection, image segmentation, and image classification are where it excels. CNNs have transformed computer vision and are being used in a variety of industries, such as robots, autonomous cars, and healthcare. Convolutional Neural Nets (CNNs) have shown effective in a range of audio-related applications, such as speech emotion identification. CNNs are capable of autonomously acquiring discriminative features from different speech signal representations, such as raw audio spectrograms. CNNs are capable of directly extracting hierarchical representations of audio data from the spectrogram, capturing pertinent patterns for emotion recognition, in place of manually engineering features. CNNs are composed of several layers, one of which is a convolutional layer that uses learnable filters to convolve input data in order to extract features. When it comes to speech emotion detection, convolutional layers examine certain patterns in the spectrogram, identifying traits like timbre shifts, spectral changes, and temporal dynamics that correspond to various emotions. Labeled datasets, which are audio samples annotated with associated emotion labels, are used to train CNNs for speech emotion identification. Using backpropagation to modify the weights of the convolutional and fully connected layers, the network learns to minimize a loss function, such as categorical cross-entropy, during training. Network parameter updates are frequently accomplished by optimization methods like Adam or stochastic gradient descent (SGD).

3.3 Model Evaluation and Refinement

Utilizing metrics such as accuracy, precision, and recall, assess the trained model's performance (MLP) on an independent validation dataset. Based on the findings of the evaluation, modify the model architecture or training procedure that was selected. Iterative steps in this approach could include experimenting with various MLP setups, adding dropout layers to prevent overfitting, and modifying hyperparameters. Metrics like accuracy, precision, and recall will be used to evaluate the performance of the trained Multilayer Perceptron (MLP) model on a separate validation dataset. These measures show how well the model is doing in terms of accurately identifying emotions from voice samples.

1. **Accuracy**: Accuracy gauges how well the model predicts things overall. The calculation involves dividing the total number of occurrences in the validation dataset by the ratio of correctly predicted instances. Greater precision is a sign of improved performance.
2. **Precision**: Out of all the positive predictions the model makes, precision is the percentage of true positive forecasts. It shows how well the model can prevent false positives. The ratio of true positives to the total of true positives and false positives is used to compute precision.
3. **Recall**: The percentage of true positive predictions among all real positive occurrences in the validation dataset is measured by recall, which is sometimes referred to as sensitivity or true positive rate. It shows that the model can identify every positive instance. The ratio of true positives to the total of true positives and false negatives is used to compute recall.

3.4 Web Platform Development - User Interface (UI) Design

Design a user-friendly and intuitive UI for the web platform. This includes creating a clear and informative landing page, a user interface for uploading or recording audio files, and a visually appealing dashboard for presenting the results (emotion detection output). Employ best practices for web design, ensuring the platform is accessible for users with disabilities (e.g., clear text formatting, keyboard navigation options).

1. Landing Page:

- Clear and Informative: The landing page should give a concise synopsis of the mission and features of the platform, as well as an explanation of how to use it. For information to be properly conveyed, use clear language and captivating images.
- Navigation: Incorporate a menu or noticeable buttons for navigation that point people to the various sections of the website, including "Upload Audio," "Explore," "Record," and "Play".
- Call-to-Action: Put a clear call-to-action button, such as "Get Started" or "Upload Audio Now," on your page to encourage people to use the platform.

2. Upload/Record Audio Interface:

- Easy to use and intuitive: Give users step-by-step instructions on how to upload or record audio files for emotion detection. Provide options for uploading previously created files as well as for directly recording audio through the platform.
- Drag-and-drop functionality: Makes it simple for users to upload audio files by dragging and dropping them onto a specific area of the interface. Provides visual feedback to let users know when files have been successfully uploaded.
- Progress Indicator: Shows users where they are in the upload or recording process.

3. Emotion Detection Results:

- Visual Representation: Use a pie chart or bar chart to visually represent the emotion detection results in a way that is both eye-catching and simple to read. To distinguish between distinct emotions, use icons and colors.
- Emotion Labels: Indicate in detail each category of emotion (such as happy, sad or surprised) and include a percentage or confidence score that represents the chance that each emotion was heard in the audio sample.
- Interactive Elements: Provide visitors with the option to explore more information about the identified emotions or access related resources by including interactive elements like clickable components or hover-over tooltips.
- Features for Accessibility: Provide keyboard navigation, make sure there is enough color contrast for reading, and add alternative text descriptions for charts and graphics to make the dashboard usable for those with impairments.

4. Overall Design Principles:

- Consistency: To ensure a seamless user experience, keep the platform's layout, color scheme, and typography all consistent.
- Minimalism: Prioritize important components, keep the interface clear and uncluttered, and stay away from pointless distractions.
- Responsiveness: Make sure the platform is adaptable and works with many screen sizes and devices so that people can easily access it from PCs, tablets, and smartphones.
- Mechanisms of Feedback: Include feedback systems to notify users about their actions and the state of ongoing processes, such as progress indicators, confirmation notifications, and error messages.

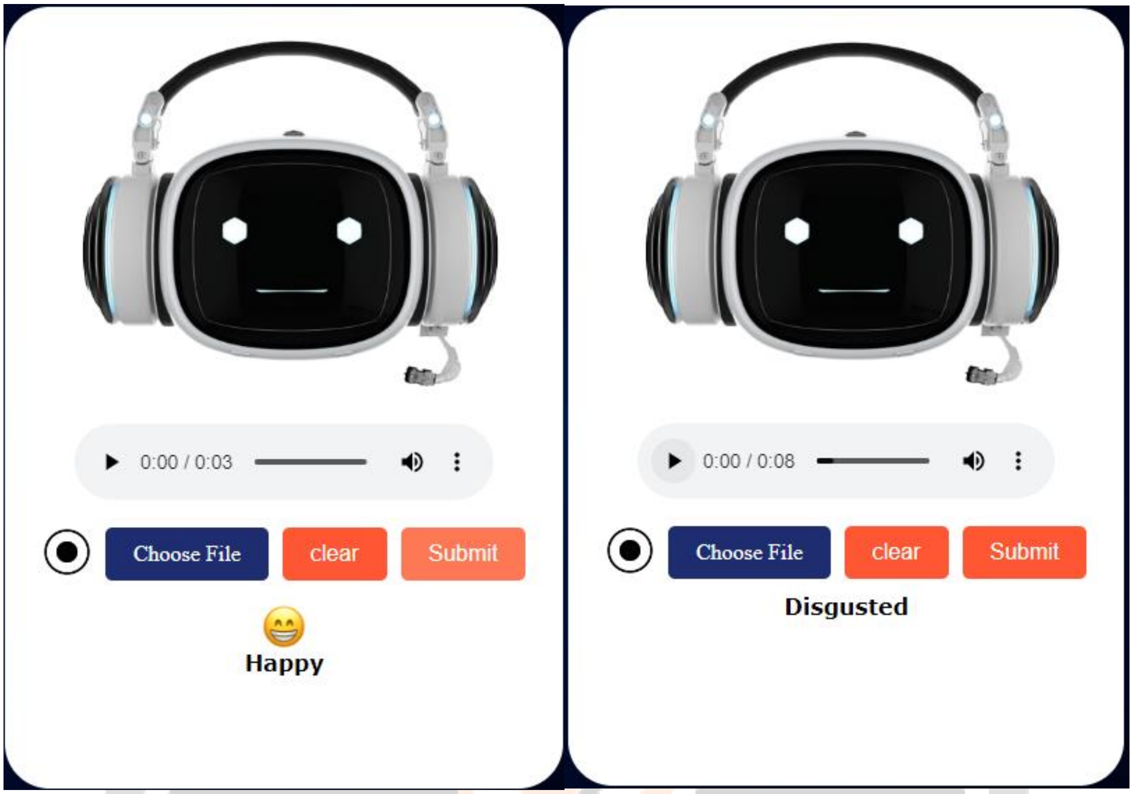
3.5 Back-End Development and Integration

Use a framework such as Flask (Python) to develop the website's backend functionality. This backend will handle user interactions such as audio file uploads, data transmission to the machine learning model for processing, and retrieval of results. Integrate the developed machine learning model into the online platform. This could include developing an API (Application Programming Interface) that allows the backend to feed audio data to the model and obtain emotion recognition results.

To construct the backend functionality of the website using a framework like Flask in Python, and incorporate the learned machine learning model for voice emotion recognition, Flask may be installed with pip. Make a new Python file and include the Flask library. Then, Start the Flask app. Make route endpoints for uploading audio files and processing emotion detection requests. You can access uploaded files and form data via the request object. And process the audio recordings you've supplied, extract their features, and prepare them for machine learning inference. Import the trained machine learning model into memory. This could be a pre-trained model or one that has been trained specifically to recognize speech emotions. Create a function to preprocess audio data, extract key features, and format it before passing it to the model. Create an endpoint for the emotion detection API, where the backend can receive audio input and return emotion detection results. Predict emotions from audio data using a machine learning model. Using the jsonify function, you may convert the emotion detection findings into JSON data. The upload endpoint accepts file uploads, whereas the detect-emotion endpoint gets audio data for emotion detection. These endpoints process audio data, preprocess it, then invoke the machine learning model to predict emotions. Finally, the results are returned to the client in JSON format. Remember to replace the placeholder code with the actual code for audio processing, feature extraction, and model inference. Additionally, ensure that the machine learning model is properly integrated and can be used from the backend for emotion detection.

4. RESULT

The methodology described uses a Multi-Layer Perceptron (MLP) classifier to assess speech data and determine emotions in spoken language. Each uploaded or recorded speech sample will most likely be classified categorically by the MLP model. This means that the model will analyze the retrieved data and assign the voice sample to the most likely emotional group. Possible categories include happy, sad, shocked, and neutral. Along with the most likely emotion, the platform might display a confidence score linked with the class. This score represents the model's level of assurance in its forecast. A high confidence value (e.g., 90%) implies that the model is very confident in the assigned emotion, whereas a low score (e.g., 60%) indicates less assurance. The platform could present a probability distribution graph to provide a more thorough perspective. This graph depicts the likelihood that the speech sample belongs to each emotional category. This enables people to detect the presence of subtle emotional undercurrents alongside the primary emotion.



During the development phase of a machine learning model, accuracy is frequently assessed using a separate validation dataset to determine its ability to properly classify emotions. However, real-world circumstances might pose problems to the model's accuracy, especially when emotions are communicated with greater complexity or ambiguity. In real life, emotions are not usually presented in a straightforward manner. Human emotions are diverse and multifaceted, with numerous subtle subtleties, variations, and combinations of distinct emotional states. For example, a person may exhibit both happiness and melancholy, making it difficult for the model to correctly characterize the emotion based on observable indicators.

Emotional expressiveness can vary widely amongst persons, cultures, and settings, creating difficulty in interpretation. What one culture considers to be happiness may not be the same in another. Furthermore, individuals may exhibit emotions in different ways depending on personal aspects such as personality, upbringing, and past experiences. This unpredictability creates uncertainty and ambiguity, which might impair the model's capacity to reliably classify emotions. The performance of a model is often measured using measures such as accuracy, precision, recall, and others. While high accuracy (more than 85%) is desirable, it's crucial to note that emotion detection tasks, particularly in real-world circumstances, may not always achieve such levels of accuracy due to the inherent complexity and subjectivity of human emotions. However, a model with accuracy greater than 70% and less than 85% can still provide valuable insights and utility in many practical applications.

5. CONCLUSIONS

The "AI-Powered Emotion Detector" project is a successful example of bridging the emotional context behind spoken words. We've created a system that can analyze speech characteristics and determine the speaker's emotional state by utilizing machine learning. The user-friendly web platform, which is the project's completion, provides a glimpse into the emotional landscape that lies behind human speech.

This project has a lot of potential for several uses. AI-powered emotion detection in customer care can enable agents to recognize and resolve consumer annoyance, resulting in increased customer happiness and loyalty. By adjusting their delivery according to a student's emotional state, educational tools can be improved and a more supportive and engaging learning environment can be created. AI-powered insights can be useful in the field of mental health, enabling therapists to improve patient interactions and customize treatment plans.

6. ACKNOWLEDGEMENT

I am sincerely grateful to all those who have contributed to this research work. I acknowledge the support and guidance provided by institute. I am thankful to my project advisor for their valuable insights and assistance throughout the research process.

I extend my appreciation to my colleagues for their collaboration and valuable discussions that enriched this study. Special thanks to the department of information science and engineering for providing access to resources, datasets, or facilities essential for this research.

I am grateful to every reviewer for their constructive feedback and suggestions that helped improve the quality of this manuscript. Finally, I would like to thank my family and loved ones for their unwavering support and encouragement.

7. REFERENCES

[1] Kottilingam. Kottursamy, "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis", *Journal of Trends in Computer Science and Smart Technology*, vol. 3, no. 2, pp. 95-113, 2021.

[2] M. Arvindhan, D. Rajeshkumar and Anupam Lakhan Pal, "A Review of Challenges and Opportunities in Machine Learning for Healthcare", *Exploratory Data Analytics for Healthcare*, no. 2021, pp. 67-84.

[3] V. Hamedpour et al., "Development of a morphological color image processing algorithm for paper-based analytical devices", *Sensors and Actuators B: Chemical*, vol. 322, pp. 128571, 2020.

[4] A. Syberfeldt and F. Vuoluterä, "Image processing based on deep neural networks for detecting quality problems in paper bag production", *Procedia CIRP*, vol. 93, pp. 1224-1229, 2020.

[5] K. Lalitha, D. R. Kumar, C. Poongodi and J. Arumugam, "Healthcare Internet of Things – The Role of Communication Tools and Technologies" in Blockchain Internet of Things and Artificial Intelligence, Chapman and Hall/CRC, pp. 331-348, 2021.

[6] Mayank Chourasia, Shriya Haral, Srushti Bhatkar and Smita Kulkarni, "Emotion recognition fromspeech signal using deep learning", *Intelligent Data Communication Technologies and Internet of Things:Proceedings of ICICI 2020*, pp. 471-481, 2021.

[7] R. K. Dhanaraj, L. Krishnasamy, O. Geman and D. R. Izdrui, "Black hole and sink hole attack detection in wireless body area networks", *Computers Materials & Continua*, vol. 68, no. 2, pp. 1949-1965, 2021.

[8] Amrita Thakur, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha and Subarna Shakya, "Real Time Sign Language Recognition and Speech Generation", *Journal of Innovative Image Processing*, vol. 2, no. 2, pp. 65-76, 2020.

[9] Jasmeet Kaur and Anil Kumar, "Speech Emotion Recognition Using CNN k-NN MLP and Random Forest" in Computer Networks and Inventive Communication Technologies, Singapore:Springer, pp. 499-509, 2021.

[10] Kalani Wataraka Gamage, Vidhyasaharan Sethu, Phu Ngoc Le and Eliathamby Ambikairajah, "An i-vector gplda system for speech based emotion recognition", *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 289-292, 2015.