**Vocal Vibe: Emotion Recognition through Speech Analysis**

Gopi Reddy Someswara Reddy, Lokesh Kollepara, Vaan Amuthu Elango

*Department of Computer Science and Engineering,*  
*Amrita School of Computing, Bengaluru,*  
*Amrita Vishwa Vidyapeetham, India*

**Abstract: In the last decade, the field of mental fitness has been developing and the potential of technology to improve control and choice as part of plan is increasing. Our project, entitled "Vocal Vibe," targets this new discipline of mental health through the use of digital systems that employs speech analysis techniques to distinguish likely mental health problems and offer the suitable support to individuals in need of help. Using the multiplicity of such vocal features, such as pitch, tone, and speech patterns, our engine is intended to target and further identify stress, depression, anxiety, and other mental health conditions Through machine learning the system gets trained to come up with existing indicators and get equipped on detecting the emerging ones which make it to continuously improve on its accuracy over time. Similarly, VocalVibe has the supportive features that include the provision of resources which are appropriate for emotional self-care, and mental health professionals will be available to talk to when they are needed. VocalVibe aims at the development of the self-help skills in the young people to ensure them adequately equipped to take control of their wellness, and consequently, allow the early intervention for those struggling with mental health issues.**

**Keywords— MFCC, Feature extraction, Speech analysis, Machine learning, Deep learning, Identity prediction**

1. INTRODUCTION

In mental health issues, medical screening followed by appropriate help in time can become the main mechanism by which the effect of ailments on mental health is reduced and the general well-being goes up. Yet, inspite of all mentioned workers and the efforts to decrease the stigma of mental health, the most of people face the big obstacles on the way of looking for help. Recognizing this critical need, our project, "VocalVibe: In the Project "Mental Health Powers Through Speech Analysis and Support," we aim to use technical means tp change the paradigms in mental health diagnosis and guidance completely.

The primary goal of VocalVibe is to make the use of facial recognition and machine learning technology possible in analyzing speech and finding non-intrusive, easily accessible, and effective means of screening mental health symptoms. Communication in the human world is considered to go beyond just a form of exchange. In fact, it is a feature that brings to the table many different aspects, including those that can be used to detect an individual's psychological state and quality of life. VocalVibe draws on those specific vocal features for a different mental health assessment technique, which cannot be compared to the nowadays clinical approaches and which is based on combining technology with everyday life.

The fundamental thing is that this project not only brings to light its technological advancement but can also act as a democracy tool for the general health of peoples. In big part of the world, there are obstacles like cost, stigma and not enough free resources to wall up those who need immediate mental health service. The essence of VocalVibe mission is to erase the social barriers by providing easily accessible and user-friendly online singing platform that can be exercised from the comfort of the participant’s home, without the need for special equipment or a professional assistance. Through using the commonness of smart phones and other internet devices, Vocal Vibe can find a way to reach out to people amid their various races and spread of location.

Besides, Vocal Vibe is provided with a full spectrum approach to the mental wellbeing and therefore it not only sees the symptoms but it offers prevention and management actions to carry out with the people. Therefore, the system provides counselling service and establishes a collection of supportive features, such as: personalized resource recommendations, coping strategies unique for every person and access to professionals in the mental health field in case there is a need for the comprehensive care. In addition to giving them tools and knowledge to travel through such a journey, Vocal Vibe encourages them to live with resilience and confidence in day-to-day life.

To summarize, it seems clear that VocalVibe marks an impressive step towards the future of technological mental health care. Through the use of logic of speech analysis, machine learning, and digital connection we assume VocalVibe will build no barrier platform for mental health support that will be easy to use as well as provide users with the necessary tools to improve their wellbeing. While we are embarking on this mission, our ultimate vision is the construction of a world which will be the one free of stigma and available in minimum availability for people in need, such freedom will be a cornerstone of holistic health care that will be ready at hand for all.

1. LITERATURE SURVEY

This paper calls for better mental health support systems for university students, especially in Sri Lanka. [1] It notes that there aren't enough official peer support programs and that stigma and a lack of knowledge about mental health services make it difficult for students to get help. The paper suggests using a mobile app for peer support sessions and a tool to predict mental health improvements based on how people sound when they talk. The paper emphasizes the importance of mental health among students and how the proposed system could help, but it also recognizes that more research and improvements to the prediction model need to be done to make it work in different languages. This paper discusses four common mental health conditions: paranoia, bipolar disorder, depression, and obsessive-compulsive disorder.[2] It describes the symptoms and effects of each disorder and highlights the importance of seeking professional help to manage them. The paper also stresses the need for further research and public awareness campaigns to better understand and address these challenges This paper uses Convolutional Neural Networks (CNNs) to detect depression through speech patterns. [3] The paper evaluates the accuracy of the model's training and provides experimental results. The model aims to predict whether speech conveys negative emotions. While speech is often used for quick emotional recognition, the paper notes the need for more precise model training. Future research plans include using Transfer Learning for depression detection and creating a comprehensive multimodal dataset on depression.

This study introduces a technique for diagnosing psychiatric illnesses by analysing speech.[4] Speech samples were obtained from individuals diagnosed with schizophrenia, depression, and a comparison group in a controlled setting. Software was used to record and evaluate acoustic features, and a fusion model combining these features with deeper ones was created. The proposed ensemble model outperformed other classifiers in accuracy, consistency, and feature reduction, demonstrating its potential for precise psychiatric diagnosis using speech analysis.This study examines how speech characteristics like pitch and pauses are used during a psychological treatment called "Precious Memories," which helps people remember happy events. [5] The researchers want to see if these speech features can measure mental health. They think that the length of the pauses, vocal pitch, and speech may change during the therapy. To test this, they will look at how people talk in four weekly sessions to see if these prosodic features change after the intervention. This research study aims to explore communication abilities and mental well-being in children, specifically those with autism and [6] language delays. Using interviews with parents and questionnaires, the study assesses children's communication and mental health challenges. The assessment tools included the Children's Communication Checklist and the Strengths and Difficulties Questionnaire. The findings reveal that some children face difficulties with using language appropriately (pragmatic language problems) and may experience mental health concerns.

This study explores the long-term effects of speech and language difficulties at age 5 on educational and mental health outcomes at age 21. [7] Using a large sample of 3193 children, researchers assessed speech concerns through the PPVT R test and collected data on academic achievement, mental health, and school performance at age 14 via the Achenbach Child Behaviour Checklist and YASR questionnaire. The analysis revealed connections between early speech concerns and later academic success, mental well-being, and school engagement. The study also accounted for factors like parenting styles, behavioural issues, and learning disabilities that might influence the findings. The research was conducted with ethical approval, and data analysis was performed using SPSS software. This paper investigates the potential of using speech emotion recognition technology in mental health applications. [8] The authors show that by effectively identifying emotions in spoken language, healthcare professionals can benefit from: - Enhanced evaluations and decisions - Improved therapy outcomes - Remote patient monitoring - More effective training for clinicians The paper presents detailed experiments and analysis on training machine learning models for speech emotion recognition. Using the TESS dataset, they evaluated various classifiers and settings, including overlapping and non-overlapping extraction of speech features. This study explores the importance of using speech recording apps to track mental health in psychiatry. It stresses the need for data privacy and security, as well as clear [9] communication with users. However, there are some limitations, such as a limited sample size, self-reported diagnoses, and uneven group sizes. Research shows that patients have different views on privacy, with schizophrenia patients expressing varied concerns. The ongoing COVID-19 pandemic highlights the need for remote mental health monitoring. A team approach involving researchers, doctors, and patients is crucial to reduce the burden of psychiatric conditions, with speech analysis using machine learning showing promise as a valuable tool.This paper explores the creation and enhancement of a speech analysis library for smartphones. It covers previous work in this field, [10] emotional speech analysis, and identifying mental stress. The paper emphasizes the efficiency of emotional awareness using voice-based emotion recognition. It highlights the application of this technology in daily conversations, emphasizing the importance of feature normalization, addressing data imbalance, and using shared emotion labels to maximize the benefits of emotion databases. Furthermore, the paper discusses healthcare applications of this technology, such as early detection of medical conditions and promoting emotional awareness for those with depressive disorders.

1. METHODOLOGY

**1. Dataset Description:**

The RML emotion database comprises files sourced from 8 distinct speaking subjects proficient in six languages: These languages offer support in English, Mandarin, Urdu, Punjabi, Persian, Italian and the local dialects. This diversity does not only let unique flavor on the skin, but more excitingly, you will find that there are also some Mandarin phrases used in the records. Moreover, the Beardown dataset has topics showing-off- facial hair that make the data even more comprehensive.

The backbone of our data set being the resemblance of human emotion, the dataset collection policy had to be cross-checked for the occurrence of any gaps. Consensus was indeed reached with the aid of the listening tests, presented by a minimum of two unbiased participants. Each participant tested their auditory synthesizing ability with sung language. Spontaneous emotional data videos from the experimental dataset were selected for the experiment if human members could precisely understand the steaming dirty emotions.

The descriptions file meticulously balances cases with a consistent number of samples, approximately 500 for both training and testing sets, each representing one of the six principal emotions: Likewise, emotions such as happiness (HA), sadness (SA), anger (AN), fear (FE), surprise (SU), and disgust (DI) serve the same purpose in simplifying the emotional vocabulary of language. In each instance, the clip, with duration of about 5 sec, follows the originals specifics.

Knowing that the input samples are produced in both audio and video formats, the two extraction channels preserve their duration alignment in terms of time series. This was for capturing video clip via a 22050 Hz sample rate of 16bit frame rate of 30fps. Using a single channel.

**2. Feature Extraction:**

a) Mel-Frequency Cepstral Coefficients (MFCC):

Mel-Frequency Cepstral Coefficients (MFCCs) are a widely used technique in speech and audio processing to extract important features from sound signals. They capture the spectral or frequency-based characteristics of a sound by representing it on a special scale called the Mel scale. This scale mimics how the human ear perceives sound frequencies. To get the MFCCs, the process involves taking the Fourier transform of the sound signal, then finding the logarithm of the resulting spectrum, and finally, taking the discrete cosine transform of that. This allows MFCCs to effectively capture the timbral and spectral properties of speech sounds, making them very useful for applications like speech recognition and emotion detection.

b) Harmonics-to-Noise Ratio (HNR):

HNR, which stands for Harmonics-to-Noise Ratio, is a measure used to evaluate the balance between the harmonic and noisy components in a speech signal. It provides insights into the consistency and tonal quality of the voice. Higher HNR values typically indicate a clearer and more periodic voice signal, suggesting a stronger harmonic content. Conversely, lower HNR values suggest a higher presence of noise or irregularity in the signal. HNR is commonly utilized in various voice analysis tasks, such as assessing voice disorders, evaluating voice quality, and identifying speakers.

c) Teager Energy Operator (TEO):

The Teager Energy Operator is a technique used to capture changes in the energy within a speech signal. It works by estimating the momentary energy of a signal based on its amplitude and frequency variations. TEO is sensitive to quick changes and non-steady parts in the signal, making it good for detecting short-lived events and capturing dynamic aspects of speech, like the start and end of speech, speech rhythm, and prosodic features.

d) Zero-Crossing Rate (ZCR):

ZCR is a method that measures how quickly a signal changes from positive to negative. In speech processing, ZCR counts how many times the sound wave crosses the zero point within a certain time frame. ZCR can show if a speech sound is periodic and if it has voiced parts. It's useful for tasks like dividing speech into segments, detecting when someone is talking, and telling the difference between voiced and unvoiced speech.

**3. Model Development:**

Model Architecture Selection:

Consequently, for our undertaking we have opted the deep learning algorithms which are particularly successful in extracting hierarchical and sequential patterns inherent in the space data.The CNNs are useful to extract spatial properties from spectrogram representations, which allow them to capture local information as well as connections involved.Unlike CNNs, RNNs are best at doing short and long memory tasks and time delays due to their capacity for modeling sequential dependencies of speech data.

a) CNN Model

The model development procedure will involve the use of convolutional neural networks (CNNs) to deploy acoustic features, namely features from the speech signals.Thus, the CNN set will incur hierarchical representations of spectrogram features including, for instance, MFCCs through convolution over input audio frames to grab local elements. With this method, the model is designed to identify a set of features inside the raw audio data and to train them directly without the manual extraction of those features.

b) BILSTM Model:

Bidirectional Long Short-Term Memory (BILSTM) models emerge as a viable solution. Fueled by features such as Mel-frequency cepstral coefficients , zero crossing rate, and others taken from the speech signal, BILSTM networks present a high level of performance in tracking temporal relationships and linguistic agglutination necessary for recognizing expression of emotions. Integration of precedential and successive contexts permits BILSTM architectures to pick up slight alterations in a speaker’s emotive state. Overall, this allows for accurate and flexible emotion identification on speech samples.

C) GRU Model:

Gated Recurrent Unit (GRU) models are a powerful choice in the field of emotion classification using speech data. Due to their ability to encode low-level extracted features like MFCC, ZCR, etc., from the speech signal, GRU models are well-suited for capturing temporal dynamics and overall context, which is crucial for differentiating emotional content. Moreover, as aforementioned, GRU is a simpler model with fewer parameters than LSTM – this ensures desirable optimization behavior during the training and makes it easier to deploy for real-time emotion classification tasks on speech data.

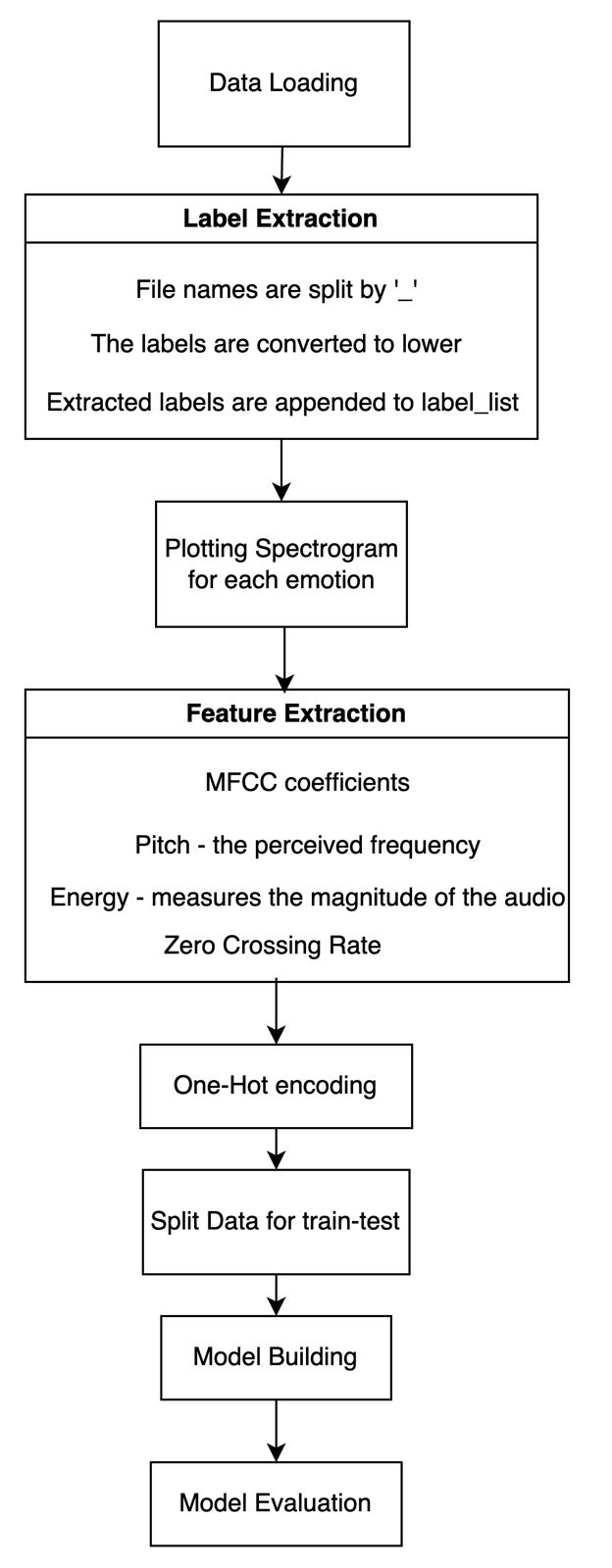


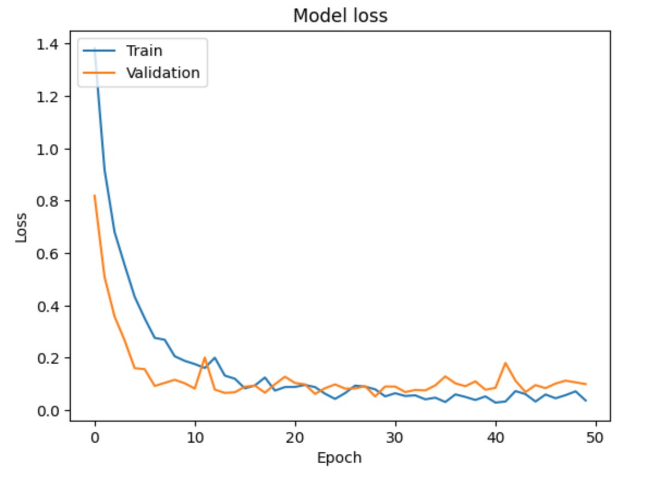
Fig 1: Flow diagram representing the proposed methodology

1. RESULTS

The performance of the classifiers such as Accuracy, Precision, Recall, F1-score are represented in the Table 1

`Table 1: Performance Metrics of Deep Learning models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| CNN | 96 | 96 | 96 | 96 |
| BI-LSTM | 99 | 99 | 99 | 99 |
| GRU | 98 | 98 | 98 | 98 |



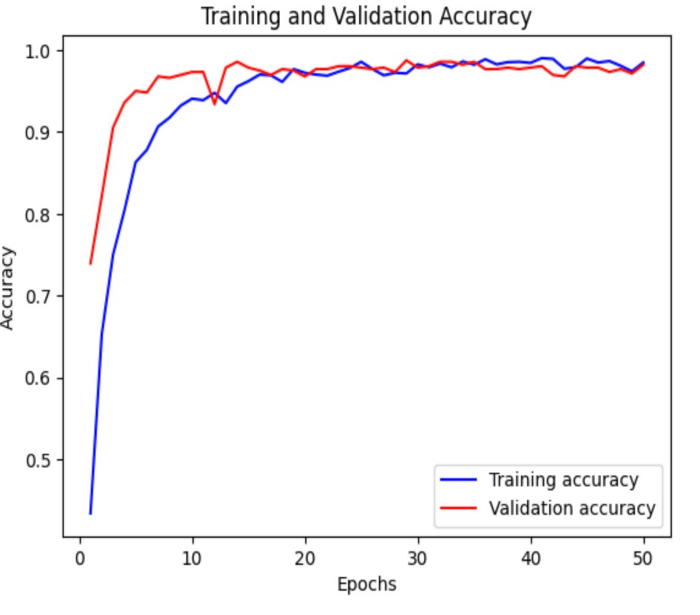
Fig 2: BI-LSTM Model Loss Graph

Fig 3: BI-LSTM Model Accuracy Graph

V. CONCLUSION

In conclusion, our project "Vocal Vibe: The application "Empowering Mental Health through Speech Analysis and Support" applied advanced algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to process speech data and mentally depressive states. We can emulate these components of our brain that are responsible for audio perception by exploiting CNNs for acoustic feature extraction and RNNs for temporal modeling, thus mimicking the way our brain processes speech data and detects emotional states. We seek to achieve a balanced model by ensuring ample training and optimization that will bring about the best system possible. The system will be able to provide meaningful and precise information on an individual's state of mind based on the patterns in their speech. Notwithstanding, this project aims to ultimately contribute to the development of mental health support systems by grabbing hold of the potential of speech analysis technology.

REFERENCES

[1] H. Chamod Welaratne and G. Shyanka Ratnayake, "Predictive Tool for Peer Support and Mental Health Improvement with Speech Emotion State Classification for University Students," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9824367.

[2] G. A. Chandra, K. K. Srinivas, P. Anudeep, S. R. Prasad, Y. Padmasai and P. Kishore, "Mental Health Disorder Analysis Using Convolution Neural Network Based Speech Signal Model With Integration Of Artificial Intelligence," 2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), Noida, India, 2021, pp. 544-547, doi: 10.1109/RDCAPE52977.2021.9633637.

[3] H. An, X. Lu, D. Shi, J. Yuan, R. Li and T. Pan, "Mental Health Detection from Speech Signal: A Convolution Neural Networks Approach," 2019 International Joint Conference on Information, Media and Engineering (IJCIME), Osaka, Japan, 2019, pp. 436-439, doi: 10.1109/IJCIME49369.2019.00094.

[4] J. Yuan, X. Lu, Y. Liu, D. Shi, T. Pan and Y. Li, "Depressive Tendency Recognition Using the Gated Recurrent Unit From Speech and Text Features," 2021 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2021, pp. 42-46, doi: 10.1109/IALP54817.2021.9675265.

[5] Lamers, Sanne & Truong, Khiet & Steunenberg, Bas & de Jong, Franciska & Westerhof, Gerben. (2014). Applying prosodic speech features in mental health care: An exploratory study in a life-review intervention for depression. 61-68. 10.3115/v1/W14-3208.

[6] Hobson H, Kalsi M, Cotton L, Forster M, Toseeb U. Supporting the mental health of children with speech, language and communication needs: The views and experiences of parents. Autism Dev Lang Impair. 2022 May 29;7:23969415221101137. doi: 10.1177/23969415221101137. PMID: 36124076; PMCID: PMC9479119.

[7] Paul L. Morgan, George Farkas, Marianne M. Hillemeier, Hui Li, Wik Hung Pun, Michael Cook, Cross-Cohort Evidence of Disparities in Service Receipt for Speech or Language Impairments, Exceptional Children, 10.1177/0014402917718341, 84, 1, (27-41), (2017).

[8] Madanian, Samaneh & Parry, David & Adeleye, Olayinka & Poellabauer, Christian & Mirza, Farhaan & Mathew, Shilpa & Schneider, Sandra. (2022). Automatic Speech Emotion Recognition Using Machine Learning: Digital Transformation of Mental Health.

[9] Brederoo SG, Nadema FG, Goedhart FG, Voppel AE, De Boer JN, Wouts J, Koops S, Sommer IEC. Implementation of automatic speech analysis for early detection of psychiatric symptoms: What do patients want? J Psychiatr Res. 2021 Oct;142:299-301. doi: 10.1016/j.jpsychires.2021.08.019. Epub 2021 Aug 16. PMID: 34416548.

[10] Chang, K. (2012). Speech Analysis Methodologies towards Unobtrusive Mental Health Monitoring. UC Berkeley. ProQuest ID: Chang\_berkeley\_0028E\_12254. Merritt ID: ark:/13030/m5xd15t3. Retrieved from https://escholarship.org/uc/item/0hb9h1qb