International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Voice to Text Emotion Detection System for Feedback Analysis

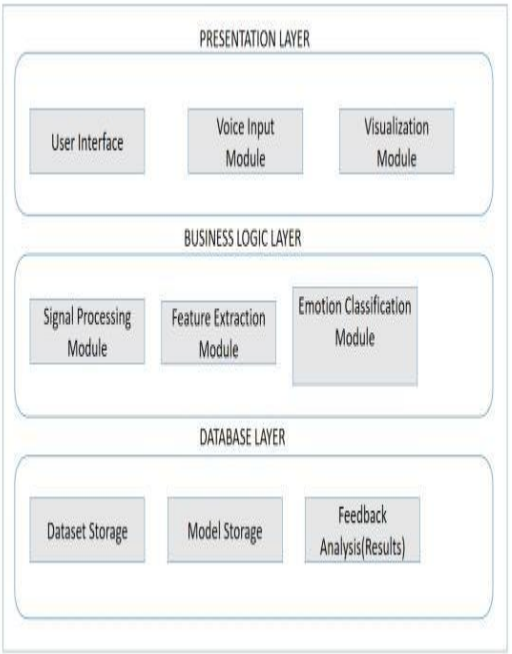
Sanjay UV[1], Dr.D.Swamydoss [2]
[1]department of Computer Application.
[2] Hod Department of Computer Application,
[1,2]adhiyamaan College of Engineering, Hosur.

ABSTRACT:

The Voice to Text Emotion Detection System for Feedback Analysis is designed to analyze emotions from spoken feedback by converting audio into text and identifying underlying sentiments. The system follows a structured pipeline: it processes the audio input using signal processing techniques, extracts key audio features such as pitch, tone, energy, and speech rate, and labels the data for training. A deep neural network (DNN) classifier is then used to detect emotions, followed by performance evaluation to ensure accuracy. The results help in understanding user sentiment, which can be applied in various domains like customer feedback analysis, mental health monitoring, and human- computer interaction. This system enhances feedback analysis by providing deeper insights into user emotions beyond textual data alone. **INTRODUCTION**

- Understanding user emotions is crucial in various fields like customer feedback, healthcare, and technology. Traditional text analysis often misses emotional nuances, making it harder to grasp true sentiment. The Voice to Text Emotion Detection System aims to fill this gap by analyzing spoken feedback and providing deeper insights.
- The system processes audio input using advanced signal processing techniques to extract essential features such as pitch, tone, and energy. These features carry emotional cues embedded in the speaker's voice, providing important context beyond words. These extracted features form the basis for accurate emotion detection.
- A Deep Neural Network (DNN) classifier is employed to interpret the emotional content of the processed audio. Trained on labeled data, the model learns to associate specific audio features with emotions like happiness, anger, and sadness. This enables the system to classify emotions effectively and accurately.
- Emotion detection in voice feedback can significantly improve customer feedback analysis, mental health monitoring, and human-computer interaction.

ARCHITECTURE DIAGRAM

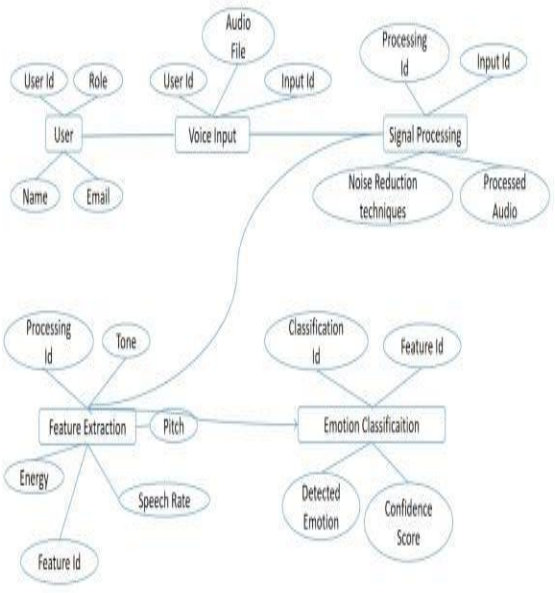


EXISTING SYSTEM

- Current emotion detection systems rely on natural language processing (NLP) to analyze textual feedback, categorizing it as positive, negative, or neutral. However, these systems often miss the emotional tone and nuances that are present in spoken language.
- Some systems convert spoken feedback into text using speech recognition and then analyze the sentiment. This approach overlooks important emotional cues in the speaker's tone, pitch, and speech rate, which are crucial for accurate emotion detection.Other systems combine basic audio signal processing with sentiment analysis, focusing on features like speech rate and volume. These systems can detect some emotional indicators but struggle with subtle emotional variations and may be less effective in noisy or diverse environments.

- Emotion detection in customer service and mental health monitoring often uses rule-based models to assess spoken feedback. While these systems can identify broad emotions, they are less accurate in real- time analysis and struggle with adapting to unexpected emotional expressions.

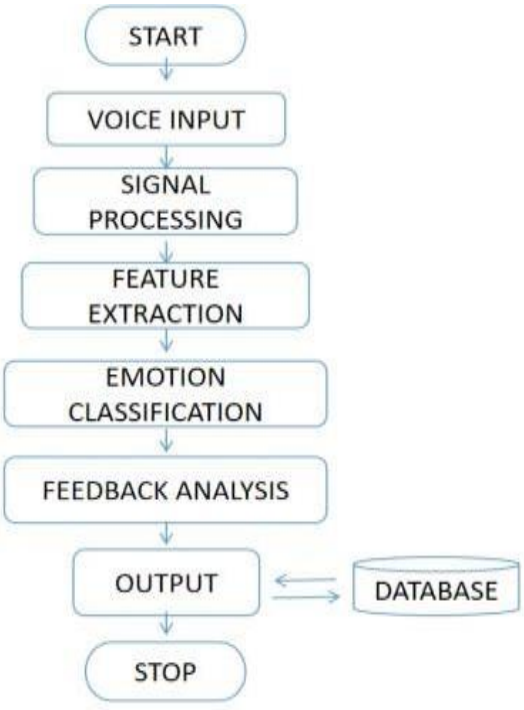
ER DIAGRAM



PROPOSED SYSTEM

- The proposed system aims to improve emotion detection by analyzing spoken feedback through a combination of advanced audio processing and deep learning techniques. Here's an overview of the proposed system:
- The system processes spoken feedback by first converting audio into text through speech recognition. Then, it extracts key features from the audio, such as pitch, tone, energy, and speech rate, which are crucial for understanding the emotional state of the speaker.
- A Deep Neural Network (DNN) classifier is used to analyze the extracted audio features and detect emotions. The model is trained on a labeled dataset, enabling it to classify emotions like happiness, anger, sadness, and surprise with high accuracy.
- The proposed system provides real-time emotion detection and feedback analysis, which can be applied in various domains such as customer service, mental health monitoring, and human-computer interaction. This real-time approach ensures more accurate insights into user emotions and sentiment.

DATA FLOW DIAGRAM



MODULE DESCRIPTION SPEECH SIGNAL ACQUISTION MODULE

- The Speech Signal Acquisition Module is responsible for capturing and preprocessing the audio input for emotion detection. Here's an overview of the module
- The module uses high-quality microphones or audio sensors to capture clear and accurate speech signals from the user. These sensors are optimized to pick up a wide range of frequencies, ensuring of the speaker and are critical for understanding underlying sentiments.that the audio data reflects the true characteristics of the speaker's voice.
- The acquired speech signal is then digitized and converted into a suitable format for further processing. This is typically done using an Analog-to-Digital Converter (ADC), which ensures that the signal is transformed into discrete data that can be analyzed.
- Noise reduction techniques are applied to the raw speech signal to remove background interference. This preprocessing step helps in obtaining cleaner audio data, which improves the accuracy of the subsequent emotion detection and feature extraction processes.

FEATURE EXTRACTION MODULE

- The Feature Extraction Module is responsible for extracting key characteristics from the processed speech signal that are essential for emotion detection. Here's an overview of the module:
- The module analyzes the speech signal to extract various features such as pitch, tone, energy, rhythm, and speech rate. These features reflect the emotional state

The system applies advanced algorithms such as neural networks to identify complex patterns in the data. By learningfrom a vast dataset of different emotional speech samples, the classifier can accurately distinguish between various emotional states based on subtle changes in pitch, tone, and energy levels

MODULE DESCRIPTION SPEECH SIGNAL ACQUISITION MODULE

- The Speech Signal Acquisition Module is responsible for capturing and preprocessing the audio input for emotion detection. Here's an overview of the module:
- The module uses high-quality microphones or audio sensors to capture clear and accurate speech signals from the user. These sensors are optimized to pick up a wide range of frequencies, ensuring that the audio data reflects the true characteristics of the speaker's voice.
- The acquired speech signal is then digitized and converted into a suitable format for further processing. This is typically done using an Analog-to-Digital Converter (ADC), which ensures that the signal is transformed into discrete data that can be analyzed.
- Noise reduction techniques are applied to the raw speech signal to remove background interference. This

preprocessing step helps in obtaining cleaner audio data, which improves the accuracy of the subsequent emotion detection and feature extraction processes.

FEATURE EXTRACTION MODULE

- The Feature Extraction Module is responsible for extracting key characteristics from the processed speech signal that are essential for emotion detection. Here's an overview of the module:
- The module analyzes the speech signal to extract various features such as pitch, tone, energy, rhythm, and speech rate. These features reflect the emotional state

EMOTION CLASSIFICATION MODULE

- The Emotion Classification Module is responsible for analyzing the extracted features from the speech signal and determining the

emotional state of the speaker. Here's an overview of the module:

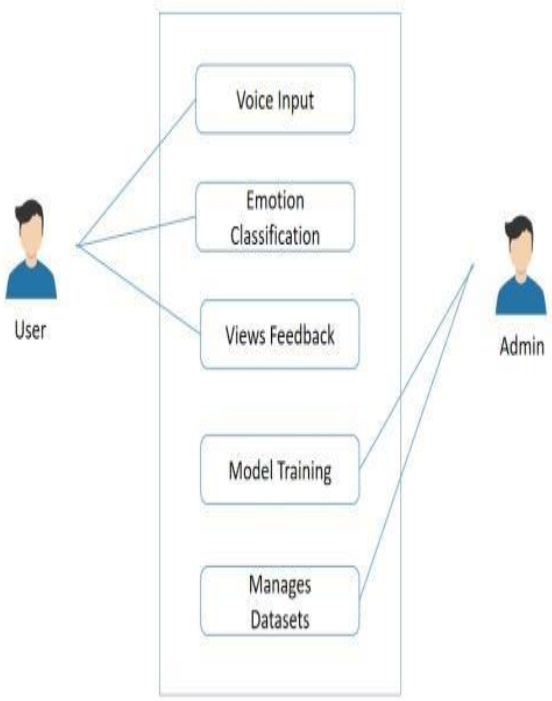
The module uses machine learning models, particularly a Deep Neural Network (DNN) or other classifiers, to analyze the extracted features and classify the emotions. These models are trained on labeled data, which helps the system learn the relationship between audio features and emotions like happiness, anger, sadness, or surprise.

- The system applies advanced algorithms such as neural networks to identify complex patterns in the data. By learning from a vast dataset of different emotional speech samples, the classifier can accurately distinguish between various emotional states based on subtle changes in pitch, tone, and energy levels.

MODEL TRAINING AND OPTIMIZATION MODULE

- The **Model Training and Optimization Module** is responsible for training the machine learning model, improving its accuracy, and ensuring it performs effectively in real-world scenarios. Here's an overview of the module:
- The module begins by preparing a large labeled dataset containing various speech samples with known emotional labels. The features extracted from these samples are used to train the model, allowing it to learn the relationship between speech characteristics (like pitch, tone, and energy) and emotions such as happiness, anger, or sadness.

USE CASE DIAGRAM



LITERATURE SURVEY

explored various deep-learning algorithms t o study the trends in text-based emotion recogn ition and found that bidirectional LSTMs outpe rform simple LSTMs. The model was found to do well when combined with bidirectional proc essing, dropout regularization, and weighted lo ss functions to manage the imbalances in the d atasets. The idea of using Bi-LSTMs in our Te xt Emotion Recognition model has been inspir ed by this paper. The authors of [2] worked on an end-to-end multimodal system that was ope rated on REmote COLlaborative and Affective (RECOLA) database to recognize spontaneous emotions using deep neural networks on raw speech and visual data. The proposed model show cased superior performance when compared to unimodal models, affirming the value of com bining speech and visual data for emotion reco gnition.In [3] the authors delved into various d eep learning-based architectures including LST M, CNNs, multi-layered Perceptron, and techn iques such as Adam, Attention-based RNN enc oders to operate on

IEMOCAP dataset for spee ch, text, and facial data. Their multimodal mod el provided an accuracy of 71.04%.The authors of [4] compared two multimodal emotion reco gnition models – deep canonical correlation an alysis (DCCA) and bimodal deep autoencoder (BDAE) across five different emotion recogniti on datasets- SEED-IV, SEED-V, DEAP, MAH
NOB-HCL, and AMIGOS. Wei Liu and their c o-auth ors observed that DCCA achieves an ave rage accurac y of 80.5% while BDAE achieved 77.3%. These resul ts were also compared with a traditional approach whi ch was 73.8% accura te. DCCA shows better robustne ss to noise and missing data than BDAE and traditiona l approa ches. The paper [5] proposes an approach to i m prove emotion recognition by using an attentio n m echanism and sequence model. The propos ed approa ch is operated on the two modalities. The results sho w that the proposed achievestat e-of-the-art performa nce on the IEMOCAP dat aset. The proposed approac h with the Oracle te xt achieves the best results in the dataset, showing that further Improvement can be ach ieved by more accurate speech recognition.

REFERENCE

[1] Kratzwald, B., Ili?, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems, 115, 24-35. https://doi.org/10.1016/j.dss.2018.09.002
[2] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017), “End-to-End Multimodal Emotion Recognition using Deep Neural Networks”,ArXiv,
https://doi.org/10.1109/JSTSP.2017.2764438
[3] Tripathi, S., Tripathi, S., & Beigi, H. (2018), “Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning”, ArXiv, /abs/1804.05788
[4] W. Liu, J. -L. Qiu, W. -L. Zheng and B. -L. Lu, \"Comparing Recognition Performance and Robustness of Multimodal Deep Learningn the two modalities. The results show that the proposed a14, no. 2, pp. 715-729, June 2022, doi: 10.1109/TCDS.2021.3071170.
[5] Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., & Li, X. (2019),” Learning Alignment for Multimodal Emotion Recognition from Speech”, ArXiv, /abs/1909.05645.Models for Multimodal Emotion Recognition,\" in IEEE Transactions on Cognitive and Developmental Systems, vol.