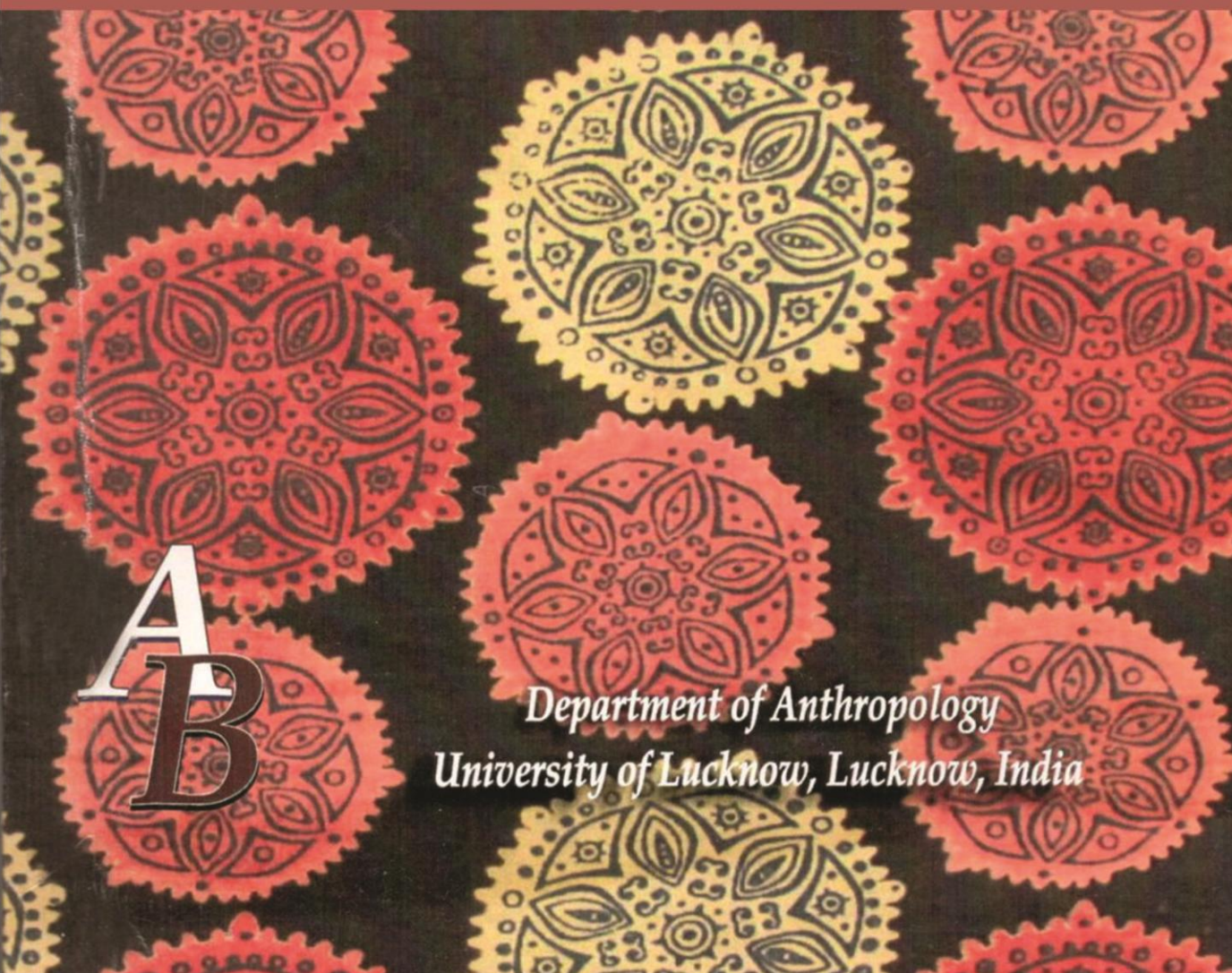
Volume 13, Issue 1
July 2019

ISSN No.:2348-4667

Anthropological Bulletin

a peer reviewed international journal



ANTHROPOLOGICAL BULLETIN
Vol. 13, Issue 1, July, 2019

CONTENTS

ISSN: 2348-4667
Anthropological Bulletin
13(1), 67-75, 2019

Speech Recognition: An Integrated Approach

Sehba Masood♣

ABSTRACT

The main aim of the proposed study is to combine speech recognition, sentiment analysis, and emotion detection into an integrated platform. This innovative web application-based system is designed to enhance user interaction through advanced AI technologies. The primary function of the proposed system is to accurately transcribe spoken language into text using Google Cloud Speech-to-Text, the application ensures precise and efficient transcription of audio data. Additionally, the system analyzes the underlying emotions conveyed through spoken words. Bidirectional LSTM networks were chosen due to their proficiency in modeling long-term dependencies and context within sequential data. This feature provides users with valuable insights into the emotional nuances of the transcribed text.

Keywords: Speech Recognition, audio, sentiment analysis, emotion detection, NLP, LSTM networks.

1. INTRODUCTION

Speech Recognition technology enables the user to operate the electronic device through spoken word rather than using different devices like keyboard. To make the use of the device through speech easy the speech recognition systems convert spoken words and phrases into machine-readable format. Speech recognition technology was designed to accelerate method of typing and was originally developed for physically or mentally challenged people who find typing difficult, tedious, or even impossible and this helps them to work by speaking. The domain of speech recognition is a fast-growing research area with broad practical application in marketing, banking, healthcare, language learning and so on. Some of the existing systems which use speech recognition are Apple Siri, Google Assistant, Microsoft Cortana and Amazon Alexa. Currently, the research in speech recognition is being carried out with various technologies like deep learning, Natural Language Processing (NLP), IOT etc.

Although speech recognition technology has made significant progress in recent years, it still faces critical challenges in accurately transcribing spoken language and understanding

♣ Assistant Professor, Department of Computer Science, Aligarh Muslim University, India
Email: sehba.masood.cs@amu.ac.in

the underlying emotional and sentiment context. Variations in accents, background noise, and complex linguistic nuances often lead to errors and misinterpretations. Additionally, current systems lack the ability to fully grasp and respond to the emotional tone of conversations, which is essential for applications such as customer service, virtual assistants, and mental health support. This paper proposes an Integrated Speech Recognition approach to address these challenges.

The main goal of the proposed system is to identify WHO is speaking, WHAT was spoken and HOW it is spoken. It aims to redefine user interactions with technology by seamlessly integrating speech recognition, sentiment analysis, and emotion detection functionalities. A key feature of this system is its ability to transcribe audio inputs into text before conducting emotion and sentiment analysis.

Major Contributions

The major contributions of the study are mentioned below:

- Convert spoken language into text for input.
- Determine the sentiment (positive, negative, or neutral) expressed in the speech using Deep Learning model.
- Recognize emotions (e.g., happiness, sadness) conveyed in the speech.
- Provide visual feedback on sentiment and emotion analysis results.

2. RELATED WORK

To develop an advanced Speech Recognition Assistant, understanding the evolution and advancements in speech transcription, sentiment analysis, and emotion detection technologies is essential. This literature survey delves into the historical development, working principles, and advantages and disadvantages of these technologies, providing a foundation for their integration into a cohesive system.

The journey of speech transcription technology began with simple rule-based systems in the early 20th century. The introduction of Hidden Markov Models (HMMs) in the 1970s marked a significant advancement, enabling better handling of time-sequenced data.

The real breakthrough in speech transcription came with the advent of deep learning techniques, mainly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. Hannun et al. in 2014 introduced Deep Speech, a deep learning-based approach using RNNs that significantly improved transcription accuracy. However, it still faced challenges with accents and background noise. Amodei et al. in 2016 further improved performance with Deep Speech 2, leveraging larger datasets and advanced training techniques.

Google's Deep Speech, utilizing RNNs, set new benchmarks by achieving high accuracy even in noisy environments. The integration of Google Cloud Speech-to-Text in modern applications leverages these advancements, providing robust real-time transcription capabilities. Despite its high accuracy, speech transcription still faces challenges with accents, dialects, and homophones.

Sentiment analysis has emerged from rule-based systems to refined machine learning models. Initial methods lacked context understanding, leading to limitations in detecting subtleties in language. Machine learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), improved sentiment classification accuracy. Socher et al. in 2013 introduced a recursive deep model capturing compositional effects in language, although requiring large, annotated datasets and facing challenges in detecting sarcasm and irony.

Emotion detection has seen advancements through deep learning, Bidirectional Long Short-term memory (LSTM) networks, and transformer models like BERT. Felbo et al. in 2017 introduced DeepMoji, leveraging transfer learning for accurate emotion detection, albeit demanding extensive training data. Poria et al. 2017 proposed a multimodal approach integrating text, audio, and visual data, enhancing emotion detection effectiveness but increasing computational costs. These sophisticated models developed using tools like TensorFlow and Keras, face challenges in distinguishing closely related emotions and understanding subtle emotional expressions.

Table 1. Comparative analysis of previous work

Reference	Technique Used	Contribution	Limitation
Hannun et al. (2014)	Deep Speech (RNN)	Significant improvement in transcription accuracy with deep RNNs	Struggles with accents and background noise
Amodei et al. (2016)	Deep Speech 2 (Bidirectional LSTM-RNN)	Enhanced performance using larger datasets and advanced training techniques	Computationally intensive and requires substantial data for training
Socher et al. (2013)	Recursive Deep Model	Improved sentiment prediction by capturing compositional effects	Requires large annotated datasets and may struggle with sarcasm and irony
Felbo et al. (2017)	DeepMoji (Transfer Learning)	High accuracy in emotion detection using transfer learning	Computationally intensive and requires extensive training data

In this study, following research gaps have been identified:

- **Accuracy in Diverse Conditions:**

Modern speech recognition systems, while improved, still face challenges with different accents, dialects, and background noise. These issues can lead to errors in transcription, making the systems less reliable in real-world scenarios. Misunderstandings caused by complex speech patterns and similar-sounding words can also occur, further complicating effective communication.

- **Lack of Emotional and Sentiment Analysis:**

Most current systems focus solely on transcribing speech to text without analyzing the emotional tone or sentiment behind the words. This gap means they can't respond appropriately to the speaker's emotional state. For example, in customer service, recognizing whether a customer is frustrated or happy could significantly enhance the interaction and response.

This paper aims to fill the identified research gaps by proposing an Integrated Speech Recognition System capable of accurately transcribing speech while simultaneously analyzing sentiment and detecting emotions.

3. THE PROPOSED SYSTEM

The proposed system presents a ground-breaking approach by integrating speech-to-text conversion, sentiment analysis, and emotion detection into a unified platform. Users have the flexibility to either upload their voice for automatic transcription or input text directly. Once the text is obtained, the system seamlessly conducts sentiment analysis to evaluate the emotional context, followed by sophisticated emotion detection algorithms to identify nuanced emotions expressed within the text. By consolidating these functionalities, the proposed system streamlines the user experience, eliminating the need for multiple tools and

providing a comprehensive solution for understanding and analyzing spoken and written communication.

The proposed system offers the following functionalities accessible to individual users who interact with the system:

- **Speech-to-Text Conversion**: Users can upload audio files or speak directly into the system, which then converts their spoken words into written text.
- **Text Input:** Users can input text directly into the system for sentiment analysis and emotion detection.
- **Sentiment Analysis:** The system analyses the text input by users to determine the sentiment expressed, such as positive, negative, or neutral.
- **Emotion Detection:** Emotion detection algorithms identify and analyze the emotions conveyed in the text, providing insights into the emotional tone of the communication.
- **Data Visualization:** Users may have access to visualizations or graphical representations of sentiment and emotion analysis results to aid in understanding and interpretation.

3.1 Dataset

A dataset for emotion detection and sentiment analysis was obtained from Kaggle. The dataset "***tweet_emotion.csv***" contains 40,000 entries and consists of four columns:

1) **tweet_id**: An integer representing the unique identifier for each tweet.
2) **Sentiment**: A categorical variable representing the sentiment expressed in the tweet. Possible sentiments include "empty", "sadness", "enthusiasm", "neutral", worry, etc.
3) **Author**: A string representing the username of the tweet's author.
4) **Content**: A string containing the text of the tweet.

The dataset can be used for tasks such as sentiment analysis, natural language processing, and emotion detection in text. This dataset contained uncleaned text samples sourced from twitter platform. The dataset served as the foundation for training the models for emotion detection and sentiment analysis.

The critical component of this study is Natural Language Processing (NLP), allowing the system to fathom and interpret human language. Following the dataset acquisition, extensive preprocessing operations were performed to clean and standardize the text data. These operations included text normalization, tokenization, and removal of stop-words, punctuation, and special characters. Tokenization breaks down the text into smaller, which are then converted into numerical representations. Additionally, techniques like stemming or lemmatization were applied to ensure consistency and uniformity in the dataset. Feature extraction techniques, capture the semantic meaning of words, allowing the model to fathom context and sentiment/emotion. The preprocessing phase was essential to prepare the data for effective training of the model and to enhance the accuracy and reliability of the emotion detection and sentiment analysis algorithms.

3.2 Modelling Architecture for Deep Learning Model

The deep learning model architecture for sentiment analysis and emotion detection was carefully crafted to achieve optimal performance. Bidirectional LSTM networks were chosen for their ability to capture past and future context and long-term dependencies in sequential data effectively. Bidirectional Long Short-Term Memory (BiLSTM) is an advanced version of the Recurrent Neural Network (RNN) which employs two independent LSTM layers: 1) Forward LSTM that processes the sequence from the first element to the last, and 2)

Backward LSTM that processes the exact same sequence in reverse, for retaining information across longer sequences. The architecture includes multiple layers of LSTM units with dropout regularization to prevent overfitting. This model was trained on the pre-processed text data to classify sentiments and detect emotions accurately. Additionally, visualizations were created to provide insights into the detected emotions and sentiment analysis results.

Figure 1: Speech Recognition System

For the successful implementation of the proposed System, the following input requirements were essential:

Audio Data Input:

- **Format:** The system should accept audio inputs in various standard formats such as WAV.
- **Sampling Rate:** The audio input should have a sampling rate of at least 16 kHz for optimal performance.
- **Channel:** Mono or stereo channels, with mono being preferable for better speech recognition accuracy.
- **Duration:** The system should handle audio inputs ranging from short clips (a few seconds) to longer recordings (up to few minutes).

Language Support:

- **Languages:** Initially, the system will support English. Additional language support may be added based on demand.
- **Accents and Dialects:** The system should recognize and accurately transcribe various accents and dialects within the supported languages.

Environment Considerations:

- **Noise Levels:** The system should be capable of handling background noise up to a certain threshold. However, extremely noisy environments may require additional noise-cancellation hardware.
- **Microphone Quality:** A good-quality microphone is recommended to ensure clear audio input. Built-in laptop microphones or external USB microphones are preferable.

User Interface Input:

- **Text Input:** The interface should allow users to input text manually for cases where audio input is not feasible.

The web application is designed to be intuitive and user-friendly. Users can input text, record or upload audio files directly through the interface. The submitted text or audio is processed, and the extracted features are fed into the machine learning model. The model's predictions are then displayed to the user.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed system has been evaluated on Kaggle Notebook which operates as a fully interactive and cloud based Jupyter environment. Following Python Libraries and Modules for preprocessing, feature engineering, and model training has been utilized for implementation:

SpeechRecognition: Python library for speech recognition functionality.

Scikit-learn: Machine learning library for sentiment analysis and machine learning tasks.

NLTK (Natural Language Toolkit): Python library for natural language processing tasks.

TensorFlow: TensorFlow for building deep learning model architectures

The Google Speech-to-Text API was integrated for speech recognition functionality. The coding process involved writing scripts for preprocessing of data, model training, and integration of various components into a cohesive system. Extensive testing and debugging were conducted to ensure smooth integration and proper functioning of the speech recognition system. VSCode has been used for developing the user interface. Flask has been utilized for integrating the machine learning model into the web application.

4.1 Evaluation Metrics

This study has evaluated the system's performance in transcribing speech to text, detecting sentiment, and identifying emotions, and analyze its effectiveness in providing accurate and meaningful insights. The metrics used for evaluation, to assess the performance of the system, include:

- **Precision:** Precision measures the proportion of correctly identified sentiments and emotions out of the total detections. It indicates the accuracy of the system in generating true positive detections.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- **Recall:** Recall measures the proportion of correctly identified sentiments and emotions out of the actual instances present in the speech data. It reflects the system's ability to detect true positive instances.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

- **Accuracy:** Accuracy measures the overall correctness of the system's transcriptions and detections, considering both true positive and true negative instances. It provides an overall assessment of the system's performance.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

- **F1 Score:** The F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both accuracy and completeness in the detection of sentiments and emotions.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 2. Performance Metrics of proposed Speech Recognition System

Accuracy	Precision	Recall	F1 Score
0.85	0.87	0.84	0.85

These results show the performance of the proposed Speech Recognition System on ***tweet_emotion.csv*** dataset, highlighting their accuracy, precision, recall and F1-score values. Accuracy value of 0.85 means that model correctly predicts about 85% of spoken words or commands. Precision score of 0.87 indicates that 87% of words or intents identified by the

model are correct. Recall value of 0.84 indicates that the model successfully identifies or transcribe 84% of all actual words or commands. F1 Score of 0.85 is the indicator of good performance of the system. The strong performance across multiple metrics demonstrates the effectiveness of this model in accurately transcribing speech to text, detecting sentiments and identifying emotions.

- **Emotion Detection:** The model distinguishes between the emotions based on the patterns it has learned during training, such as specific words, phrases, and context that are indicative of each emotion label (Neutral, Happy, Sad, Love, Anger). When the model processes an input text, it outputs a probability distribution over these five emotion labels using the SoftMax activation function. The label with the highest probability is selected as the predicted emotion. This probabilistic approach allows the model to account for the nuances and complexities of human language, making it capable of accurately detecting and classifying emotions based on the textual input.

- **Sentiment Analysis:** The model maps detected emotions to corresponding sentiments. Emotions are categorized into positive, negative, or neutral sentiments based on predefined groups. Positive sentiments include emotions like 'Happy' and 'Love,' while negative sentiments cover 'Sad' and 'Anger.' Neutral sentiment corresponds to the 'Neutral' emotion. This categorization is implemented using a function that takes the predicted emotion as input and returns the corresponding sentiment. By classifying emotions into broader sentiment categories, the system provides a more generalized understanding of the emotional tone of the text.

The following histograms represent the distribution of detected sentiments and emotions, providing a clear visual overview of the system's performance.

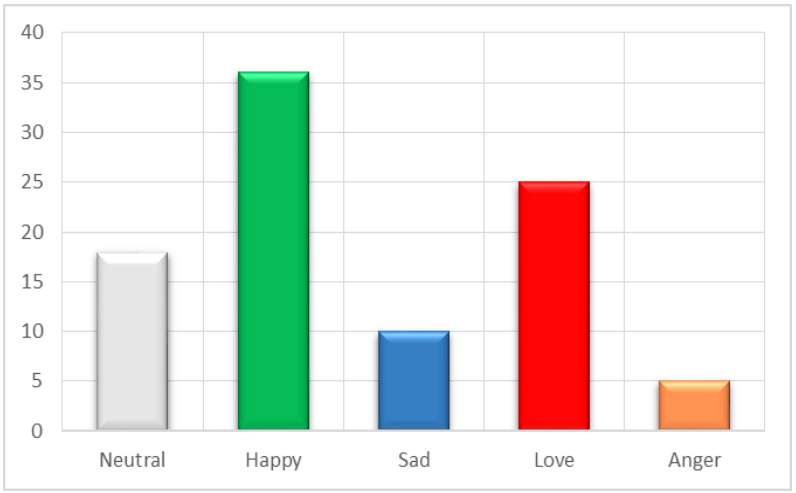


Figure 2. Histogram representing the distribution of detected emotions

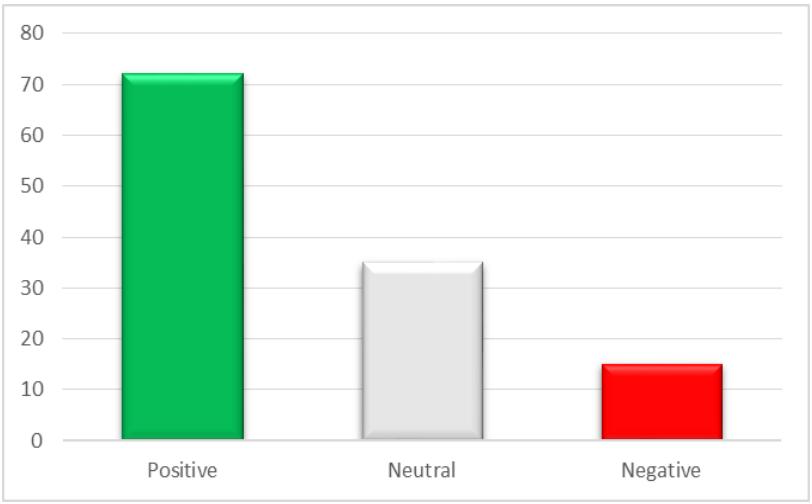


Figure 3. Histogram representing the distribution of detected sentiments

5. LIMITATIONS AND FUTURE SCOPE

Limitations:

While the system supports various accents and dialects, it may struggle with highly uncommon or very strong accents, leading to reduced transcription accuracy. Performance can be affected by high levels of background noise, which can interfere with speech recognition and sentiment analysis. The system may have difficulty understanding complex or context-dependent expressions, potentially leading to inaccurate sentiment or emotion detection. Emotion detection can sometimes be ambiguous, especially in cases where the context is neutral or mixed with multiple emotions. Initially, the system primarily supports English, with limited capabilities for other languages. Expanding language support requires additional development and training.

Future Scope:

To further enhance the capabilities of this system, several future extensions have been proposed. Develop advanced models and training techniques to better handle a wider range of accents and dialects, improving transcription accuracy. Implement more sophisticated noise reduction algorithms to minimize the impact of background noise on speech recognition and analysis. Integrate additional data sources such as facial expressions and body language to provide a more comprehensive analysis of emotions and sentiments. Extend support to multiple languages and dialects by training models on diverse linguistic datasets, making the system more accessible globally. Improve the system's ability to understand and analyze context-dependent speech by incorporating context-aware natural language processing techniques. Enhance integration capabilities with other applications and platforms, such as customer relationship management (CRM) systems, educational tools, and healthcare management systems. Develop advanced analytics and reporting features that provide deeper insights into sentiment trends, emotional patterns, and overall communication effectiveness implement machine learning models that continuously learn and adapt based on user interactions and feedback, ensuring the system remains up to date with evolving speech patterns and expressions. Create mobile applications for both iOS and Android platforms, allowing users to access the system's features on-the-go.

6. CONCLUSION

The Integrated Speech Recognition System offers several advantages and special features that set it apart from other similar systems. The proposed system utilizes advanced speech recognition technology, to provide highly accurate transcriptions even in diverse and noisy environments. It is capable of real-time speech-to-text conversion and sentiment analysis, ensuring immediate feedback and insights. It employs sophisticated Bi-LSTM networks to analyze sentiments (positive, negative, neutral) and detect a wide range of emotions (happiness, sadness, anger, etc.) from speech data. This system displays results with clear visualizations, including emotion detection templates, sentiment types, and histograms, to facilitate easy interpretation of data. Its intuitive and easy-to-use interface allows users to effortlessly record, upload, and analyze speech data. It is suitable for a wide range of applications across various industries, such as customer service, healthcare, education, business meetings, media, legal, and market research.

7. REFERENCES:

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng, 2014, "Deep Speech: Scaling up end-to-end speech recognition" Published in arXiv.org, https://doi.org/10.48550/arXiv.1412.5567

Bjarke Felbo, Alan Mislove, Anders Sogaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, etc., 2016, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin", published in *Proceedings of 33rd International Conference on Machine Learning*, PMLR 48:173-182.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, Amir Hussain, 2017, "A review of affective computing: From unimodal analysis to multimodal fusion", An International Journal on Multi-Sensor, Multi-Source Information Fusion, Volume 37, Pages 1-156.

Web-based resources:

1. https://app.diagrams.net/
2. https://codebasics.io/
3. https://monkeylearn.com/
4. https://stackoverflow.com/
5. https://www.analyticsvidhya.com/
6. https://www.geeksforgeeks.org/
7. https://www.kaggle.com/
8. https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text/
9. https://www.techtarget.com/

****