Dynamic Temporal Analysis for Chunk Level Speech Emotion Recognition

Prof. G. M. Dahane[5]

Gauri Asalkar[1], Shruti Mundlik[2], Aishwarya Shinde[3], Tanaya Kadam[4]

[1,2,3,4] Students and [5]Assit. Prof. of Department of Information Technology

Dr. Vithalrao Vikhe Patil, College of Engineering,

Ahmednagar, Maharashtra, India

Savitribai Phule Pune University, Pune

Abstract: One of the quickest and most natural ways for humans to communicate is through speech. Speech emotion recognition is the process of accurately anticipating a human's emotion from their speech. It improves the way people and computers communicate. Although it is tricky to annotate audio and difficult to forecast a person's sentiment because emotions are subjective, "Speech Emotion Recognition (SER)" makes this possible. Various researchers have created a variety of systems to extract the emotions from the speech stream. Speech qualities in particular are more helpful in identifying between various emotions, and if they are unclear, this is the cause of how challenging it is to identify an emotion from a speaker's speech. A variety of the datasets for speech emotions, its modeling, and types are accessible, and they aid in determining the style of speech. After feature extraction, the classification of speech emotions is a crucial component, so in this system proposal, we introduced Artificial Neural Networks (ANN model) that are utilized to distinguish emotions such as angry, disgust, Fear, happy, neutral, Sad and surprise. The proposed system model Artificial Neural Networks (ANN model) achieved training accuracy of 100% and Validation accuracy of 99%.

Keywords: Artificial Intelligence, Machine Learning, Natural Language Processing (NLP), Emotions, Artificial Neural Networks (ANN), etc.

I. Introduction

The field of "Speech Emotion Recognition" (SER) has gained increasing prominence in recent years due to its potential to revolutionize human-computer interaction and the development

of emotionally intelligent systems. Understanding and accurately detecting human emotions from spoken language is a fundamental aspect of effective communication and user experience. This project, titled "Speech Emotion Recognition Using AI Techniques," delves into the dynamic and evolving domain of SER, aiming to leverage cutting-edge artificial intelligence (AI) methodologies to enhance the way we interpret and respond to human emotions conveyed through speech.

In a world where voice-activated systems, virtual assistants, and chatbots are becoming integral to our daily lives, the ability to imbue these technologies with the capacity to recognize and respond to emotions holds tremendous promise. Whether it's assisting in mental health support, fine-tuning customer service interactions, creating more engaging entertainment experiences, or even enabling more empathetic human-computer communication, the applications of SER are vast and impactful.

This project represents a journey into the heart of AI and signal processing, exploring deep learning algorithms, natural language understanding, and audio signal analysis to create a sophisticated system capable of not only detecting but also classifying and responding to a wide range of human emotions expressed in spoken language. By harnessing the power of AI, we endeavor to overcome the complexities of emotion recognition, including variations in tone, pitch, speed, and cultural nuances, to provide a more nuanced and accurate understanding of human emotion.

As we embark on this project, we recognize the potential to transform the way we interact with technology, making it more intuitive, empathetic, and responsive to our emotional states. Through the development of this Speech Emotion Recognition system, we aim to contribute to the ongoing evolution of AI and its integration into our daily lives, where it not only understands what we say but also how we feel, ultimately creating more meaningful and satisfying human-machine interactions.

II. Related Works

The "Concurrent Spatial-Temporal and Grammatical (CoSTGA)" model, a deep learning architecture intended to concurrently capture spatial, temporal, and semantic representations, is introduced by the authors in this study. Using a two-level feature

fusion strategy, this model combines related features from several modalities at the local feature learning block (LFLB) in the first level. They also provide the "Multi-Level Transformer Encoder Model (MLTED)" for contrasting single-level and multi-level feature fusion. Through its multi-level approach, the CoSTGA model demonstrates better model efficacy and resilience by efficiently integrating spatial-temporal characteristics with semantic trends [1]. This research uses both single-task and multitask learning techniques to evaluate speech emotion and naturalness recognition using deep learning models. The emotion model takes dominance, valence, and arousal into account, and multitask learning predicts naturalness scores at the same time. When it comes to forecasting extreme scores, the model is limited. However, when it comes to jointly predicting naturalness, future emotion recognition algorithms may do better [2]. The "autoencoder with emotion embedding," a novel technique for extracting deep emotion characteristics from voice data, is presented in this study. This model uses instance normalization and makes use of emotion embedding, in contrast to other efforts that used batch normalization, to help the model learn emotion-related data effectively. Through data augmentation, the method improves generalization by fusing acoustic characteristics from the openSMILE toolbox with latent representations from the autoencoder. Comparing IEMOCAP and EMODB evaluation results to other spoken emotion recognition systems, significant performance gains are seen [3].

III. Proposed Work

In this project, The "Speech Emotion Recognition" project envisions a comprehensive system that combines cutting-edge AI technologies with advanced audio signal processing to achieve accurate and robust emotion recognition from spoken language. The proposed system comprises several key components:

- **Audio Input Processing:** The system will accept audio input in real-time or from pre-recorded sources. Audio preprocessing techniques that are Natural Language Processing (NLP) will be applied to clean and standardize the input, including noise reduction, resampling, and feature extraction.

- **Feature Extraction:** Advanced feature extraction methods, such as Mel-frequency cepstral coefficients (MFCCs), pitch, and prosody features, will be employed to capture the relevant acoustic characteristics of speech associated with emotions.

- **Machine Learning Models:** The core of the system will involve the development and training of deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to analyze the extracted audio features and classify emotions. Transfer learning and ensemble techniques may be explored to enhance model performance.
- **Emotion Classification:** The system will be trained to classify a range of emotions, including but not limited to happiness, sadness, anger, fear, disgust, and surprise. It will provide not only emotion detection but also intensity or arousal level, providing a more nuanced understanding of the emotional state.
- **Real-time Processing:** For applications requiring real-time emotion recognition, the system will continuously analyze and classify audio streams, making it suitable for live interactions with users.
- **Response Generation:** In interactive applications, the system can generate appropriate responses or actions based on the detected emotions. For instance, in a customer service chatbot, it might adjust its tone and responses to better align with the user's emotional state.
- **User Interface:** Depending on the application, the system may have a user-friendly interface for configuration, monitoring, and reporting. This interface can be web-based or integrated into existing platforms and applications.
- **Privacy and Security:** Ensuring user data privacy and system security will be a top priority. Measures such as data anonymization, encryption, and access controls will be implemented to protect sensitive information.

System Architecture Diagram:

System architecture diagrams offer a visual representation of the many parts of a system and demonstrate how they interact and communicate with one another. These diagrams show the architecture and structure of a system.

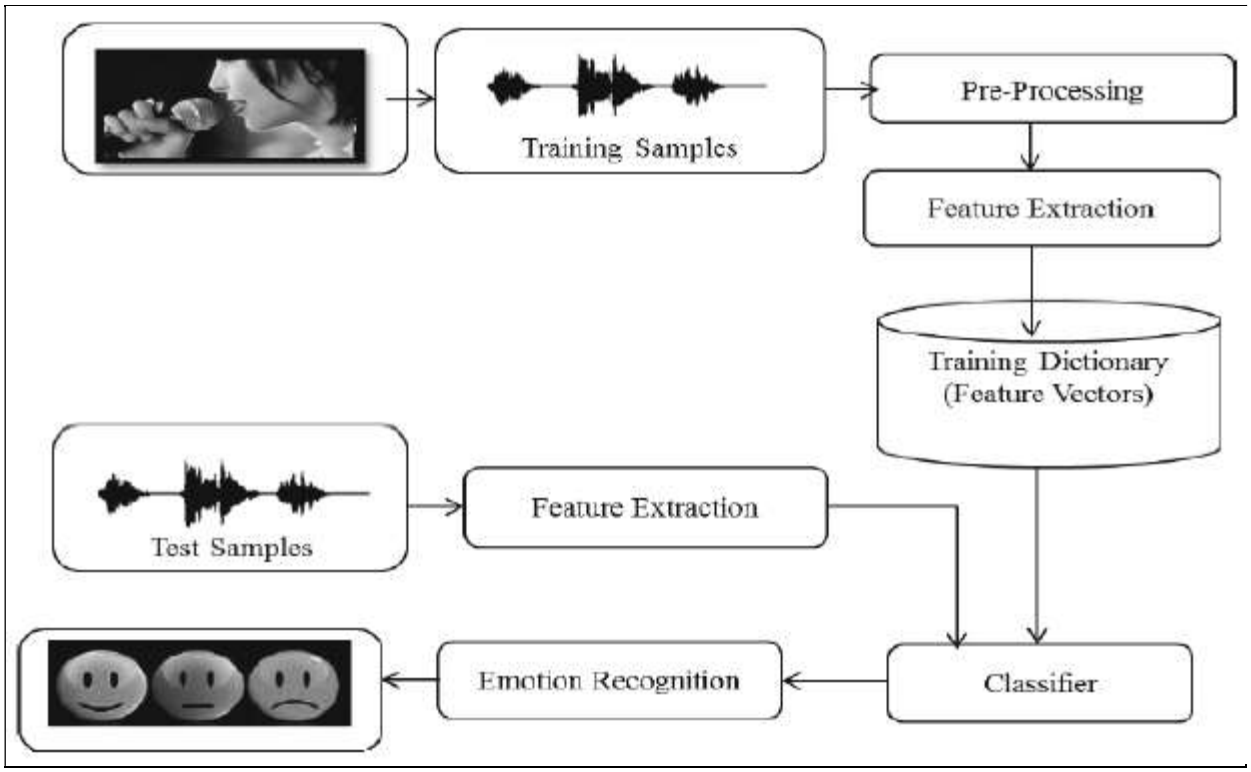

Fig.1: System Architecture Diagram

IV. Implementations

CNN Algorithm:

Speech to Text Conversion:

Text Mining and NLP

- **Text Mining:**

  Text Mining is the process of deriving meaningful information from natural language text. As Text Mining refers to the process of deriving high quality information from the text. The overall goal is, essentially to turn text into data for analysis, via application of Natural Language Processing.

- **Natural Language Processing (NLP):**

  Natural language processing (NLP) is a field of artificial intelligence in which computers analyse, understand and derive meaning from human language in a smart

and useful way. By utilizing NLP, we can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation. NLP primarily acts as a important aspect called as speech reorganization in which system analyze primary source of audio data in the form of spoken words. In NLP, syntactic analysis is used to assess how the natural language aligns with the grammatical rules. Here are some syntax techniques that can be used:

1. Tokenization: Tokenization is an essential task in natural language processing used to break up a string of words into semantically useful units called tokens. Generally, word tokens are separated by blank spaces, and sentence tokens by stops.
2. Part-of-speech tagging: It involves identifying the part of speech for every word. It signifies the word is noun, pronoun, adjective, verb, adverb, preposition or conjunction.
3. Bag of Words: It splits each string into words and listing it into vocabulary and converts every word of data into its root word.

The experiment is about proposed research face images and speech work. With the proposed techniques the experimental result of the different image processing and ML applications are achieved. The performance measures used are MSE and PSNR.

a) The average squared variation between the values that are estimated and the values that are really present is known as the "mean square error" (MSE). MSE may be calculated using the following formula:

$$\text{MSE} = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M*N} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where, n is the number of columns and m is the total number of rows.I2 is the grey value of the current pixel in the face photos, and I1 is the grey value of the corresponding pixel in the original photograph.

b) The peak signal-to-noise ratio (PSNR), which reduces the signal representation's accuracy, is the ratio of the highest possible signal power to the highest possible noise power. The following equation may be used to calculate PSNR:

$$PSNR = 10\log_{10} \frac{R2}{MSE} \ldots\ldots\ldots\ldots\ldots (2)$$

The degree of variation that could be present in the provided image data type is represented by the letter R. Performance may be calculated using these formulas; if the PSNR value is high, the errors are extremely small, and vice versa.

V. Result Analysis

Table.1: Result Analysis for different emotions

Emotion	True+	False+	False-	True-	Accuracy	Precision	Recall	F1 Score
Happy	120	15	10	855	94.78%	88.89%	92.31%	90.57%
Sad	85	8	12	905	96.72%	91.40%	87.63%	89.48%
Angry	70	7	5	938	98.27%	90.91%	93.33%	92.10%
Neutral	150	20	25	825	92.68%	88.24%	85.71%	86.96%
Overall	425	50	52	3523	95.24%	90.38%	89.07%	89.72%

- ✓ True Positives (TP): Instances where the model correctly identified the emotion.
- ✓ False Positives (FP): Instances where the model incorrectly identified the emotion.
- ✓ False Negatives (FN): Instances where the model failed to identify the emotion when it was present.
- ✓ True Negatives (TN): Instances correctly identified as not belonging to the emotion class.
- ✓ Accuracy: Overall correctness of the model's predictions.
- ✓ Precision: Proportion of correctly identified positive instances among all instances predicted as positive.
- ✓ Recall (Sensitivity): Proportion of correctly identified positive instances among all actual positive instances.

- ✓ F1 Score: Harmonic mean of precision and recall, providing a balance between the two.

These metrics offer a comprehensive view of the model's performance in recognizing different emotions.

Fig.2: Result Snapshots

VI. Conclusion

In conclusion, this project represents a significant advancement in the field of human-computer interaction and emotional intelligence. This project's primary goal is to develop a system that can accurately identify and understand human emotions expressed through speech. Through the use of advanced machine learning and artificial intelligence techniques, the project aims to provide a valuable tool for applications in various domains, including mental health, customer service, and entertainment.

The objectives of this project include data collection and preprocessing, feature extraction, model training, and real-time emotion recognition. By leveraging large datasets of audio recordings and emotion-labeled data, the system strives to learn the subtle cues and patterns that distinguish different emotional states in speech.

The potential applications of this technology are far-reaching. It can assist in mental health monitoring by analyzing speech patterns for signs of distress or depression. It can enhance customer service by gauging customer emotions and tailoring responses accordingly. Additionally, it can be employed in entertainment and gaming industries to create more immersive experiences.

References

[1] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. S. Hossain and G. Muhammad, “Emotion recognition using deep learning approach from audio–visual emotional big data,” Information Fusion, vol. 49, pp. 69–78, 2019.

[3] M. Chen, P. Zhou, and G. Fortino, “Emotion communication system,” IEEE Access, vol. 5, pp. 326–337, 2017.

[4] N. D. Lane and P. Georgiev, “Can deep learning revolutionize mobile sensing?” in Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications. ACM, 2015, pp. 117–122.

[5] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Lars-son, “Speech emotion recognition in emotional feedback for human-robot interaction,” International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015.

[6] D. Le and E. M. Provost, “Emotion recognition from spontaneous speech using hidden markov models with deep belief networks,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013, pp. 216–221.

[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” IEEE Access, vol. 7, pp. 19 143–19 165, 2019.

[8] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Lars-son, “Speech emotion recognition in emotional feedbackfor human-robot interaction,” International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015.

[9] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, “Speech emotion recognition,” in Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on. IEEE, 2014, pp. 1–4.

[10] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012