Speech Emotion Recognition using NLP Algorithm

Shaila Pawar[1], Pratiksha Atkari[2], Atul Lamkhade[3]
[1,2&3] *Students, Samarth Group of Institutions College of Engineering, Belhe, Maharashtra, India*
E-mail: [1]pawarshaila883@gmail.com, [2]atkaripratiksha44@gmail.com, [3]atulklamkhade016@gmail.com

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 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 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 ArtifIicial 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.**

1.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, finetuning 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 humanmachine interactions.

2.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 multilevel 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

This is an open source journal

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 open SMILE 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].

3.Proposed Works

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

1.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.

2.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.

3.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.

4.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.

5.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.

6.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.

7.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.

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.

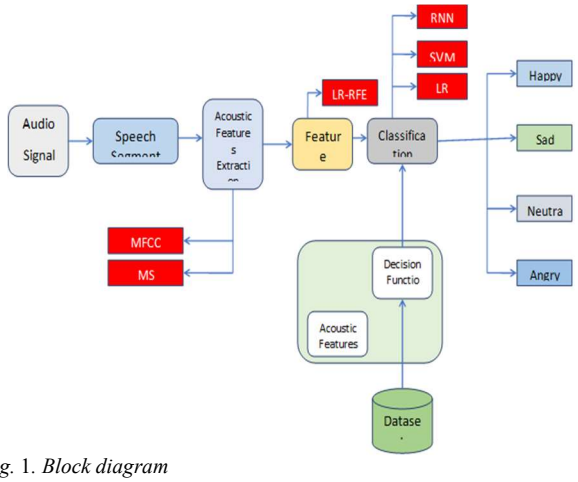
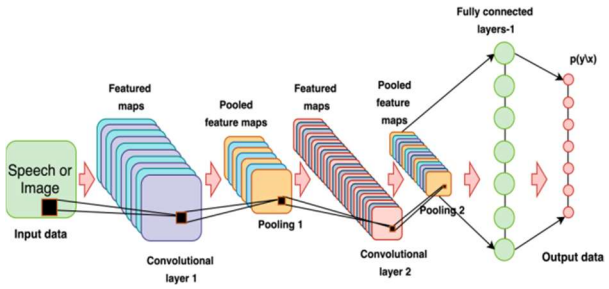


Fig. 1. *Block diagram*

4.Implementations

1.CNN Algorithm:

CNN is one of the main categories to do image recognition, image classification. Object detection, face recognition, emotion recognition etc., are some of the areas where CNN are widely used. CNN image classification takes an input image, process it and classify it under certain categories (happy, sad, angry, fear, neutral, disgust). CNN is a neural network that has one or more convolutional layers.



Steps:

Step 1: Dataset containing images along with reference emotions is fed into the System. The name of dataset is Face Emotion Recognition (FER)which is an open – source data set that was made publicly available on a Kaggle.

Step 2: Now import the required libraries and build the model.

Step 3: The convolutional neural network is used which extracts image features f pixel by pixel.

This is an open source journal

Step 4: Matrix factorization is performed on the extracted pixels. The matrix is of m x n.
Step 5: Max pooling is performed on this matrix where maximum value is selected and again fixed into matrix. **Step 6:** Normalization is performed where the every negative value is converted to zero.
Step 7: To convert values to zero rectified linear units are used where each value is filtered and negative value is set to zero.
Step 8: The hidden layers take the input values from the visible layers and assign the weights after calculating maximum probability.

2. Recurrent Neural Networks (RNNAlgorithm: -:- RNN-RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations

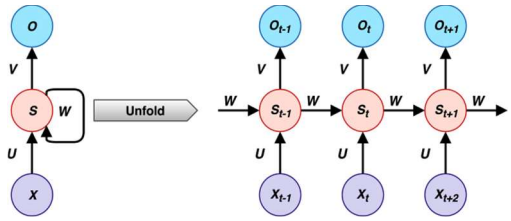


Fig :- Recurrent Neural Networks (RNN)

3.SVM (Support Vector Machine):

Support Vector Machine (SVM) is a very powerful machine learning algorithm. SVM is used for linear or nonlinear classification and regression. SVM is used for many tasks, such that text classification, image classification, face detection.SVM algorithm is a very effective for as we try to find maximum separate the hyperplane in between different classes.

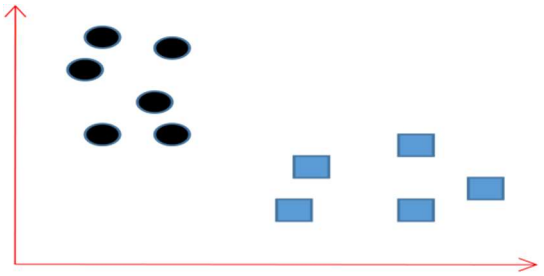


Fig:-A. Draw a line that separates black circles and blue squares

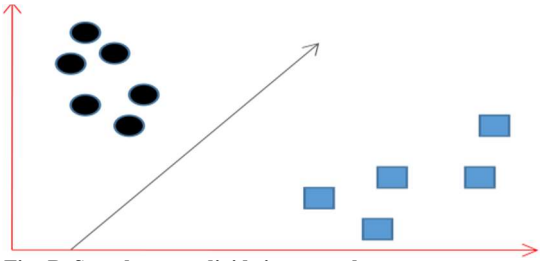


Fig:-B. Sample cut to divide into two classes.

4.MFCC(Mel-frequency cepstral coefficients):

MFCC is a feature extraction technique .it is widely used in speech and audio processing. This algorithm are used to represent the spectral characteristics of sound in a many way that issuited for various machine learning tasks, such as speech recognition and music analysis.

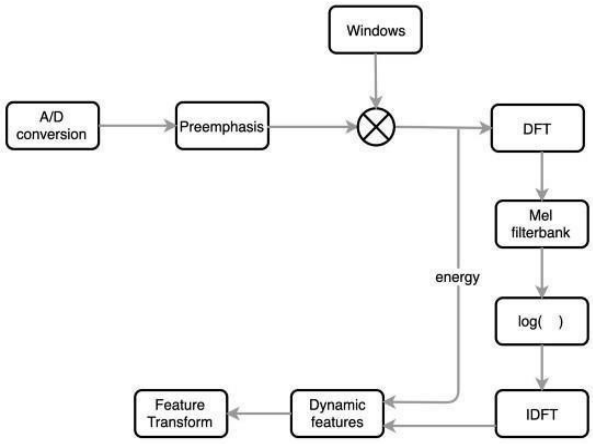


Fig.3. Mel-Frequency Cepstral Coefficients

Speech to Text Conversion:
Text Mining and NLP.

1.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.

2.Natural Language Processing (NLP):
Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, 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.

This is an open source journal

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:

$$MSE=\sum M,N[I_1(m,n)-I_2(m,n)]^2 M * N \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

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:

$$PSNR = 10\log_{10} R2MSE \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

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.

5.Result

Emotion	Happy	Sad	Angry	Neutral	overall
True +	120	85	70	150	425
False +	15	8	7	20	50
False -	10	12	5	25	52
True -	855	905	938	825	3523
Accuracy	94.7 %	96.7 %	98.27 %	92.68 %	95.2 %
Precision	88.8 %	91.4 %	90.91 %	88.71 %	90.3 %
Recall	92.3 %	87.63 %	93.33 %	85.71 %	89.0 %
F1 score	90.5 %	89.4 %	92.10 %	86.96 %	89.7 %

1. **True Positives (TP):** Instans **(TP):** Instances where thee model correctly identified the emotion..
2. **False Positives (FP):** Instances where the model incorrectly identified the emotion.
3. **False Negatives (FN):** Instances where the model failed to identify the emotion when it was present.
4. **True Negatives (TN):** Instances correctly identified as not belonging to the emotion class.
5. **Accuracy:** Overall correctness of the model's predictions.
6. **Precision:** Proportion of correctly identified positive instances among all instances predicted as positive.
7. **Recall (Sensitivity):** Proportion of correctly identified positive instances among all actual positive instances.
8. **F1 Score:** Harmonic mean of precision and recall, providing a balance between the two.

6.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. Using 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.

7.References

[1] Huiyun Zhang 1,2,3, Heming Huang 1,2,3, And Henry Han 1,4 Attention-Based Convolution Skip Bidirectional Long Short-Term Memory Network for Speech Emotion Recognition Received November 21, 2020, accepted December 14, 2020, date of publication December 25, 2020,date of current version January 11, 2021.

[2] Hao Meng1,Tianhao Yan1, Fei Yuan1, And Hongwei Wei1,Speech Emotion Recognition from 3D Log- Mel Spectrograms with Deep Learning Network.

[3] Taiba Majid Wani 1 , Teddy Surya Gunawan 1,3,, Syed Asif Ahmad Qadri 1 , Mira Kartiwi 2 , (Member, Ieee), And Eliathamby Ambikairajah 3 A Comprehensive Review of Speech Emotion Recognition Systems Received February 25, 2021, accepted March 10, 2021, date of publication March 22, 2021, date of current version April 1, 2021.

[4] Mohan Ghai, Shamit Lal, Shivam Dugga l and Shrey Manik Delhi Emotion Recognition On Speech Signals Using Machine Learning 2017 International Conference On Big Data Analytics and computational Intelligence.

[5] Raoudha Yahia Cherif, Abdelouahab Moussaoui, Nabila Frahta ,Mohamed Berrimi Effective Speech Emotion Recognition Using Deep Learning Approaches For Algerian Dialec Tauthorized Licensed Use Limited To: University Of Liverpool. Downloaded On July 03,2021 At 07:08:56 Utc From IEEE Xplore.

This is an open source journal

[6] Weijian Zhang, Peng Song, Member, IEEE, Dongliang Chen, Chao Sheng, Wenjing Zhang Cross-

[7] Speech Emotion Recognition Based on Joint Transfer Subspace Learning and Regression Journal Of Latex Class Files, September 2020.

[8] Youngdo Ahn , Student Member, IEEE, Sung Joo Lee, , Member, IEEE Cross- Corpus Speech Emotion Recognition Based on Few- Shot Learning and Domain Adaptation IEEE Signal Processing Letters, Vol. 28, 2021.

[9] J. Ancilin , A. Milton Improved speech emotion recognition with Mel frequency magnitude coefficient Received 1 July 2020 Received in revised form 11 February 2021Accepted 9 March 2021 Available online 26 March 2021

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] . Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015.

This is an open source journal