Speech Emotion Recognition
With sentiment analysis using deep learning

BRINDA DAVDA
LABDHI GATHANI
PRIYAL PANARA

Department of Information & Technology
V.V.P. Engineering College , Rajkot , India

I. ABSTRACT

In today's revolutionary world time is one of the most precious thing as said "time and tide waits for none". Deep learning has rapidly emerged as a powerful technology that has potential to save time, provide comfort and enable a wide range of visualization. This research paper proposes a combined approach for speech emotion recognition with sentimental analysis using deep learning. Speech emotion recognition (SER) is the process of identifying human emotions from spoken language. Sentimental analysis (SA) is the process of determining the emotional tone of a piece of text. The proposed system uses a deep learning model that takes both the audio signal and textual content as inputs to predict the emotional state of the speaker. The model is trained using a large dataset of speech samples labelled with corresponding emotions and their textual transcriptions. The combination of speech and textual features allows the model to better capture the nuances of human emotions and improve the overall recognition performance. The proposed approach has potential applications in various fields, such as healthcare, education, and customer service. The ability to accurately recognize human emotions can lead to the development of more effective communication systems that can better understand and respond to the needs of users. In conclusion, the proposed combined approach of speech emotion recognition with sentimental analysis shows promising results in accurately identifying human emotions from real time speech, and then converting that speech into text. Finally sentimental analysis will be carried out considering that text which has the potential to be used in various practical applications.

Keywords: speech emotion recognition, sentimental analysis, speech to text, deep learning,

II. INTRODUCTION

In the revolutionary world of today, deep learning is empowering machines to tackle complex problems and make decisions with greater accuracy and speed than ever before.
Deep learning is a key technology driving the ongoing revolution in artificial intelligence and machine learning [1]. With its ability to learn from large and complex datasets, deep learning is enabling machines to perform tasks that were previously thought to be the exclusive domain of human intelligence, such as image and speech recognition, natural language processing, and autonomous decision-making. These breakthroughs are transforming many industries, from healthcare and finance to transportation and entertainment, and are likely to have a profound impact on society in the years to come.

Deep learning is a type of artificial intelligence (AI) that involves training artificial neural networks to learn from vast amounts of data[1]. Neural networks are computing

systems inspired by the structure and function of the human brain, composed of layers of interconnected nodes or "neurons." The term "deep" in deep learning refers to the depth of the neural network, which can have many layers, each one processing a different level of abstraction. By combining multiple layers, deep learning models can learn increasingly complex representations of data and make more accurate predictions. This makes deep learning particularly well-suited for tasks such as image and speech recognition, natural language processing, and even game playing. Deep learning algorithms are designed to analyze and process complex, multi-layered data sets, such as images, videos, and speech, and extract features and patterns to make predictions or decisions.

One of the key advantages of deep learning is its ability to learn from unstructured or unlabeled data, which is a significant improvement over traditional machine learning methods. Deep learning has enabled breakthroughs in a wide range of fields, from computer vision and speech recognition to natural language processing and robotics. Some examples of deep learning applications include self-driving cars, facial recognition systems, virtual assistants, and medical image analysis.

A. APPLICATIONS

The applications of SER are as follows.[1][5][8]

- **Marketing**: SER can be used in marketing to analyze customer feedback and sentiment on social media. This information can be used to improve products and services and create more targeted marketing campaigns.
- **Social media monitoring**: Sentiment analysis is used to monitor social media conversations and detect trends and sentiments related to brands, products, or topics. This information can be used for reputation management, customer service, or market research.
- **Customer feedback analysis**: Sentiment analysis is used to analyze customer feedback and reviews to determine customer satisfaction levels and identify areas for improvement. This information can be used to improve products and services and enhance customer experiences.
- **Market research**: Sentiment analysis is used to analyze customer feedback and social media conversations to identify emerging trends, consumer preferences, and potential product opportunities.
- **Customer service**: SER can be used in customer service to analyze the tone of a customer's voice during a call to determine their emotional state. This information can be used to provide a personalized response and improve the customer experience.
- **Healthcare**: SER can be used in healthcare to detect emotional distress in patients. For example, it can be used to detect signs of depression, anxiety, or stress in a patient's voice during a telemedicine session.
- **Human-robot interaction**: SER can be used in human-robot interaction to improve the emotional intelligence of robots. For example, it can be used to detect the emotional state of a human interacting with a robot and adjust the robot's response to better match the human's emotional state.

III. SPEECH EMOTION RECOGNITION

A. EMOTION RECOGNITION

Emotion recognition is a technology [2]that uses machine learning algorithms to recognize and interpret human emotions. It involves analyzing various features, such as facial expressions, voice tone, body language, or physiological signals, to identify emotional states such as happiness, sadness, anger, fear, or surprise.

Following are the different methods of emotion recognition.

- **Mel-Frequency Cepstral Coefficients (MFCC):** This is a technique used to extract features from speech signals that are relevant to SER.[1] MFCC involves converting the speech signal into a logarithmic frequency scale, dividing it into frames, and calculating the spectrum of each frame.
- **Deep Neural Networks (DNNs**): This is a machine learning algorithm used to train models for speech recognition. DNNs involve creating a layered neural network that learns to recognize patterns in the input data.
- **Support Vector Machines (SVMs**): This is a machine learning algorithm that can be used to classify speech signals into different emotional states. SVMs involve creating a hyperplane that separates the input data into different categories.
- **Hidden Markov Models (HMMs**): This is a statistical model used to recognize speech signals. HMMs involve modeling the probability distribution of speech signals and calculating the most likely sequence of emotional states.
- **Convolutional Neural Networks (CNNs):** This is a type of deep neural network that is particularly effective for image and speech recognition.[2] CNNs involve creating a network of layers that extract features from the input data and classify it into different emotional states.

B. SPEECH TO TEXT

Speech-to-text technology, also known as speech recognition or voice recognition technology, is a technology that enables a computer or device to recognize and interpret spoken language and convert it into text [3]. This technology uses various algorithms and models to analyze speech and convert it into a written form that can be understood by a computer. Below given are the different approaches.

- **Automatic Speech Recognition (ASR**): This is a technique that involves converting speech into text using machine learning algorithms. ASR involves breaking down the speech signal into individual phonemes and using statistical models to predict the most likely transcription.
- **Deep Neural Networks (DNNs**): This is a machine learning algorithm that is commonly used in ASR systems. DNNs involve creating a layered neural network that learns to recognize patterns in the input speech signal.
- **Hidden Markov Models (HMMs):** This is a statistical model that is often used in ASR systems. HMMs involve modeling the probability distribution of speech signals and calculating the most likely sequence of words.

- **Language Models**: These are statistical models that are used to predict the probability of different word sequences based on their frequency in a given language. Language models are often used in combination with ASR or HMMs to improve the accuracy of speech recognition.
- **Spectral Analysis**: This technique involves analyzing the spectral characteristics of the speech signal to identify the phonemes and words being spoken. Spectral analysis involves calculating features such as Mel-Frequency Cepstral Coefficients (MFCCs)[3] or Linear Predictive Coding (LPC) coefficients.

C. SENTIMENTAL ANALYSIS

Sentiment analysis [4], also known as opinion mining, is the process of analyzing text data to determine the sentiment or emotion expressed within it. This can involve using machine learning algorithms to classify text as positive, negative, or neutral, or to identify more specific emotions such as joy, anger, or sadness.

Sentiment analysis can be applied to various types of text data, such as social media posts, customer reviews, news articles, or survey responses. The goal of sentiment analysis is to extract insights from text data and understand the opinions and attitudes of individuals or groups toward particular topics, products, brands, or events.

There are many different approaches.

- **Lexicon-based approach**: This approach involves using sentiment lexicons, which are lists of words that are associated with positive or negative sentiment. Each word in the lexicon is assigned a sentiment score, and the sentiment score of the text is determined by summing the sentiment scores of the individual words in the text.
- **Machine learning**: Machine learning techniques involve training a model to classify text into positive, negative, or neutral sentiment. The model is trained using a set of labeled data, and the accuracy of the model can be improved by adjusting the parameters of the model or by providing it with more training data.
- **Deep learning**: Deep learning techniques use deep neural networks to identify patterns in text data and to determine sentiment. Deep learning models can be trained on large amounts of data and can achieve high levels of accuracy.
- **Rule-based approach**: Rule-based approaches involve creating a set of rules that define how certain words or phrases are associated with positive or negative sentiment. These rules can be based on the presence of specific words or phrases that are associated with positive or negative sentiment.
- **Hybrid approach**: Hybrid approaches combine multiple techniques, such as lexicon-based and machine learning, to improve the accuracy of sentiment analysis.
-

IV. RELATED WORK

SR.NO / YEAR	Title	Methods	Future Goal

[1] 2022	Modulation Spectral feature for SER Using DNN	Dataset : (EmoDB) & (RAVDESS) Feature Extraction : CQT-MSF and MFSCfeatures Classification : two different machine learning algorithm (1) convolutional neural network with fully connected layer for emotion classification (termed henceforth as DNN). (2) Convolutional layers to extract emotion embeddings and SVM to classify embeddings into emotion classes (termed as DNN-SVM).	They want to work on joint spectral and temporal domain for analysing it's suitability for SER.
[2] 2020	SER using Neural Network and MLP Classifier	Dataset : RAVDESS Feature Extraction : MFCC , MEL , Chroma , Tonnetz Classification : Neural Network and Multi-Layered Perceptron classifier.	To archive more than 80% accuracy.
[3] 2019	Voice Emotion Recognition using CNN and Decision Tree	Dataset : (Customized Kannada Dataset) & (RAVDESS) Feature Extraction : MFCC Classification: CNN	Identification. Of depression and mood swings.
[4] 2022	SER using Machine Learning	Dataset : RAVDESS Feature Extraction: MFCC , Mel , chroma, Tonnetz Classification : MLP classifier	They try to improve accuracy By increase the dataset size
[5] 2021	Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus	Dataset : KSUEmotions corpus Feature Extraction : zero-crossing rate , short – term energy, MFCC's and delta features Classification : KNN / SVM	They want to implement SER using deep learning corpus dataset.

- **The main goal of this paper are as below:**
  1. to identify and analyze the emotional state of a speaker based on their vocal tone, intonation, and other speech-related features in real time or from test data
  2. to convert real time human speech or recorded test data into written text .
  3. to determine the sentiment or emotional tone of the language.

  Thus main overall aim of this paper is to save time and to provide comfort to users.

V. PROPOSED WORKFLOW

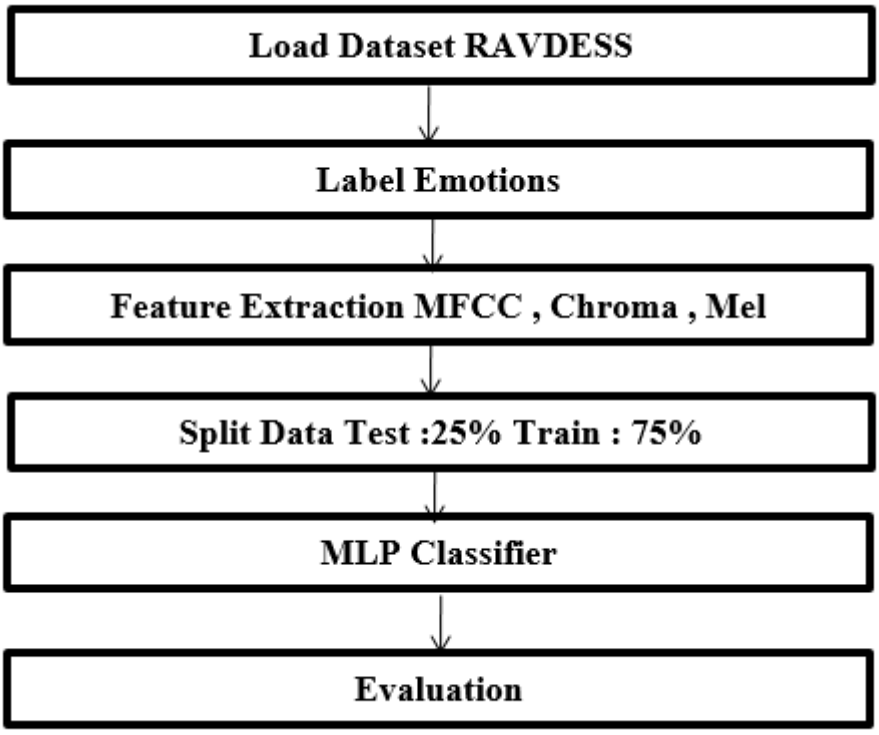
General workflow followed in this paper is preprocessing, feature extraction and classification.

1. **Pre-processing**: The speech signal is first pre-processed to remove noise and other artifacts that may affect the accuracy of emotion recognition.
2. **Feature extraction**: Various features are extracted from the pre-processed speech signal. These features may include pitch, energy, and spectral features, among others.
3. **Classification**: The selected features are used to train a classifier model, which can be based - on various machine learning algorithms such as support vector machines (SVM), artificial neural networks (ANN), or decision trees.

VI. IMPLEMENTATION

In this paper emotion recognition, speech to text and opinion mining is implemented using python(3.9.6) programming language. This python is very popular tool for deep learning area ,In addition to that Jupyter notebook (6.4.3 version) as an environment. hereRyerson Audio-Visual Database of Emotional Speech and Song Dataset (Ravdees) has been utilized For training the model.

A. WORKFLOW



B. Workflow Description

i. **Data Loading :**

1. RAVDESS :The Ryerson Audio-Visual Database of Emotional Speech and Song Dataset (Ravdees) which contain 1440 files.[4] From that per actor 60 trials they give us. Like wise there are total 24 actors so by calculating you will get total 1440 data. Which is sufficient for training our model. In the 24 Actors there are 12 female and 12 male actor. They used neutral North American accent for prepare this data. In this they mainly include 8 most important emotion and this are the emotions calm, happy, sad, angry, fearful, surprise, and disgust.

b. Label Emotion:

i. In this section we are creating an observed emotions array which contain mainly this 8 emotion. Neutral, calm, happy, sad , angry , fearful , disgust , surprised.

ii. So from this we can archive to detection of emotion.

c. Feature Extraction :

i. Feature extraction used to extract relevant features from audio signals and emotions. In this mainly audio frequently reflects hidden feeling through tone and pitch. From the given signal information there are mainly three feature are extracted. The three features are MFCC, Mel, Chroma [6].

1. **MFCC :**
   MFCC stands for Mel Frequency Cepstral Coefficients, a technique to compress spectral content of speech signals into a set of coefficients. These coefficients are often used as features in speech emotion recognition systems to classify the emotional state of a speaker based on their speech signal. MFCCs are effective in capturing the patterns [5]of spectral content associated with different emotions in speech.
2. **Chroma:**
   Chroma is a feature extraction technique that captures the tonal content of speech signals in speech emotion recognition. It maps the distribution of energy across the 12 pitch classes of the chromatic scale. Chroma features are used along with [8]other features, such as MFCCs, to improve the accuracy of emotion classification.
3. **Mel:**
   Mel refers to Mel-frequency cepstral coefficients, which are commonly used in speech signal processing to represent the spectral features of speech signals. These coefficients are calculated by mapping the frequency spectrum of the speech signal onto a Mel-scale, which is a logarithmic scale that mimics the way human ears perceive sound.

d. Split data:

The RAVDESS dataset passed to MLP Classifier. Then we split the dataset in 25:75 ration i.e testing and training dataset.

e. MLP classifier:

MLP (Multilayer Perceptron) is a type of neural network that can be used in speech emotion recognition to classify speech signals into different emotional categories based on their extracted features, such as [8]Mel-frequency cepstral coefficients. MLP classifiers consist of multiple layers of nodes and can learn complex relationships between input features and emotional categories. They are trained on a labeled dataset of speech samples and can predict the emotional category of new, unlabeled speech samples.

f. Evaluation:

In brief and concise terms, the process of speech to text, speech emotion recognition, and sentimental analysis is a powerful tool for analyzing the emotional sentiment expressed in speech signals. By using Mel-frequency cepstral coefficients and MLP classifiers, speech signals can be classified into different emotional categories. Once the speech signal has been converted to[9] text using speech recognition technology, sentimental analysis can be applied to determine the emotional tone of the text data.

VII. CONCLUSION

In conclusion, the integration of speech emotion recognition using MLP classifier, speech to text using ASR, and sentimental analysis can provide several advantages in understanding and analyzing spoken content. MLP classifier can accurately classify speech signals into different emotional categories. When combined with speech to text using ASR, the resulting textual data can be analyzed to determine the emotional sentiment expressed by the speaker using sentimental analysis. This integration can lead to an accurate and efficient method for extracting and analyzing valuable insights from spoken data. Additionally, the combination of these technologies can save time and resources while also improving the accuracy of analysis.

Overall, this process can be useful in a variety of applications, such as customer feedback analysis, market research, and opinion mining. The ability to extract valuable information from spoken content can provide a significant advantage for organizations seeking to optimize business operations, improve customer satisfaction, and make data-driven decision.

VIII. REFERENCES

[1].NavyaDamodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019.

[2].Jianfeng Zhao, Xia Mao, Lijiang Chen. Learning Deep features to Recognise Speech Emotion using Merged Deep CNN. IET Signal Process., 2018

[3] UrbanoRomeu, Á. (2016). Emotion recognition based on the speech, using a Naive Bayes classifier (Bachelor's thesis, UniversitatPolitècnica de Catalunya).

[4] Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNNdecision tree SVM model. Speech Communication, 115, 29-37

[5] Meftah, A., Qamhan, M., Alotaibi, Y. A., &Zakariah, M. (2020). Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus. International Journal of Simulation--Systems, Science & Technology, 21(2).

[6].Ayush Kumar Shah ,MansiKattel,Araju Nepal. Chroma Feature Extraction using Fourier Transform. Chroma_ Feature_ xtraction. January 2019

[7] J. Auguste, D. Charlet, G. Damnati, F. Bechet, and B. Favre, “Can ´ we predict self-reported customer satisfaction from interactions?” in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2019, pp. 7385–7389.
[8] B. Li, D. Dimitriadis, and A. Stolcke, “Acoustic and lexical sentiment analysis for customer service calls,” in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2019, pp. 5876–5880. [3] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” IEEE Access, vol. 7, pp. 100 943–100 953, 2019.
[9] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “Dialoguernn: An attentive rnn for emotion detection in conversations,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 6818–6825.
[10] C. Huang, A. Trabelsi, and O. R. Za¨ıane, “Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert,” in Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACLHLT, 2019, pp. 49–53.
[11] Prof. Guruprasad G, Mr. Sarthik Poojary, Ms. Simran Banu, Ms. Azmiya Alam, Mr. Harshith K R “EMOTION RECOGNITION FROM AUDIO USING LIBROSA AND MLP CLASSIFIER” International Research Journal of Engineering and Technology (IRJET), vol8, issue 7, 2021.
[12] Monorama Swain, Aurobinda Routray, Prithviraj Kabisatpathy,”Databases, features and classifiers for speech emotion recognition: a review”, I. J. Speech Technology 2018.
[13] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” International Journal of Speech Technology, vol. 15, no. 2, pp. 84–115, 2012.
[14]. Awni Hannun, Ann Lee, Qjantong Xu and Ronan Collobert, Sequence to sequence speech recognition with time-depth deperable convolutions, interspeech 2019, Sep 2019.
[15]. Li, J., Deng, L., Gong, Y. (2014). An Overview of Noise-Robust Automatic Speech Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing, Vol.22, No.4, pp.745-777.