Enhancing Human-Computer Interaction through Emotion Recognition in Real-Life Speech

S Anjan, Eshwari Rao K P,
East West Institute of Technology,
Bengaluru, India
anjan2251@gmail.com,
eshwarirao2001@gmail.com

Prof. T. N Nagabhushana
East West Institute of Technology
Bengaluru, India
ed@ewgi.edu.in

Kirill Krinkin, Yulia Shichkina
Saint-Petersburg Electrotechnical
University 'LETI'
Saint-Petersburg, Russia
kirill@krinkin.com, shichkina@etu.ai

***Abstract*—Extracting the data from real-life speech and recognizing emotions from them is one of the challenging tasks. This has gained popularity over the past few years. The goal of this study is to have a direct human-computer interaction (HCI) to determine the person's condition or emotion through analyzing voice. The main purpose of this study is to extract emotion from the recorded audio as well as from the analysis of text extracted from the audio which is integrated together in a precise manner. The multimodal integration of speech and text gives good results by observing the emotional state of a person. Machine learning and deep learning algorithms are used to determine the emotional state of a person. The result of this study shows the accuracy around 62%.**

I. Introduction

We have come a long way in the field of artificial intelligence, though there is human computer interaction taking place, the machines cannot understand the emotional state of humans [4]. Humans use multiple modalities to express their emotions. In psychology, the emotional states are based on the observation of visual and auditory non-verbal cues. The emotions are expressed explicitly i.e. through facial expression or speech and implicitly through text or image[8]. In the absence of visual features, the emotion is recognized through the voice medium. For example, telephone line, call center, E-learning, gaming experience improvement, sentiment analysis, stress analysis, voice message, recognition of lies, audio surveillance, healthcare centers and clinical studies. The emotion recognition through voice in a call center is used for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on which helps companies to improve their services.

The real-life emotion extraction through speech is one of the challenging tasks. Even though there are many speech emotion recognition (SER) systems, they show poor performance and most of them use only speech signals so the research is being carried out by researchers. The speech signal may seem complicated as it comprises a message from the speaker with several types of feelings in it. For example, the speaker can perceive emotion not only from the words spoken but also from the speech being heard. In addition, usually people tend to use some specific words while expressing certain emotions. The present study shows the integration of multimodal features to increase the performance of the existing system.

This paper presents simple methods to increase the accuracy of speech emotion recognition system. To determine the performance of the proposed system both speech features and text features are considered. Initially, the live audio is recorded in the system which is stored in the system as WAV format. Speech feature also defined as acoustic feature is formed from speech segments where silence is removed and converted into frames. In the text feature, the recorded speech converts audio transcripts into text. By using machine learning algorithms, audio processing and natural language processing techniques this model is built. The various python modules are imported to set up this model.

II. State of art

There are many studies and research going on in the field of continuous speech emotion recognition. The researchers have used different algorithms and dataset to train the model and acquired good results. But they lack some factors to implement it in a real time situation. Usually the speech emotion recognition system comprises three main segments i.e Feature Abstraction, Feature Collection and Classification. In this section, we describe the review of the literature and theoretical foundation of our work.

The primitive stage of recognizing emotions through speech used rustic speech features and low level descriptors to train the model. Recently, studies have drawn increasing attention towards SER using Automatic Speech Recognition(ASR) features and Deep Neural Network(DNN) methods. Although, there are two major issues found in Deep Learning approaches: (1) adequate sum of labelled speech data and (2) extracting emotion related features from audio. Transfer Learning is the method within deep learning where a neural network is trained for a particular task. This method mainly focuses on storing knowledge gained while training one model and applying it to another task. It has been successfully applied in various domains[11]-[13].

Paper[5] uses Two unidirectional LSTM layers to connect text and acoustic emotion recognition. Both networks are merged by fully connected networks to predict emotions. IEMOCAP consists of 9 categories of emotions , out of which 4 categories of emotions produce results. The accuracy of the model is around 70% by combining both speech and text segments[5]. Here the model shows whether the person is satisfied or frustrated so it is better to be used with other classification of emotions. To improve the continuous speech emotion recognition system, paper[4] uses wav2vec and camemBERT

as self-supervised learned models to represent our data in order to perform continuous emotion recognition from speech (SER) on AlloSat, a large French emotional database. This model is evaluated by concordance correlation coefficient (CCC). The CCC value can reach up to 0.825. The word error rate in wav2vec is comparatively higher.

In one study[1], it was proposed to conduct SER based on BLSTM and self-attention mechanisms using ASR results while fine-tuning ASR simultaneously, which results in SER performance improvement and ASR robustness to emotions. The main drawback of this method is that the fine-tuning of ASR performance alongside SER is computationally costly. Moreover, it is difficult to collect data with emotional labels and transcriptions that are essential to perform fine-tuning properly.

III. Dataset

In our study, to extract textual features from audio, we are using a twitter dataset which has 34792 text data classified into different emotions such as joy, sadness, fear, anger, surprise and neutral. Fig. 1. shows the twitter dataset which has been loaded into the system to train the model. Fig. 2. shows the values of different emotions present in the dataset. Fig. 3. displays the graph of emotions and their count. By looking at the graph we can conclude that joy is having the highest count and shame is having lowest count.

	Unnamed: 0	Emotion	Text
0	0	neutral	Why ?
1	1	joy	Sage Act upgrade on my to do list for tommorow.
2	2	sadness	ON THE WAY TO MY HOMEGIRL BABY FUNERAL!!! MAN ...
3	3	joy	Such an eye ! The true hazel eye-and so brill...
4	4	joy	@lluvmiasantos ugh babe.. hugggzzz for u .! b...

Fig. 1. Twitter dataset which is loaded into the system

```
joy         11045
sadness      6722
fear         5410
anger        4297
surprise     4062
neutral      2254
disgust       856
shame         146
Name: Emotion, dtype: int64
```

Fig. 2. Values in the dataset

In this section we discuss the possible solution for extracting emotion through speech. Fig. 4. displays the workflow of the model. (a)The numerous python libraries are used to construct this model. (e)The audio file is received as input in this model. The live audio is recorded and this recorded audio is converted to frames. By using the below formula:

frames = RATE/FRAMES_PER_BUFFER*seconds

Then it is saved in the system as a WAV file. To analyze the audio file and to extract audio data a python library named librosa is used. (f)As computers need to understand human language, a speech recognition module is used. So that the recorded speech converts audio transcripts into text.

(b)The twitter dataset is loaded for this study. There are basically 8 types of emotions such as joy, sadness, fear, anger, surprise, neutral, disgust, shame observed from this data. After loading the dataset, the NLP technique is used for data preprocessing. Natural Language Processing is the ability of a computer program to understand human language (i.e natural language such as English) as it is spoken and written.

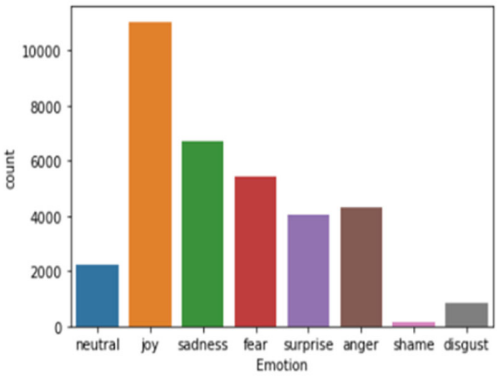

Fig. 3. Graph plotted against the emotions and count of emotions

IV. Implementation

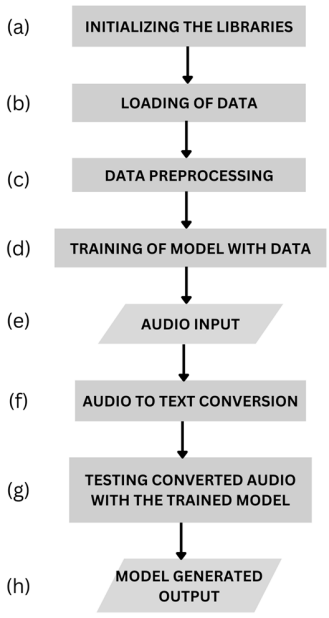

Fig. 4. Flowchart of model workflow

There are mainly two phases of natural language processing: data preprocessing and algorithm development. (c)Data Preprocessing involves analyzing and cleaning of text data. The text data in the form of statements are broken down into words, then on further process special characters will be

removed. Preprocessing highlights features in the text that an algorithm can work with. (d)The clean data will be loaded into the model for training. The process of converting text data into numbers is called feature extraction from text. It is also called text vectorization. Here CountVectorizer is one such tool used to convert a collection of text documents to a vector on the basis of frequency of each word that exists in the entire text. By using pipeline it helps independently executable workflow of complete machine learning tasks. Logistic regression is a machine learning, supervised learning algorithm used to train the model to predict the accurate value. (g)Logistic regression can be used with CountVectorizer in a pipeline. In a multi-class classification logistic regression model, you can take in cleansed text descriptions of an object and classify the object into a category using a basic sklearn pipeline. In section IV. the result is displayed.

V. Result

In this study, our model is trained with a twitter dataset which contains 34792 tweets recorded from real life scenarios. Then our trained model will request the user for input. Microphone is the device used to give input to the model. There is a certain time gap specified in the code during which the user must speak into a microphone to give an input. The audio will be recorded in .wav format. The recorded audio will be converted to text. These words are now analyzed by the trained model. Based on the feature and meaning of each word in the input suitable emotion is given as an output from the model. Our model was able to detect emotions like joy, sadness, fear, surprise, anger and neutral. As shown in Fig. 5. we were able to precisely determine these emotions based on the trained model with an accuracy of 62%.

```
SPEAK
STOP
Converting audio transcripts into text ...
I got the job
Emotion :  ['joy']
Accuracy   : 62.00421536692853
```

Fig. 5. Result

VI. Conclusion and Future Scope

The emotion extraction through audio is a promising research area with potential applications in various fields such as healthcare, entertainment, and education. This research paper aimed to explore different methods and techniques used for emotion recognition from audio signals. Despite the progress made in this field, there are still some challenges that need to be addressed, such as the lack of large annotated datasets, the variability of emotional expression across cultures, and the need for more sophisticated algorithms that can handle different types of emotions and their complexities. Overall, emotion extraction through audio is a promising research area that has the potential to revolutionize the way we interact with technology and improve our understanding of human emotion.

The future scope of our model is to implement larger datasets for boosting performance and for the model to understand the actual precise emotion of the given input statement. To achieve the highest accuracy and reduce error margin multimodal integration is required. Our study is based on extracting emotion through text which was converted from an audio file. Integration of a model where emotion is extracted from the audio based on pitch, frequency, wavelength, cepstral coefficients will be a greater innovation towards achieving the highest accuracy and getting the actual emotion which is in the input given by the user. The integrated models can be implemented with CHI - coevolutionary hybrid intelligence, where a machine remembers the last stated input given by the user so that the whole model would not need to execute if the same input is given.

VII. References

[1] H. Feng, S. Ueno and T. Kawahara, "End-to-end speech emotion recognition combined with Acoustic-to-Word ASR model", *Proc. Interspeech*, pp. 501-505, Oct. 2020.

[2] Deepak Bharti, Poonam Kukana(2020), “A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals”.

[3] Kun Han, Dong Yu and Ivan Tashev ,“Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine”, 2014.

[4] Manon Macary, Marie Tahon and Yannick Est`eve, Anthony Rousseau, “On the use of Self-Supervised Pre-trained acoustic and linguistic features for Continuous Speech Emotion Recognition”, 2021.

[5] Bagus Tris Atmaja , Kiyoaki Shirai, and Masato Akagi, “Speech Emotion Recognition Using Speech Feature and Word Embedding”, 2019.

[6] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, Vahid Tarokh, “Speech Emotion Recognition with Dual-Sequence LSTM Architecture”, 2020.

[7] Ling Cen, Fei Wu, Zhu Liang Yu, Fengye Hu “Chapter 2 – A Real-Time Speech Emotion Recognition and its Application in Online Learning”, 2016.

[8] Tanja Banziger, “Measuring Emotion Recognition Ability”.

[9] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, Kurt Keutzer, “Emotion Recognition from Multiple Modalities: Fundamentals and Methodologies”.

[10] LivingstoneR., Steven. 2020. "RAVDESS Emotional SpeechAudio".Kaggle.Com https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speechaudio.

[11] Goodfellow, Y. Bengio, and A. Courville, Deep learning, MIT press, 2016.

[12] M. Ezzeldin A. ElShaer, S. Wisdom and T. Mishra, "Transfer learning from sound representations for anger detection in speech", 2019.

[13] S. Latif, R. Rana, S. Younis, J. Qadir and J. Epps, "Cross corpus speech emotion classification—An effective transfer learning technique”, 2018.