Speech Emotion Recognition Using Librosa

Computer Science Engineering Department, Chandigarh University

Abhinav Kapoor	Aditya Sharma	Aryan	Radhika Patel	Shakir Hussain
20BCS9877	20BCS9872	20BCS9854	20BCS1212	20BCS9719

Abstract - One of the most popular marketing techniques nowadays is emotion detection, in which the consumer's mood is key. Speech Emotion Recognition, or SER, is an example of an effort to recognise both human disposition and behaviour. Normal voice tone and pitch convey basic emotions. Many animals other than humans exhibit behaviours that are synchronised with humans. In this essay, we will examine music tones, speech sounds, and a python-based library called Librosa. In this regard, a collection of libraries is being put together in order to create a detection model that makes use of an MLP (Multilayer Perceptron) classifier.

Keywords - Speech Emotion Recognition (SER), Machine Learning, Multi-layer Perceptron (MLP) classifier, CNN, Deep learning.

I INTRODUCTION

Speech Emotion Recognition (SER) is the method or approach for identifying human emotions in a spoken input. This takes advantage of the fact that a person's tone and voice pitch occasionally reflect the emotion they are going through. For example, ponies and dogs use this oddity to determine human propensity. Because utterances are abstract, it is difficult to explain speech tone in SER [1]. However, speech analysis also provides a major advantage. As a part of shrewd living, they can be used to create clever frameworks that seamlessly integrate into our daily activities. It serves as the foundation for innovations from tech goliaths such as voice recognition, voice control, order capabilities, and much more like Google and Microsoft as Alexa, Cortana, and Samsung's Bixby, as well as other man-made brainpower (AI) applications.

Feelings can be evoked via a variety of techniques or tactics, including appearance, calligraphy analysis, and mental appraisals of the subject. Despite this, the most important mode of communication between any two people at any moment is speech [2]. This has prompted various experts from a variety of professions to examine, test, and come to constructive conclusions in the field of speech examination, leading to the development of numerous models and ideas over time. The 1950s saw widespread awareness of this. The researchers had the opportunity to create a theoretical strategy or recipe to translate speech into a collection of words for obviously design expectation in a variety of uses with the aid of previous people's information and distributions.

By examining the acoustic characteristics of the audio data of recordings, we try to detect underlying emotions in speech in this work. Speech features can be divided into three categories: lexical, visual, and acoustic elements. Analysis of one or more of these features can be used to address the issue of speech emotion recognition.

II LITERATURE SURVEY

The literature has produced some important works on voice recognition systems. According to Vincius Maran et al., learning speech is a laborious process where the infant's processing of criteria is highlighted by their unpredictability on the path to the modern production of ambient language segments and structures [3]. It makes sense that this concept should be applied to computers as G. Tsontzos et al. highlighted the importance of feelings in improving human communication [4]. Mehmet Berkehan Akçay et al. claim that industrial control and robotics applications are the main uses of neural networks for their intended purpose [5]. But with subsequent improvements, this technology has spread more accurately across a variety of fields.

By systematically comparing assessments of perceived emotions using two alternative theoretical frameworks—the discrete emotion model and the dimensional model of affect—the present study's main goal is to add to the theoretical discussion that presently dominates music and emotion research. The comparison is significant due to the prevalence of these models in music and emotion studies, as well as the categorically constrained affect space that the excerpts have up to this point and the possible neurological differences involved in emotion categorization and the evaluation of emotion dimensions.

Additionally, the different alternative formulations of the dimensional model have not previously been examined in research on music and emotion. A secondary goal is to present a new, enhanced collection of stimuli for the research of music-mediated emotions, made up of unfamiliar, thoroughly examined, and validated non-synthetic music snippets. Additionally, this set of stimuli has to have moderate examples that enable the study of more subtle variations in emotion in addition to the best examples of the goal emotions. [6]

Data gathered from several modalities were used to classify the arousal, valence, and liking scales in a single trial. The results were discovered to be significantly better than those attained using other categories. The fact that there was a modest performance improvement after decision fusion of these outcomes shows that the modalities are at least somewhat complementary. We expect that other academics will utilise the database to test their methods and algorithms now that it is open to the public [7].

Speaker identification, according to Peng et al., is the process of identifying people by their speech. This technology is gradually being accepted and employed as a kind of biometrics due to its simplicity and lack of involvement, and it has fast developed into a research hotspot in the field of biometrics [8]. The final emotion label for the provided input voice signal is obtained by fusing the out-turned emotion labels using majority voting. SVM and its combination approaches have also been extensively employed in research.

However, the state can be combined with different detection models in the context of ensemble learning. Furthermore, the ideal SER model has not yet been identified.

III METHODOLOGY

The main goal of this work is to develop a simple, effective, and usable model that incorporates machine learning techniques as its fundamental component and for which we can have confidence that the system will produce accurate and error-free results. Librosa, a Python programming framework, makes it relatively easy to plan and carry out this job. This phrase is crucial to the process. A few other aspects of this work, besides writing the code, are thoroughly examined. The method is simple, and we've done our best to be thorough in our work. This comprises integrating a few fresh features and functionalities into the current execution procedure.

(1) The Ryerson Audio-Visual database of emotional speech and song (RAVDESS) dataset is the one that will be worked on in the planned research. A total of 7356 files with 10 times on emotional validity and sincerity are included in this collection. The total dataset is 24.8GB in size and consists of 24 actors, 12 male and 12 female, whose numbers range from 01 to 24. The number of male performers is odd, whereas the number of female actors is even. The dataset includes expressions for sadness, joy, happiness, pleasure, anger, disgust, surprise, fear, and calmness. It is initially downloaded from the Kaggle repository and saved in a local system as the main location where it can be examined and updated versions are saved. This is so that the system may easily access the dataset when compiling.

(2) The necessary libraries are then imported into the code file for quick execution. Despite the fact that they are already installed on our PC, we should import them in the program.

(3) We can now access the audio files that were previously obtained for analysis after saving it to our local system. The dataset can be read in the very next step by building functions with different arguments like chroma, filename, mfcc, and mel. Due to limited computing power, analysing such a huge dataset is exceedingly challenging. As a result, we will read each of the files using the looping idea. The programme reads all of the data in float32 format using a variable called X.

(4) A network made up of perceptrons is known as a multi-layer perceptron (MLP). The layers present between the input and output layers are referred to as hidden layers. It has an input layer that receives the input signal and an output layer that generates predictions or choices based on a given input. The MLP network will contain one input layer, 300,40,80,40 hidden layers, and one output layer in the suggested methodology for speech emotion recognition. There will be a lot of hidden layers, and the number can be altered depending on the situation.

(5) The information was obtained directly from the collection of audio files, and it was then translated into 264 feature vectors. A speech signal's content can be parametrically represented as a vector in a variety of ways with the goal of extracting important information from it.

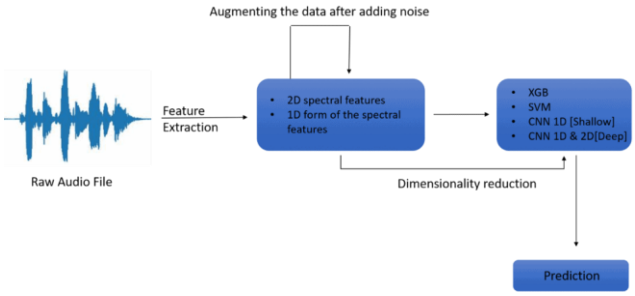

Fig 1: Algorithm Flow

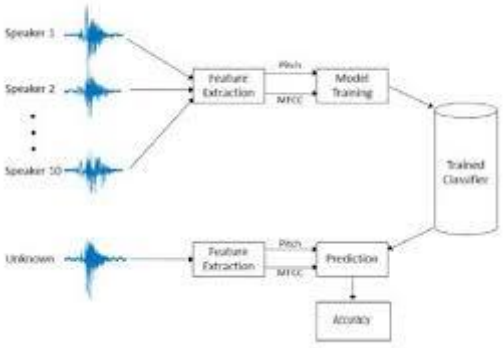

Fig 2: Speech Emotion Recognition

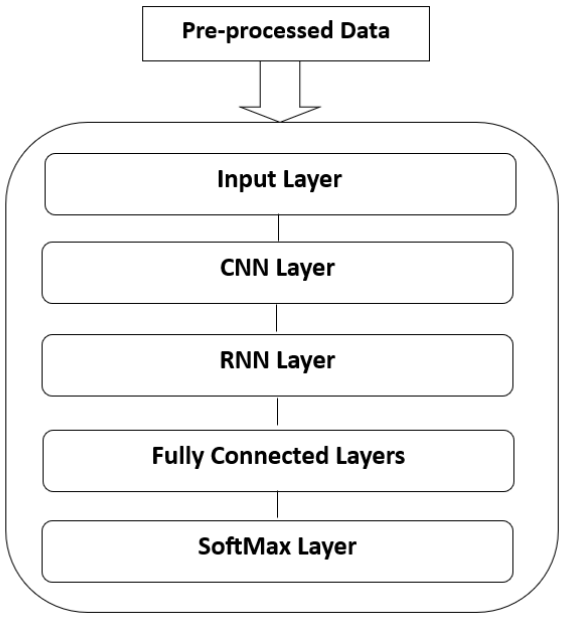

Fig 3: Layer Architecture of Model

IV. ISSUES AND CHALLENGES

Even while voice recognition systems have come a long way, there are still a few obstacles to overcome before we can achieve truly excellent recognition. Possibly the biggest problem is the process utilised to create the dataset that the learning system uses. In designated silent rooms, the majority of the SER informative collections are acted out, invoked, and recorded. Compared to other data, real-world data is raucous and has essentially more unique characteristics. There are additionally accessible but fewer regular informational collections.

These days, there are ethical and legitimate concerns raised by the recording and utilisation of common emotions. The majority of the expressions in typical informational indexes originate from syndicated programmes, contact focus accounts, and other situations when the included groups are aware of the recording.

This compilation of data could leave out all the associated emotions and not correctly portray the situation. The marking of the expressions has problems as well. Therefore, it might be very difficult to tell the speaker's true intentions when they are speaking.

The human annotator can identify around 90% of the samples, but not more. However, as individuals, we believe that when choosing a speech, we should also consider the content and setting of the communication milieu. Social and phonological factors also have an impact on SER. There are several accessible techniques for analysing cross-language SER. The findings, however, show that the current guidelines and features are insufficient. For instance, different dialects may have different sounds for emotions in speech. The SER framework should decide which signal to focus on and which signal to ignore when there are several signal flags. The current frameworks, however, fail to recognise these problems since they typically handle them using a speech division method during the pre-processing stage.

V. CONCLUSION

The entire system provides some insights into how human emotions are represented verbally as well as how machine learning can be used to extract the most important emotions from audio data. As a result, this technology can be applied in a variety of contexts, such as voice-based virtual assistants, contact centres for customer support in marketing, and linguistic research.

Some of the steps that can be taken to make sure that the models are accurate and well-formed include the following:

• Some of the model's shortcomings can be fixed by investigating the correctness of an accurate implementation of the speaking speed.

• Trying to figure out how to get rid of the aimless immobility in the audio sample.

• Different acoustic aspects of sound data are being explored to see if they may be applied to the field of speech emotion recognition.

• Applying a lexical features-based approach and the ensemble method to SER.

ACKNOWLEDGMENT

This Research is conducted by student of Chandigarh University from Computer science Engineering Branch under the guidance and support of Prof. Shiwali Yadav(E13277).

REFERENCES

[1] Puri, Tanvi, et al. "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network." Journal of Healthcare Engineering 2022 (2022).

[2] Xu, Mingke, Fan Zhang, and Wei Zhang. "Head fusion: improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset." IEEE Access 9 (2021): Pp. 74539-74549.

[3] Franciscatto, Maria Helena, et al. "Towards a speech therapy support system based on phonological processes early detection." Computer speech & language 65 (2021): 101130.

[4] Tsontzos, Georgios, et al. "Estimation of general identifiable linear dynamic models with an application in speech recognition." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 4. IEEE, 2007.

[5] Akçay, Mehmet Berkehan, and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers." Speech Communication 116 (2020): Pp. 56-76.

[6] Tuomas Eerola and Jonna K. Vuoskoski, “A comparison of the discrete and dimensional models of emotion in music”, Psychology of Music, 1–32, The Author(s) 2010.

[7] Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." Biomedical Signal Processing and Control 59 (2020): 101894.

[8] Peng, Shuping, et al. "Remote speaker recognition based
on the enhanced LDV-captured speech." Applied Acoustics
143 (2019): 165-170.

[9] Varghese, Ashwini Ann, Jacob P. Cherian, and Jubilant J. Kizhakkethottam. "Overview on emotion recognition system." 2015 International Conference on Soft-Computing and Networks Security (ICSNS). IEEE, 2015.

[10] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." Sensors 21.4 (2021): 1249.

[11] Mao, Shuiyang, et al. "Revisiting hidden Markov models for speech emotion recognition." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2019.

[12] Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018.

[13] Wen, Guihua, et al. "Random deep belief networks for recognizing emotions from speech signals." Computational intelligence and neuroscience 2017 (2017). [14] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern recognition 44.3 (2011): 572-587.