EasyChair Preprint

№ 5592

Speech Emotion Recognition: Sentiment Analysis on Speaker Specific Speech Data

Aditya Kulgod, Shubham Shinde and Chinmay Phadke

May 23, 2021

Speech Emotion Recognition: Sentiment Analysis On Speaker Specific Speech Data

Aditya Kulgod, Shubham Shinde, Chinmay Phadke

Student, Department of Computer Science, Smt.Kashibai Navale college of Engineering, Pune.(akulgod@gmail.com)
Student, Department of Computer Science, Smt.Kashibai Navale college of Engineering, Pune(shubham.shinde245@gmail.com)
Student, Department of Computer Science, Smt.Kashibai Navale college of Engineering, Pune(phadkecu23@gmail.com)

ABSTRACT Sentiment analysis has evolved over the past few decades. Most of the work revolves around text sentiment analysis using text mining techniques, but audio sentiment analysis is still in its infancy. Recognizing emotions from speech signals is an important but challenging component of human-computer interaction (HCI). Many techniques have been used in the Speech Emotion Recognition (SER) literature to extract emotions from signals, including many well-established speech analysis and classifications. Techniques. In this proposed model, we perform sentiment analysis for speech discriminated by speaker protocols to capture the emotions of each speaker involved in the conversation. We analyze various techniques for performing speaker discrimination and sentiment analysis to find efficient algorithms to perform this task.In this attempt to extract the human emotions from affective states of speech, this literature uses many different techniques to make machines understand the human emotions. This paper covers the different datasets, extracted emotions & different attempts made towards analysis of human sentiments.

INDEX TERMS: Speech emotion recognition, deep learning, deep neural network, convolutional neural network.

INTRODUCTION

Recognizing speech emotions, abbreviated as BEING, is the attempt to recognize human emotions and affective language states, taking advantage of the fact that the voice often reflects the underlying emotions through pitch and tone.. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon in different animals like dogs and horses who are employed to be able to understand human emotion. SER is difficult because emotions are subjective and annotating audio is a challenge. Sentiment Analysis is the study of human emotion or attitude towards an event, conversation on topics or in general Sentiment analysis is used in various applications. Here we use them to understand the way people think through their conversations with one another.

Emotion recognition from speech has become an important component for Human-Computer Interaction (HCI). These systems aim to facilitate the natural interaction with machines by direct voice interaction instead of using traditional devices as input to understand verbal content listeners to react. Some applications include dialogue systems for telephonic conversation such as call center conversation. Understanding the mood of humans can be very useful in many instances. For example, computers that possess the ability to understand and respond to human communication such as emotions. In such a case, after detecting humans' emotions, the machine could customize the settings according to his/her needs and preferences. Determining the emotional state of humans is a difficult task and may be used as a standard for any emotion recognition model.Amongst the multiple models used for categorization of these emotions, a discrete emotional approach is considered as one of the fundamental approaches. It uses several emotions such as anger, disgust, surprise, scared, happiness, neutral, sadness, calm. Another important model that is used is a 3-D continuous space with parameters such as arousal, valence, and potency.

LITERATURE WORK

Sentiment Analysis identifies the sentiment expressed in a text then analyses it to find whether the input expresses positive or negative sentiment. Majority of work on sentiment analysis is done by algorithms such as Naive Bayesian, decision tree, support vector machine, etc.

Our models uses Mel Frequency Cepstrum Coefficient (MFCC) to identify human based voices on the variations and characteristics unique to humans. Figure 1 shows how a normal MFCC algorithm works.

The approach for speech emotion recognition (SER) mainly consists of two phases called feature extraction and feature classification. The second phase includes feature classification using linear and non linear classifiers.

This model uses MLP(Maximum Likelihood Principle) & MFCC(Mel Frequency cepstrum coefficient).

MFCC— Which is Mel frequency cepstrum coefficient. Humans perceive audio on a nonlinear scale, MFCC tries to replicate the human audiostic as a mathematical model. The actual acoustic frequencies are mapped to Mel frequency scales which typically range between 300Hz to 5KHz. The Mel scale is linear below 1KHz and logarithmic above 1KHz.

Our review goes through several stages such as: a) Enhancement of input speech data, b) Feature extraction &
selection in emotion recognition, c)Measuring acoustics in SER, d) Classification of feature

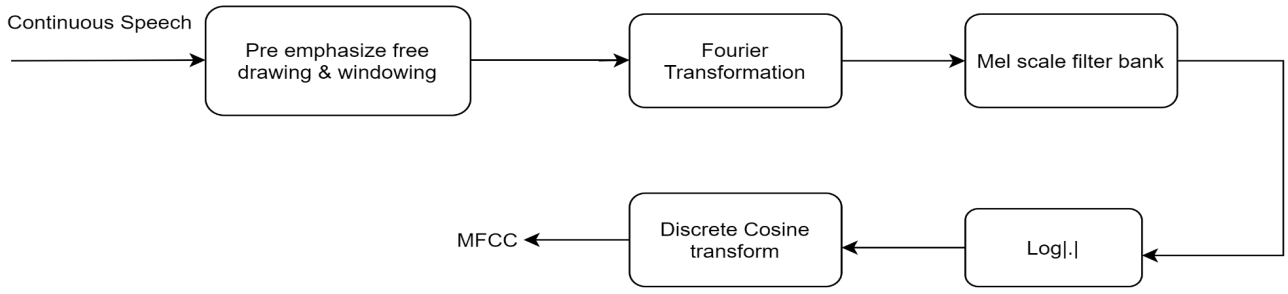

FIg.1. Working of MFCC model

Proposed System

In this research paper, we have proposed a model for sentiment analysis that utilizes features extracted from the speech signal using different algorithms to detect the sentiments of the speakers involved in the conversation. The process involves four steps:
1)Pre-processing
2) Speech Recognition System
3) Speaker Discrimination System
4) Sentiment Analysis System.

The input signal to the model is given in the audio form which then is preprocessed to clean the noise & to increase the efficacy of audio. This dataset is then stored in short term memory models.

Dataset: Our data set comprises RAVDESS which includes 1500 audio inputs. 12 male & 12 female actresses performed in this exercise to give 5 different emotions. Another dataset is SAVEE which contains another 500 audio inputs with 4 different male actors.

Our proposed model extracts the following emotions
1)Happy 2) Angry 3) Sad 4) Calm 5) Neutral

Dynamic Time Wrap (DTW) This algorithm measures

the similarity between two time series that differ in speed or time. This technique is also used to find the optimal alignment between time series when a time series cannot be linearly "deformed" by stretching or shrinking. This deformation between two time series can be used to find the corresponding areas between the two time series or to determine the similarity between the two time series.

Then the next important phase is building a model for which our choice was CNN(convolution neural network) which is used to build multilayer perceptrons & long in chunks as a dataset. Such dataset is then passed to the speech recognition & speaker discrimination system. Where the algorithm such as MFCC(Mel frequency cepstrum coefficient) is used to emulate the human ear & to make machines understand the human sentiments by feature extraction. Another library in Python which we use for feature extraction is libROSA which analyses the audio.

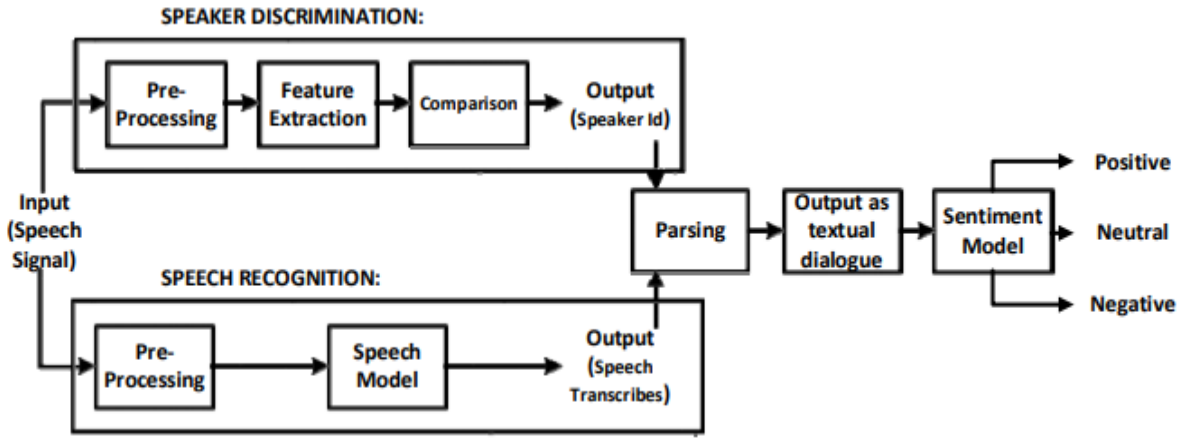

Fig. 2. Proposed Structure for the Sentiment Analysis System

GAP ANALYSIS

In the papers, many algorithms were used such as MFCC, PLP, but since the project is a classification problem, it led us to use CNN(Convolutional Neural Network). Following is the analysis of CNN along with its functioning.

Improvise the system so that it can handle a conversation between two speakers and in the conversation more than one speaker should be able to talk at a given time.

CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is another type of Deep learning technique based solely on feed-forward architecture for classification. CNNs are commonly used for pattern recognition and provide better data classification. These networks have small size neurons present.on every layer of the designed model architecture that process the input data in the form of receptive fields. Following figure provides the layer-wise architecture of a basic CNN network layer.

It should be noted that inputs are very small regions of the original volumes as depicted in Figure 3.. In the final stage, the layers need to be fully connected as in other neural networks. These later layers take the previous low level and mid-level features and generate high-level abstraction form the input speech data. The last layer also known as SVM or Softmax is utilized to further generate the score of classification in probabilistic terms to relate to a certain class.

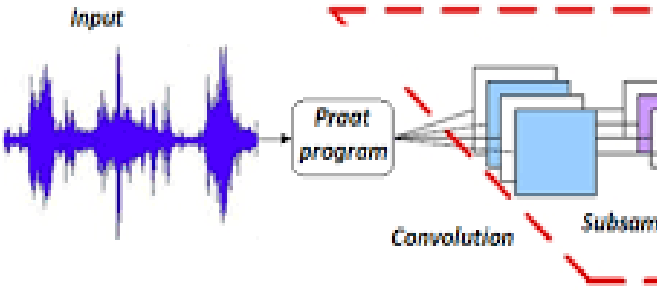

Fig. 3 CNN Algorithm

CONCLUSION

In this work a model is presented that uses audio that contains a short sentence spoken by the speaker and examines the content and identity of the speaker by performing speaker recognition. In this study, we proposed a simple system to accomplish the above task. The system works well with the data generated by actual events. These deep learning methods and t9.heir layered architectures are briefly elaborated based on the classification of various natural emotions such as happiness, sadness, fearful, angry and calm.

FUTURE WORK

Although the system is accurate to understand the sentiment of the speakers in dialogue, we are working on collecting a larger dataset and improving the scalability of the system. It has some shortcomings. Currently, the system can handle one speaker at a time, but it cannot. understand when two people are speaking at the same time. Our future work would address these issues and improve the accuracy and scalability of the system.

REFERENCES

[1]JonathanHui: https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9#:~:text=PLP%20is%20very%20similar%20to,instead%20of%20the%20log%20compression.&text=It%20also%20uses%20linear%20regressive,and%20slightly%20better%20noise%20robustness

[2]Jonathan Hui- Speech recognition Weighted Finite State

[3] Ruhul Amin Khalil , Edward Jones , Mohammad Inayatullah Babar , Tariqullah Jan , Mohammad Haseeb Zafar , And Thamer Alhussain (August 19, 2019)'Speech Emotion Recognition Using Deep Learning Techniques: A Review'

[4] Maghilnan S, Rajesh Kumar M (2017), 'Sentiment Analysis on Speaker Specific Speech Data'

[5]L. Deng and D. Yu, "Deep learning: Methods and applications,"Found. Trends Signal Process., vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[6] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,"Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, 2015.

[7]J. A. Coan and J. J. Allen, Handbook of Emotion Elicitation and Assessment. London, U.K.: Oxford Univ. Press, 2017.

[8]Sentiment Analysis on Speaker Specific Speech Data 2017 International Conference on Intelligent Computing and Control (I2C2).

[9]Speech Emotion Recognition Using DeepLearning Techniques:ReviewRUHUL AMIN,KHALILEDWARD JONES,1Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Engineering and Technology

[10]Herbig, T., Gerl, F., & Minker, W. (2010, July). Fast adaptation of speech and speaker characteristics for enhanced speech recognition in adverse intelligent environments. In Intelligent Environments (IE), 2010 Sixth International Conference on (pp. 100-105). IEEE