WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition

Prof. Kavita Namdev, Urvashi Goswami, Vanshika Rajput, Vartika Joshi

Department of Computer Science and Engineering, Acropolis Institute of Technology And Research , Indore, Madhya Pradesh , India.

urvashigoswami44@gmail.com, rvanshika24@gmail.com, vartikajoshi2208@gmail.com

ABSTRACT: -

Emotion detection through speech enables us to detect and understand emotions of the person at that point of time. It is a very successful implementation of machine learning and NN (Neural Network). After finding out the emotions of the person a variety of suggestions will be presented to the person accordingly. There are few universal emotions- including Neutral, Anger, and Happiness, Sadness in which the system with finite computational resources will be trained to identify or synthesize emotions as required. Additionally, person can also choose the emotion directly without speech and view the suggestions.

Key-Words: - SER, CNN, SVM, RAVDESS, ML, MFCC.

I.Introduction

Emotions play an extremely important role in human mental life. It is a medium of expression of one's perspective or one's mental state to others. Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions- including Neutral, Anger, and Happiness, Sadness in which the system with finite computational resources can be trained to identify or synthesize emotions as required. In this work spectral and prosodic features are used for speech emotion recognition because both of these features contain the emotional information. Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features which are used to model different emotions.

II.Problem Formulation

Emotion detection through speech recognition helps in extracting and providing optimized suggestion for the same.

- Weak marketing strategies
- Lack of motivation
- Inappropriate content.

Speech is the most natural way of expressing ourselves as humans. It is only natural then to extend this Communication medium to computer applications. Speech emotion recognition (SER) system is a collection of methodologies that process and classify speech signals to detect the embedded emotions. The main objective is to understand human emotions with maximum accuracy and as a result suggest the best they could get. It will benefit the users with numerous choices, as well as be profitable for marketing industries and other industries like call centers where communication is of utmost importance.

III. Literature Review

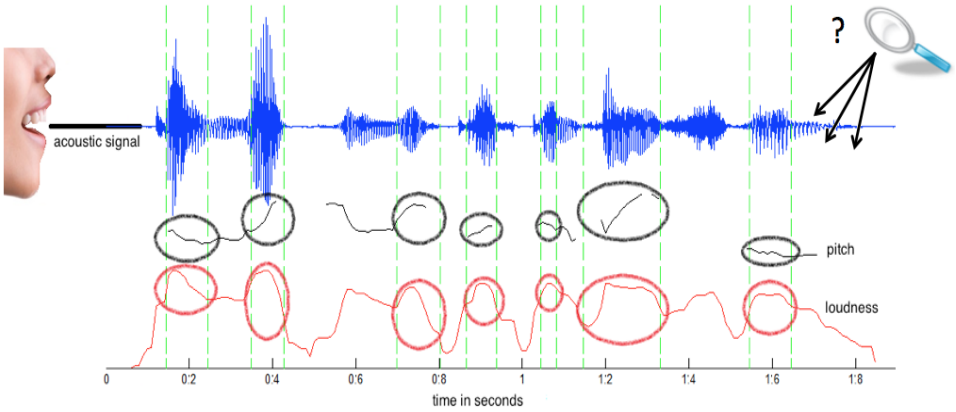
Over the last few years, a lot of research has gone into using speech statistics to discern emotions. To overcome the challenge of binary classification, Cao et al. suggested a ranking SVM approach for synthesizing information on emotion recognition. This strategy instructs SVM algorithms for certain emotions, treating data from each utterer as a separate query, then combining all ranker forecasts to apply multi-class prediction. Ranking SVM has two advantages: first, it gathers speaker-specific data for the training and testing processes in a speaker-independent manner. Second, it takes into account the fact that each speaker may convey a range of emotions in order to determine the dominating emotion. In two public datasets of acted emotional speech, Berlin and LDC, ranking techniques produce a significant gain in terms of accuracy when compared to standard SVM. Ranking-based SVM

outperformed standard SVM algorithms in detecting emotional utterances in both acted and spontaneous data, which includes neutral powerful emotional utterances. Balance accuracy or unweight average (UA) was 44.4 percent.

IV.Methodology

Feature Extraction:

When we do Speech Recognition tasks, MFCCs is the state-of-the-art feature since it was invented in the 1980s. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.



Default Model Architecture:

We developed the CNN model with Keras and constructed with 7 layers — 6 Conv1D layers followed by a Dense layer. The model only simply trained with batch_size=16 and 200 epochs without any learning rate schedule, etc. Its loss function is categorical_crossentropy and the evaluation metric is accuracy.

Exploratory Data Analysis:

In the RADVESS dataset, each actor has to perform 8 emotions by saying and singing two sentences and two times for each. As a result, each actor would induce 4 samples for each emotion except neutral, disgust and surprised since there is no singing data for these emotions. Each audio wave is around 4 second, the first and last second are most likely silenced.
The key features of the audio data are namely, MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram and Chroma.

- **MFCC (Mel Frequency Cepstral Coefficients)**- MFCC is taken on a Mel scale which is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear. The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.
- **Mel Spectrogram**- A Fast Fourier Transform is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram. This is just a spectrogram that depicts amplitude which is mapped on a Mel scale.
- **Chroma**- A Chroma vector is typically a 12-element feature vector indicating how much energy of each pitch class is present in the signal in a standard chromatic scale.

For the EDA we have used MFCC and Mel Spectogram

Observation:

After the selection 1 actor and 1 actress's dataset and listened to all of them. We found out male and female are expressing their emotions in a different way. Here are some findings:

1. Male's Angry is simply increased in volume.
2. Male's Happy and Sad significant features were laughing and crying tone in the silenced period in the audio.
3. Female's Happy, Angry and Sad are increased in volume.
4. Female's Disgust would add vomiting sound inside.

Result:

This excluded the class neutral, disgust and surprised to do a 10 class recognition for the RAVDESS dataset. We tried to replicate his result with the model provided.

However, we found out there is a data leakage problem where the validation set used in the training phase is identical to the test set. So, we re-do the data splitting part by isolating two actors and two actresses data into the test set which make sure it is unseen in the training phase.
Actor no. 1–20 are used for Train / Valid sets with 8:2 splitting ratio.
Actor no. 21–24 are isolated for testing usage.
Train Set Shape: (1248, 216, 1)
Valid Set Shape: (312, 216, 1)
Test Set Shape: (320, 216, 1) — (Isolated)
We re-trained the model with the new data-splitting setting and here is the result:

Augmentation:

After we tuned the model architecture, optimizer and learning rate schedule, we found out the model still cannot converge in the training period. we assumed it is the data size problem since we have 800 samples for train valid set only. Thus, we decided to explore the audio augmentation methods. Let's take a look at some augmentation method with code. we simply augmented all of the datasets once to double the train / valid set size.

Validation:

We can see that the augmentation can jack up the Validation Accuracy a lot, 70+% in general. Especially that adding white noise can achieve 87.19% Validation Accuracy, however, the Testing Accuracy and Testing F1-score dropped more than 5% respectively. That's how we achieved 82% accuracy.

V. Result & Discussions

In this project we expect to learn about emotions of a person from the speech or choices, and help them to enhance it from providing suggestions. As human emotion can be complex and can have many categories, we expect it work with accuracy up to 70-80%. With help of this project one can get insights of the feeling and improve if possible. Along with the test outcomes, we expect that if we use a new sound file ourselves, we will get the desired output. We will use the MLPClassifier which has an internal neural network for the purpose of classification. This is a feed forward ANN model. After that we train and test our model which gives us the output as emotions we mentioned based on its working and accuracy.

VI. Conclusion

A lot of uncertainties are still present for the best algorithm to classify emotions. Different combinations of emotional features give different emotion detection rate. The researchers are still debating for what features influence the recognition of emotion in speech. Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc..

Acknowledgment

We thank the almighty Lord for giving me the strength and courage to sail out through the tough and reach on shore safely. There are number of people without whom this projects work would not have been feasible. Their high academic standards and personal integrity provided me with continuous guidance and support. We owe a debt of sincere gratitude, deep sense of reverence and respect to our guide and mentor Prof. Kavita Namdev, Professor, AITR, Indore for his motivation, sagacious guidance, constant encouragement, vigilant supervision and valuable critical appreciation throughout this project work, which helped us to successfully complete the project on time.
We express profound gratitude and heartfelt thanks to Dr. Kamal Kumar Sethi, HOD CSE, AITR Indore for his support, suggestion and inspiration for carrying out this project. I am very much thankful to other faculty and staff members of CSE Dept, AITR Indore for providing me all support, help and advice during the project. We would be failing in our duty if do not acknowledge the support and guidance received from Dr. S C Sharma, Director, AITR, Indore whenever needed. We take opportunity to convey my regards to the management of Acropolis Institute, Indore for extending academic and administrative support and providing me all necessary facilities for project to achieve our objectives.

References

[1] https://librosa.org/librosa/tutorial.htmlhttps://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

[2] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[3] Mel Spectrogram- https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0

[4] https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3

[5] https://www.kaggle.com/kritika4142/speech-emotion-recognizer/notebook#LIVE-DEMO

[6] https://www.frontiersin.org/articles/10.3389/fcomp.2020.00014/full