ML Project Report

Anonymous CVPR submission

Paper ID *****

Abstract

One of the most fundamental components of humans to convey what they feel is via emotion. The best thing would be if machines were able to understand human feelings, it will improve human-machine interaction and emotions will be automatically recognized by machines. This is the reason why this particular topic has recently been gaining steadily more attention. The first problem is selecting adequate features for speech representation, the second problem is properly setting up a dataset of emotional speeches, and the third problem is designing a suitable classification technique.

1. Introduction

Humans most frequently and quickly communicate with one another through speech. This fact has prompted academics to view voice as a quick and effective method of human-machine communication. However, for this to work, the computer must be capable of comprehending human speech. Speech recognition, or the process of turning human speech into a series of phrases, has been the subject of extensive research. We are still far from developing a typical human-machine interaction since the machine cannot discern the speaker's emotional state, despite substantial advancements in speech recognition.

This has given rise to a comparatively new study area called Speech Emotion Recognition (SER), that is described such as extraction of a subject's emotional situation using their speech. Speech emotion recognition is thought to enhance the effectiveness of recognition system by allowing an extraction of meaningful contextual information from voice.

The method of speech emotion recognition is quite challenging for the reasons listed above. First off, it's not entirely obvious which aspects of speech are best at differentiating various sentiments. Another challenge is the auditory diversity brought about by the integration of varied terms, persons, expressing modes, and speaking speeds because these characteristics significantly affect the most of the key characteristics of derived audio, like tone and power levels. Furthermore, multiple emotions may be sensed while listening to the exact word; each emotion correlates to a distinct section of the uttered word. Furthermore, it is quite challenging to define the limits among these segments.

Furthermore, the problem is that the speaker, their culture, and surroundings all play a role in the way a particular feeling is communicated. The majority of research has concentrated on categorizing emotions in homogenous contexts, implying that people are from the same ethnic heritage.

An essential topic in speech emotion recognition is the requirement to give a list of significant emotions that must be recognized by a computerized sentiment identification system. Languages have created catalogues of the most prevalent feelings and emotions that humans encounter on a daily basis. It's challenging to categorize such a wide range of feelings, though.

Anger, disgust, fear, joy, sadness, and surprise are among the basic emotions. These feelings are the clearest and most vivid feelings we encounter throughout our lives. The term "archetypal emotions" refers to them.

2. Problem Statement

The factors that make speech emotion recognition a highly difficult endeavour. First, it is unclear which speech characteristics are effective at differentiating various emotions. The second problem is that a person's culture, geography, and linguistic all play a role in how they portray various emotions. The majority of the research has concentrated on categorizing emotions in solitary contexts, presuming that there are no regional variations among individuals. This study use machine learning approaches to accurately anticipate speech emotions.

3. Dataset

The extent of genuineness of the dataset being used to evaluate an emotional speech recognizer's efficiency is a crucial factor to take into account. Using a poor-quality data could result in the development of false conclusions.

Furthermore, for the classification model under consideration, the dataset is crucial.

There ought to be standards by which one can assess how well a given emotional database replicates the world today. It is more useful to use voice data gathered from real-world events. These files contain sounds that very organically describe sentiments. Consequently, they might not be appropriate for use in research due to moral and ethical issues. Similar to the modern bookstores, acoustic labs can likewise produce sensitive phrases. This has traditionally been maintained that enacted emotions are distinct from true ones.

3120 files can be found in the RAVDESS database, which is maintained by Ryerson University. 10 professional actors, five of whom are women and five of whom are men, are recorded in the database performing two phonemically related sentences with neutral North American accents. Both song and speech can display emotions such as calmness, happiness, sadness, anger, fear, surprise, and disgust. With the inclusion of a neutral expression, each facial expression is produced at two different emotional intensity levels (normal and strong). The three modality forms are Audio-only (16bit, 48kHz.wav), Audio Video (720p H.264, AAC 48kHz,.mp4), and Video-only.

4. Problem Statement

All actors (01–10) have audio-only files available:

• There are 600 files in the speech file (Audio Speech Actors 01- 10.zip): 60 trials divided by 10 actors equals 600.

• There are 440 files in the song file (Audio Song Actors 01- 10.zip): Ten performers times 44 trials each equals 440.

The RAVDESS collection consists of 3120 files in total (600+440+1200+880 files).

File naming standards:

The 3120 RAVDESS files each have a distinct filename. A seven-part number identification (for example, 02-01-06-01-02-01-10.mp4) makes up the name. These descriptors specify the features of the trigger:

5. Database Design Criteria

There ought to be a criteria for judging how well an emotional data replicates an external world. The aforementioned are said to be the most crucial factors to take into account: of staged emotions or those found in the real world? Utilizing voice data that has been gathered from actual life scenarios makes the system more lifelike. Such audios include sounds that communicate sentiments in a very conventional manner.

Who displays emotions? Many libraries for emotional expression ask skilled actors to deliver pre-written words while evoking the necessary emotions. To avoid emphasizing emotions and to be more authentic, some of them, meanwhile, use sub performers as an alternative.

How will the expressions be reproduced? The sounds that were not produced in a social environment can be found in several emotional speech collections. As a reason, since it is considered that most emotions are the outcome of our responses to diverse situations, speech can have no sense of genuineness. There are typically two methods for evoking emotional statements. The very first method involves trained communicators acting as though they were in a certain mental situation, such as being happy, mad, or sad. Many advanced organizations did not have access to such seasoned performers and instead recruited moderately or novice players to deliver the emotive lines.

Are speech distributed throughout sentiments equally? In order to more effectively assess categorization accuracy, desire that the amount of sentences for every feeling be nearly equal. In contrast, many other academics believe that the ratio of emotions in the dataset must correspond to how frequently they arise in reality. The neutral emotion, for instance, is the most prevalent in daily life. Therefore, in the corpus of emotional experience, the proportion of statements expressing neutral feeling should be the largest.

6. Data Visualisation

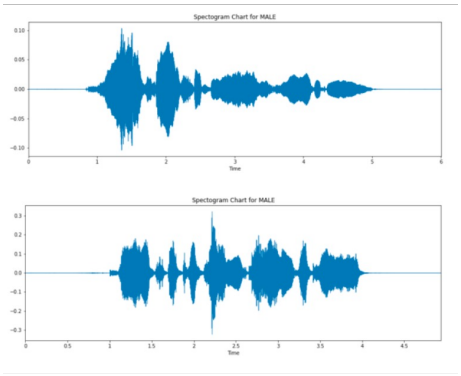

Figure 1. Male spectogram

7. Methodology

For this prediction, we've utilized the Python soundfile, librosa, and sklearn packages. We built a model using MLP-Classifier. The log-loss function is optimized by the MLP classifier using stochastic gradient descent. Given that the data is somewhat vast, we have also employed the Adam optimizer here. A feedforward artificial neural network is a multilayer perceptron. Backpropagation is used by MLP as

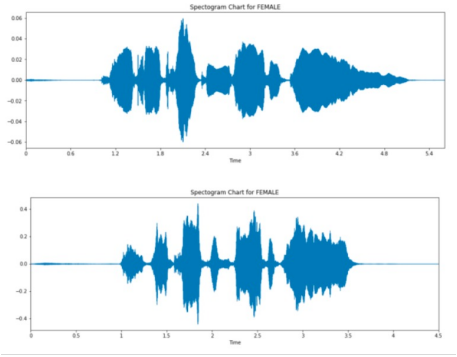

Figure 2. Female Spectogram

supervised learning during training. In a perceptron, learning happens by adjusting connection weights based on mistake following data processing. MLP has a built-in neural network for classification purposes.

MFCC: Mel Frequency Cepstral Coefficient, represents short term power spectrum of a sound.

• Chroma: Pertains to 12 pitch class. • Mel: Mel Spectrogram Frequency.

In order to distinguish language features from noisy environment, the very first stage in a speech recognition system is to determine the elements of a voice file that are useful for doing so.

Chroma: The amount of energy in each pitch class contained in the signal is indicated by a chroma vector, which is typically a 12-element feature vector.

Mel: The frequency scale underwent some sort of nonlinear alteration to produce the Mel Scale. This Mel Scale is designed such that noises that are equally spaced apart on the Mel Scale also ”sound” to people since they are equally spaced apart.

As opposed to the Hz scale, where the distinction between 500 and 1000 Hz is clear, the distinction between 7500 and 8000 Hz

The Cepstrum:

The Fourier transform can be used to analyse periodic features in a signal on various scales. To be more specific, we can use the log-spectrum can be transformed using the discrete cosine transform (DCT) or the Fourier transform (DFT) to get the cepstrum, a representation. Since this representation is a complex rearrangement of time-frequency transforms, the name tries to be a humorous reflection of that fact. Technically speaking, the cepstrum for a temporal signal x(t) is defined as

$F(x(t)) = F\text{-}1\ [LOG\ (\ F\ [\ x(t)\ ]\ )\ ]$

In this case, F stands for the Fourier Transform while F inverse stands for the Inverse Fourier Transform. The harmonic makeup of the log-spectrum is an important aspect of the cepstrum.

The MEL frequency CEPSTRAL COEFFICIENT is: The frequency-to-mel transform function for a frequency f is defined as M = 2595 * log10(1 + f/700) in a classical approximation. It is simple to derive the inverse transform as F = 700 (10m/2595-1).

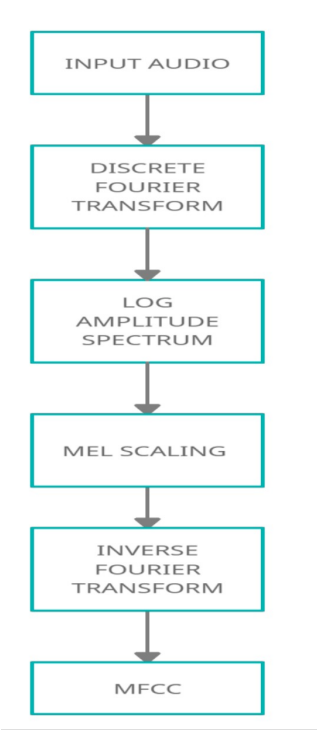

Figure 3. Representing MFCC flowchart

8. Speech Emotion Recognition

A neural network-based system for language-free emotion recognition has been implemented. Ten subjects were chosen after considering eight emotional classes. The spoken words were captured in English

For training and testing, 3120 spoken utterances that were each delivered with one of eight distinct emotions were used. 936 utterances were chosen from these samples for the network’s training, while the remaining utterances were used for testing. The voice spectrogram is analyzed to derive a total of 180 prosodic features. On the test set, the Neural Network with chosen features achieved an overall classification accuracy of 63.31

The MLP Classifier was used to train the model, and it produced the best accuracy of any classifier I tested—nearly 78 percent. In addition, I utilized Random Forest Classifier.

9. Applications

Applications that call for genuine human-machine communication, such web movies and voice-overs, can benefit greatly from SER and computer instructional programs where the user's answer is based on the emotion that is being sensed. It is also helpful for the in-board systems, which can be informed of the driver's cognitive state to start the protection process. Doctors can use it as a clinical instrument as well. Additionally, machine translations may benefit from the methods in which the subject's psychological response significantly influences how the participants communicate. Speech processing systems that are trained to recognize pressured speech perform better in airplane cockpits than those that are trained to recognize everyday conversation. Speech emotion detection is mostly used to modify the system's reaction when it detects anger in the human voice.

10. Conclusion

his paper provides an overview of recent research in the field of speech emotion recognition. The characteristics of various emotions, study categorization methods, and crucial design requirements of emotional speech datasets significant issues—have all been investigated. Out of this work, a number of inferences can be taken. The first is that while good classification accuracy among high and low aroused sentiments has been attained. Furthermore, there is still much room for enhancement in the effectiveness of the available tension sensors. Most strategies have an moderate classification accuracy of less than 80 percent for person speaking speech emotion identification systems.

Three classifiers have been tried for speech emotion recognition such as the Neural Networks, MLP, LSTM and Random Forest. However, we see MLP performs better than rest.

11. Future Work

The majority of recent studies concentrates on investigating various speech qualities and how they relate to the emotional resonance of spoken occurrences. As well as the TEO-based functionality, additional features were also developed. In an effort to locate the optimal features for this purpose, there are additional attempts to use other attribute selection strategies. However, there is a lack of coherence in the results we got from the many investigations. The primary explanation for this could be that every experiment only looks at one collection of emotional speech. The majority of the current databases are not ideal for assessing the effectiveness of a SER system.

Other issues for some datasets include the poor caliber of the captured sounds, the dearth of usable expressions, and the absence of phonological notations. As a result, it's possible that a few of the results obtained from certain studies can't be applied to certain other datasets. More collaboration between research institutions is required to create standard emotional speech files in order to solve this issue. The various potential modifications are suggested to improve the effectiveness of current SER systems. The first expansion is based on the idea that categorization that is person speaking is typically simpler than categorization that is speaker-independent.

Additionally, it should be emphasized that the bulk of classification methods in use today do not simulate the sequence of the learning algorithm. The HMM might be the lone exception, as temporal dependency can be described using its states. In actuality, this presumption is false. The performance of the classifier is hoped to be improved by simulating of the relationship between extracted features, such as by the use of regression models.

References

1. W. Campbell, "Databases of emotional speech", in Proceedings of the International Speech Communication and Association (ISCA) ITRW on Speech and Emotion, 2000, pp. 34-38.

2. C. Lee, S. Narayanan, "Toward detecting emotions in spoken dialogs", IEEE Transactions on Spee

3. Tomas Pfister and Peter Robinson,"Real- Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis",IEEE Transactions A

4. W. Campbell, "Databases of emotional speech", in Proceedings of the International Speech Communication and Association (ISCA) ITRW on Speech and Emotion, 2000, pp. 34-38.

5. T. Nwe, S. Foo, L. De Silva, "Speech emotion recognition using hidden Markov models", Speech Commun., 2003, Vol. 41, pp. 603-623