JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Speech Emotion Recognition using Machine Learning

Mrs. T.Veda Reddy,

tvedareddycse@cvsr.ac.in

Assistant professor, Department of Computer Science & Engineering, Anurag University, Telangana, India.

B.Saicharan,

18h61a05c9@cvsr.ac.in

P.Sandesh,

18h61a05g1@cvsr.ac.in

M. Muralidhar Reddy,

18h61a05f9@cvsr.ac.in

A.Saketh Reddy

18h61a05c5@cvsr.ac.in

Abstract:

Emotions are the most common way on how express ourselves. Emotion detection from voice can be used for various applications like Digital assistants like Siri, Cortona, Alexa, Google Assistant. For business marketing, where they can recommend products based on user emotions and machine customer support, machines can detect the customer's emotions and replay according to that situation based on emotions. Similarly, if we use emotion detection, it can help detect the user's exact emotion as well, i.e., the user's mood, and respond according to the user's perspective. This work mainly involves recognizing the emotion of the speaker from a voice sample and discovering multi-level representation of the signal. We are considering three datasets for accomplishing this task, namely – RAVDESS, TESS, and SAVEE datasets. We have 5732 unique voice files with these three datasets combined. Our voice contains many features like Energy, Pitch, Rhythm, Loudness, and so on.

Keywords:

Speech Emotion Recognition, Machine Learning, Librosa.

I. Introduction:

In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications. At present, speech emotion recognition is an emerging crossing field of artificial intelligence and artificial psychology; besides, it is a popular research topic of signal processing and pattern recognition. The research is widely applied in human-computer interaction, interactive teaching, entertainment, security fields, and so on. Speech emotion processing and recognition system is generally composed of three parts, the first being speech signal acquisition, then comes the feature extraction followed by emotion recognition. The most propitious technique for speech recognition is the neural network based approach. Artificial Neural Networks, (ANN) are biologically inspired tools for information processing. Speech recognition modelling by artificial neural networks (ANN) doesn't require any prior knowledge of speech process and this technique quickly became an attractive substitute to HMM. RNN can learn the sublunary relationship of Speech – data & is capable of modelling time dependent phonemes. The conventional neural networks of Multi- Layer Perceptron (MLP) type have been increasingly in use for speech recognition and also for various other speech processing applications. Speech recognition is the process of converting an acoustic signal, captured by microphone or a telephone, to a set of characters. They can also serve as the input to further linguistic processing to achieve speech understanding, a subject covered in section. As we know, speech recognition performs tasks that similar with human brain.

II. Literature Survey:

Guihua Wen et al. [1][2] have proposed a random Deep Belief Networks for Recognizing Emotions from Speech Signals, which describes about ensemble learning technique of Random Deep Belief Networks (RDBN) method for recognizing emotions from speech signals. Where they firstly extracted the low-level features of the given input speech signal and then applied the method of Random subspaces. Where each Random subspace is fed into the input of DBN to extract the higher-level features of the given input speech signal and provided these higher-level features as the input to the base classifier to output a predicted emotion label. Furthermore, each outputted emotion label is fused through the majority voting to decide the given input speech signal's final emotion label. M. Shamim Hossain and Ghulam Muhammad have proposed an emotion recognition system using deep learning approach from Audio-Visual emotional big data, which describes how emotions can be detected using speech and video as input. For this purpose, have used two datasets; one is a Big Data database containing both speech and video input files, and another one is the eNTERFACE database. In this method, they first extracted the given input speech signals features to obtain a Mel-spectrogram, which can be considered an image. This Mel-spectrogram is fed to 2D CNN followed by extreme learning machines (ELMs) for the fusion of scores. Similarly, in the case of video signals, some representative frames from a video segment are extracted and fed to the 3D CNN, followed by extreme learning machines (ELMs) for the fusion of scores. The output of both these speech and video fusions is given to the SVM for the final classification of the emotions of given input speech and video signals. Mingke Xu et al have proposed a speech emotion recognition with multiscale area attention and data augmentation, where they have applied multiscale area attention to SER and designed an attention-based convolutional neural network. In the work, firstly, they used the Librosa library to extract the log Mel spectrogram as features; these features are fed into two parallel convolutional layers to extract textures from these features for the time axis and frequency axis, respectively. The output of this is fed into four consecutive convolutional layers and generates an 80-channel representation. Followed by attention layer attends on the representation and sends the results to the fully connected layer for final classification of emotion. Furthermore, there are lot of work done using SVM and its combination methods. However, the state of the art do not combine random subspace, MLP and CNN for speech emotion recognition in the framework of ensemble learning. Also, the selection of a best model for SER is not yet studied.

III. Methodology:

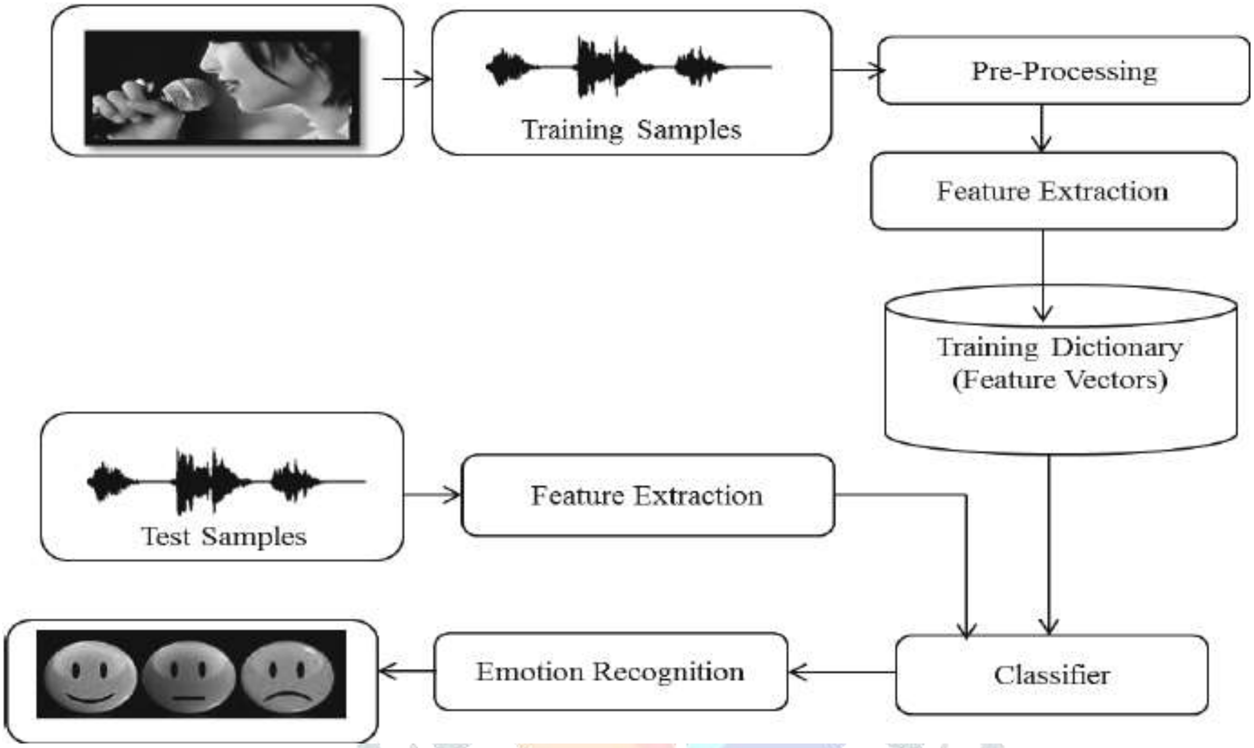


Fig 1.suggests the block diagram of a Speech Emotion Recognition

☒ We are going to build a classification model using Multi-layer perceptron classifier also known as MLP classifier.Multi layer perceptron (MLP) is a supplement of feed forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer.The input layer receives the input signal to be processed. The required task such as prediction and classification is performed by the output layer.

☒ **The two main reasons to use MLP classifier :**

MLP Classifier relies on an underlying Neural Network to perform the task of classification.

Model Performance

A. System Architecture:

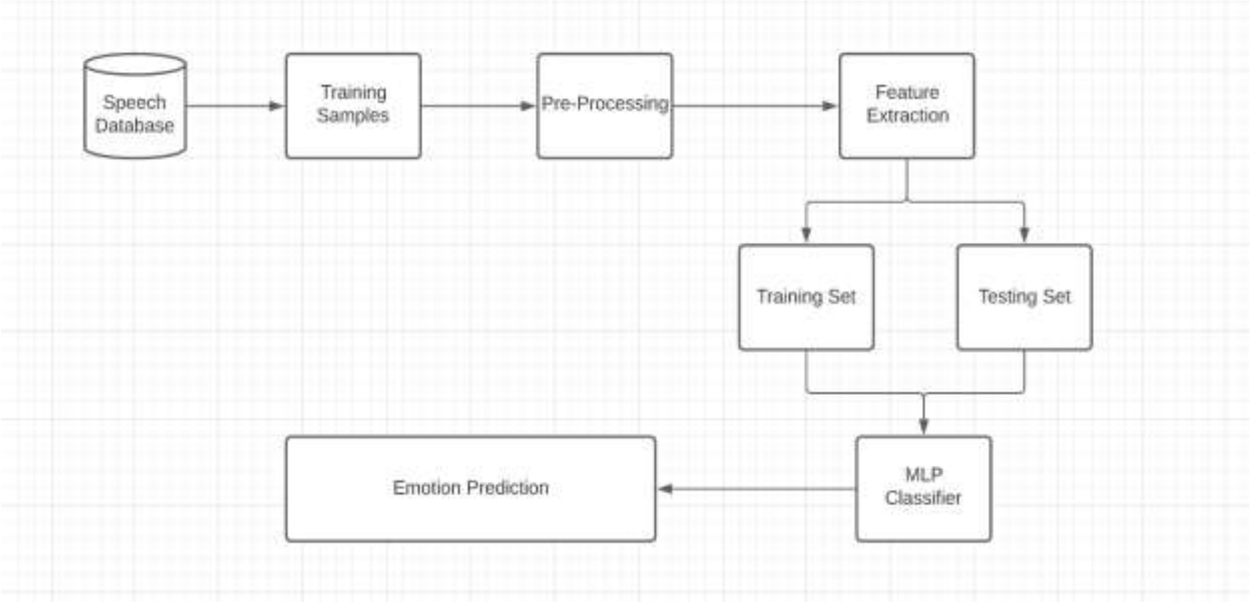


Figure 2: Architecture

Collecting facts is step one of the way drift which includes defining the project, installing location the device environment suitable for the development requirements and later statistics the facts using wonderful Python libraries and device studying techniques. Data Cleaning wants to be achieved on the facts accumulated just so the assessment be very accurate for satisfactory results.

B. Algorithm:

MLP Classifier: MLP Classifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLP Classifier relies on an underlying Neural Network to perform the task of classification

Feature Extraction:

In Feature extraction, we are converting our given input audio files to digital data since the models will not be able to understand the audio data so that we are transforming our given input audio file to digital data so that the model will be able to understand this digital data for this we are using librosa module which contains various libraries for extracting the features from the given input audio file by using sample rate and sample data. Firstly, we are considering Mel Frequency Cepstral Coefficient (MFCC) feature, which is the most important and effective method for performing this feature extraction task, it analyses the signal based on short term power spectrum i.e. by segmenting the speech sample into number of frames by framing, then by applying certain window to reduce the signal discontinuities at the beginning and end of each frame, after that by using FFT to identify frequency spectrum of each frame by analysing which frequency is present in the particular frame, this frequency spectrum is passed to map the powers of the frequency spectrum onto the Mel scale, and then by applying log of these powers, followed by the discrete cosine transform for eliminating overlapping of filter banks and then finally applying a cepstral mean correction by subtracting the cepstral mean of a frame from the cepstral coefficients for providing increased robustness in recognition.

Second feature, we are considering is that Mel feature, which will be used to capture characteristic of the frequency of the given input audio file signal represented on the Mel scale. T

Third feature, we are considering is that Chroma feature which is used to capture melodic and harmonic characteristics of sound based on pitch of the given input audio file, Zcr feature, which is used to specify the rate of sign changes of the particular signal during the duration of the particular frame, and Rms feature, which is used to analyse the loudness in the given input audio file since changes in loudness are important for extracting features in new input file.

IV. Implementation:

Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation(using LSTM's), Automatic Speech Recognition.It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques.

A. Introduction to Technologies Used :

Python

Python is presently the maximum extensively used multi-motive, immoderate-degree programming language. Python allows programming in any element-oriented and procedural paradigm. Python packages are typically smaller than precise programming languages consisting of Java. Programmers need to kind plenty less, and the language's indentation requirement makes them readable all of the time. The Python language is used by almost all companies of the time such as Google, Amazon, Facebook, Instagram, Dropbox, Uber...etc. The largest power of Python is its big series of modern libraries which may be used for the subsequent –

- Machine Learning
- Sklearn
- Librosa
- PyAudio
- SoundFile

Scikit-learn

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. Think of any supervised machine learning algorithm you might have heard about and there is a very high chance that it is part of scikit-learn. Starting from Generalized linear models (e.g Linear Regression), Support Vector Machines (SVM), Decision Trees to Bayesian methods – all of them are part of scikit-learn toolbox.

PyAudio

PyAudio provides Python bindings for PortAudio, the cross-platform audio I/O library. With PyAudio, you can easily use Python to play and record audio on a variety of platforms. Recording Audio the python-sounddevice and pyaudio libraries provide ways to record audio with Python. python-sounddevice records to NumPy arrays and pyaudio records to bytes objects. Both of these can be stored as WAV files using the scipy and wave libraries, respectively.

Soundfile

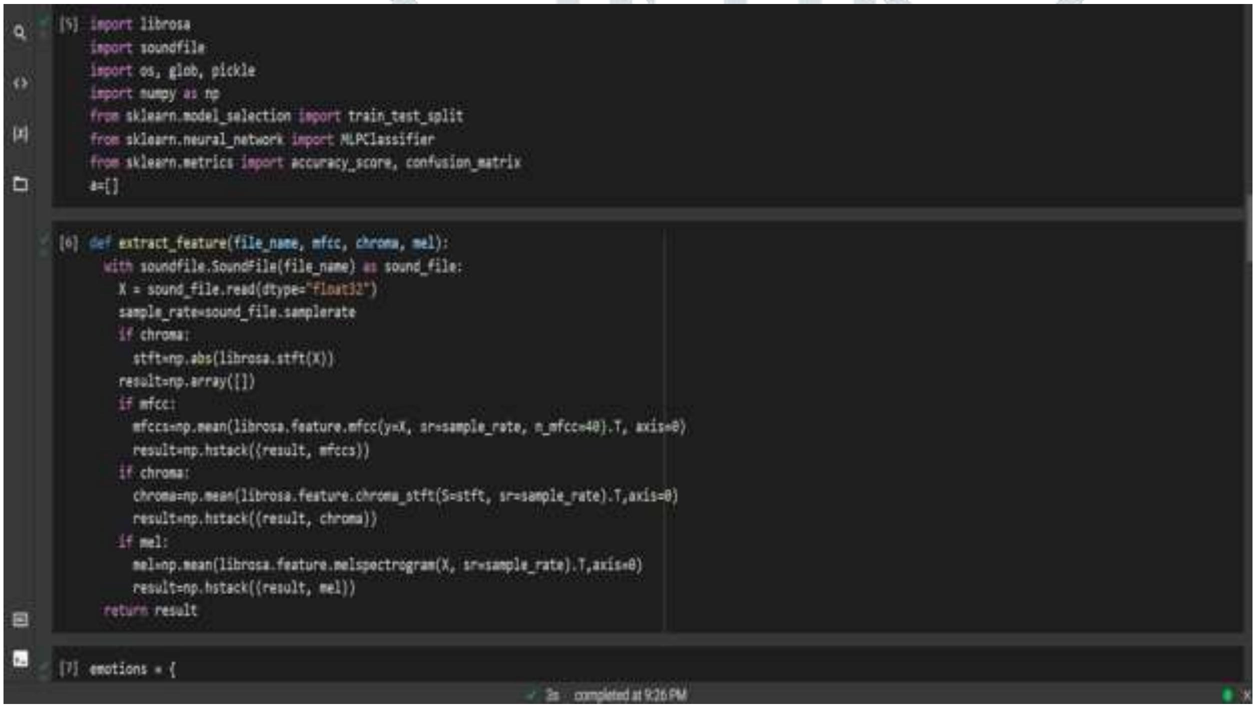
SoundFile is an audio library based on libsndfile, CFFI and NumPy.SoundFile can read and write sound files. File reading/writing is supported through libsndfile, which is a free, cross-platform, open-source (LGPL) library for reading and writing many different sampled sound file formats that runs on many platforms including Windows, OS X, and Unix.

Sklearn

Scikit-take a look at (Sklearn) is the maximum beneficial and sturdy library for system reading It offers a preference of green system for system reading and statistical modeling together with elegance, regression, clustering, and dimensionality good deal via a regular interface in Python. This library, which is primarily written in Python, is built on **NumPy, SciPy, and Matplotlib.**

V. Results and Discussion:

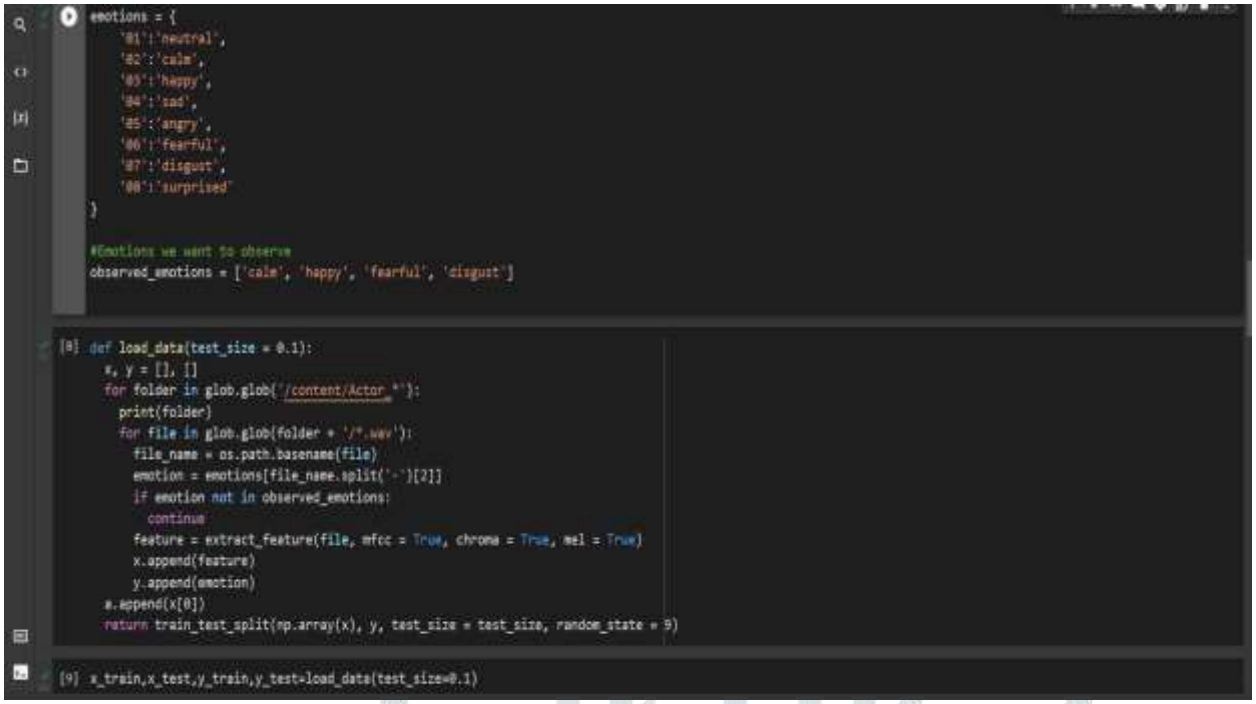
Sample code



```
[5] import librosa
    import soundfile
    import os, glob, pickle
    import numpy as np
    from sklearn.model_selection import train_test_split
    from sklearn.neural_network import MLPClassifier
    from sklearn.metrics import accuracy_score, confusion_matrix
    a=[]

[6] def extract_feature(file_name, mfcc, chroma, mel):
      with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
          stft=np.abs(librosa.stft(X))
        result=np.array([])
        if mfcc:
          mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
          result=np.hstack((result, mfccs))
        if chroma:
          chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
          result=np.hstack((result, chroma))
        if mel:
          mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
          result=np.hstack((result, mel))
      return result

[7] emotions = {
```



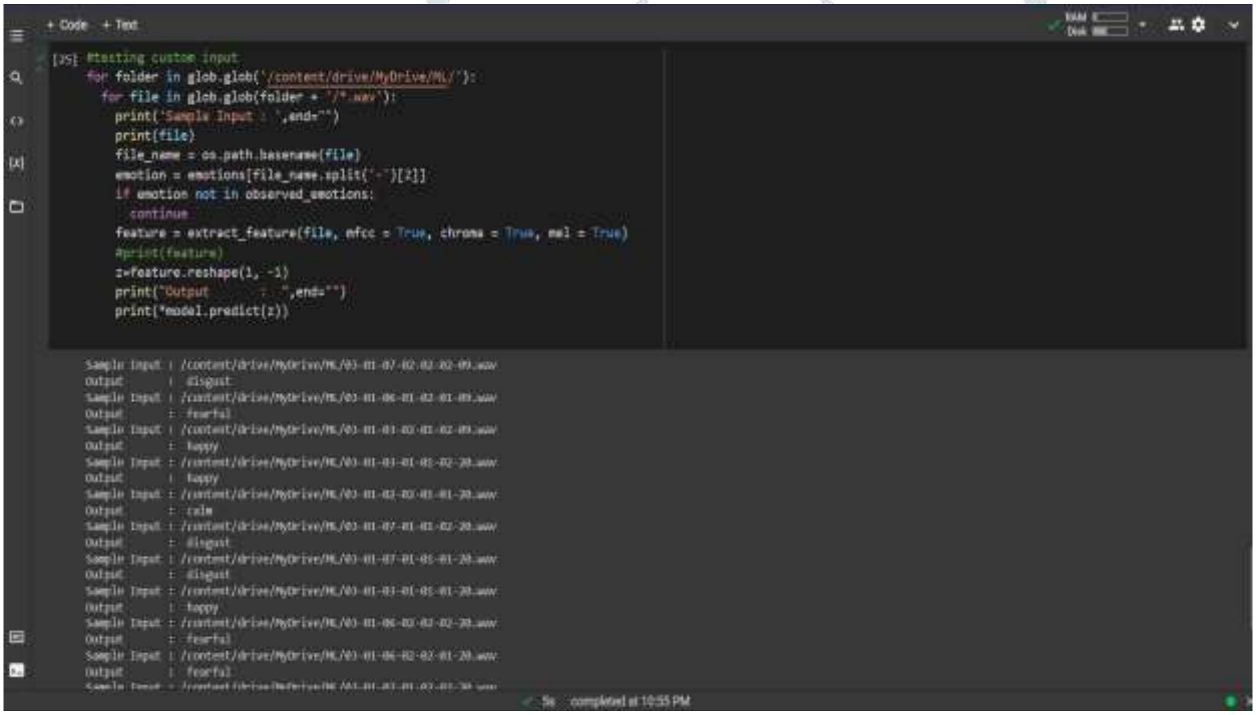
```
emotions = {
    '01':'neutral',
    '02':'calm',
    '03':'happy',
    '04':'sad',
    '05':'angry',
    '06':'fearful',
    '07':'disgust',
    '08':'surprised'
}

#Emotions we want to observe
observed_emotions = ['calm', 'happy', 'fearful', 'disgust']
```

```
[8] def load_data(test_size = 0.1):
      x, y = [], []
      for folder in glob.glob('/content/Actor_*'):
        print(folder)
        for file in glob.glob(folder + '/*.wav'):
          file_name = os.path.basename(file)
          emotion = emotions[file_name.split('-')[2]]
          if emotion not in observed_emotions:
            continue
          feature = extract_feature(file, mfcc = True, chroma = True, mel = True)
          x.append(feature)
          y.append(emotion)
      x.append(x[0])
      return train_test_split(np.array(x), y, test_size = test_size, random_state = 9)
```

```
[9] x_train,x_test,y_train,y_test=load_data(test_size=0.1)
```

Test cases :



```
[25] #testing custom input
     for folder in glob.glob('/content/drive/MyDrive/ML/'):
       for file in glob.glob(folder + '/*.wav'):
         print('Sample Input : ',end="")
         print(file)
         file_name = os.path.basename(file)
         emotion = emotions[file_name.split('-')[2]]
         if emotion not in observed_emotions:
           continue
         feature = extract_feature(file, mfcc = True, chroma = True, mel = True)
         #print(feature)
         z=feature.reshape(1, -1)
         print("Output       : ",end="")
         print(*model.predict(z))
```

```
Sample Input : /content/drive/MyDrive/ML/03-01-07-02-02-02-09.wav
Output       :  disgust
Sample Input : /content/drive/MyDrive/ML/03-01-06-01-02-01-09.wav
Output       :  fearful
Sample Input : /content/drive/MyDrive/ML/03-01-03-02-01-02-09.wav
Output       :  happy
Sample Input : /content/drive/MyDrive/ML/03-01-03-01-01-02-20.wav
Output       :  happy
Sample Input : /content/drive/MyDrive/ML/03-01-02-02-01-01-20.wav
Output       :  calm
Sample Input : /content/drive/MyDrive/ML/03-01-07-01-01-02-20.wav
Output       :  disgust
Sample Input : /content/drive/MyDrive/ML/03-01-07-01-01-01-20.wav
Output       :  disgust
Sample Input : /content/drive/MyDrive/ML/03-01-03-01-01-01-20.wav
Output       :  happy
Sample Input : /content/drive/MyDrive/ML/03-01-06-02-02-02-20.wav
Output       :  fearful
Sample Input : /content/drive/MyDrive/ML/03-01-06-02-02-01-20.wav
Output       :  fearful
```

Conclusion:

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.In this Python mini project, we learned to recognize emotions from speech. We used an MLPClassifier for this and made use of the soundfile library to read the sound file, and the librosa library to extract features from it.

Future Enhancement:

A speech emotion recognition algorithm based on multi-feature and Multi-lingual fusion is proposed in order to resolve low recognition accuracy caused by lack of large speech dataset and low robustness of acoustic features in the recognition of speech emotion. First, handcrafted and deep automatic features are extracted from existing data in Chinese and English speech emotions. Then, the various features are fused respectively. Finally, the fused features of different languages are fused again and trained in a classification model. Distinguishing the fused features with the unfused ones, the results manifest that the fused features significantly enhance the accuracy of speech emotion recognition algorithm. The proposed solution is evaluated on the two Chinese corpus and two English corpus, and is shown to provide more accurate predictions compared to original solution. As a result of this study, the multi-feature and Multi-lingual fusion algorithm can significantly improve the speech emotion recognition accuracy when the dataset is small.

VI. Bibliography:

[1] Ronald, M. Baecker, “Readings in human-computer interaction: toward the year 2000”, 1995.

[2] Melanie, Pinola, “Speech Recognition Through the Decades: How We Ended Up With Siri”, PCWorld.

[3] Ganesh Tiwari, “Text Prompted Remote Speaker Authentication: Joint Speech and Speaker Recognition/Verification System”.

[4] Dr.Ravi Sankar, Tanmoy Islam, Srikanth Mangayyagari, “Robust Speech/Speaker Recognition Systems”.

[5] Bassam A.Q.Al-Qatab and Raja.N.Aninon, “Arabic Speech Recognition using Hidden Markov Model ToolKit (HTK)”, IEEE Information Technology (ITSim), 2010, pp. 557-562.

[6] Ahsanul Kabir, Sheikh Mohammad Masudul Ahsan, ”Vector Quantization in Text Dependent Automatic Speaker Recognition using Mel-Frequency Cepstrum Coefficient”, 6th WSEAS International Conference on circuits, systems, electronics, control & signal processing, Cairo, Egypt, dec 29-31, 2007, pp. 352-355

[7] Lindasalwa Muda, Mumtaj Begam and Elamvazuthi.,”Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Techniques “,Journal of Computing, Volume 2, Issue 3, March 2010

[8] Mahdi Shaneh and Azizollah Taheri, ”Voice Command Recognition System based on MFCC and VQ Algorithms”, World Academy of Science, Engineering and Technology Journal, 2009.

[9] Remzi, Serdar Kurcan, “Isolated word recognition from in-ear microphone data using hidden Markov models (hmm)”, Master’s Thesis, 2006.

[10] Nikolai Shokhirev, ”Hidden Markov Models “, 2010.

[11] L.R. Rabiner, “A tutorial on Hidden Markov Models and selected applications in Speech Recognition”, Proceedings of the IEEE Journal, Feb 1989, vol. 77, Issue: 2. [12] Four phonetically balanced words list, Available from: http://www.meyersound.com/support/papers/speech/pblist.htm