Speech Emotion Recognition

Prithvi J (1BM19CS122)
S Shree Lakshmi (1BM19CS136)
Sohan R Kumar (1BM19CS159)
Suraj Nair (1BM18CS163)

Guide: **Dr. Kavitha Sooda**
Designation: Associate Professor
Department of Computer Science & Engineering
B.M.S. College of Engineering

OUTLINE

1. Abstract
2. Introduction
3. Problem Statement
4. Literature Survey
5. References

ABSTRACT

- Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech. Selection of suitable feature sets, design of a proper classifications methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems.
- In a voice-based system, a computer agent is required to completely comprehend the human's speech percept in order to accurately pick up the commands given to it. This field of study is termed as Speech Processing and consists of three components:
- Speaker Identification
- Speech Recognition
- Speech Emotion Detection
- Speech Emotion Detection is challenging to implement among the other components due to its complexity.

INTRODUCTION

- Speech is the fast and best normal way of communicating amongst human. This reality motivate many researchers to consider speech signal as a quick and effective process to interact between computer and human. It means the computer should have enough knowledge to identify human voice and speech.
- The recognition of emotional speech aims to recognize the emotional condition of individual utterer by applying his/her voice automatically.Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction.
- We are going to develop a machine learning project to automatically classify the given input sample (audio) into a few predefined emotions.
- In this project we transform the audio data into numeric include Mel Spectrograms that visualize audio signals based on their frequency components which can be plotted as an audio wave and fed to train a CNN as an image classifier. We can capture this using Mel-frequency cepstral coefficients (MFCCs)

PROBLEM STATEMENT

- In this project, We transform the audio data into numeric include Mel Spectrograms that visualize audio signals based on their frequency components which can be plotted as an audio wave and fed to train a CNN as an image classifier. We can capture this using MFCCs. After extracting features from the audio, the popular choice of model architecture has changed over time. Some of the commonly used ones are:
- **RNN/LSTMs**:Numeric features are fed to a neural network that generates an output logit vector.
- **Attention-based models**:use previously predicted sequences and learn the mapping of new ones using an encoder-decoder approach.
- **Listen-Attend-Spell (LAS)**: This was one of the first approaches to combine the above two methods by creating an encoder that learns features using bidirectional LSTMs.

LITERATURE SURVEY

Dataset overview:

MU-Multimodal (CMU-MOSI) is a benchmark dataset used for multimodal sentiment analysis. It consists of nearly 65 hours of labeled audio-video data from more than 1000 speakers and six emotions: happiness, sadness, anger, fear, disgust, surprise.

EMOVO Corpus: Containing over 500 samples showing six emotional states: disgust, fear, anger, joy, surprise, sadness, and the neutral state. The original dataset can be found here - EMOVO Corpus.

TESS: Toronto Emotional Speech: Recordings contain the same set of phrases spread over seven emotions (anger, disgust, fear, happiness, surprise, sadness, and neutral) totaling 2800 samples.

Libraries Installation:

Before we move to load dataset and model building it's important to install certain libraries. We will use the python speech feature library to extract features.to load the dataset in the WAV format we will use the scipy library.

Methodology:

The speech emotion detection system is implemented as a Machine Learning (ML) model.The flowchart represents a pictorial overview of the process.

1.data collection:The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data.

2.feature engineering:is a collection of several machine learning tasks that are executed over the collected data.

3.algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to
respond to any new data it is exposed to.

4.is to evaluate the functioning of the built model.

Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms.

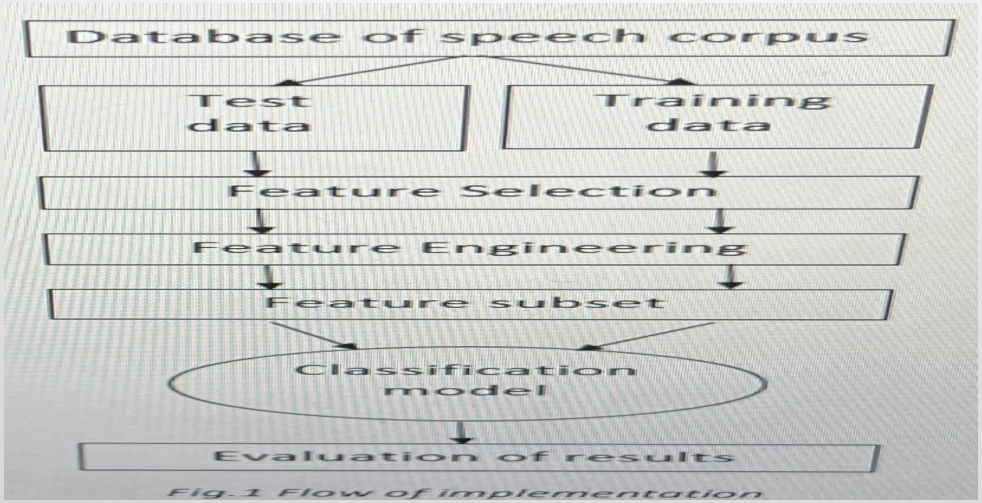


Fig.1 Flow of implementation

DATA COLLECTION

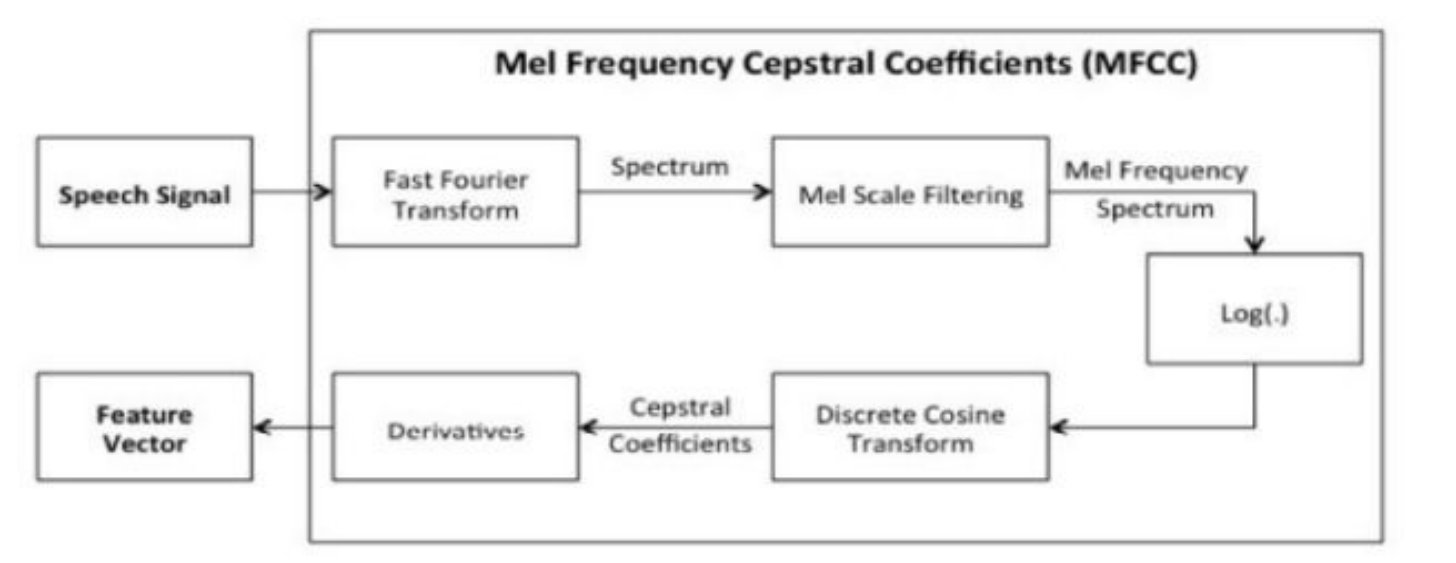
The first step in implementing the Speech Emotion Recognition system is to collect audio samples under different emotional categories which can be used to train the model. The audio samples are usually wav or mp3 files and publically available for download. The following steps are explained relative to the experiments performed on the TESS dataset. The next step after data collection was to represent these audio files numerically, in order to perform further analysis on them. This step is called feature extraction, where quantitative values for different features of the audio is obtained. The pyAudioAnalysis library was used for this purpose. pyAudioAnalysis is an open Python library that provides a wide range of audio-related functionalities focusing on feature extraction, classification, segmentation, and visualization issues.

FEATURE ENGINEERING

- the process of transforming, reducing or constructing features for the dataset. As mentioned earlier in the raw data, each feature has multiple values for each frame of the audio signal.
- By the frame blocking and windowing techniques, the frame size and frame overlap values can be tuned to obtain accurate values of the audio signal. Further, using the averaging technique, average values of different features for the audio signals are obtained. Now the transformed data contains 34 discrete values representing each audio signal. Reducing the number of features is a crucial decision to take. Considering features to be removed is generally based on subject knowledge and hence can affect the performance of the system.
- Next, a series of experiments are performed with this prepared dataset in order to analyze the important features.

Mel Frequency Cepstral Coefficients(MFCC):

- Since the audio signals are constantly changing, first we divide these signals into smaller frames. Each frame is around 20-40ms long.
- We identify different frequencies present in each frame.
- Then separate linguistic frequencies from the noise.
- To discard the noise, it then takes discrete cosine transform (DCT) of these frequencies using which we keep only a specific sequence of frequencies that have a high probability of information.



K-Nearest Neighbors Algorithm (KNN):

K-Nearest Neighbor (KNN) is the simplest classification algorithm. The approach is to plot all data points on space, and with any new sample,observe its k nearest points on space and make a decision
based on majority voting. Thus, KNN algorithm involves no training and it takes the least calculation time
when implemented with an optimal value of k. The steps of KNN algorithm is as follows:

1. For a given instance, find its distance from all other data points. Use an appropriate distance
metric based on the problem instance.
2. Sort the computed distances in increasing order. Depending on the value of k, observe the nearest
k points.
3. Identify the majority class amongst the k points, and declare it as the predicted class. Choosing an optimal value of k is a challenge in this approach. Most often, the process is repeated for a number of different trials of k. The evaluation scores are then observed using a graph to find the optimal value of k.

There is no training in the model of KNN and hence there is no training time complexity value. While testing, the number of nearest samples to be looked up for decides the complexity of the algorithm and
is controlled by the value of k.

EVALUATION

- The most important characteristic of machine learning models is its ability to improve. Once the model is built, even before testing the model on real data, machine learning experts evaluate the performance of the model.
- Evaluation metrics reveal important model parameters and provides numeric scores that will help judge the functioning of the model.
- The most important metric needed to evaluate the model is the confusion matrix.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.

[2] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression.,” Pers. Soc. Psychol, vol. 70, no. 3, pp. 572–587, 1996.

[3] V. Hozjan and Z. Kačič, “Context-Independent Multilingual Emotion Recognition from Speech Signals,” Int. J. Speech Technol., vol. 6, no. 3, pp. 311–320, 2003..

[4]F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech.,” in Interspeech, 2005, vol. 5, pp. 1517–1520.

[5] A. Schuller, B. , Steid, S. l, and Batliner, “The interspeech 2009emotion challengee,” Interspeech, pp. 312–315, 2009.

[6]W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion Recognition and Affective Computing on Vocal Social Media,” Inf. Manag., Feb. 2015

[7]S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

[8] H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[9] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.

[10] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

THANK YOU