A Project Report

on

Speech emotion recognition

Submitted in partial fulfillement of the
Requirement for the award of the degree

Bachelor of Technology in Computer Science and Engineering

GALGOTIAS UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Mr.Rajkamal Kishor Gupta
Associate Professor
Department of Computer Science and Engineering

Submitted By
19SCSE1180028 - Nikita Sirohi
19SCSE1180027 - Shivank

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE’S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **“ SPEECH EMOTION RECOGNITION”** in partial fulfillment of the requirements for the award of the BTech submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of July, 2021 to December and 2021, under the supervision of Mr.Rajkamal Kishor Gupta Designation, Department of Computer Science and Engineering/ Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida .

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

19SCSE1180028 Nikita Sirohi
19SCSE1180027 - Shivank

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor
(Mr.Rajkamal Kishor Gupta, Assistant Professor)

CERTIFICATE

The Project Review 2 Viva-Voce examination of Nikita Sirohi ,19SCSE1180028, Shivank ,19SCSE1180027, has been held on 22-12-21 and there work is recommended for the award of BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:
Place:

Abstract

Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterise speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognise a particular emotion accurately.

Several modalities have been explored to recognize the emotional states such as facial expressions, speech , physiological signals ,etc. But in this project we work on speech dataset. To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed.Feature selection (FS) was applied in order to seek for the most relevant feature subset.

These extracted feature vectors were sent to the classifier model. The test dataset will undergo the extraction procedure following which the classifier would make a decision regarding the underlying emotion in the test audio. Several machine learning paradigms were used for the emotion classification task. A recurrent neural network (RNN) classifier is used first to classify seven emotions. Their performances are compared later to multivariate linear regression (MLR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals.

Speech is the most natural and environment friendly way of verbal exchange between humans. Lots of efforts have been made to boost a human computer interface so that one can easily have interaction and talk in an unskilled way.

Keyword

Speech Emotion Recognition, emotion recognition, feature extraction neural network

Table of Contents

List of Figures

List of Tables

Acronyms

SVM	Support Vector Maching
ML	Machine Learning
DL	Deep Learning
CNN	Convolution Neural Networks
RNN	Recurrent Neural Network
MLR	Multivariate Linear Regression

Chapter 1 Introduction

1.1 Problem

Speech emotion recognition (SER) systems identify emotions from the human voice in the areas of smart healthcare, driving a vehicle, call centers, automatic translation systems, and human-machine interaction.

In the classical SER process, discriminative acoustic feature extraction is the most important and challenging step because discriminative features influence the classifier performance and decrease the computational time. Nonetheless, current handcrafted acoustic features suffer from limited capability and accuracy in constructing a SER system for real-time implementation.

Therefore, to overcome the limitations of handcrafted features, in recent years, variety of deep learning techniques have been proposed and employed for automatic feature extraction in the field of emotion prediction from speech signals. However, to the best of our knowledge, there is no in-depth review study is available that critically appraises and summarizes the existing deep learning techniques with their strengths and weaknesses for SER.

Hence, this study aims to present a comprehensive review of deep learning techniques, uniqueness, benefits and their limitations for SER. Moreover, this review study also presents speech processing techniques, performance measures and publicly available emotional speech databases. Furthermore, this review also discusses the significance of the findings of the primary studies. Finally, it also presents open research issues and challenges that need significant research efforts and enhancements in the field of SER systems.

1.2 Emotion and classification

Emotion is one of the most difficult concepts to define in psychology. In fact, there are different definitions of emotions in the scientific literature. In everyday speech, emotion is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure. Scientific discourse has drifted to other meanings and there is no consensus on a definition. Emotion is often entwined with temperament, mood, personality, motivation, and disposition. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes.

These changes influence thought and behavior. According to other theories, emotions are not causal forces but simply syndromes of components such as motivation, feeling, behavior, and physiological changes .

Categorization of emotions

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous). In the first approach, emotions are described with a discrete number of classes. Six basic emotions, which are anger, disgust, fear, happiness, sadness, and surprise. Each emotion acts as a discrete category rather than an individual emotional state. In the second approach, emotions are a combination of several psychological dimensions and identified by axes. The categorical approach is commonly used in SER [30]. It characterizes emotions used in everyday emotion words such as joy and anger. In this work, a set of six basic emotions (anger, disgust, fear, joy, sadness, and surprise) plus neutral, corresponding to the six emotions of Ekman’s model, were used for the recognition of emotion from speech using the categorical approach.

1.3 Sensory modalities for emotion expression

There is vigorous debate about what exactly individual can express nonverbally. Humans can express their emotions through many different types of nonverbal communication including facial expressions, quality of speech produced, and physiological signals of the human body. In this section, we discuss each of these categories.

Facial expressions

The human face is extremely expressive, able to express countless emotions without saying a word .And unlike some forms of nonverbal communication, facial expressions are universal. The facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same across cultures.

Speech

In addition to faces, voices are an important modality for emotional expression. Speech is a relevant communicational channel enriched with emotions: the voice in speech not only conveys a semantic message but also the information about the emotional state of the speaker.

Physiological signals

The physiological signals related to autonomic nervous system allow to assess objectively emotions. These include electroencephalogram (EEG), heart rate (HR), electrocardiogram (ECG), respiration (RSP), blood pressure (BP), electromyogram (EMG), skin conductance (SC), blood volume pulse (BVP), and skin temperature (ST) [32]. Using physiological signals to recognize

Chapter 2 Literature Survey

Machine learning offers different techniques to extract information from data that are translated into knowledge. Machine learning algorithms are categorized into supervised, unsupervised and semi-supervised learning. Supervised learning is learning from labeled data that provides corrective information to the algorithm. Unsupervised learning is the learning of patterns without labeled training data. Whereas semi-supervised learning is the learning with partially labeled data or by receiving a reward from the environment.

Emotions are the affective state which influences the behavior and cognitive processes. As a result of external and internal stimuli they appear. Different types of stimuli used for emotion elicitation. Using existing data results of the analysis were 31 to 81 percent correct and by using Fuzzy logic 72 to 81 percent for two classes of emotions.

The proposed system depends on human brain activities or emotions. Speech recognition involves capturing and digitizing sound waves, converting to basic language units, constructing words and analyzing words to ensure correct spelling for words that sound alike. Neural networks are naturally discriminative. Attributes of the neural network are set of the processing units, set of connections, computing procedure, and training procedure.

Chapter 3 Functionality/Working of Project

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, Happiness, Sadness in which any intelligent system with finite computational resources can be trained to identify or synthesize as required.

Requirements:

(1) Choice of a good emotional speech database

(2) Extracting effective features

(3) Designing reliable classifiers using machine learning algorithms.

Both feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage. The last step of speech emotion recognition is classification. It involves classifying the raw data in the form of utterance or frame of the utterance into a particular class of emotion on the basis of features extracted from the data.

Data Analysis

The combined data set from the original 5 sources is thoroughly analysed with respect to the following aspects

- Emotion distribution by gender
- Variation in energy across emotions
- Variation of relative pace and power across emotions

We checked the distribution of labels with respect to emotions and gender and found that while the data is balanced for six emotions viz**.** *neutral, happy, sad, angry, fear* and *disgust*, the number of labels was slightly less for *surprise* and negligible for *boredom*. While the slightly fewer instances of surprise can be overlooked on account of it being a rarer emotion, the imbalance against boredom was rectified later by clubbing sadness and boredom together due to them being similar acoustically. It’s also worth noting that boredom could have been combined with neutral emotion but since both *sadness* and *boredom* are negative emotions, it made more sense to combine them.

Emotion Distribution of Gender

Regarding the distribution of gender, the number of female speakers was found to be slightly more than the male speakers, but the imbalance was not large enough to warrant any special attention.

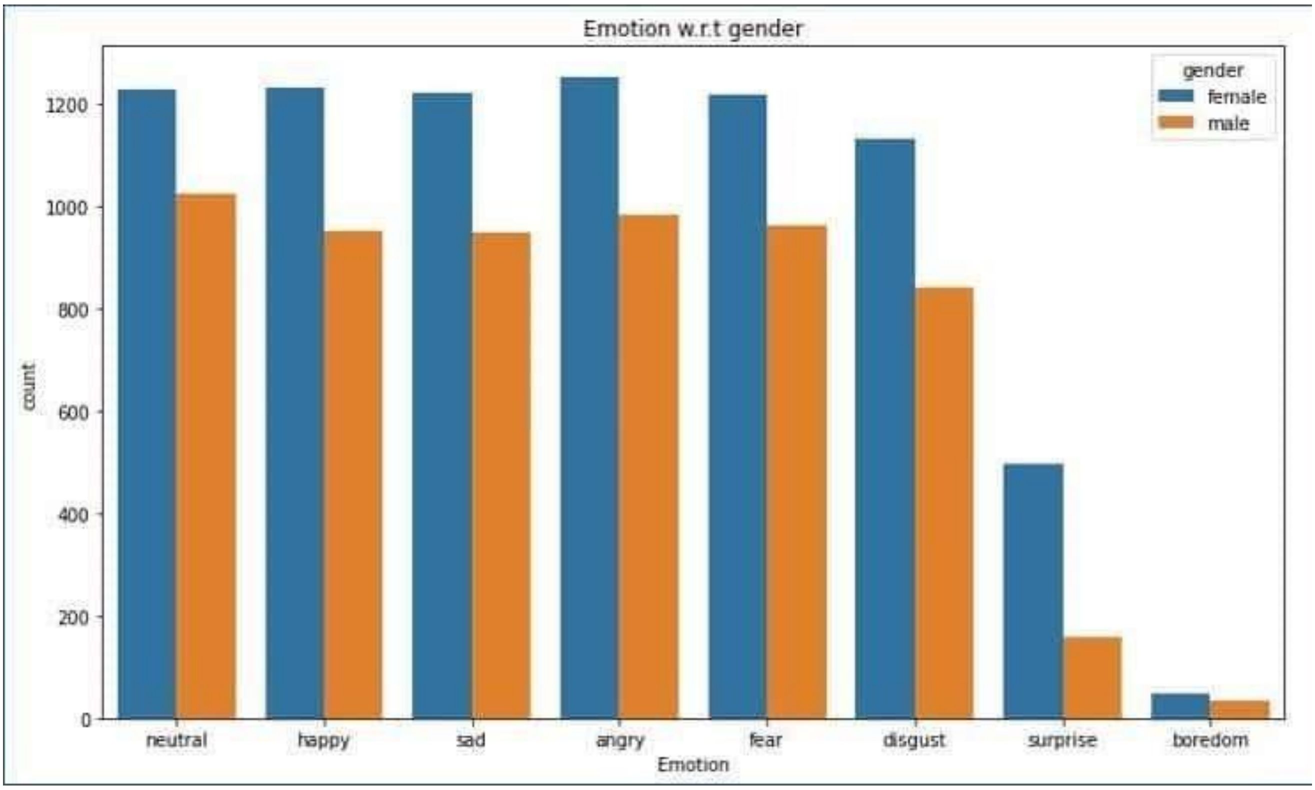

Fig.1 Distributions of emotion with respect to gender

Variation in Energy Across Emotions

To ensure uniformity in our study of energy variation as the audio clips in our dataset were of different lengths, a power which is energy per unit time was found to be a more accurate measure. This metric was plotted with respect to different emotions. It is quite evident that the primary method of expression of anger or fear in people is a higher energy delivery. We also observe that disgust and sadness are closer to neutral with regards to energy although exceptions do exist.

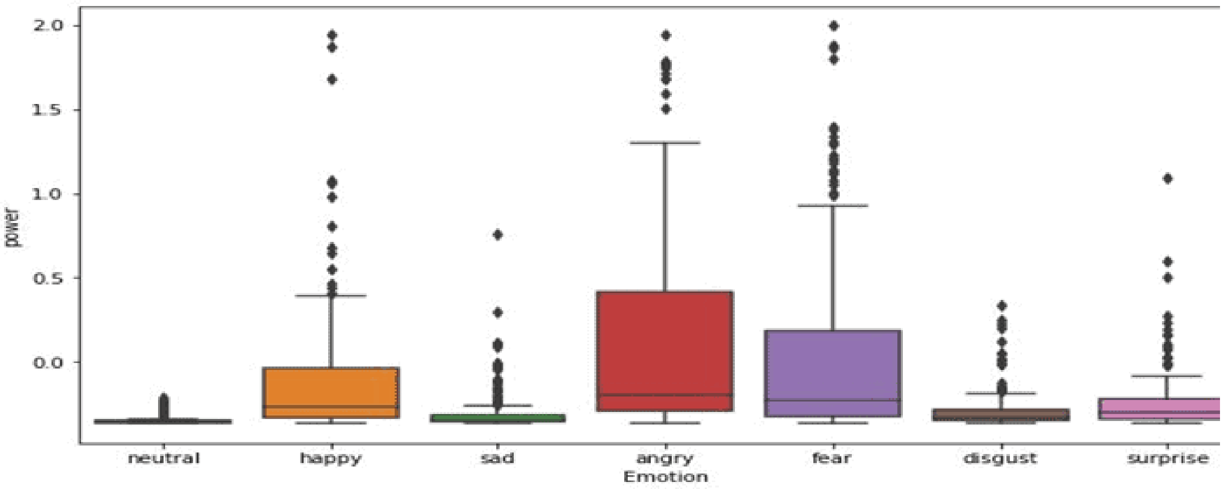

Figure 2 Distributions of emotion with respect to gender

Variation of Relative Pace and Power with respect to Emotions

A scatter-plot of power vs relative pace of the audio clips was analysed and it was observed that the ‘disgust’ emotion was skewed towards the low pace side while the ‘surprise’ emotion was skewed more towards the higher pace side. As mentioned before, anger and fear occupy the high power space and sadness and neutral occupy the low power space while being scattered pace-wise. Only, the RAVDESS dataset was used for plotting here because it contains only two sentences of equal length spoken in different emotions, so the lexical features don’t vary and the relative pace can be reliably calculated.

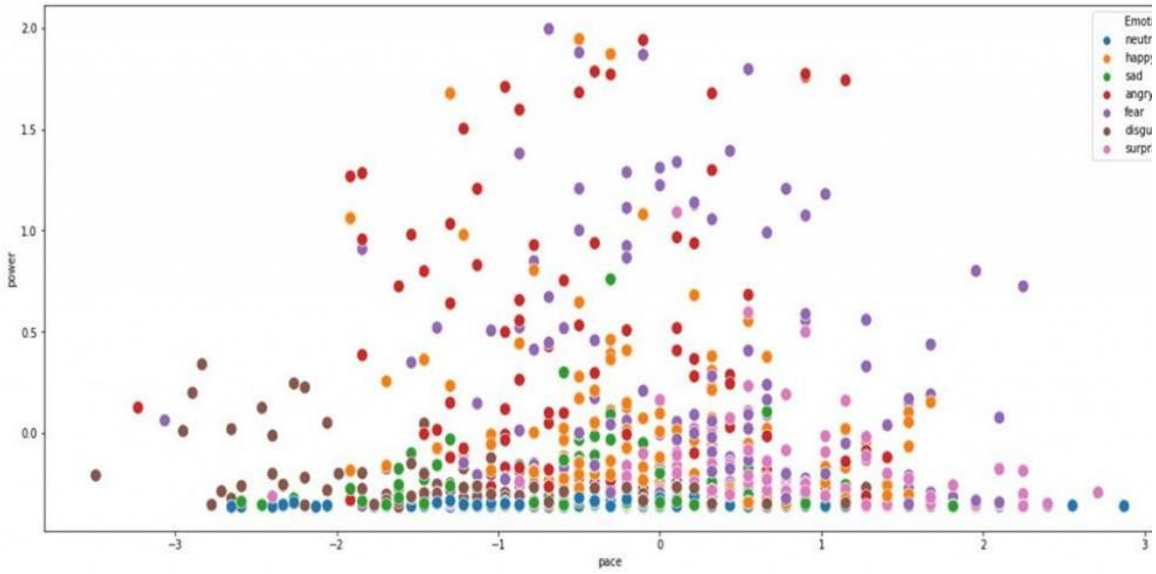

Figure 3 Scatter of power Vs relative pace of audio clips

Block Diagram

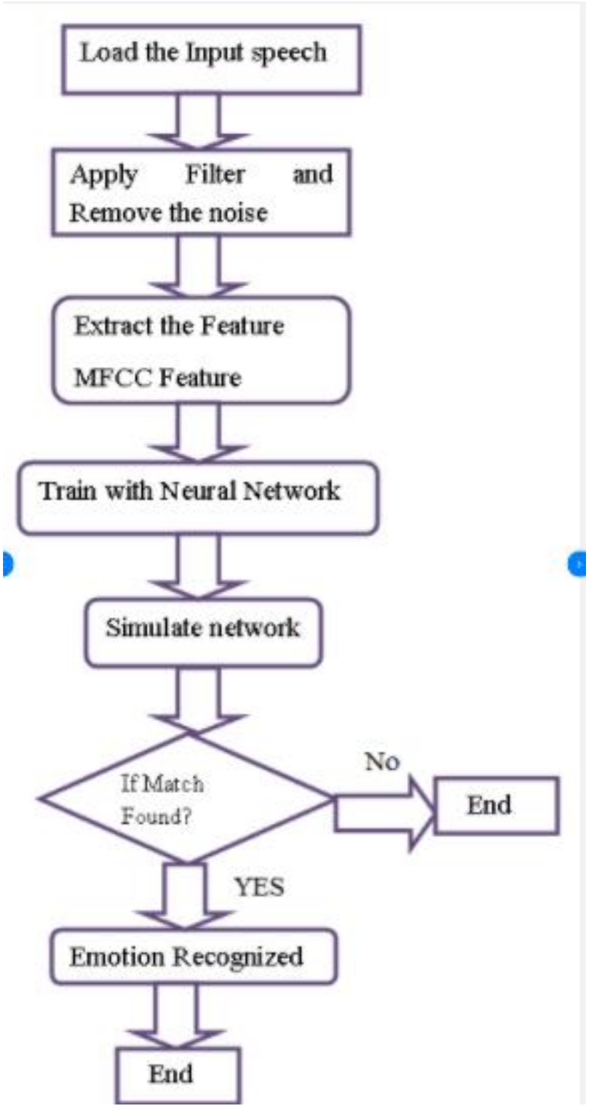

Fig 4 : Emotion recognition through speech

Sequence Diagram

Sequence Diagrams are interaction diagrams that detail how operations are carried out. They capture the interaction between objects in the context of a collaboration. Sequence Diagrams are time focus and they show the order of the interaction visually by using the vertical axis of the diagram to represent time what messages are sent and when.

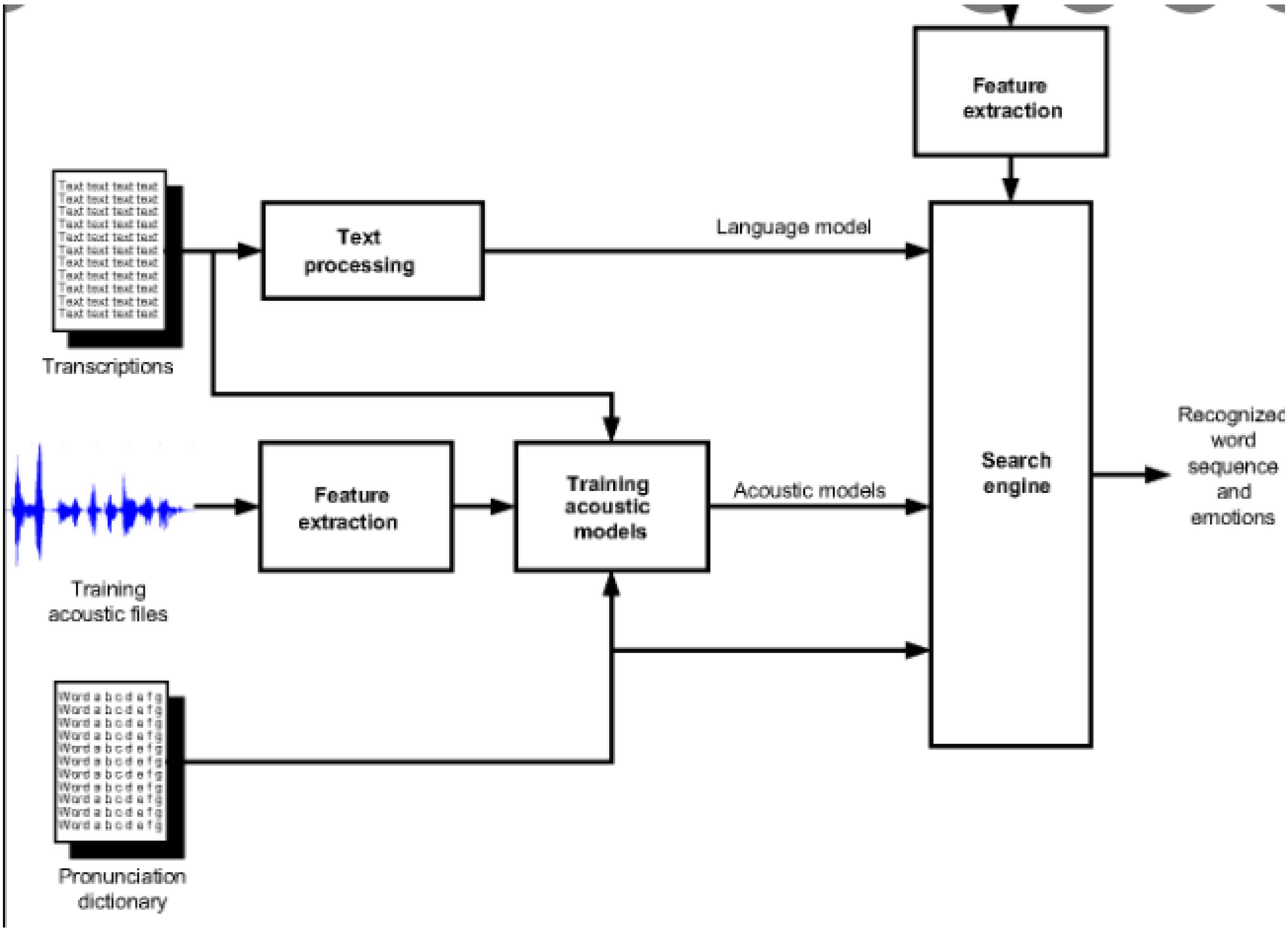

Fig 5: Sequence Diagram

FEATURE EXTRACTION FOR SPEECH EMOTIONS RECOGNITON

There are number of methods for feature extraction like Linear predictive cepstral coefficients (LPCC), Power spectral analysis (FFT), First order derivative (DELTA), Linear predictive analysis (LPC), Mel scale cepstral analysis (MEL), perceptual linear predictive coefficients (PLP) and Relative spectra filtering of log domain coefficients (RASTA).

- Linear predictive coding (LPC): In encoding quality speech at a low bit rate LPC Method is useful that is one of the most powerful techniques of speech analysis.
At current time specific speech sample can be approximated as a linear combination of past speech samples is the basic idea behind linear predictive analysis. It is a human speech production base model that utilizes a conventional source filter model. Vocal tract acoustic are simulated by Lip radiation, vocal tract and glottal transfer functions that are integrated into one all pole filter.
Over a finite duration the sum of squared differences between estimated and original speech signal is minimized using LPC that helps in having unique sets of predictor coefficients.
In real recognition actual predictor coefficients are not used as a high variance is shown by it. There is transformation of predictor coefficient to a cepstral coefficients more robust set of parameters. Some of the types of LPC are residual excitation, regular pulse excited, pitch excitation, voice excitation and coded excited LPC.

- Mel frequency cepstral coefficients (MFCC): It is considered as one of the standard method for feature extraction and in ASR most common is the use of 20 MFCC coefficients. Although for coding speech use of 10-12 coefficients are sufficient and it depend on the spectral form due to which it is more sensitive to noise. This problem can be overcome by using more information in speech signals periodicity although aperiodic content is also present in speech. Real cesptal of windowed short time fast Fourier transform (FFT) signal is represent by MFCC [21]. Non linear frequency is use. The parameters similar to humans used for hearing speech are used to extracts parameters using audio feature extraction MFCC technique. Other information is deemphasizes and arbitrary number of samples contain time frames are used to divide speech signals. Overlapping from frame to frame is used to smooth the transition in most systems and then hamming window is used to eliminate the discontinuities from each time frame.

- Perceptual linear prediction (PLP): Hermansky developed a PLP model that uses psychophysics concept of hearing to model a human speech. The speech recognition rate gets improved by discarding irrelevant information by PLP. Spectral characteristics are transformed to human auditory system match is the only thing that makes PLP different from LPC. The intensity-loudness power-law relation, equalloudness curve and critical-band resolution curves are three main perceptual aspects approximates by PLP.

- Mel scale cepstral analysis (MEL): PLP analysis and MEL analysis is similar to each other in which psychophysically based spectral transformations is used to modify the spectrum. According to the scale of MEL a spectrum is wrapped in this method on other hand according to bark scale a spectrum is warped in PLP. So output cepstral coefficients are the main different between scale cepstral analysis of PLP and MEL. The modified power spectrum is smooth using all pole model in PLP and then on the basis of this model a output cepstral coefficients are computed. On other hand modified power spectrum is smooth using cepstral smoothing in MEL scale cesptral analysis. In this Discrete Fourier Transform (DFT) is used to convert log power spectrum is directly transform into capstal domain.

- Relative Spectra filtering (RASTA): The ability to perform RASTA filtering is provided by analysis library to compensate for linear channel distortions. It can be used either in cepstral or log spectral domains and in both of them linear channel distortions is appear as an additive constant. Each feature coefficient is band passes by RASTA filter and convolutional introduced noise in the channel effect is alleviated by band pass filter equivalent high pass portion. Then frame to frame spectral changes are smoothened with the help of low pass filtering.

- Power spectral analysis (FFT): This is the more common techniques of studying speech signal and over the frequency content of the signal over time is described by speech signal power spectrum. Fourier Transform (DFT) of the speech signal is the first step to compute power spectrum that computes time domain signal equivalent frequency information. Real point values consist speech signal can use Fast Fourier Transform (FFT) to increase the efficiency.

Speech Database

A database is an organized collection of data. It is the collection of schemas, tables, queries, reports, views, and other objects.The data are typically organized to model aspects of reality in a way that supports processes requiring information.

RAVDESS

This dataset includes around 1500 audio file input from 24 different actors, 12 male and 12 female where these actors record short audios in 8 different emotions.

1=Neutral
2=Calm
3=Happy
4=Sad
5=Angry
6=Fearful
7=Disgust
8=Surprised

Working

The task of speech emotion recognition (SER) is traditionally divided into two main parts: feature extraction and classification, as depicted in Figure 5.

During the feature extraction stage, a speech signal is converted to numerical values using various front-end signal processing techniques. Extracted feature vectors have a compact form and ideally should capture essential information from the signal. In the back-end, an appropriate classifier is selected according to the task to be performed.

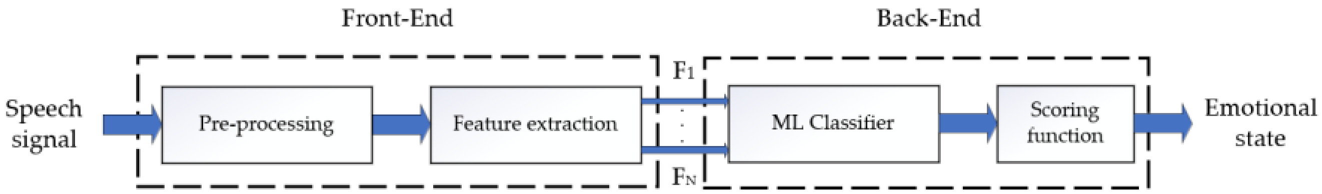

Figure 5. Block scheme of general speech emotion recognition system.

Architecture of Speech Emotion Recognition System

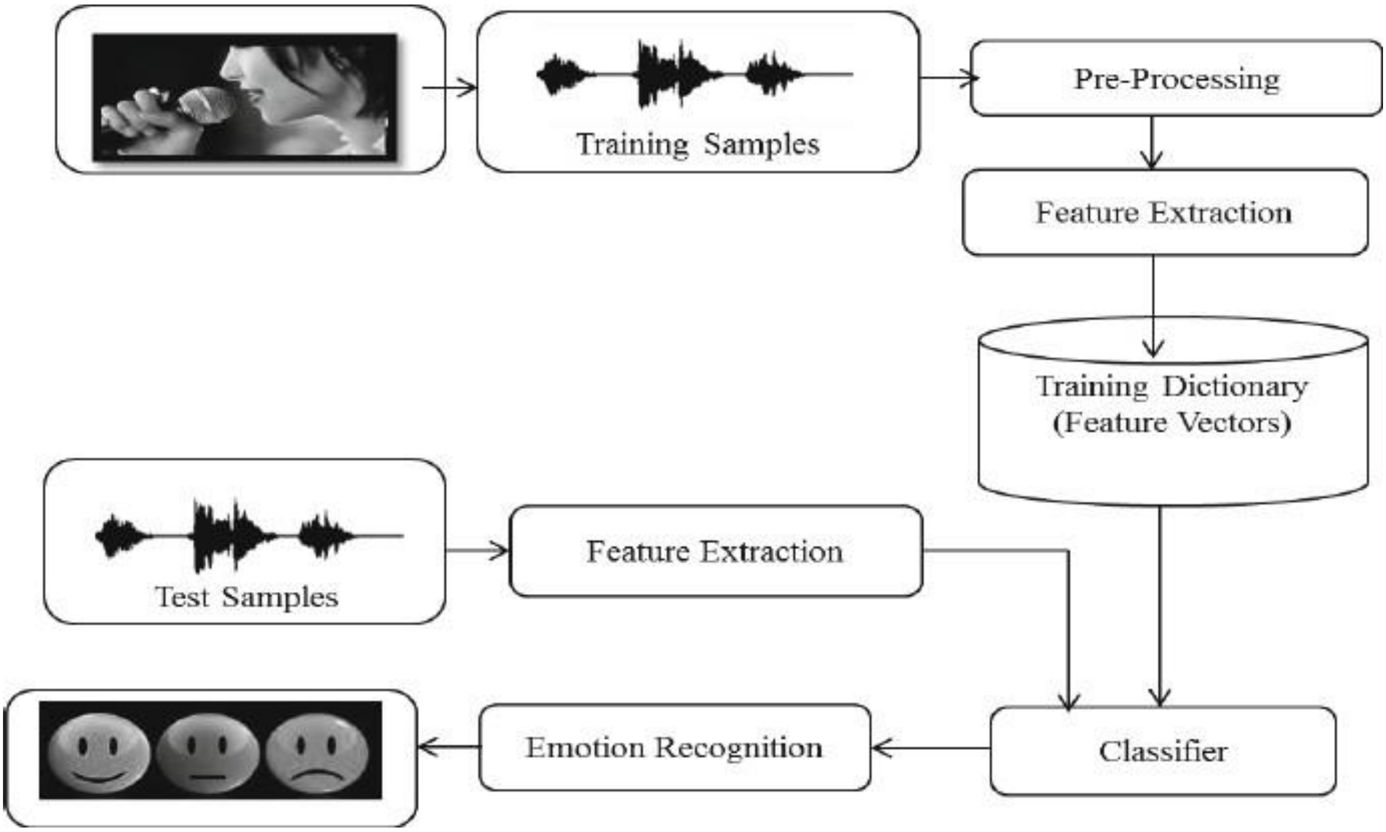

Fig 6: Architecture of Speech Emotion Recognition System

System

Our SER system consists of four main steps.

Steps:

1. **Sample Collection**: The voice sample collection.

2. **Feature Extraction**: The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch.

3. **Feature Selection**: The objective of feature selection in ML is to “reduce the number of features used to characterize a dataset so as to improve a learning algorithm’s performance on a given task.” The objective will be the maximization of the classification accuracy in a specific task for a certain learning algorithm; as a collateral effect, the number of features to induce the final classification model will be reduced.

4. **Classification Method**: Many machine learning algorithms have been used for discrete emotion classification. The goal of these algorithms is to learn from the training samples and then use this learning to classify new observation. **Multivariate linear regression classification (MLR**) is a simple and efficient computation of machine learning algorithms, and it can be used for both regression and classification problems.

Support vector machines (SVM) are an optimal margin classifier in machine learning. SVM are an optimal margin classifier in machine learning.

It is also used extensively in many studies that related to audio emotion recognition which can be found. It can have a very good classification performance compared to other classifiers especially for limited training data .

Recurrent neural networks (RNN) are suitable for learning time series data, and it has shown improved performance for classification task While RNN models are effective at learning temporal correlations, they suffer from the vanishing gradient problem which increases with the length of the training sequences. To resolve this problem, long short-term memory (LSTM) RNNs were proposed.

Radial Basis Function :A radial basis function (RBF) is a function that assigns a real value to each input from its domain (it is a real-value function), and the value produced by the RBF is always an absolute value**;** i.e. it is a measure of distance and cannot be negative.

Principal Component Analysis :PCA is a widely covered method on the web, and there are some great articles about it, but many spend too much time in the weeds on the topic, when most of us just want to know how it works in a simplified way.

Code

```
# IMPORT NECESSARY LIBRARIES
import librosa
%matplotlib inline
import matplotlib.pyplot as plt
import librosa.display
from IPython.display import Audio
import numpy as np
import tensorflow as tf
from matplotlib.pyplot import specgram
import pandas as pd
from sklearn.metrics import confusion_matrix
import IPython.display as ipd  # To play sound in the notebook
import os # interface with underlying OS that python is running on
import sys
import warnings
# ignore warnings
if not sys.warnoptions:
    warnings.simplefilter("ignore")
warnings.filterwarnings("ignore", category=DeprecationWarning)
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import tensorflow.keras
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv1D, MaxPooling1D, AveragePooling1D
from tensorflow.keras.layers import Input, Flatten, Dropout, Activation,
BatchNormalization, Dense
from sklearn.model_selection import GridSearchCV
from tensorflow.keras.wrappers.scikit_learn import KerasClassifier
from tensorflow.keras.optimizers import SGD
from tensorflow.keras.regularizers import l2
import seaborn as sns
from tensorflow.keras.callbacks import EarlyStopping, ModelCheckpoint
from tensorflow.keras.utils import to_categorical
from sklearn.metrics import classification_report
```

Male Neutral

```
# LOAD IN FILE
x, sr = librosa.load('/Users/murielkosaka/Desktop/capstone_project/audio/audio
_speech_actors_01-24/Actor_01/03-01-01-01-01-01-01.wav')
# DISPLAY WAVEPLOT
plt.figure(figsize=(8, 4))
librosa.display.waveplot(x, sr=sr)
plt.title('Waveplot - Male Neutral')
plt.savefig('Waveplot_MaleNeutral.png')
# PLAY AUDIO FILE
librosa.output.write_wav('ipd.Audio Files/MaleNeutral.wav', x, sr)
Audio(data=x, rate=sr)
# CREATE LOG MEL SPECTROGRAM
spectrogram = librosa.feature.melspectrogram(y=x, sr=sr, n_mels=128,fmax=8000)
spectrogram = librosa.power_to_db(spectrogram)
librosa.display.specshow(spectrogram, y_axis='mel', fmax=8000, x_axis='time');
plt.title('Mel Spectrogram - Male Neutral')
plt.savefig('MelSpec_MaleNeutral.png')
plt.colorbar(format='%+2.0f dB')
```

Female Calm

```
# LOAD IN FILE
x, sr = librosa.load('/Users/murielkosaka/Desktop/capstone_project/audio/audio
_speech_actors_01-24/Actor_02/03-01-02-01-01-01-02.wav')
# DISPLAY WAVEPLOT
plt.figure(figsize=(8, 4))
librosa.display.waveplot(x, sr=sr)
plt.title('Waveplot - Female Calm')
plt.savefig('Waveplot_FemaleCalm.png')
# PLAY AUDIO FILE
librosa.output.write_wav('ipd.Audio Files/FemaleCalm.wav', x, sr)
Audio(data=x, rate=sr)
# CREATE LOG MEL SPECTROGRAM
spectrogram = librosa.feature.melspectrogram(y=x, sr=sr, n_mels=128,fmax=8000)
spectrogram = librosa.power_to_db(spectrogram)
librosa.display.specshow(spectrogram, y_axis='mel', fmax=8000, x_axis='time');
plt.title('Mel Spectrogram - Female Calm')
plt.savefig('MelSpec_FemaleCalm.png')
plt.colorbar(format='%+2.0f dB');
```

Application

1.Web-based E-learning

2.Commercial applications

3.Clinical studies, entertainment

4.Call centers

5.Computer games

Chapter 4 Result

Here we trained model to recognize emotions from speech. In this section, we describe the experiment environment and report the recognition accuracy of using MLR, SVM and RNN classifiers on two emotional speech database. We used Berlin database and spanish database for network training and validation. To evaluate the classification error 10-cross validation test were used. We used 70% of data for training and 30 % a simple LSTM.

It consists of two consecutive LSTM layers with hyperbolic tangent activations followed by two classification dense layers. More detailed diagrams are shown in figure 4, 5 and 6 and can be found in appendix A. Table 1, 2 and 3 show the recognition rate for each combination of various features and classifiers based on Berlin and spanish databases.

As shown in table 1, MLR classifier performed better results with feature combination of MFCC and MS for both databases. And under the conditions of limited training data (Berlin database), it can have a very good classification performance compared to other classifiers. A high dimension can maximize the rate of MLR.

Chapter 5 Conclusion and Future Scope

5.1 Conclusion

Speech is the most natural and environment friendly way of verbal exchange between humans. Lots of efforts have been made to boost a human computer interface so that one can easily have interaction and talk in an unskilled way. With the dawn of AI human like Automatic Speech Recognition (ASR) accuracy is doable in voice primarily based shrewd agents. Emotion recognition from speech statistics can make the machine extra human like and it can resource the robot to respond according to the tone and sentiments of the user's voice.

5.2 Future Scope

A few possible steps that can be implemented to make the models more robust and accurate are the following:

- An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
- Figuring out a way to clear random silence from the audio clip.
- Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of
- MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum.
- Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.
- Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

Reference

1. https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/
2. https://www.frontiersin.org/files/Articles/497007/fcomp-02-00014-HTML/image_m/fcomp-02-00014-g002.jpg
3. Han, K., Yu, D., and Tashev, I. (2014). "Speech emotion recognition using deep neural network and extreme learning machine," in Interspeech (Singapore).
4. Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural Comput. *18*, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

5 . Zhao, C. Jiang, C. Zou, and Z. Wu, "Study on emotional feature analysis and recognition in speech," *Acta Electronica Sinica*, vol. 32, no. 4, pp. 606–609, 2004.View at: Google Scholar