Islamic University of Technology (IUT)

Study on the Accent Independent Features for Speech Emotion Recognition

Authors

Nowshin Tabassum, 170041033

Tasfia Tabassum, 170041037

Tahiya Sultana Safa, 170041063

Supervisor

Dr. Hasan Mahmud

Assistant Professor, Department of CSE, IUT

Systems and Software Lab (SSL)

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc.

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

A Subsidiary organ of the Organization of Islamic Cooperation (OIC)

Academic Year: 2020-2021
April, 2022

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Nowshin Tabassum, Tasfia Tabassum and Tahiya Sultana Safa under the supervision of Dr. Hasan Mahmud, Assistant Professor of Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Authors:

Nowshin Tabassum	Tasfia Tabassum	
Nowshin Tabassum	Tasfia Tabassum	Tahiya Sultana Safa
170041033	170041037	170041063

Supervisor:

Dr. Hasan Mahmud

Assistant Professor

Department of Computer Science and Engineering

Islamic University of Technology

Acknowledgment

We would like to express our grateful appreciation for our supervisor Assistant Professor Dr. Hasan Mahmud sir, Department of Computer Science and Engineering (CSE), IUT, Professor Dr. Md. Kamrul Hasan sir, Department of Computer Science and Engineering (CSE), IUT, and Fardin Saad sir, Department of Computer Science and Engineering (CSE), IUT for being our adviser and mentor, for their valuable inspection and suggestions on our proposal. Their motivation, suggestions and insights for this research have been invaluable. Without their support and proper guidance this research would never have been possible. Their valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project finalization helped us to do our thesis work in proper way. We are really grateful to them.

Abstract

Great progress have been made in speech recognition but we still have a long way to go to have a smooth human-computer interaction because the computer still finds it difficult to understand the emotional state of the speaker. This has introduced a brought into light a relatively recent research field,namely **Speech Emotion Recognition**. There are some implicit information about the emotions in every speech signal, which can be extracted through speech processing methods. There are many systems proposed in the literature to identify the emotional state through speech. Extraction of features from speech,Selecting a suitable feature set, designing a proper classifications method and preparing an proper dataset are the main points of designing a Speech Emotion Recognition (SER) systems.However despite significant progress in this area there still remains many things which are not well understood, specially, when attention was given to the cultural differences of people.Emotions Recognition in speech can vary from person to person based on their age, gender, language, accents and many other factors. To explore how much accents affect SER, we looked into how the feature varies for different accents in the domain of Speech Emotion Recognition. This paper focuses on the issue if Speech Emotion Recognition is Accent Independent or not. Study on different speech features, experiments on their extraction process and reduction techniques and experiments on selection of accent independent features are carried out. Which will be used to train a model and will lead us to a conclusion if SER depends on accents or not and which of the features of Speech help identify the emotions more accurately despite of the accent.

Keywords — Speech Emotion Recognition, Feature Selection, Prosodic, Spectral, Voice Quality

Contents

List of Tables

Chapter 1

Introduction

1.1 Overview

Emotion is a means of expressing one's perspective or state of mind to others. Great progress has been made in emotion recognition, nevertheless, having a natural interaction between human and device is a long way to go due to the fact that the devices is unable to understand the emotional state or condition of the speaker. This has given rise to a relative new research field known as Speech Emotion Recognition. The extraction of the speaker's emotional state or condition from his or her speech signal is defined as speech emotion recognition (SER).

While a person's emotional state can be demonstrated in a variety of approaches, including facial expressions, motions, gestures and postures, researchers have developed a keen interest in the automated recognition of emotion from speech. After years of research, however, effectively and precisely recognizing an individual's emotional state from their speech in an automatic way continues to be a significant challenge. This prompts an investigation into the characteristics that affect Speech Emotion Recognition (SER), such as gender, age and accent because each independent speaker has their own style and characteristics.

There are a few universal emotions that any intelligent system with finite processing resources can be trained to recognize as needed, including Anger, Sadness, Happiness, and Neutral.[28] Spectral, prosodic, and voice quality features are the features that are utilised in the emotion identification process. Because both spectral and prosodic features include emotional statistics information, they are used for speech emotion recognition. One of the spectral feature is mel-frequency cepstral coefficients (MFCC).Prosodic features such as fundamental frequency, amplitude, pitch, speech intensity, and voiced factors are utilized to model various emotional responses. The features are selected in a way so that they are prominent to classify emotion.

The goal of our research is a smooth experience of human-computer interaction has led to the design of the an Accent Independent Speech Emotion Recognition system (SER).Recently SER systems are being used in customer services,Call centers,educational institutions and many other places where everyday a huge and diverse group of people come with their inquiries or problems each one having their own way of expressing emotions,own accents.Just like language has an effect on Speech Emotion Recognition.Many Studies have found how language effects the speech emotion recognition systems,In the same way there are many variations of accents for the same language.This has led to our research of whether the Speech Emotion Recognition Systems can correctly identify emotions despite of such diverse accents.

In order to look into how variations of accents affect the Speech Emotion Recognition tasks. We will be using datasets of 4 different accents of English.When people conversate,talk or express emotions there are some cues or patterns which reveal the type of emotion being expressed.Through this pattern, some cues(features) are generated which helps a machine learning model detect an emotion from a speech.Our aim is to find a subset of such features that has a common pattern among all accents.

1.2 Research Challenges

Our research to find accent independent features for SER comes with several challenges, There weren't enough datasets and samples for several accents in English. If present most of them didn't have both Male and Female samples.Most of the dataset had a huge bias for some particular emotions.It was very challenging to remove the bias and collect equal number of samples of each emotion for each accents.

From several feature selections techniques, we had to select those which can capture unique qualities in a feature to classify emotions with higher accuracy. Checking every possible combination of features was very time-consuming, but was necessary to find a common set of accent independent features.

1.3 Research Contribution

The main contributions of our research are -

- A robust Speech Emotion Recognition System that can correctly identify emotions despite of such diverse accents.
- Determine features that are Accent-Neutral

1.4 Report Layout

The report is structured as follows: Chapter 2 and 3 gives a brief description on the Background Study and Related Works for Speech Emotion Recognition.Chapter 4 illustrates the proposed approach of our experiments. Chapter 5 describes the experimental setup analyses and draws conclusion from the results obtained from our experiments.

Chapter 2

Background study

In order to successfully build a speech emotion recognition system, we must precisely define and model emotion. Basic emotions are divided into six categories: sadness, happiness, fear, anger, disgust, and surprise. Other emotions are obtained by combining the basic ones. The majority of existing SER systems emphasize on these basic emotional categories.

2.1 Databases and their types

Because the classification process relies on labeled data, databases are an essential part of speech emotion recognition. The success of the recognition process is influenced by the data quality. Data that is incomplete, low-quality, or flawed can lead to inaccurate predictions; thus, data should be properly prepared and collected.

Three types of databases for speech emotion recognition can be investigated:[1]

- **Acted (Reproduced) speech emotion datasets**

  Actors and Actresses spoked in different emotions for the same sentence. Berlin emotional database [4], The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [5], Surrey Audio-Visual Expressed Emotion

(SAVEE) [8], Toronto Emotional Speech Database (TESS).[21]

- **Elicited (Influenced) speech emotion datasets**

  Different emotional situations are created in the surroundings which influenced people to speak in different emotions.

  eNTERFACE'05 Audio-Visual Emotion Database [16], Speech Under Simulated and Actual Stress Database (SUSAS). [7]

- **Natural speech emotion datasets**

  Natural emotional speeches of different people are used as samples in the datasets.

  Chinese Annotated Spontaneous Speech corpus (CASS) [13], RECOLA Speech Database [24], SAMAINE Database [17].

2.2 Speech Preprocessing Techniques

Preprocessing is the initial stage in a SER system after collecting the data that will be utilized to train the classifier. Some of these techniques are used to extract features, while others are used to normalize features so that variations in speakers and recordings do not impact the recognition process.[1].

Framing

Signal framing, or speech segmentation, is the act of splitting continuous speech signals into fixed-length segments in order to overcome a variety of SER challenges.[1]

Windowing

After the framing of speech signals, the next step is applying a window function to the frames. The windowing function is used to reduce the effects of leakages generated by discontinuities at the edges of signals during the Fast Fourier Transform (FFT) of data.For this purpose usually A Hamming window is used.[1].

Normalization

Feature normalization is a necessary step that reduces speaker and recording variability while maintaining the features' discriminative strength. The ability of features to generalize is improved by using feature normalization.[1].

Noise Reduction

In real life, noises in the surroundings is also captured along with the speech signal. As noises has an effect on recognition rate, noise reduction techniques must be used to eliminate or reduce noise.

The most commonly used noise reduction techniques are the minimum mean square error (MMSE) and log-spectral amplitude MMSE (LogMMSE) estimators.[1].

2.3 Speech Features

Speech emotion recognition relies heavily on features. Variety of feature sets help a lot in classifying speech emotions; however, there is no generally agreed set of features for accurate and distinct classification.

Speech is a variable-length continuous signal that conveys both information and emotion. As a result, depending on the technique required, global or local features can be extracted.Global features, also known as long-term or supra-segmental features include the gross statistics such as mean, minimum and maximum values, and standard deviation. The temporal dynamics are usually represented by local features, also known as short-term or segmental features, with the goal of approximating a stationary state. Because emotional features are not equally distributed across the whole speech signal, these stationary states are important. For example, anger is more recignized in the start of an utterance, whereas surprise is usually expressed at the end. As a result, local features are used to extract the temporal information from speech.

In terms of classification accuracy and time, the majority of studies agree that

global features outperform local features.
These local and global features of Speech Emotion Recognition systems are divided into the following four categories.

- Prosodic Features
- Spectral Features
- Voice Quality Features
- Teager Energy Operator (TEO) Based Features

Prosodic and spectral features are used more commonly in SER systems. TEO features are not that much popular in speech emotion recognition.[1]

2.3.1 Prosodic Features

Prosodic features are those that can be perceived by humans, such as Intensity and rhythm. For speech emotion recognition, these features have been identified to transmit the most distinguishing qualities of emotional content. The most widely used Prosodic features are:

1. **Pitch** The Pitch or fundamental frequency, F0, is the vibrations created in the vocal cord during a conversation/speech. It determines the speech's rhythmic and tonal properties. The change of pitch over the course of an utterance yields its pitch contour whose statistical properties can be used as features.

2. **Energy** The energy of the speech signal, also known as volume or intensity, serves as a representation of the amplitude change of speech signals across time. According to studies, high arousal emotions like anger, happiness, or surprise lead to an increase in energy, but disgust and sadness lead to a decrease in energy.

3. **Duration** The length of time it takes to build vowels, words, and other comparable entities in speech is referred to as duration. The most commonly employed duration features include speech rate, duration of silent regions, rate of duration of voiced and unvoiced regions, and duration of longest voiced speech. The time it takes to express anger is shorter than the time it takes to express sadness.[12]

When prosodic features are employed, several studies show that SER systems achieve similar or better results than human judges. The fundamental frequency is an important prosodic feature for SER, as previously stated. The F0 contour can be used to derive a variety of features.

Local and global prosodic features, as well as their combination, were compared by Rao et al.[23] Gross statistics of prosodic features are used to calculate global features. The sequence of syllable duration, frame level pitch, and energy values are used to extract local prosodic features. When local and global prosodic features are merged, performance is somewhat improved when compared to the performance of the local features.

2.3.2 Spectral Features

When a speech comes from a person, it is usually filtered by the shape of their vocal tract. The sound that comes out is determined by this shape. An precisely reproduced shape may result in an accurate representation of the vocal tract and the sounds produced. In the frequency domain, vocal tract characteristics are strongly reflected. When time domain signal is transformed into the frequency domain signal using the Fourier transform, it gives us the Spectral features. They are extracted from speech segments of length 20 to 30 milliseconds that is partitioned by a windowing method. [1]

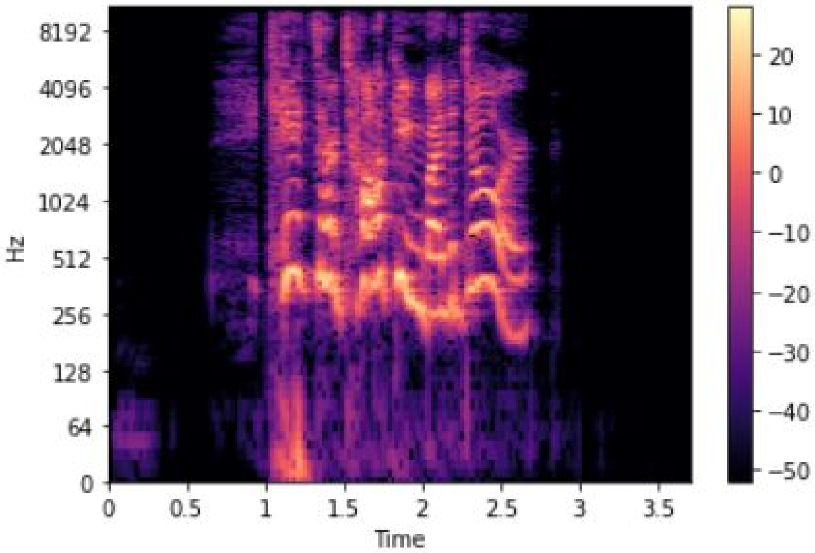


Figure 2.1: Spectogram of a Speech signal

Some of the spectral features are quite popular in SER.

- **Mel Frequency Cepstral Coefficients (MFCC)** the feature that represents the short term power spectrum of a speech signal.

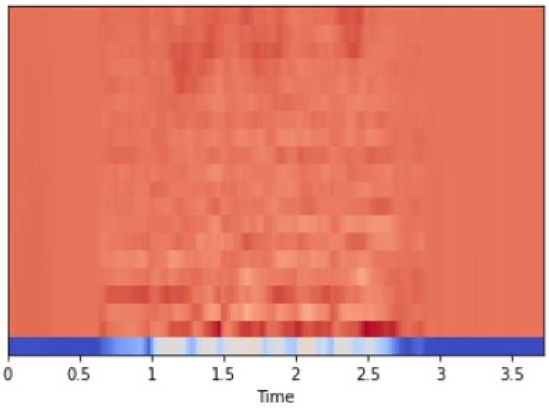


Figure 2.2: MFCC coefficients of a Speech signal

- **Linear Prediction Cepstral Coefficients(LPCC)** also represents the vocal tract characteristics of a speaker. Those characteristics varies with different emotions. LPCC can be directly extracted with a recursive method from **Linear Prediction Coefficient(LPC)**. LPC is the coefficients of allpole filters and is equivalent to the smoothed envelope of the log spectrum of the speech.[1]

- **Gammatone Frequency Cepstral Coefficients (GFCC)** is also a spectral feature obtained in a similar way as MFCC is extracted. But instead of applying the Mel filter bank to the power spectrum, Gammatone filter-bank is applied here.[1]

- **Log-Frequency Power Coefficients (LFPC)**,measures the spectral band energies using Fast Fourier Transform to simulate the logarithmic filtering characteristics of the human auditory system.[1]

- **Formants** the frequencies of the acoustic resonance of the vocal tract are known as formants. They are calculated as amplitude peaks in the frequency spectrum of the sound. They determine the phonetic quality of a vowel, hence used for vowel recognition. [1]

2.3.3 Voice Quality Features

Voice quality is determined by the physical properties of the vocal tract. Involuntary changes may result in a speech signal with properties such as jitter, shimmer, and the harmonics to noise ratio that can be used to distinguish emotions (HNR).

Jitter is the measure of frequency instability.
shimmer is the measure of amplitude instability.
Harmonics to Noise Ratio is the quantification of the relative level of noise in the frequency spectrum of vowels of a speech signal. In voiced speech signals, it is the ratio of periodic to aperiodic components. Changes in voice quality are detected as a result of these variances.[1]
Zhang used prosodic and voice quality features like shimmer, jitter, HNR, and the

first three formants. When prosodic and voice quality features were combined, they achieved a 10% higher recognition rate than when prosodic features were used alone. [30]

2.4 Machine Learning Models

Before deep learning approach , the popular classification models used in this speech emotion recognition tasks were **Support Vector Machine (SVM)**[6] and Bayesian Networks. But more research found that since speech signal tends to be non-stationary, so non-linear classifiers work effectively for SER. Most popular non-linear classifiers used for SER were- the **Hidden Markov Model (HMM)** [19], the **K-nearest neighbor** and the **Gaussian mixture model (GMM)** [3]. For a long time this machine learning models was in use for classifying emotions from speech.
Bjorn Schuller et al.[27] did a revolution in Speech Emotion Recognition in the early stages of this research fields by using a continuous HMM model to get a good recognition rate for the seven emotions. Besides this several other findings have been done using the machine learning models in speech emotion recognition and different boosting algorithms like **Catboost**, **XGBoost** also gives standard results for speech emotion and are able to classify the emotions correctly.

With the emergence of Deep-Learning in machine learning it has gained more attention in recent years. Deep Learning techniques for Speech Emotion Recognition have several advantages over traditional methods, because they are able to detect the complex structure and features without the need for manual feature extraction and they can also deal with un-labeled data. So with the evolution of different Deep learning models it became easier to carry on more detailed researches as complex patterns from any speech signal can be detected by those Deep learning models. Recurrent architectures such as the **Recurrent Neural Networks (RNNs)** and the **Long-Short-Term Memory (LSTM)** are effective approaches in speech-based classification such as the natural language processing (NLP) and Speech Emotion Recognition.[26] [11]. Different researches are going on to check Language Independence, Culture Independence and Gender Independence using the Deep learning models in the recent times.

Chapter 3

Literature Review

Speech Emotion Recognition have been researched, and worked on by a number of researchers. To understand the research gap, we looked at existing systems designed for speech emotion recognition. By looking into those system, we discovered that there has been little research into accent reliance in such recognition systems. So from there, we identified an area where we might make a significant contribution. We categorized the related research we looked at into three categories- Features selection methods, Classifiers that are used and Independency characteristics captured for SER.

3.1 Features Selection Methods and Salient Features used for SER Systems

M Kächele et. al. [10] proposes to find a feature set through forward-selection/backward elimination algorithm.

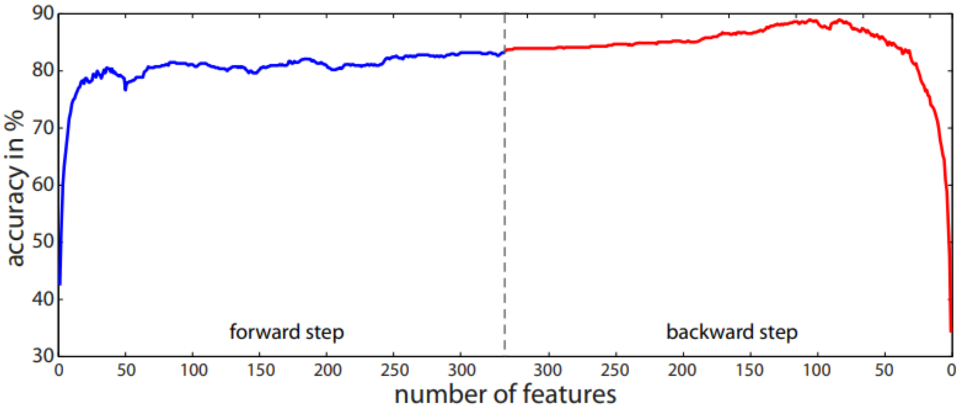


Figure 3.1: Feature Selection Using forward-selection/backward elimination algorithm[10]

They proposed a method where they relied on a popular heuristic algorithm which is forward-backward feature selection. This algorithm is used to prevent convergence to suboptimal minima. To extract the feature two steps were followed-segmental features extracted from short time windows and segmental features integrated over larger time windows. Then these features are selected via forward selection wherein each iteration the most promising features are added to the feature set.The final obtained feature set from the algorithm is used as the initial set for the backward elimination step. In the backward step,least promising features are removed from the feature set. When consecutive runs of the forward and backward step did not result in an accuracy improvement, the algorithm terminates. The final feature set that we find by the algorithm yields the best accuracy. This feature selection algorithm was able to yield an accuracy of 88.97% on the Berlin Database with SVM classifier. [10]

Turgut Özseven et. al.[20] proposed a statistical feature selection method based on the changes in emotions on acoustic features to increase emotional recognition success.[20] They found out that a reduction in the number of features increases the classification success.

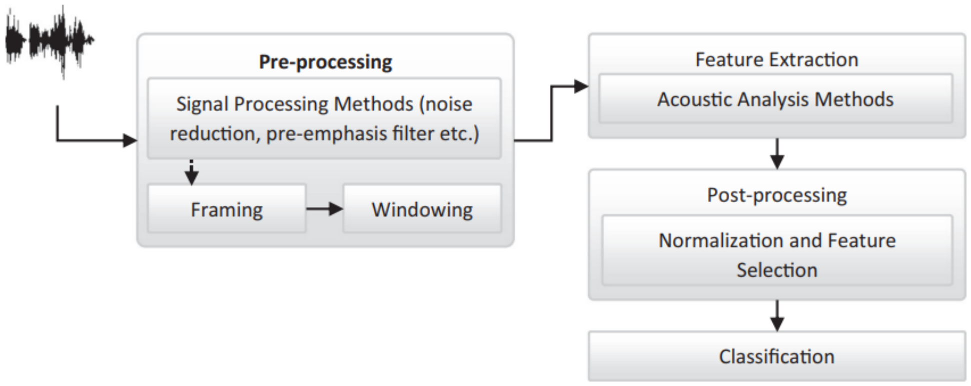

Pre-processing
Signal Processing Methods (noise reduction, pre-emphasis filter etc.)
Framing
Windowing
Feature Extraction
Acoustic Analysis Methods
Post-processing
Normalization and Feature Selection
Classification

Figure 3.2: Proposed SER flow diagram[20]

Here those features are obtained that can characterize emotional content of the speech and doesn't depend on the speaker. As the change in acoustic feature is different by the emotional state, the proposed feature selection method is based

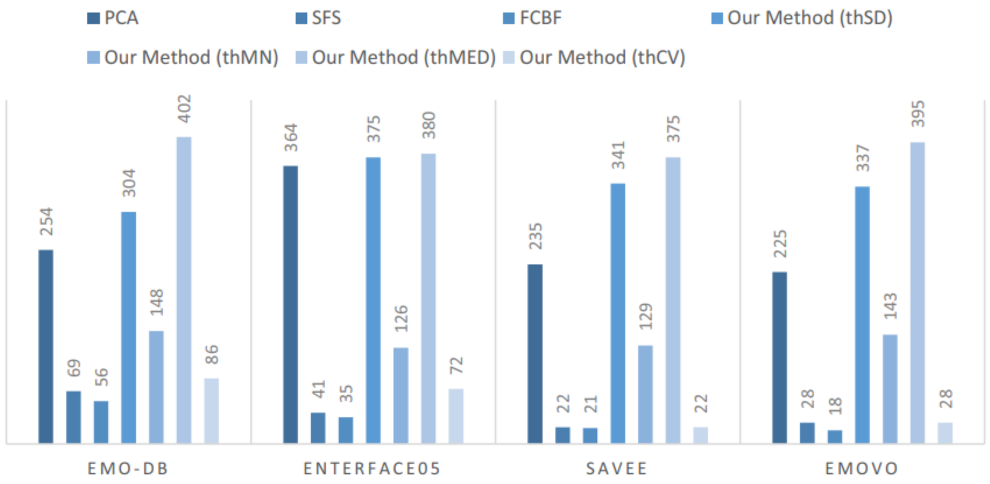
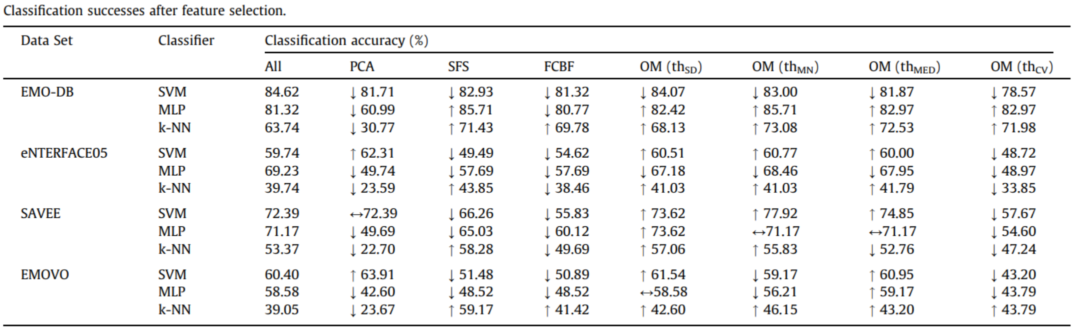


Figure 3.3: Change in feature dimensions after feature selection[20]

on the changes of the feature set according to the emotions.After the selection the feature set is divided into training and test set of 10-fold cross-validation. The experiment was carried on 4 datasets with 3 classifiers-SVM, MLP and k-NN with different feature selection methods. All of the experiments with the reduced and increased success rate was noted. After all the experiments it was proposed that MLP classifier should be chosen if success rate was more important. And if



Classification successes after feature selection.

Data Set	Classifier	Classification accuracy (%)							
		All	PCA	SFS	FCBF	OM (th_{SD})	OM (th_{MN})	OM (th_{MED})	OM (th_{CV})
EMO-DB	SVM	84.62	↓ 81.71	↓ 82.93	↓ 81.32	↓ 84.07	↓ 83.00	↓ 81.87	↓ 78.57
	MLP	81.32	↓ 60.99	↑ 85.71	↓ 80.77	↑ 82.42	↑ 85.71	↑ 82.97	↑ 82.97
	k-NN	63.74	↓ 30.77	↑ 71.43	↑ 69.78	↑ 68.13	↑ 73.08	↑ 72.53	↑ 71.98
eNTERFACE05	SVM	59.74	↑ 62.31	↓ 49.49	↓ 54.62	↑ 60.51	↑ 60.77	↑ 60.00	↓ 48.72
	MLP	69.23	↓ 49.74	↓ 57.69	↓ 57.69	↓ 67.18	↓ 68.46	↓ 67.95	↓ 48.97
	k-NN	39.74	↓ 23.59	↑ 43.85	↓ 38.46	↑ 41.03	↑ 41.03	↑ 41.79	↓ 33.85
SAVEE	SVM	72.39	↔72.39	↓ 66.26	↓ 55.83	↑ 73.62	↑ 77.92	↑ 74.85	↓ 57.67
	MLP	71.17	↓ 49.69	↓ 65.03	↓ 60.12	↑ 73.62	↔71.17	↔71.17	↓ 54.60
	k-NN	53.37	↓ 22.70	↑ 58.28	↓ 49.69	↑ 57.06	↑ 55.83	↓ 52.76	↓ 47.24
EMOVO	SVM	60.40	↑ 63.91	↓ 51.48	↓ 50.89	↑ 61.54	↓ 59.17	↑ 60.95	↓ 43.20
	MLP	58.58	↓ 42.60	↓ 48.52	↓ 48.52	↔58.58	↓ 56.21	↑ 59.17	↓ 43.79
	k-NN	39.05	↓ 23.67	↑ 59.17	↑ 41.42	↑ 42.60	↑ 46.15	↑ 43.20	↑ 43.79

Table 3.1: Classification successes after feature selection[20]

calculation load was more important, SVM should be chosen.

3.2 Classifiers Used for Speech Emotion Recognition Systems

Jain et al.[9] proposed a research they aimed to classify 4 emotions from speech namely: sadness, anger, fear and happiness using the SVM classifier which uses two classification strategies: One-Against-All and Gender dependent classification. Then they conducted a comparative analysis between these two strategies and between the result obtained using LPCC and MFCC.

Their project contains four modules: input speech signal, MFCC and LPC feature extraction, Classification on SVM and output. Features used are pitch, energy, MFCC coefficients, LPCC coefficients and speaker rate. Pitch is chosen since it has all the information about emotions present in it. The mean of pitch value varies from sample to sample and the contours are different in different basic emotional states. MFCC mimics the human auditory system by its calculation of nonlinear frequency unit. LPCC removes the effects of formants and estimates the intensity and frequency of the remaining signal. Since Energy is an important feature in speech they extracted the value of energy from every speech frame to get the overall statistics of energy. Speech rate is also an important feature as it strongly correlated with happy, fear, sad, angry. They used two datasets LDC and UGA. UGA contains 100 samples spoken by students. LDC also contains 100 samples spoken by actors. From each of them 70 used for training 30 for testing.

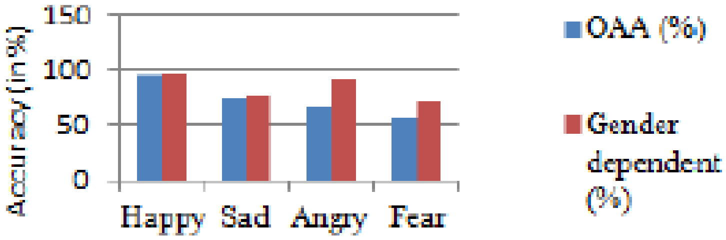
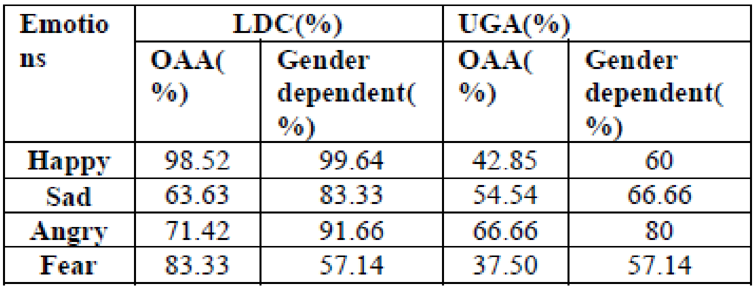


Figure 3.4: OAA vs Gender dependent classifier[9]

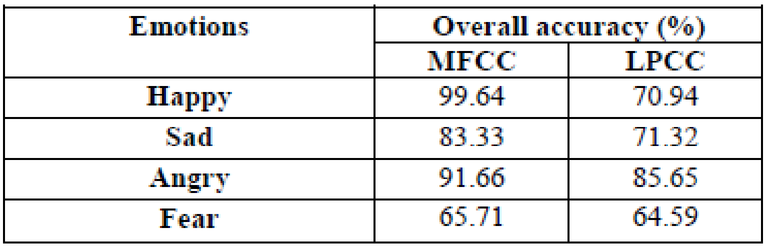
The dataset LDC showed an overall accuracy of 90.08% which is much higher than the accuracy given by UGA dataset which is 65.95%. Reason given for such difference is LDC dataset is performed by trained actors in a noise less environment. But UGA dataset is performed by students in a noisy environment, so LDC has higher accuracy.

For the Gender dependent classifier accuracy was quite higher 84.42% than the One-Against-All classifier which was 72.28%. Also the classifier performs better for MFCC features 85.08% than the LPCC features 73.13%.[9]



Emotions	LDC(%)		UGA(%)	
	OAA(%)	Gender dependent(%)	OAA(%)	Gender dependent(%)
Happy	98.52	99.64	42.85	60
Sad	63.63	83.33	54.54	66.66
Angry	71.42	91.66	66.66	80
Fear	83.33	57.14	37.50	57.14

Table 3.2: Accuracy for both datasets using MFCC[9]



Emotions	Overall accuracy (%)	
	MFCC	LPCC
Happy	99.64	70.94
Sad	83.33	71.32
Angry	91.66	85.65
Fear	65.71	64.59

Table 3.3: Overall accuracy comparison with MFCC AND LPCC algorithms[9]

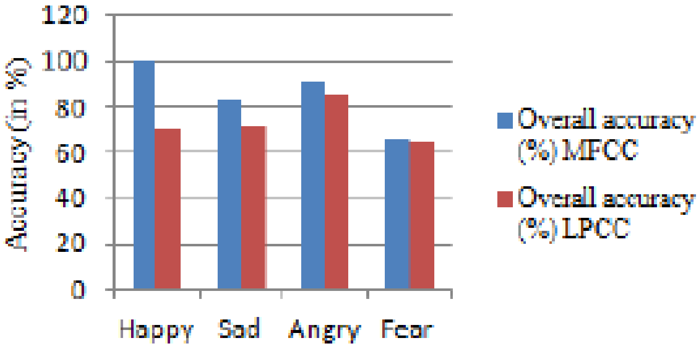


Figure 3.5: MFCC vs LPCC accuracy comparison[9]

Sun et al. [29] proposed a speech emotion recognition method based on a decision tree support vector machine along with the Fisher feature selection method to improve the classification accuracy. The fisher criterion here is used to find out the speech features of higher differentating ability. The decision tree SVM here has a 2-step classification, first rough classification and then fine classification. This removes the unnecessary features and improves the performance of the speech emotion recognition system.

As confusion between emotions increases in multi-class emotion recognition, it reduces the recognition rate. this paper solves this problem by establishing a decision tree SVM by calculating the degree of emotional confusion. The decision tree SVM is used because the generalization capability of SVM is stronger. Among the two strategies (One-to-All and One-to-One) this paper adopts the One-to-One strategy since it's faster and uses the best RBF kernel function.

The speech signals are pre-emphasized and framed, then MFCC is extracted. The

MFCC features are used to train a traditional SVM and calculate the emotional confusion matrix. Then decision tree SVM is constructed. For that, an appropriate initial threshold is selected. The emotions whose confusion degree exceeds that threshold is classified into the same group. Then again the confusion degree is calculated between the ungrouped emotions and divided into existing groups or new groups. In the end, all emotions are categorized. After constructing the tree, the Fisher discriminant coefficient can be obtained by mean and variance calculation of each dimension feature parameter. According to the feature discriminant coefficient, feature parameters having higher discriminative ability are selected for each SVM in the decision tree which is then used for training. This reduces the feature dimension and complexity of the system.

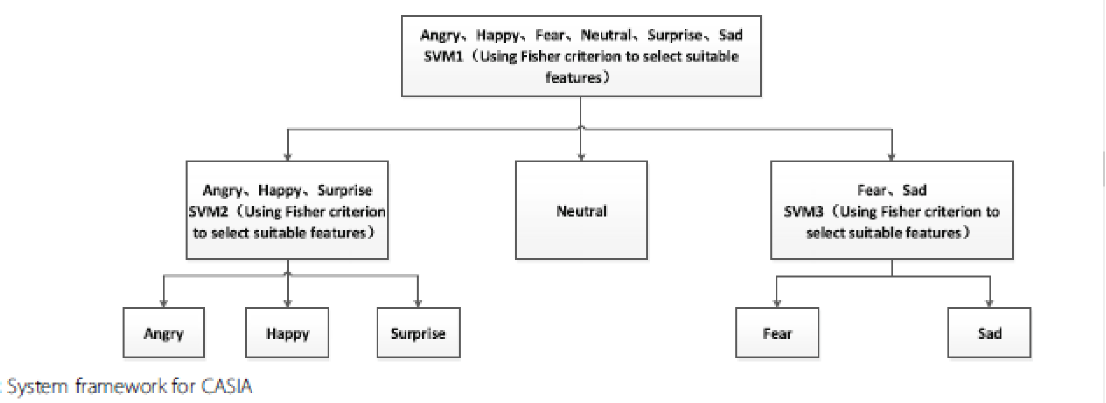

Angry、Happy、Fear、Neutral、Surprise、Sad
SVM1（Using Fisher criterion to select suitable features）
Angry、Happy、Surprise
SVM2（Using Fisher criterion to select suitable features）
Neutral
Fear、Sad
SVM3（Using Fisher criterion to select suitable features）
Angry
Happy
Surprise
Fear
Sad
: System framework for CASIA

Figure 3.6: Framework for Decision tree SVM model [29]

The two datasets used are CASIA Chinese speech emotion corpus (contains 6 emotions) and EMO-DB Berlin speech corpus (contains 7 emotions). All the experiments carried out are tenfold cross-validation. The experiment is carried out 10times and the final result is the average of them.

From several experiments, it is seen that this proposed method gives 83.75% accuracy on the CASIA dataset which is 9% higher than traditional SVM and 8.08% higher than the decision tree without feature selection.[29] This shows the importance of appropriate feature selection methods for speech emotion recognition systems.

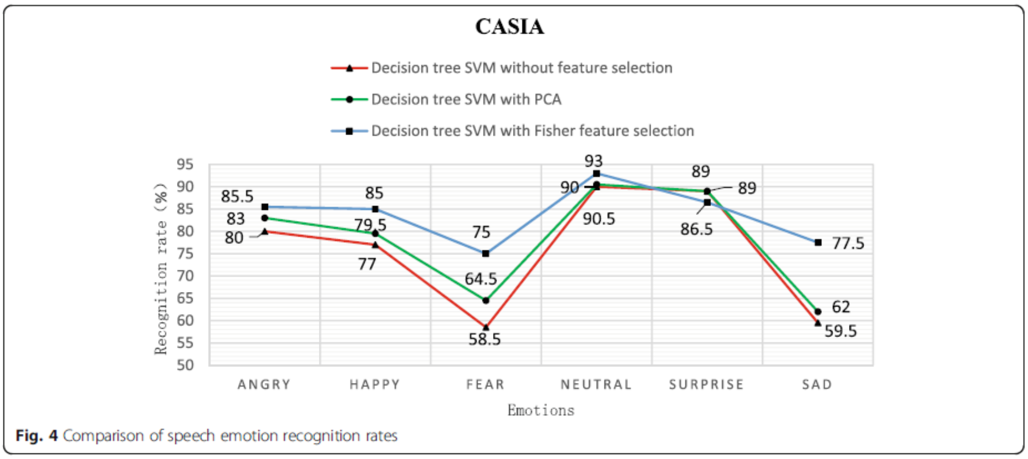


Figure 3.7: Result analysis Decision tree SVM model [29]

3.3 Independency characteristics captured for SER

Liu Z T et al. [14] proposes to find a feature set that is selected through correlation analysis and Fisher Score Selection Algorithm that performs well for speeches from different speakers i.e they aim to find a speaker independent Feature set for SER.

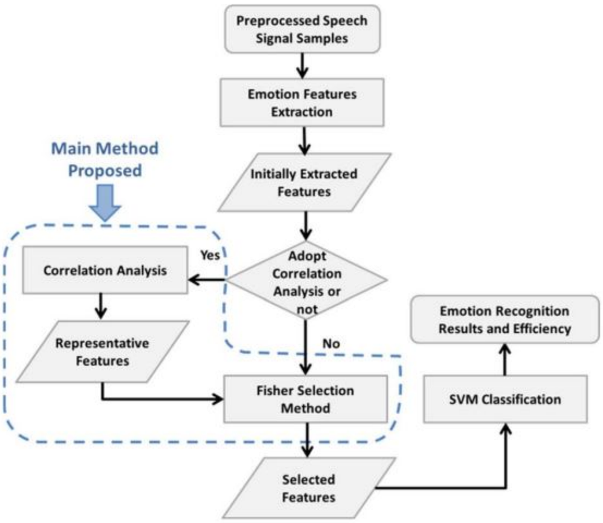
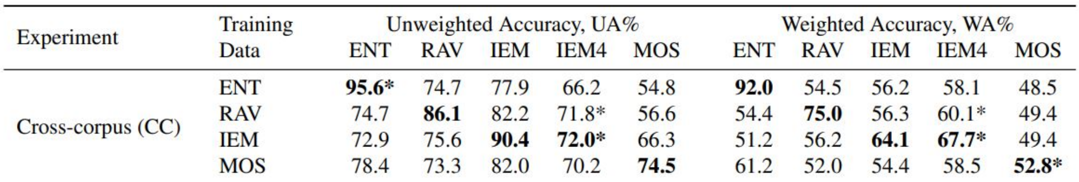

Preprocessed Speech Signal Samples
Emotion Features Extraction
Initially Extracted Features
Main Method Proposed
Adopt Correlation Analysis or not
Yes
Correlation Analysis
Representative Features
No
Fisher Selection Method
Selected Features
SVM Classification
Emotion Recognition Results and Efficiency

Figure 3.8: Feature Analysis framework for speaker Independent SER [14]

Their proposed method is extracting features from speech samples at first and then adopting co relation analysis on the initially extracted features which removes the unnecessary features/features that affects other features and keep the representative features and then apply Fisher Selection Method on the representative features which finally gives the selected features which are used to train the classifier SVM.Through the whole process they find a set of features that gives good performance for speech emotion recognition for different speakers.So through result analysis they found the set of features that are speaker independent are Fundamental Frequency,Formant frequency ,MFCCs and short-time Energy and found an accuracy of 70% in average for speaker independent Speech Emotion Recognition.[14]

Milner et al. [18] performs several different experiments to study the effects of cross-corpus Speech Emotion Recognition. Firstly, they conducted experiments to compare of the efficacy of different features in SER. Secondly, cross-corpus experiments were performed to see the effectiveness of SER using matched and unmatched data. This work only involves English speaking adult datasets-, eNTERFACE, RAVDESS, IEMOCAP and MOSEI.

They designed BLSTM model which contains two hidden layers of 512 nodes each. The output layer of size 1024 feeds in to the attention mechanism computing a context vector of size 128 which is projected to 1024 nodes. This is then passed to the predictor stage which linearly projects to the number of emotion classes. Through background study they claim that for deep learning,log-Mel filterbanks (LMFB) yield better performance over Mel frequency cepstral coefficients (MFCCs). So the remaining experiments were performed with 23 dimensional LMFBs.

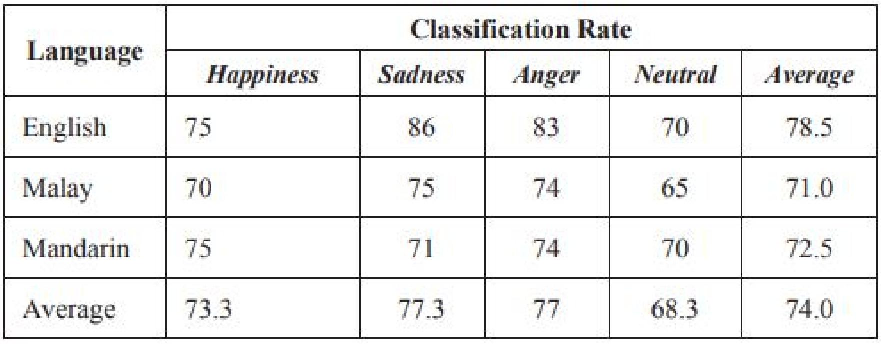


Experiment	Training	Unweighted Accuracy, UA%					Weighted Accuracy, WA%				
	Data	ENT	RAV	IEM	IEM4	MOS	ENT	RAV	IEM	IEM4	MOS
Cross-corpus (CC)	ENT	**95.6***	74.7	77.9	66.2	54.8	**92.0**	54.5	56.2	58.1	48.5
	RAV	74.7	**86.1**	82.2	71.8*	56.6	54.4	**75.0**	56.3	60.1*	49.4
	IEM	72.9	75.6	**90.4**	**72.0***	66.3	51.2	56.2	**64.1**	**67.7***	49.4
	MOS	78.4	73.3	82.0	70.2	**74.5**	61.2	52.0	54.4	58.5	**52.8***

Table 3.4: A Cross-Corpus results on Speech Emotion Recognition[18]

The cross-corpus (CC) results shown in above Table where the performances in the bold diagonal refer to the same dataset test-train and the non-bold performances refer to different dataset train-test. Their Study shows their proposed system is set up well for recognising emotions from speech and acceptable to use for cross-corpus SER. The best performance is found in the same dataset test-train condition as would be expected but the model trained on the elicited dataset achieves best performance in cross-corpus train-test in most cases. [18]

Raju et al. [22] proposes a study is to see if a SER system can detect an individual's emotional state regardless of the language they speak. This paper adopts a k-Nearest Neighbor (kNN) classifier to recognize four discrete emotions using acoustical features.



Language	**Classification Rate**				
	Happiness	***Sadness***	***Anger***	***Neutral***	***Average***
English	75	86	83	70	78.5
Malay	70	75	74	65	71.0
Mandarin	75	71	74	70	72.5
Average	73.3	77.3	77	68.3	74.0

Table 3.5: Classification rate for SER on different languages[22]

SER is language independent, but the result analysis as shown in figure from the paper reveals that when it comes to emotion recognition, there are language-specific variances, with English having a greater recognition rate than Malay and Mandarin. The accuracy rates of native speakers are frequently greater. One probable explanation could be that words spoken in a second or foreign language have less impact than words stated in one's native tongue. Studies claim that When speaking a second language, speakers tend to feel less since there are fewer deeply rooted memories and associations. [22] This paper's approach on studying the effects of language on Speech emotion Recognition inspired us for our study.Since there are many accent variations to a language,our idea was to focus on the effects of accents on speech emotion recognition despite the same language.

Chapter 4

Proposed Approach

Our study primarily focused on finding the dependency of Speech Emotion Recognition over accent. As each user has their distinct native accent and voice characteristics, it was hard to identify if speech emotion recognition depends on accent or not. We tried to determine the dependency by training on a dataset of a specific accent and testing on some other. The overall accuracy of that experiment was very low. It proved that there were some amount of dependency as a result we got a poor result. So to implement this speech recognition system practically it was necessary to implement it irrespective of whichever person uses it no matter their accent. So we tried to focus on the identification of the features that make SER accent independent. The proposed method consists of 5 parts-

- Data Collection
- Speech Preprocessing
- Feature Extraction
- Feature Selection
- Classify and Compare Emotions for Accents

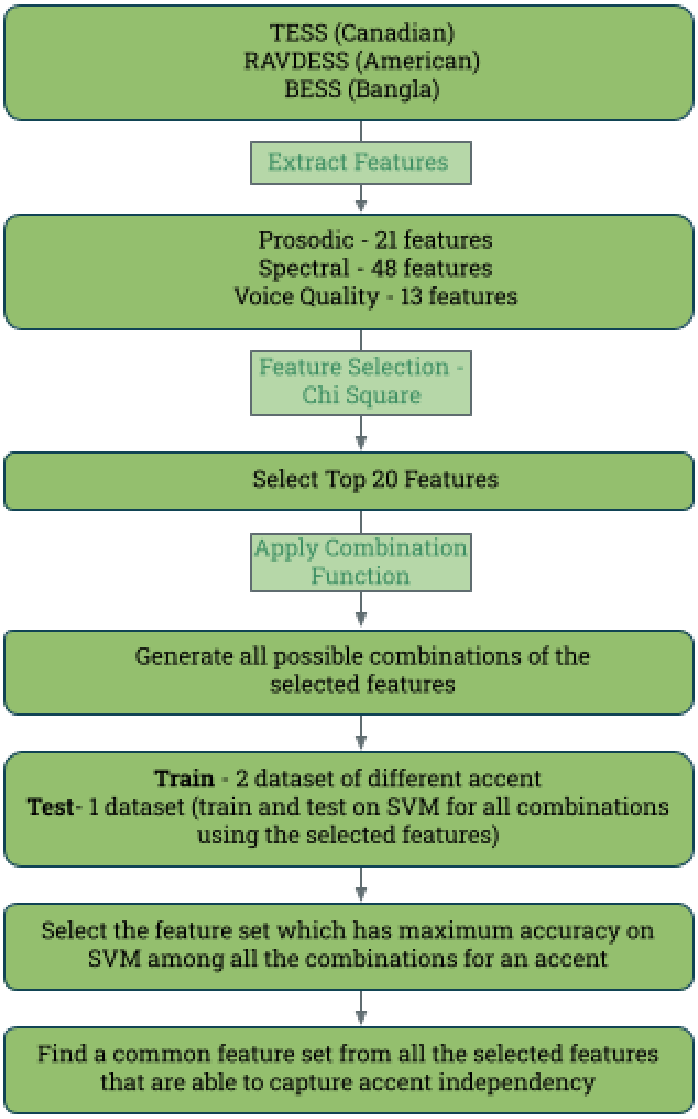
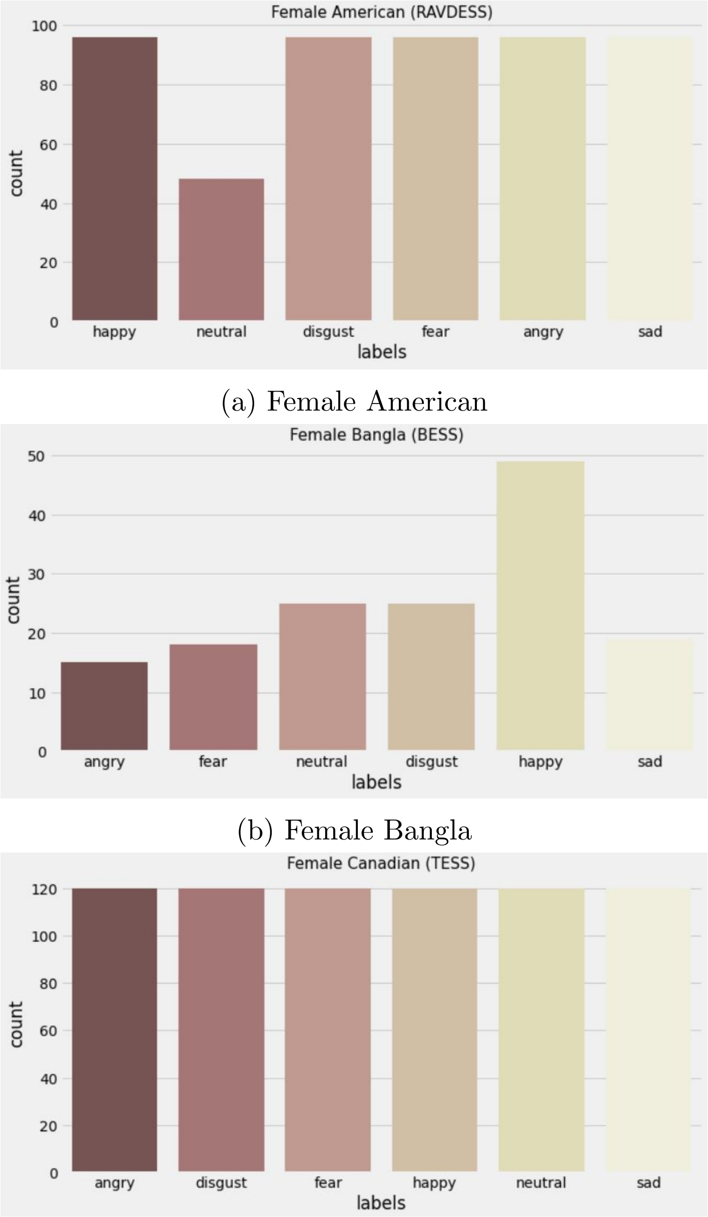

TESS (Canadian)
RAVDESS (American)
BESS (Bangla)
Extract Features
Prosodic - 21 features
Spectral - 48 features
Voice Quality - 13 features
Feature Selection - Chi Square
Select Top 20 Features
Apply Combination Function
Generate all possible combinations of the selected features
Train - 2 dataset of different accent
Test- 1 dataset (train and test on SVM for all combinations using the selected features)
Select the feature set which has maximum accuracy on SVM among all the combinations for an accent
Find a common feature set from all the selected features that are able to capture accent independency

Figure 4.1: Proposed Methodology

4.1 Data Collection

As we focused on studying the dependency of SER over accent, we had to collect datasets of different accent. Moreover, to eliminate gender dependency we needed datasets of male and female separately. We collected 4 datasets for 4 types of Accents

- American English (RAVDESS) [15]
- British English (SAVEE) [8]
- Bengali (BESS) [25]
- Canadian English (TESS) [21]



(a) Female American

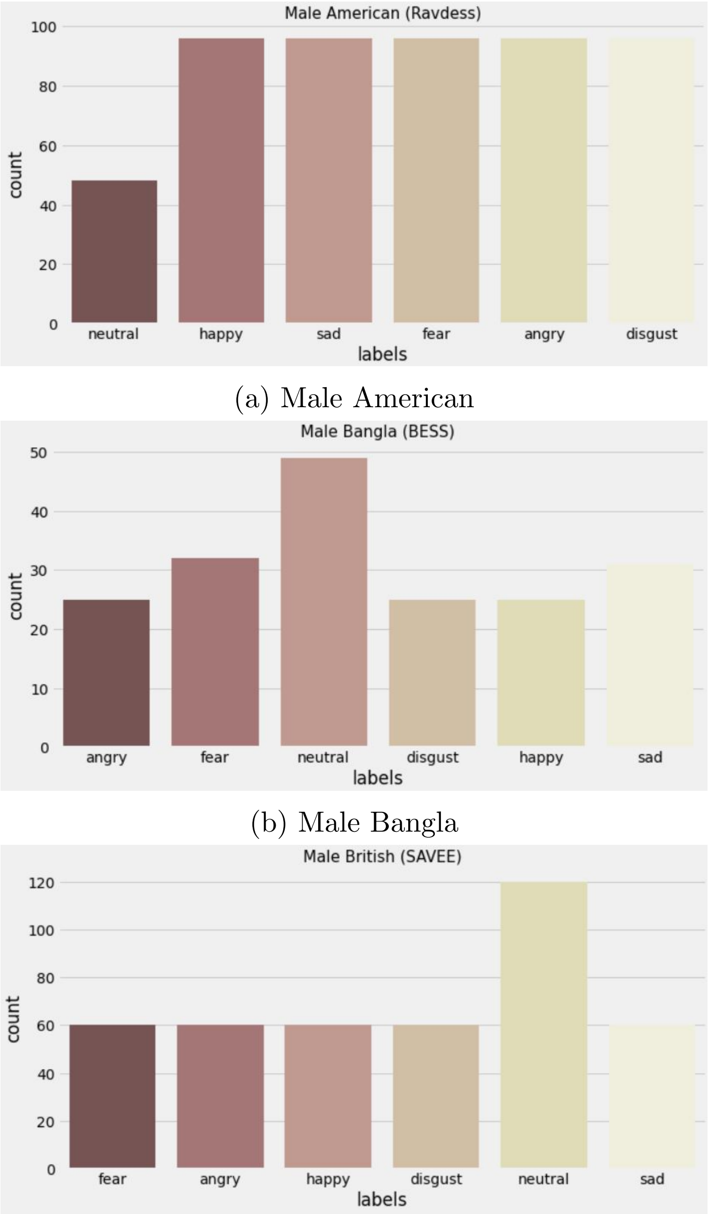
(b) Female Bangla

(c) Female Canadian

Figure 4.2: Female Sample count

For the American English Accent, we chose The RAVDESS(Ryerson Audio-Visual Database of Emotional Speech and Song) which consists of 8 emotions: calm, happy, surprised, angry, sad, disgust, fearful, neutral and 1440 recordings by 24 actor and actresses.

For the British English Accent, we chose Savee which consists of 7 emotions:



(a) Male American

(b) Male Bangla

(c) Male British

Figure 4.3: Male Sample count

happiness, fear, disgust, anger, surprise, sadness, and neutral and 480 British English utterances by 4 males actors.

And for the Bengali English Accent we chose BESS (Bangla Emotional Speech Set) which consists of 6 emotions: happiness, fear, disgust, anger, sadness, and neutral.

For the Canadian English Accent, we chose The TESS(TORONTO EMOTIONAL SPEECH SET) which consists of 7 emotions: happiness, fear, disgust, anger, surprise, sadness, and neutral.and 2800 recordings by 2 actresses.

Among these 4 datasets, TESS has only female samples and SAVEE only male samples. BESS and RAVDESS has both samples. So we have experimented with TESS, RAVDESS and BESS for female and SAVEE, RAVDESS and BESS for male separately.

4.2 Dataset Preprocessing

We will preprocess the 3 datasets to take equal no. of samples for all emotions of all accents and also undersize the dataset to a dataset which has less samples so that the model will not overfit to a specific accent which has more data.

4.2.1 Speech Preprocessing

- Pre Emphasis
- Framing
- Windowing
- Overlapped Frames
- Speech Normalization

Because an audio signal is non-stationary,so as to make things easier we presume that it will remain stationary for a brief period of time (statistically stationary). This is why we use 20-40ms short time frames to frame the signal. We don't have enough samples to get a reliable spectrum approximation if the frame is too short, and the signal fluctuates too much throughout the frame if it is too lengthy. That is why the Framing Windowing and Overlapping of Frames are done. Mostly the Framing, Windowing ,and overlapping of Frames are done automatically while extracting Features through Librosa or Praat Libraries.However , We will add some noise reduction preprocessing techniques like Pre Emphasis and Speech Normalizations for better generalization of the emotion detection

4.3 Feature Extraction

We will Extract all the types and kinds of Features that we listed down through Background study till now.The types of Features are-

- Global Features
- Local Features
- Prosodic Features
- Spectral Features
- Voice Quality Features

In the above features prosodic, spectral and voice quality features are largely used for speech emotion recognition. There are 21 prosodic features such as pitch, intensity, zero crossing rate. These are the features which represent stress, tone or word juncture that is added over consonants or vowels.There exists 47 spectral features such as mfcc, lpc, spectral density. The spectral features are the features which are obtained by converting time domain features into frequency domain by fourier transform. To identify rhythm, pitch, notes, melody, these features are used. Finally, there are 13 voice quality features such as shimmer, jitter. These are the features which depends on an individual and derived from laryngeal and supralaryngeal features. Total there are 81 features. But these 81 features altogether don't contribute on speech emotion recognition. So we have focused on finding the specific features which will contribute on speech emotion recognition irrespective of people with their different accent.

4.4 Feature Selection And Analysis

When building a predictive model, feature selection is the process of minimizing the number of input variables before training a model to prevent overfitting or to improve generalization to data i.e to choose the most representative features and remove the unnecessary ones. Reducing the number of input variables is desired since it lowers the computational time of modeling and, in some situations, enhances the model's performance. For feature selection we will use Correlation Based Feature Selection which will decrease the feature to feature correlation and Recursive Feature Elimination which will find out the most relatable features for the emotion labels thus increasing feature to label correlation and through techniques we will find out which Features are more Accent neutral for emotion Classification.

In the case of speech emotion recognition, it is more necessary to detect features which will contribute in emotion recognition because all the extracted 81 features do not provide a good accuracy altogether. So we have to experiment via trial an error to detect a specific feature set that will be able to give a good result. So feature selection is the most vital part in our study.

After Feature Extraction and Selection , now its time to run experiments.We collected 3 datasets on 3 different accents. We will train a model with 2 accents and then we will test on the 3rd Dataset. For example if we train with Bangla and British English Accent, then we will test with Canadian accent. Following this method, we will experiment all the possible way by combination and note all the features that is giving good result for every experiment. After the experiment we will analyse the features that we've obtained and infer the features to detect the ones that are found common in every experiment. This are the features we can say is contributing to emotion recognition no matter the accent as we have detected them by training and testing on different datasets.

4.5 Classification

For classification we have used LSTM, SVM, Catboost and some other classifiers. All of them performed quite well for classifying the emotions after feature selection. But they took a lot of time for just 1 combination, except SVM. Since we had lots of combination to test, we carried our further experiments using SVM.

SVM is a supervised machine learning algorithm which is mostly used for classification task.Here the data items are plotted in n-dimentional plane where n is the feature number. Then the classification is performed by finding the hyperplane that differentiates the classes. There are some hyperparameters in SVM. one of the hyperparameter is kernel. 'rbf', 'poly' are non-linear kernel. There are gamma values which determines how the classifier will fit the data. Higher gamma values can cause overfitting. And there is also penalty parameter C which determines how smooth the decision boundary will be.
For our experiments we used 'rbf' kernel, C = 10, gamma = 0.001.

Chapter 5

Experimental Design & Result Analysis

This reserach aims to find out if Speech Emotion Recognition is affected by various accents and if so which features which help the SER System be more accent neutral.For this we designed some experiments for Feature Anlaysis to find out the accent neutral features.In this chapter we will discuss about the experiments conducted and the results concluded from the experiments.

5.1 Dataset

Since our study is on the impact of accents on Speech Emotion Recognition.4 different Datasets for 4 different accents on the same English Language were taken-

- SAVEE - British Accent
- RAVDESS - American Accent
- TESS - Canadian Accent
- BESS - Bengali Accent

Among them The dataset SAVEE has samples of only male actors and the dataset TESS has samples of only female actors.So the dataset collected were divided seperated among Male and Female samples to conduct experiments on Male and Female samples seperately, which will also help to get rid of any gender biasness in the experiments. Among the datasets,

SAVEE had 7 emotion labels - 'angry', 'disgust', 'fear', 'happy', 'neutral', 'sad', 'surprise'.

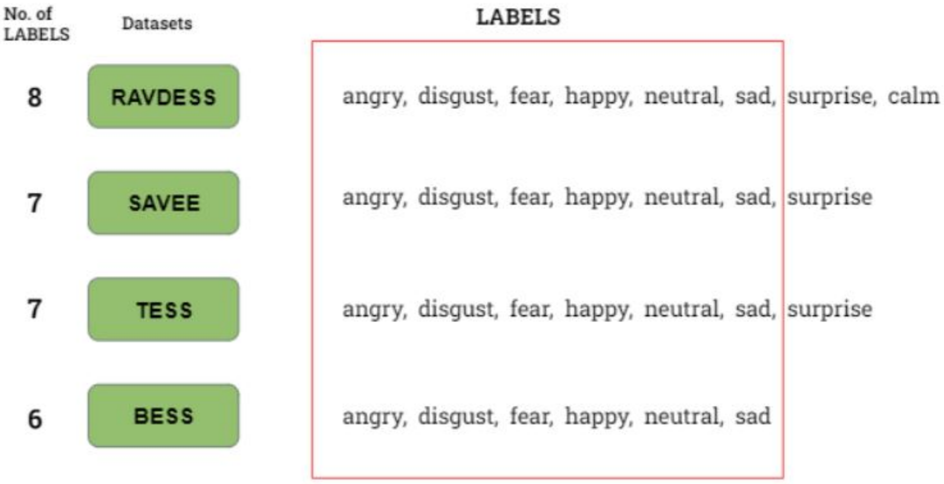
Ravdess had 8 emotion labels - 'angry', 'calm', 'disgust', 'fear', 'happy', 'neutral', 'sad', 'surprise'.

TESS had 7 emotion labels. 'angry', 'disgust', 'fear', 'happy', 'neutral', 'sad', 'surprise'

and

BESS had 6 emotion labels -'angry', 'disgust', 'fear', 'happy', 'neutral', 'sad'.

So we preprocessed the datasets to take the common 6 emotions from all 4 datasets for the ease of experimentations.



No. of LABELS	Datasets	LABELS
8	RAVDESS	angry, disgust, fear, happy, neutral, sad, surprise, calm
7	SAVEE	angry, disgust, fear, happy, neutral, sad, surprise
7	TESS	angry, disgust, fear, happy, neutral, sad, surprise
6	BESS	angry, disgust, fear, happy, neutral, sad

Figure 5.1: Dataset Preprocessing

5.2 Feature Extraction Tools

We are extracting 81 features from the speech samples.And the Libraries and softwares we used to extract these features are-

- Praat Software
- Praat-parselmouth library
- Librosa library

5.2.1 Librosa

To extract features, we broke down the audio file into small time frames(windows), about 20-100 milliseconds. We then extract these features per window.All of these are done directly through the Python Library Librosa. We mainly used the library Librosa for Extracting Spectral Features.

Here is a brief Description of what is happening inside the library while extracting the spectral Features:

MFCC

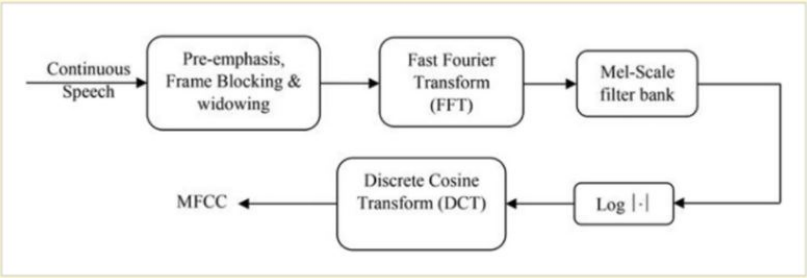

Continuous Speech
Pre-emphasis, Frame Blocking & widowing
Fast Fourier Transform (FFT)
Mel-Scale filter bank
Log |·|
Discrete Cosine Transform (DCT)
MFCC

Figure 5.2: MFCC feature extraction

1. Framing the signal into short frames. 2. Converting the signal to the frequency domain through FFT 3. Applying the mel filterbank to the spectra

and summing up the energy in each filter. 4. Then taking the logarithm from all filterbank energies. 5. Taking the DCT of the log filterbank energies. 6. And finally keeping DCT coefficients 2-13, discarding the rest. [2]

LPC

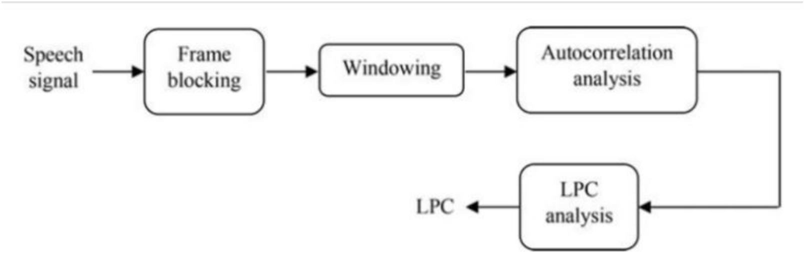


Figure 5.3: LPC feature extraction

Each frame of the windowed signal is autocorrelated, while the highest autocorrelation value is the order of the linear prediction analysis. This is followed by the LPC analysis, where each frame of the autocorrelations is converted into LPC parameters set which consists of the LPC coefficients.[2]

LPCC

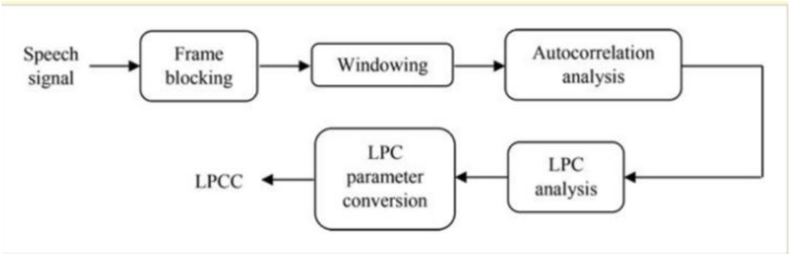


Figure 5.4: LPCC feature extraction

Linear prediction cepstral coefficients (LPCC) are cepstral coefficients derived from the spectral enveloped calculated form LPC analysis.[2]

5.2.2 Praat Software and Praat-Parselmouth Library:

The prosodic , statistical and voice quality features are extracted using the software PRAAT and the library built from this software Praat-Parselmouth.

5.3 Feature Selection

For Feature Selection, We used Chi2, ANOVA feature selection(fclassif()), Information Gain, Fisher Score to get the top 20 features that has the most impact on speech emotion recognition task from the 81 features we extracted.

5.3.1 Experiments

To look into if Accents have an impact on speech emotion recognition, we designed a series of experiments. For the experiments that we ran after Feature Extraction and Feature Selection-

5.3.1.1 Same Accent Experimentation

For the first Experiment, our aim was to investigate how SER performs on the same accent. So , we took each dataset of a particular accent and split the dataset into training and testing samples and then we trained the classifier we chose -SVM with the training samples and tested the classifier on the testing samples from the dataset of same accent using all the 81 features extracted.

5.3.1.2 Cross Accent Experimentation

For the next experiment or aim was to focus on how Speech Emotion Recognition performs on different accents of a language.So, we took datasets on different accents and we trained and tested our classifier on different accent but on the same language - English. For this part we split the datasets into Male and Female samples and separated them.

For Female Samples - RAVDESS, TESS and BESS were used.

For Male Samples - SAVEE, RAVDESS and BESS were used.

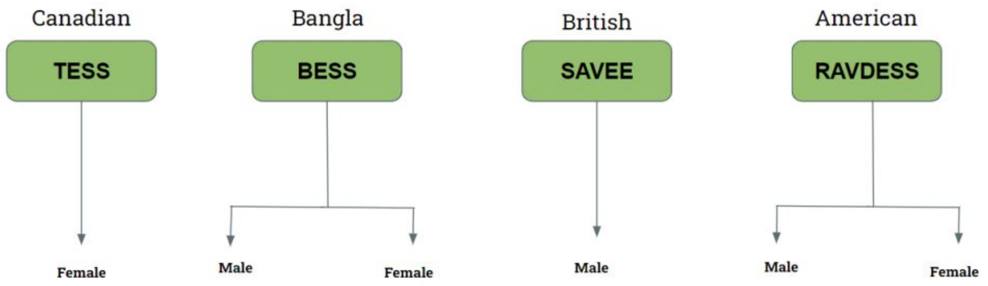


Figure 5.5: Dataset Split

5.3.1.3 Experimentation on Female Samples

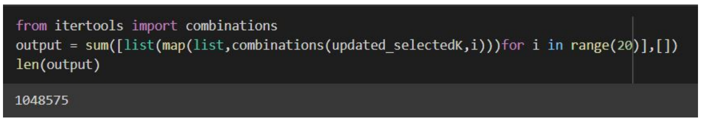
With only female samples, We ran some One vs All experiments, where we took one dataset(one accent) as a test set and rest of the datasets(accents) as training set. i.e we trained our classifier(SVM) with two accents and tested the trained classifier on a different accent.

Exhaustive Feature Search :

Here at first we use all 81 features as input to our classifier and train and test on different accents to find out how our model performs on different accents of same language for the speech emotion recognition task using all 81 extracted features. Then to find out if speech emotion recognition is accent independent and to find out the best subset of features that might capture the accent indepedency for this task, we conducted experiments in two ways.

- Using Feature Selection Algorithm
- Dividing features into Prosodic , Spectral and Voice Quality without using Feature Selection Algorithm

For the first approach we took top 20 features from the extracted 81 features selected by the Feature Selection Algorithms and formed all possible combinations from the top 20 features using the python functions itertools.combinations().



```
from itertools import combinations
output = sum([list(map(list,combinations(updated_selectedK,i)))for i in range(20)],[])
len(output)

1048575
```

Figure 5.6: Combinations of features

And the second approach was, dividing the 81 features into Spectral (47) , Prosodic(21) and Voice Quality(13) features , and forming combinations of all the features seperately from each of the feature division.Then finding the best set of features from each feature division seperately and combining the best sets to get a combined best feature set among 81 features.

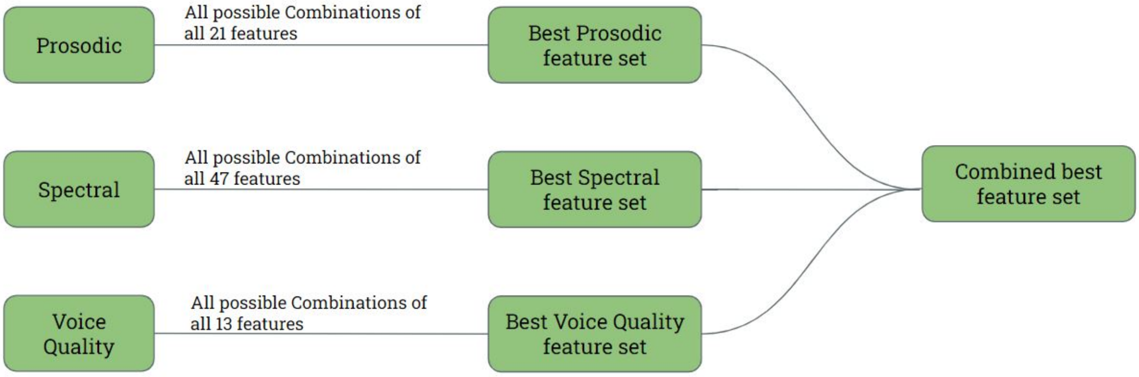

Prosodic
All possible Combinations of all 21 features
Best Prosodic feature set
Spectral
All possible Combinations of all 47 features
Best Spectral feature set
Combined best feature set
Voice Quality
All possible Combinations of all 13 features
Best Voice Quality feature set

Figure 5.7: Finding best set of features

3 experiments were conducted in this ways-

- Train : American + Canadian Test: Bangla
- Train : Canadian + Bangla Test : American
- Train : Bangla + American Test : Canadian

5.3.1.4 Experimentation on Male Samples

With only male samples, We also ran some One vs All experiments similar to the female samples experimentations.And 3 experiments were conducted in this ways-

- Train : American + British Test: Bangla
- Train : British + Bangla Test : American
- Train : Bangla + American Test : British

In this way the experiments were conducted to find out which features make speech emotion recognition accent independent for the same language or if it makes speech emotion recognition accent independent at all,the results are discussed in the next section.

5.4 Result Analysis

To investigate whether or not SER system can identify the emotional state of a person from their speech signals regardless of different accents in the same language English we conducted several experiments.
Since we didn't have samples for both males and females in every accent, our experiment was conducted separately for Males and for Females.

5.4.1 Training and testing with Same accent:

In the first experiment, we can see that individually the datasets have a good enough recognition rate, when the same accent is used for both training and testing the accuracy is good enough. Ravdess has a little less accuracy since it's American English some of the speakers couldn't pronounce the sentence clearly and SVM couldn't classify those emotions properly. For the other accents Canadian, Bangla and British the result was satisfactory.

Female samples	
Train Test with same accent	**Accuracy**
TESS(Canadian)	99.54%
BESS(Bangla)	95.65%
Ravdess(American)	52.83%
Male samples	
Train Test with same accent	**Accuracy**
SAVEE(British)	65.87%
BESS(Bangla)	78.95%
Ravdess(American)	50.31%

Table 5.1: Results on same accent experiments.

5.4.2 Training and testing with different accents:

But when we used Different accents for training and testing using the features we extracted accuracy dropped by a large scale. SER system couldn't recognize the emotions from speech if trained with one accent in English and tested with the other accent in English.

We conducted One Vs ALL experiments (One accent/dataset for testing Vs all other accents/datasets for training) separately for Male and Female samples, so that gender dependency doesn't become an issue. Without any kind of feature selection when such experiments are conducted accuracy dropped by a large scale indicating that accent variation matters in Speech Emotion Recognition systems.

Female samples		
Train	**Test**	**Accuracy**
Canadian+American	Bangla	32.45%
Bangla+American	Canadian	16.67%
Bangla+Canadian	American	18.18%
Male samples		
Train	**Test**	**Accuracy**
British+American	Bangla	26.2%
Bangla+American	British	14.29%
Bangla+British	American	9.09%

Table 5.2: Results on cross accent experiments without feature selection.

5.4.3 Training and testing with different accents by selected features:

Our aim was to find such a subset of features for which SER system will have a good enough accuracy and will predict the emotions correctly despite variation in accents for the same language.

So, in the next One-Vs-All experiments, we used different feature selection techniques, Chi-Square, Fisher score, F_classify, information gain and selected top 20 features, also used some trivial approach to select some best features, applied combination function to generate all possible combinations of those features, then trained and tested SVM for each combination and found a common pattern of features for which even if we train and test with different accents accuracy is higher and the system can recognize emotions properly.

5.4.3.1 Ablation Study

The table shows the results found **after applying the selected features**, when we keep one dataset for testing and the others for training, the feature sets for which we have maximum accuracy are noted.

Female samples		
Train	**Test**	**Accuracy**
Canadian+American	Bangla	62.91%
Bangla+American	Canadian	64.44%
Bangla+Canadian	American	35.04%
Male samples		
Train	**Test**	**Accuracy**
British+American	Bangla	60.43%
Bangla+American	British	44.52%
Bangla+British	American	32.01%

Table 5.3: Best results on cross accent experiments after feature selection.

The best feature sets for which we are having this high accuracies:

For Female samples

1. Test with **Bangla**: *q3_pitch, median_intensity, q1_intensity, mfcc 7, mfcc 2, mfcc 6* gives accuracy of 62.91%.

2. Test with **canadian**: *min_pitch, stddev_pitch, pitch_slope_without_octave_jumps, max_intensity, relative_min_intensity_time* gives accuracy of 64.44%.

3. Test with **American**: *min_pitch, max_intensity, relative_min_pitch_time, relative_max_intensity_time, q3_intensity* gives accuracy of 35.04%.

.

For Male samples

1. Test with **Bangla**: *’median_intensity’, ’min_intensity’, ’q1_intensity’, ’mfcc 9’, ’mfcc 12’* gives accuracy of 60.43%.

2. Test with **British**: *’fitch_vtl’, ’min_intensity’, ’q1_pitch’, ’stddev_hnr’* gives accuracy of 44.52%.

3. Test with **American**: *mfcc 4, mfcc 8, min_pitch, stddev_intensity, pitch_slope_without_octave_jumps, relative_min_intensity_time ,max_intensity* gives accuracy of 32.01%.

From all those sets a common pattern of **Pitch, Intensity and MFCC** is found for which SER is having high accuracy despite 3 different accents in English. Thus it may be deduced that these features are able to capture the accent independence for a common language(English) in SER.

Chapter 6

Conclusion and Future Works

Speech Emotion Recognition (SER) seems to have a very poor performance when there is a variation of accents in the same language. Since with different accents human speech sounds different, it is hard for a machine to recognition the emotion if there is an accent variation in the same language. Our research is to investigate and find a set of features for which SER has a better performance and recognizes emotions correctly despite having variations in accents for the English language.

Our analysis focused on a total of 4 accents in the English language which are Canadian, British, Bangla, and American. 3 of them were for Female samples (Canadian, Bangla, American) and 3 of them for Male samples (British, Bangla, American). Total of 6 experiments were designed, 3 experiments for each gender where we saw when trained and tested an SVM classifier with the same dataset (accent) i.e. for intra-accent experiments SER system performs in an impressive way. When One-Vs-All Cross accent experiments are performed by testing with one accent and training with the other accents without any feature selection the performance of the system drops due to variation of accents in the samples. Then after an exhaustive feature search by different feature selection approaches, we selected such sets of features for which despite having variation of accents in the same language Speech Emotion Recognition gives better performance and is able to detect emotions correctly. We noted down the best feature sets for which we are having maximum accuracies in the cross-accent experiments. Observing the selected best feature sets, a common pattern is found which consists of Pitch, Intensity and MFCC features. Performance of the system seemed to improve by 20% to 30% when trained and tested by our selected features. From this it may be concluded that when this 3 features Pitch, Intensity and MFCC are used together for classifying emotions in Speech Emotion Recognition, then despite having accent variations in the same language the system will perform in an accent independent way.

Our future research focuses on including more accents of the English language and finding a feature set that gives decent performance for all those accents. And to find out which emotion is more accent dependent or independent and to find those features responsible for such qualities. This research will be able to remove the accent variation for a language and will help to develop an accent neutral speech emotion recognition system.

Bibliography

[1] Mehmet Berkehan Akçay and Kaya Oğuz. “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers”. In: *Speech Communication* 116 (2020), pp. 56–76.

[2] Sabur Ajibola Alim and N Khair Alang Rashid. *Some commonly used speech feature extraction algorithms*. IntechOpen, 2018.

[3] Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray. “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models”. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* 4 (2007), pp. IV-957-IV–960.

[4] Felix Burkhardt et al. “A database of German emotional speech”. In: *INTERSPEECH*. 2005.

[5] Carlos Busso et al. “IEMOCAP: Interactive emotional dyadic motion capture database”. In: *Language Resources and Evaluation* 42 (Dec. 2008), pp. 335–359. DOI: `10.1007/s10579-008-9076-6`.

[6] Houwei Cao, Ragini Verma, and Ani Nenkova. “Speaker-sensitive Emotion Recognition via Ranking: Studies on Acted and Spontaneous Speech”. In: *Computer Speech Language* 29 (Feb. 2014). DOI: `10.1016/j.csl.2014.01.003`.

[7] John HL Hansen et al. “Getting started with SUSAS: a speech under simulated and actual stress database.” In: *Eurospeech*. Vol. 97. 4. 1997, pp. 1743–46.

[8] Philip Jackson and SJUoSG Haq. "Surrey audio-visual expressed emotion (savee) database". In: *University of Surrey: Guildford, UK* (2014).

[9] Manas Jain et al. *Speech Emotion Recognition using Support Vector Machine.* 2020.

[10] Markus Kächele et al. "Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition". In: *2014 22nd International Conference on Pattern Recognition.* IEEE. 2014, pp. 803–808.

[11] Ruhul Amin Khalil et al. "Speech Emotion Recognition Using Deep Learning Techniques: A Review". In: *IEEE Access* 7 (2019), pp. 117327–117345. DOI: 10.1109/ACCESS.2019.2936124.

[12] Shadi Langari, Hossein Marvi, and Morteza Zahedi. "Efficient speech emotion recognition using modified feature extraction". In: *Informatics in Medicine Unlocked* 20 (2020), p. 100424.

[13] Aijun Li et al. "CASS: A phonetically transcribed corpus of Mandarin spontaneous speech". In: *Sixth International Conference on Spoken Language Processing.* 2000.

[14] Zhentao Liu et al. "Emotional feature selection of speaker-independent speech based on correlation analysis and Fisher". In: *2015 34th Chinese Control Conference (CCC)* (2015), pp. 3780–3784.

[15] Steven R Livingstone and Frank A Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PloS one* 13.5 (2018), e0196391.

[16] Olivier Martin et al. "The eNTERFACE'05 audio-visual emotion database". In: *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE. 2006, pp. 8–8.

[17] Gary McKeown et al. "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent". In: *IEEE transactions on affective computing* 3.1 (2011), pp. 5–17.

[18] Rosanna Milner et al. "A Cross-Corpus Study on Speech Emotion Recognition". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019, pp. 304–311. DOI: 10.1109/ASRU46091.2019.9003838.

[19] Tin Nwe, S.W. Foo, and Liyanage De Silva. "Speech Emotion Recognition Using Hidden Markov Models". In: *Speech Communication* 41 (Nov. 2003), pp. 603–623. DOI: 10.1016/S0167-6393(03)00099-2.

[20] Turgut Özseven. "A novel feature selection method for speech emotion recognition". In: *Applied Acoustics* 146 (2019), pp. 320–326.

[21] M. Kathleen Pichora-Fuller and Kate Dupuis. *Toronto emotional speech set (TESS)*. Version DRAFT VERSION. 2020. DOI: 10.5683/SP2/E8H2MF. URL: https://doi.org/10.5683/SP2/E8H2MF.

[22] Rajesvary Rajoo and Ching Chee Aun. "Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages". In: *2016 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*. 2016, pp. 35–39. DOI: 10.1109/ISCAIE.2016.7575033.

[23] K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. "Emotion recognition from speech using global and local prosodic features". In: *International journal of speech technology* 16.2 (2013), pp. 143–160.

[24] Fabien Ringeval et al. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, pp. 1–8.

[25] Fardin Saad et al. *Is Speech Emotion Recognition Language-Independent? Analysis of English and Bangla Languages using Language-Independent Vocal Features*. Nov. 2021.

[26] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[27] B. Schuller, G. Rigoll, and M. Lang. "Hidden Markov model-based speech emotion recognition". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).* Vol. 2. 2003, pp. II–1. DOI: 10.1109/ICASSP.2003.1202279.

[28] Maheshwari Selvaraj, R Bhuvana, and S Padmaja. "Human speech emotion recognition". In: *International Journal of Engineering & Technology* 8 (2016), pp. 311–323.

[29] Linhui Sun, Sheng Fu, and Fu Wang. "Decision tree SVM model with Fisher feature selection for speech emotion recognition". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2019.1 (2019), pp. 1–14.

[30] Shiqing Zhang. "Emotion Recognition in Chinese Natural Speech by Combining Prosody and Voice Quality Features". In: Sept. 2008, pp. 457–464.