Emotion Detection Through Speech Analysis

MSc Research Project
Data Analytics

Alfred Johnson
Student ID: X17170141

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Alfred Johnson
Student ID:	X17170141
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	12/08/2019
Project Title:	Emotion Detection Through Speech Analysis
Word Count:	4544
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Emotion Detection Through Speech Analysis

Alfred Johnson
X17170141

Abstract

In today's day and age of digital assistants there is a whole new avenue of data that is not being tapped, the audio signal that is being spoken to these assistants. This can be used to great effect be a variety of industries that face regular challenges in identifying the emotional makeup of their clients. Institutions like hospitals, emergency service centers would find such a decision support system invaluable in their day to day working.
Objective: Create a multi-label classification model that will identify the emotion from speech samples.
Methodology: We employ the use of various different classification models and compare and contrast their outputs using robust mathematical evaluation metrics to try and find the most optimal model for the use case.
Results: We can see from the table 2 that in this study the Deep Neural Network (DNN) based model performs the best among the various classification models employed with an overall accuracy of almost 78%.

1 Introduction

Over the years a lot of research has gone into understanding speech emotion recognition from a multi-disciplinary standpoint. Researchers from neuroscience area understand the way brain perceives the input stimuli by processing the raw data and applying the knowledge stored in the brain to reflex the various actions. Computational intelligence researchers provide tools to justify the knowledge gained from the neuroscientists in a mathematical solution while linguists would rather view speech that can provide emotional knowledge through semantic and syntactic analysis of the speech. These combinations of interdisciplinary researches have shown tremendous potential in understanding speech emotion. Scherer (2003) provides a an overview of the various design paradigms on the subject using a modified Brunswick's functional lens model of perception. There have been many other attempts as well at using the spectral analysis and Hidden Markov Model for identifying and recognizing emotions as described in Ververidis and Kotropoulos (2006), Womack and Hansen (1999) and Zhou et al. (2001).

Humans interact with their environment using many different sensing mechanisms. Detecting an unpleasant state during the task and intervening the process is possible with real time affective systems. In human computer interaction, the main task is to keep users level of satisfaction as high as possible. A computer with affective properties could detect the users emotion and could develop a counter response to increase user satisfaction. Speech and gesture recognition are the most popular affective computing topics. Speech and gesture recognition are possible with passive sensors. While in the new age of information security there is an unease in granting full video sensor access but audio based devices are already well entrenched in the market and hence an excellent avenue for exploring emotion detection.

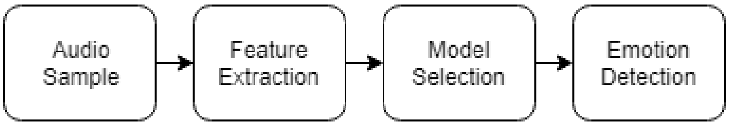

Figure 1: Data flow Model

In a typical model it is important to decide the features for extraction. These are mostly based on a series of robust mathematical models. The same exists for speech analysis. Vocal tract information like formant frequency, bandwidth of formant frequency and other values may be linked to a sample. There is a wide variety of options for parametrically representing the speech signal in a machine understandable way so that statistical analyses can be performed on it to arrive at informed results. Some of these techniques are : Linear Prediction Coding (LPC); Mel-Frequency Cepstral Coefficients (MFCC); Linear Predictive Cepstral Coefficients (LPCC); Perceptual Linear Prediction (PLP); and Neural Predictive Coding (NPC). Mel Frequency Cepstral Coefficients (MFCC) is a popular technique because it is based on the known variation of the human ear's critical frequency bandwidth. MFCC coefficients are obtained by de-correlating the output log energies of a filter bank which consists of triangular filters, linearly spaced on the Mel frequency scale. Conventionally an implementation of discrete cosine transform (DCT) known as distributed DCT (DCT - II) is used to de-correlate the speech as it is the best available approximation of the Karhunen-Loeve Transform (KLT). (Sahidullah and Saha; 2009).MFCC data sets represent a melodic cepstral acoustic vector (Barbu; 2009),(Wang et al.; 2006). The acoustic vectors can be used as feature vectors. It is possible to obtain more detailed speech features by using a derivation on the MFCC acoustic vectors. You can obtain higher order MFCC coefficients as well as log energy scale values for them which are all valid features to aid in classification.

The next step after the feature selection is the model selection. There have been a vast array of models that have been explored for emotion detection. We will focus on 3 main models namely : Support Vector Machine (SVM),K Nearest Neighbours (K-NN) and DNN.

The rest of this paper is arranged in the following order. The Related work section deals with relevant papers from the community which inspire us in our current paper. The Methodology and design section will explain in detail the experimental setup for the purpose of reproducibility. The implementation will detail the exact steps undertaken to clean and create the data as well as the models that are going to be used. The evaluation and future work section will expand on the results obtained from the models and how to expand them for better performance in the future.

2 Related Work

Any consumer specific device can tailor make the interaction with its user if it can get access to their current emotional make-up. This will help create a more seamless and enriching interactive experience. This is why we see a host of new research in this area. Speech based emotion recognition has a few core areas that are important :

- Feature extraction.
- Classification algorithm.

We will focus the related work in each of these areas and showcase the advantages and disadvantages.

2.1 Feature extraction

In their paper Gupta et al. (2014) enumerate that in recent times various speech feature extraction methods have been proposed. Diverse methods are differentiated by the ability to use most information about human speech processing perception by considering distortions and by the length of the observation window. The speech is highly redundant due to human physiology and has a variety of speaker-dependent features such as pitch, speaking rate, frequency and accent. In the paper they have employed a pitch energy and MFCC based feature selection.

Zeng et al. (2008) in their work, also showcase the MFCC based features. They use it because of its low complexity, better ability to extract the feature from speech, efficient technique and also has the advantage like anti-noising etc.

After suppressing vocal tract (VT) characteristics, excitation source signal is obtained from speech. This is achieved by first predicting the vocal tract information using linear prediction coefficients from speech signal and then separating it by inverse filter formulation. The resulting signal contains mostly the information about the excitation source and is known as linear prediction residual(Makhoul; 1975). The paper by Shashidhar et al. (2012) explores the concept of using LPC for detection of emotions. There are few other papers that use this possibly because it is seen as an error signal.

The concept of using pitch and energy has been explored in the work written by Schuller et al. (2003).An analysis is done on the contours of pitch and energy since they have a well-known capability to carry a large amount of information considering a users emotion and. In order to calculate the contours, a Hamming window function is used every 10 seconds to analyse the frames of the speech signal. Energy value is calculated by the logarithmic mean energy within a frame. By using average magnitude difference function (AMDF), the pitch contour is achieved.

2.2 Classification Algorithm

Tarunika et al. (2018) has applied two promising algorithms in the paper. The paper has employed a deep neural network as well as a K-NN model to carry out the emotion detection. They apply multiple hidden layers in their architecture to increase the accuracy of their predictions.

Before the deep learning era people have come with many different methods which mostly extract complex low-level handcrafted features out of the initial audio recording of the utterance and then apply conventional classification algorithms. One of the approaches is to use generative models like Hidden Markov Models or Gaussian Mixture Model to learn the underlying probability distribution of the features and then to train a Bayessian classifier using maximal likelihood principle. Variations of this method was introduced by Schuller et al. (2003) in and by Lee et al. (2004).

Another common approach is to gather a global statistics over local low-level features computed over the parts of the signal and apply a classification model. Eyben et al. (2009) and Mower et al. (2010) used this approach with SVM as a classification model. Lee et al. (2011) used Decision Trees and Kim and Provost (2013) utilized K-NN instead of SVM.

2.3 Summary

Based on the research that we have read about in the sections 2.1 and 2.2 in this paper we will aim to compare and contrast the K-NN,SVM and DNN based techniques over the features extracted from MFCC, cepstrum ,pitch and energy. We will aim to present a easily understandable comparison between the different methods based on different factors like features extracted, gender of speaker, intensity of emotion of the speaker etc.

3 Methodology

In this paper we aim to evaluate side by side three popular algorithms for detecting emotions from speech and analyze them with respect to the selected features. In this proposed model we use the Cross Industry Standard Process for Data Mining (CRISP-DM). It is a data driven reiterative process which is described by fig. 2.

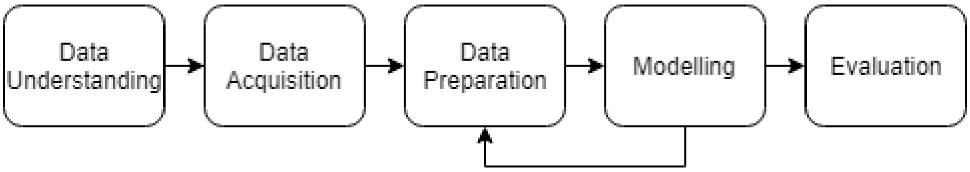

Figure 2: Crisp-DM work Flow

3.1 Data acquisition

For the purpose of emotion recognition using speech there is always a question of the type of data that will be used. Databases for usual speech recognition task are relatively easy to collect: one can take dialogues from the films, Youtube blogs, news, etc. and annotate them. Almost the only requirement is the high quality of the audio recording. These very same sources do not provide the same quality of data when it comes to emotion detection databases as they are dramatically biased.

A different approach to it is artificially create the databases. This leads to the question how to record in a way that does not corrupt the data quality. These issues have been tackled by Douglas-Cowie et al. (2003) in their paper. They suggest using a set of professional actors to make a corpus of recordings in different emotions.

Using this as the background for our paper as well we have selected the Ryerson Audio-Visual Database of Emotional Speech and Song Databse (RAVDESS). As described by Livingstone and Russo (2018) in their paper RAVDESS contains 7356 files. Each file was rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained adult research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity, interrater reliability, and test-retest intrarater reliability were reported. The entire dataset contains three formats for 24 different actors. The formats are audio only, audio-video and video only. Since we do not need and video format data for this paper we will be only making use of the audio only speech tracks of the 24 actors. This gives us a final dataset of 1440 samples.

3.2 Data Pre-Processing

To conduct our research we needed to segregate the data according to two main criteria:

- Gender.
- Intensity of emotion

The segregation by gender and intensity is to be able to compare and contrast the models under different conditions to be able to better understand the workings of the given algorithm. To this end we use a python script to get all the data from the download file and neatly segregate them. An example of the resultant data-frame would be:

Filename	**Gender**	**Emotion Name**	**Emotion value**	**Intensity**
`01-01-01-01.wav`	1	Anger	4	0

Table 1: Example data sample before feature extraction.

Once this is done we can move onto feature extraction for creating our final training data-set on basis of which criteria we are going to train the model.

3.3 Feature Extraction

Feature extraction is the next crucial step in our process. It is imperative to gather the right features as we cannot work directly with the audio file. This means the set of features we will select will be the representative of the original data in our models and this greatly enhances the role of the features in our research. Keeping this factor in mind we aim to extract almost all aspects which can give us information regarding the vocal characteristics of the speakers and samples. Below are the set of features we will be using and table shows and example row of the features.

- **MFCC**: MFCC's are derived from the cepstral representation of the audio clip. They are derived by taking the fourier transform of the signal and mapping it to the mel scale. The take the discrete cosine transform of the powers of the mel log powers to obtain the MFCC signal (Ramirez et al.; 2018).
- **Chromagram from the waveform**: Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. Since, in music, notes exactly one octave apart are perceived as particularly

similar, knowing the distribution of chroma even without the absolute frequency (i.e. the original octave) can give useful musical information about the audio – and may even reveal perceived musical similarity that is not apparent in the original spectra.[1]

- **Mel scale Spectrogram**: Computes a spectrogram on the basis of the Mel-scale.
- **Spectral contrast of waveform**: Each frame of a spectrogram S is divided into sub-bands. For each sub-band, the energy contrast is estimated by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy). High contrast values generally correspond to clear, narrow-band signals, while low contrast values correspond to broad-band noise(Jiang, Lu, Zhang, Tao and Cai; 2002).
- **Tonal Centroid features**: Computes the tonal centroid features as explained in Harte et al. (2006). This helps in understanding the harmonic change in audio signal.

3.4 Model building

In this paper three popular classification algorithms are being used, namely SVM,K-NN,DNN to carry out detection of emotion from speech samples.

3.4.1 Support Vector Machine

SVM's are particularly useful where high dimensionality exists in the dataset. By applying a different kernel you can convert any linear model to a non-linear one to achieve better results. The SVM is able to able to find a hyperplane to divide the classes in the best way possible as explained in Jain et al. (2018).

3.4.2 K Nearest Neighbours

K-NN is very computationally quick.K-NN classifies the features using nearest neighbor interpolation method. Using different methods of distance calculation we can affect the accuracy of the model as described in the work done by Gowda et al. (2017).

3.4.3 Deep Neural Network

The DNNis and extension of the Artificial Neural Network (ANN) with more than one hidden layer between the input and output layer. The DNN is a very resource intensive and black box process. We cannot understand the underlyting workings within the model. Only the results can be interpreted using the various metrics like accuracy and F-scores. The model has been put to use with great success in the research done by Han et al. (2014).

[1] `https://labrosa.ee.columbia.edu/matlab/chroma-ansyn/`

3.5 Evaluation

Evaluation will be performed using various metrics like classification accuracy, precision, recall, F1-score. The F1 score reaches its best value as it tends to 1. We will compare and contrast the scores across our experiments to present conclusive evidence on which method gives the best results for detection of speech from audio samples.

4 Implementation

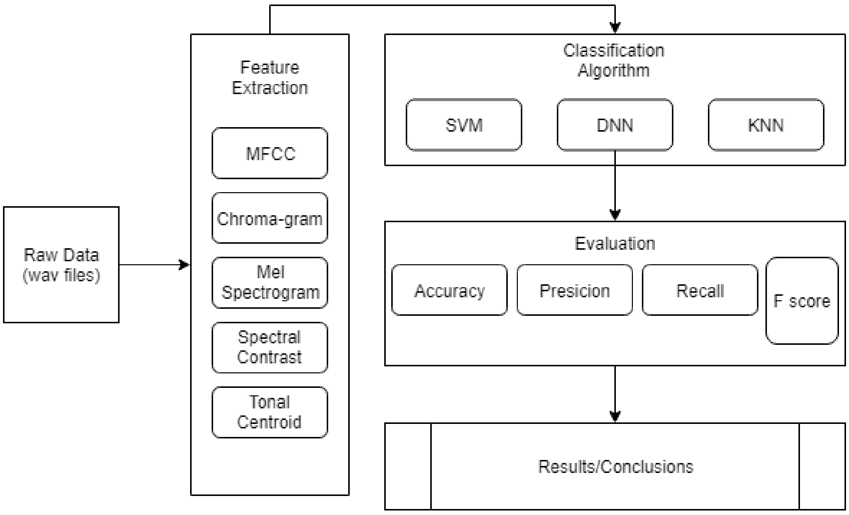

Figure 3: Process Architecture

4.1 Data collection

The Original dataset can be downloaded from the download link available on the RAVDESS website[2]. This gives us the 1440 speech audio samples that we will be using for feature extraction and model training.

4.2 Data Pre-processing

Since we are not working with the audio files directly we will be extracting all the information possible from the audio to serve as a representation for us during the modelling phase.

1. **Identifying emotion labels**: We will first use a python script to label each file with its corresponding emotion.

2. **Segregate data on gender and intensity**: using the same script the next function labels each sample with its corresponding gender and intensity of utterance.

[2] https://zenodo.org/record/1188976/files/Audio_Speech_Actors_01-24.zip?download=1

3. **Feature Extraction**: We will then use the *Librosa* [3] package to extract the features as described in section 3.3. Each of the features extracted are a matrix in their own right. This would make the resultant vector a 3 dimensional vector which is incompatible with certain models. To prevent this we use python to take the mean of the vector along their rows to collapse the matrix into 2 dimensions.

Once these steps are complete we get a resultant dataframe which contains all our samples along with their labels, features , gender and intensity. We Then subset this dataframe to obtain smaller sets which contain :

- **Data containing only feature information**
- **Data with features and gender information**
- **Data with feature, gender and intensity information**

We will be using the above three subsets of the dataframe for training, testing and comparing our models.

4.3 Classification Algorithms

A total of three different classification algorithms are implemented in this paper. They have all been implemented in python using specialized libraries that offer parameter optimization and evaluation metrics.

4.3.1 Support Vector Machine

For creating the SVM model we use the *Scikit-learn* [4] package. There are many methods to implement SVM's. At the heart of the model is the kernel function. The main objective of the kernel function is to take the given input and transform it to the required form. In this model we propose to use the polynomial kernel function. This is because the polynomial function takes into consideration not only the features to gauge their similarity but also the combinations of the features.

4.3.2 K Nearest Neighbours

The K-NN model was also created using the *Scikit-learn* package. Before running the data through the model we apply a scaler function to the data. The function standardizes features by removing the mean and scaling to unit variance. This is important in the K-NN as outliers will tend to skew the data. All points in the dataset are weighted equally. The distance metric used in the implementation is the *Minkowski distance* metric.This distance is calclulated using :-

$$d_p(\mathbf{p},\mathbf{q}) = \sqrt[p]{|x_1 - x_2|^p + |y_1 - y_2|^p}$$

[3] https://librosa.github.io/librosa/

[4] https://scikit-learn.org/stable/

4.3.3 Deep Neural Network

The DNN is implemented using the *Keras* [5] package using *Tensorflow* [6] as the back-end engine. A DNN is defined as a neural network that has more than one hidden layer between the input layers and output layers. We first have to one hot encode our label. One hot encoding converts the labels into a binary representation which will be used for classification. We have 4 Hidden layers in our model. The first layer has an input shape equal to the input data. The second layer has roughly twice the input shape of the first layer. The third layer has half of the input shape of the second layer and the fourth layer has an input shape half of the third layer. We use the Rectified Linear Unit (RelU) function as our activation function. Our optimization function is *adamax*. We train the model for 1000 epochs.

5 Evaluation

The evaluation of the experiments are carried out using the following metrics:

- **Precision**: The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.
- **Recall**: The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.
- **F-score**: The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.
- **Support** : The support is the number of occurrences of each class in `y_true`.
- **Accuracy** : the overall accuracy of the model over the multiple classes.

Model	**Feature only Accuracy**	**Feature and Gender Accuracy**	**Feature, Gender and Intensity Accuracy**
SVM	45.88%	49.35%	71.28%
K-NN	38.96%	45.02%	57.72%
DNN	55.89%	57.53%	77.92%

Table 2: Evaluation of the various models.

In table 2 all the various models are displayed with their respective accuracy's. As can been seen from the table the best accuracy across all the models comes when the we use the features extracted along with the gender and intensity information. Even among

[5] `https://keras.io/`

[6] `https://www.tensorflow.org/`

the models where we passed all of three of the information the DNN performed the best with an overall accuracy of 77%. The lowest performing model is the K-NN with an accuracy of 57%. The SVM model accuracy is just behind the DNN with an accuracy of 71%.

All the models show a marked decrease in prediction capacity when we strip the gender and intensity information. The K-NN shows almost a 20% drop in accuracy as does the DNN. The SVM shows a 30% drop in accuracy when the aforementioned features are not passed.

5.1 Experiment 1 : Support Vector Machine with only feature data.

	F-score	**Precision**	**Recall**	**Support**
angry	0.61	0.58	0.64	28.0
calm	0.46	0.41	0.53	30.0
disgust	0.43	0.46	0.41	34.0
fearful	0.51	0.60	0.45	31.0
happy	0.41	0.36	0.48	25.0
sad	0.35	0.34	0.37	24.0
surprised	0.5	0.54	0.46	28.0
average	0.45	0.46	0.46	231.0

Table 3: SVM with only feature data.

The SVM model metrics presented in table: 3 are trained on the dataset with only information of the features extracted as explained in section 3.3. Using this data the average precision and recall obtained is 0.46. The F-score obtained is 0.45.

5.2 Experiment 2 : Support Vector Machine with feature and gender data.

	F-score	**Precision**	**Recall**	**Support**
angry	0.68	0.68	0.68	32.0
calm	0.43	0.39	0.48	27.0
disgust	0.56	0.56	0.56	28.0
fearful	0.64	0.62	0.67	34.0
happy	0.42	0.42	0.42	28.0
sad	0.45	0.46	0.44	29.0
surprised	0.31	0.36	0.28	25.0
average	0.49	0.48	0.49	231.0

Table 4: SVM with feature and gender data.

The SVM model metrics presented in table: 4 are trained on the dataset with the information of the features extracted as explained in section 3.3 as well as gender information. Per emotion accuracy is detailed in table 4. Using this data the average precision is 0.48 and recall obtained is 0.49. The F-score obtained is 0.49.

5.3 Experiment 3 : Support Vector Machine with feature, gender and intensity data.

	F-score	**Precision**	**Recall**	**Support**
angry	0.85	0.85	0.87	33.0
calm	0.75	0.74	0.76	34.0
disgust	0.66	0.63	0.70	27.0
fearful	0.72	0.7	0.75	28.0
happy	0.58	0.66	0.52	23.0
sad	0.72	0.78	0.66	33.0
surprised	0.58	0.55	0.62	24.0
average	0.71	0.71	0.71	231.0

Table 5: SVM with feature, gender and intensity data.

The SVM model metrics presented in table: 5 are trained on the dataset with the information of the features extracted as explained in section 3.3 along with noth gender and intensity information as well. Per emotion accuracy is detailed in table 5. Using this data the average precision and recall obtained is 0.71. The F-score obtained is 0.71.

5.4 Experiment 4 : K Nearest Neighbours with only feature data.

	F-score	**Precision**	**Recall**	**Support**
angry	0.55	0.62	0.5	30.0
calm	0.39	0.41	0.38	34.0
disgust	0.51	0.47	0.56	30.0
fearful	0.36	0.44	0.30	26.0
happy	0.34	0.44	0.28	28.0
sad	0.36	0.38	0.34	29.0
surprised	0.38	0.47	0.32	31.0
average	0.40	0.43	0.38	231.0

Table 6: K-NN with only feature data

The table : 6 details the per emotion score for the K-NN model which is only trained on extracted features as explained in the section 3.3. The average precision obtained is 0.43 and recall is 0.38. The F-score for this model is 0.40.

5.5 Experiment 5 : K Nearest Neighbours with feature and gender data.

	F-score	Precision	Recall	support
angry	0.66	0.64	0.68	29.0
calm	0.29	0.30	0.28	38.0
disgust	0.48	0.40	0.59	22.0
fearful	0.52	0.59	0.46	28.0
happy	0.37	0.53	0.28	28.0
sad	0.58	0.60	0.56	30.0
surprised	0.44	0.52	0.38	31.0
average	0.45	0.47	0.45	231.0

Table 7: K-NN with feature and gender data.

The table : 7 details the per emotion score for the K-NN model which is trained on both extracted features as well as the gender information of each sample. The average precision obtained is 0.47 and recall is 0.45. The F-score for this model is 0.45.

5.6 Experiment 6 : K Nearest Neighbours with feature, gender and intensity data.

	F-score	Precision	Recall	Support
angry	0.79	0.94	0.69	26.0
calm	0.81	0.73	0.91	24.0
disgust	0.55	0.51	0.61	31.0
fearful	0.69	0.81	0.6	30.0
happy	0.41	0.35	0.48	31.0
sad	0.48	0.48	0.48	31.0
surprised	0.28	0.33	0.24	29.0
average	0.56	0.58	0.56	231.0

Table 8: K-NN with feature,gender and intensity data.

Table : 8 details the per emotion score for the K-NN model which is trained on both extracted features as well as the gender and intensity of utterance related information of each sample. The average precision obtained is 0.58 and recall is 0.56. The F-score for this model is 0.56.

5.7 Experiment 7 : Deep Neural Network with only feature data.

	F-score	Precision	Recall	Support
angry	0.76	0.67	0.87	24.0
calm	0.43	0.37	0.52	25.0
disgust	0.53	0.58	0.5	34.0
fearful	0.67	0.62	0.73	30.0
happy	0.44	0.48	0.40	32.0
sad	0.55	0.65	0.47	40.0
surprised	0.61	0.57	0.65	29.0
average	0.55	0.55	0.55	231.0

Table 9: DNN with only feature data

In table 9 we see the per eomtion classification as performed by the DNN model. This model is trained only on extracted feature data as explained in section 3.3. The average precision obtained is 0.55 and the recall is 0.55. The F-score for the model is 0.55. The model is trained over 400 epochs and the fig 4 shows the accuracy over the number of epochs.

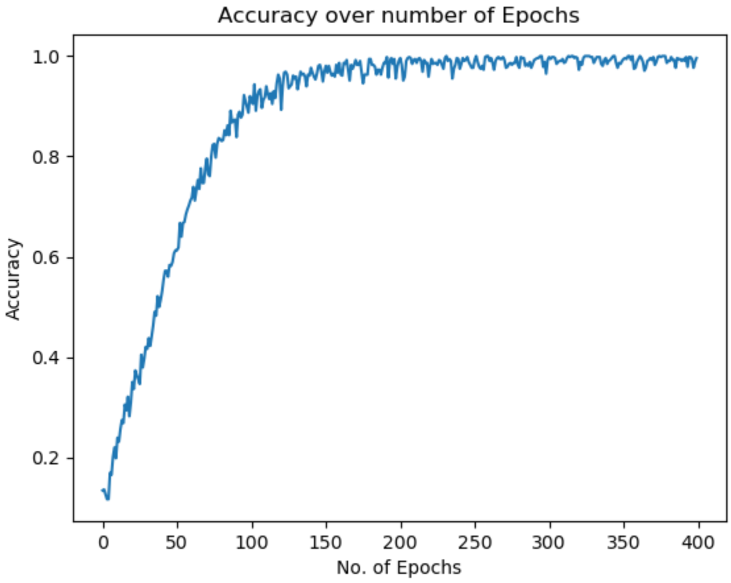

Figure 4: Feature only accuracy over number of epochs.

5.8 Experiment 8 : Deep Neural Network with feature and gender data.

	F-score	Precision	Recall	Support
angry	0.77	0.85	0.70	24.0
calm	0.4	0.31	0.56	25.0
disgust	0.64	0.58	0.73	34.0
fearful	0.67	0.73	0.63	30.0
happy	0.49	0.48	0.5	32.0
sad	0.51	0.72	0.4	40.0
surprised	0.65	0.62	0.68	29.0
average	0.61	0.64	0.61	231.0

Table 10: DNN with feature and gender data.

In table 10 we see the per eomtion classification as performed by the DNN model. This model is trained only on extracted feature data as explained in section 3.3. The average precision obtained is 0.64 and the recall is 0.61. The F-score for the model is 0.61. The model is trained over 400 epochs and the fig 5 shows the accuracy over the number of epochs.

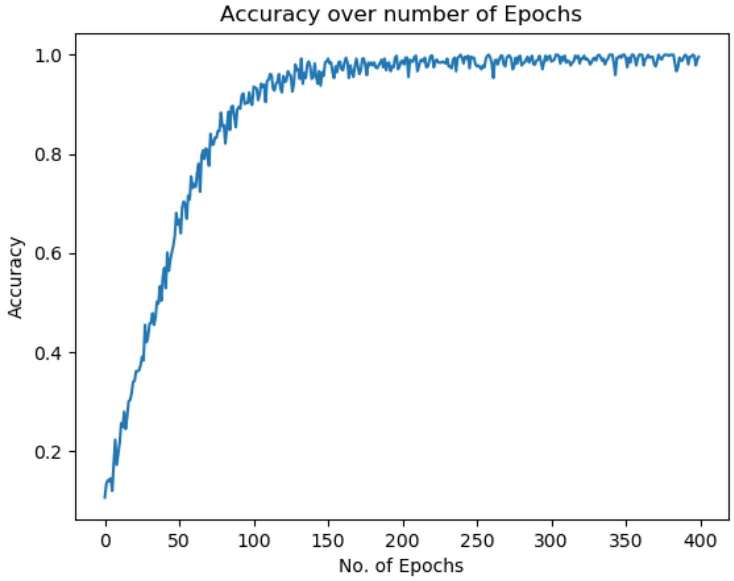

Figure 5: Feature and gender data accuracy over number of epochs.

5.9 Experiment 9 : Deep Neural Network with feature, gender and intensity data.

	F-score	Precision	Recall	Support
angry	0.95	0.92	1.0	35.0
calm	0.87	0.94	0.81	44.0
disgust	0.69	0.65	0.73	23.0
fearful	0.71	0.72	0.69	23.0
happy	0.65	0.70	0.60	28.0
sad	0.75	0.65	0.84	31.0
surprised	0.70	0.75	0.65	38.0
average	0.77	0.78	0.77	231.0

Table 11: DNN with feature,gender and intensity data.

In table 11 we see the per eomtion classification as performed by the DNN model. This model is trained only on extracted feature data as explained in section 3.3. The average precision obtained is 0.78 and the recall is 0.77. The F-score for the model is 0.77. The model is trained over 400 epochs and the fig 6 shows the accuracy over the number of epochs.

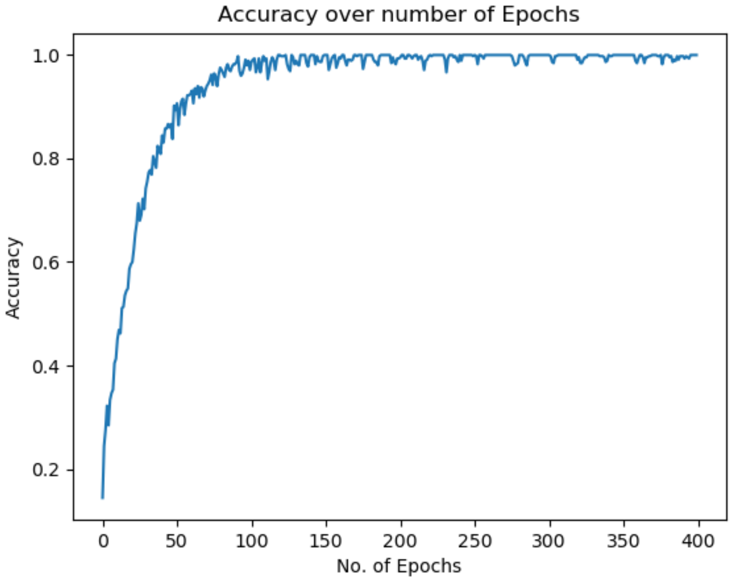

Figure 6: Feature,gender,intensity data accuracy over number of epochs.

5.10 Discussion

In this paper we have tried to define three types of datasets from our original user speech data. Of the three across all the models we can that the dataset with extracted feature info, gender info and intensity information has performed well across all three models.

This can be seen in Table 2. The models trained with only feature data and feature and gender data both suffer a drop in accuracy rates at around 20% and 10% respectively. This can bring us to conclude that greater the intensity of the emotion of the speaker it becomes easier for the models to classify the emotions.

Of the three models that we have implemented the K-NN shows the lowest accuracy. This does show that the K-NN model as it is without any modifications is not the best suited for this application. It would have been better to implement a custom distance metric as has been described by Kim and Provost (2013). Even using this distance metric they have achieved a accuracy rate of 64% which is significantly lower than that of the other models that being considered. It may be extrapolated by this research as well as past research that the K-NN model may not be optimal for this use case.

The SVM model employed in this research with a polynomial kernel gives an overall accuracy of 71%. which is 3% greater that the accuracy that Mower et al. (2010) achieved in their implementation. The time complexity of this algorithm though makes it less appealing to employ. Due to the number of datapoints the time taken for training the model is in quite high compared to the other models. A better approach than the one taken would be to try and parallelize the training. This will help in countering the issue of time complexity that this model faces.

The DNN is the model that performs the best among the three models. Even with just feature data it is almost 10% and 20% more accurate than the SVM and K-NN models respectively. We can see from the accuracy over epoch data from figs 4,5,6 that with the increase in information about the audio samples the number of epochs taken to tend to high accuracy becomes better. The model implemented here is a simple one with just 4 hidden layers. Better accuracy may be obtained by using a a convolutional layer or a Long term Short Memory (LSTM) Recurrent Neural Network (RNN) as described by Deng et al. (2017).

6 Conclusion and Future Work

We have successfully detected the emotions from speech samples and compared three models on basis of various evaluation metrics as described in section 5. We can conclude from our research that the DNN performed the best compared to the K-NN and SVM. Even though the SVM has a comparatively good accuracy its time complexity for training is a serious drawback.

We can greatly improve on this drawback by implementing a parallelized model for performing the training of the SVM model. We can also greatly increase the accuracy of the DNN by using more sophisticated layers as described in the section 5.10. These have not been implemented in the current paper due to time constraints.

This research can be extended to a real world scenario to be put to great use. The trained models can be saved and served as an Application Programmer Interface (API) over the web for real time identification of emotions. This can be used as a service. There is a great need of such kind of detection for calls to emergency numbers to confirm intention of the caller. It can also be used in hospitals to gauge the emotional make-up of the caller which can be helpful to build a patient profile. It can also be without a doubt used by marketing professionals to target the users on basis of emotional make-up too.

In conclusion we have successfully achieved the objective of detecting emotions and comparing the models. While there are certain drawbacks to the methods implemented in this paper all the findings submitted are accurate.

7 Acknowledgement

I would like to thank my guide and mentor Dr. Muhammad Iqbal for encouraging me through the duration of the thesis. He responded to me queries and doubts patiently and helped in guiding the report to its present condition.

Abbreviations

ANN Artificial Neural Network. 6

API Application Programmer Interface. 16

CRISP-DM Cross Industry Standard Process for Data Mining. 4

DNN Deep Neural Network. 1, 2, 4, 6, 9, 10, 13–16

K-NN K Nearest Neighbours. 2–4, 6, 8–12, 16

LSTM Long term Short Memory. 16

MFCC Mel Frequency Cepstral Coefficients. 2, 3, 5

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song Databse. 4, 7

RelU Rectified Linear Unit. 9

RNN Recurrent Neural Network. 16

SVM Support Vector Machine. 2, 4, 6, 8–11, 16

References

Barbu, T. (2009). Comparing various voice recognition techniques, *2009 Proceedings of the 5-th Conference on Speech Technology and Human-Computer Dialogue*, IEEE, pp. 1–6.

Deng, J., Eyben, F., Schuller, B. and Burkhardt, F. (2017). Deep neural networks for anger detection from real life speech data, *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, pp. 1–6.

Douglas-Cowie, E., Campbell, N., Cowie, R. and Roach, P. (2003). Emotional speech: Towards a new generation of databases, *Speech communication* **40**(1-2): 33–60.

Eyben, F., Wöllmer, M. and Schuller, B. (2009). Openearintroducing the munich open-source emotion and affect recognition toolkit, *2009 3rd international conference on affective computing and intelligent interaction and workshops*, IEEE, pp. 1–6.

Gowda, R. K., Nimbalker, V., Lavanya, R., Lalitha, S. and Tripathi, S. (2017). Affective computing using speech processing for call centre applications, *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, pp. 766–771.

Gupta, S., Mehra, A. et al. (2014). Gender specific emotion recognition through speech signals, *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, pp. 727–733.

Han, K., Yu, D. and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine, *Fifteenth annual conference of the international speech communication association*.

Harte, C., Sandler, M. and Gasser, M. (2006). Detecting harmonic change in musical audio, *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, ACM, pp. 21–26.

Jain, U., Nathani, K., Ruban, N., Raj, A. N. J., Zhuang, Z. and Mahesh, V. G. (2018). Cubic svm classifier based feature extraction and emotion detection from speech signals, *2018 International Conference on Sensor Networks and Signal Processing (SNSP)*, IEEE, pp. 386–391.

Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H. and Cai, L.-H. (2002). Music type classification by spectral contrast feature, *Proceedings. IEEE International Conference on Multimedia and Expo*, Vol. 1, IEEE, pp. 113–116.

Kim, Y. and Provost, E. M. (2013). Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 3677–3681.

Lee, C.-C., Mower, E., Busso, C., Lee, S. and Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach, *Speech Communication* **53**(9-10): 1162–1171.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. and Narayanan, S. (2004). Emotion recognition based on phoneme classes, *Eighth International Conference on Spoken Language Processing*.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, *PloS one* **13**(5): e0196391.

Makhoul, J. (1975). Linear prediction: A tutorial review, *Proceedings of the IEEE* **63**(4): 561–580.

Mower, E., Mataric, M. J. and Narayanan, S. (2010). A framework for automatic human emotion classification using emotion profiles, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(5): 1057–1070.

Ramirez, A. D. P., de la Rosa Vargas, J. I., Valdez, R. R. and Becerra, A. (2018). A comparative between mel frequency cepstral coefficients (mfcc) and inverse mel frequency cepstral coefficients (imfcc) features for an automatic bird species recognition system, *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, IEEE, pp. 1–4.

Sahidullah, M. and Saha, G. (2009). On the use of distributed dct in speaker identification, *2009 Annual IEEE India Conference*, IEEE, pp. 1–4.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms, *Speech communication* **40**(1-2): 227–256.

Schuller, B., Rigoll, G. and Lang, M. (2003). Hidden markov model-based speech emotion recognition, *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 2, IEEE, pp. II–1.

Shashidhar, G., Koolagudi, K. and Sreenivasa, R. (2012). Emotion recognition from speech: a review, *Springer Science+ Business Media* **15**: 99–117.

Tarunika, K., Pradeeba, R. and Aruna, P. (2018). Applying machine learning techniques for speech emotion recognition, *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, pp. 1–5.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods, *Speech communication* **48**(9): 1162–1181.

Wang, W., Zhang, Y., Li, Y. and Zhang, X. (2006). The global fuzzy c-means clustering algorithm, *2006 6th World Congress on Intelligent Control and Automation*, Vol. 1, IEEE, pp. 3604–3607.

Womack, B. D. and Hansen, J. H. (1999). N-channel hidden markov models for combined stressed speech classification and recognition, *IEEE Transactions on Speech and Audio Processing* **7**(6): 668–677.

Zeng, Z., Pantic, M., Roisman, G. I. and Huang, T. S. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE transactions on pattern analysis and machine intelligence* **31**(1): 39–58.

Zhou, G., Hansen, J. H. and Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress, *IEEE Transactions on speech and audio processing* **9**(3): 201–216.