Recognition of Speech Emotion in Call Centre Conversations in a Multilingual Environment

By

Kudakwashe Zvarevashe

(21752377)

Submitted in fulfilment of the requirements of the

Degree of Doctor of Philosophy

in the

Department of Information Technology

in the

Faculty of Accounting and Informatics

at the

Durban University of Technology

APRIL 2021

Declaration

I Kudakwashe Zvarevashe hereby declare that the work presented in this thesis has not been submitted for any other degree or professional qualification and that it is the result of my independent work. I further declare that all the sources of information used in this dissertation have been acknowledged and a list of references is provided.

08/04/2021

K. Zvarevashe Date

Approved for final submission

Promoter: _____ _____ 15/10/2021

Professor Oludayo O. Olugbara (PhD) Date

Publications associated with this research

Published conference papers

i. Zvarevashe, K. and Olugbara, O. O. 2018a. A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews. In: Proceedings of the *International Conference on Information Communications Technology and Society*: 1–4.

ii. Zvarevashe, K. and Olugbara, O. O. 2018b. Gender Voice Recognition Using Random Forest Recursive Feature Elimination with Gradient Boosting Machines. In: Proceedings of the *International Conference on Advances in Big Data, Computing and Data Communication Systems*:1–6.

Published journal articles

i. Zvarevashe, K. and Olugbara, O. O. 2020a. Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition. *Algorithms*, 70(13): 1–24.

ii. Zvarevashe, K. and Olugbara, O. O. 2020b. Recognition of Cross-Language Acoustic Emotional Valence Using Stacked Ensemble Learning. *Algorithms*, 13(10): 1–21.

iii. Zvarevashe, K. and Olugbara, O. O. 2020c. Recognition of Speech Emotion Using Custom 2D-convolution Neural Network Deep Learning Algorithm. *Intelligent Data Analysis*, 24(5): 1065–1086.

Journal articles under review

i. Zvarevashe, K. and Olugbara, O. O. 2021. Cross-language gender classification using DMLP feature extraction and majority voting ensemble learning.

Acknowledgements

“I will praise God’s name in song and glorify him with thanksgiving.” - Psalms 69:30.

I give thanks to God who has made it possible for me to complete this doctoral degree.

First and foremost, I would like to thank my supervisor, Prof. Oludayo O. Olugbara for his expert guidance throughout the study. This man saw the potential in me that no other person could see. He patiently brought the best out of me with his advice and constructive criticism. Thank you, Sir, I can never thank you enough.

I would like to thank my loving wife Chiedza, for her unwavering support. She encouraged me to keep on working especially when the experiments were becoming difficult to conduct. Thank you Mabhebheza, I love you so much.

I would like to express my deepest appreciation to my brothers Taonga, Zivanai, Runesu, and Simbarashe for their immense support throughout my studies. Thank you guys, for understanding my absence from crucial family gatherings.

I would like to thank my beloved aunt Tete Ephy Vhomo for continuously encouraging me to finish my studies. Sometimes I felt as if I was drowning in this academic Bermuda triangle, but she kept on motivating me.

I would like to thank my in-laws, the Kasinganetis for their unwavering support and prayers.

Baba naMai Nadine “Handigoni kukutendai zvakakwana.”. You went out of your way to do everything you could to help me. Thank you so much.

Finally, I would like to thank the DUT Information Technology department, the ICT Society research group for their professional support including my fellow postgraduate students in our safe house, the post-graduate laboratory. Thank you, guys.

Table of contents

Declaration ii

Publications associated with this research iii

Acknowledgements iv

Table of contents v

List of figures ix

List of Abbreviations xiii

Abstract xvii

Chapter 1: Introduction 1

1.1 Preamble 1

1.2 Speech Emotion Recognition 2

1.3 Challenges in Speech Emotion Recognition 3

1.4 Research Problem 4

1.5 Research Questions 6

1.6 Research Objectives 6

1.7 Methodology 7

1.8 Summary of contributions 8

1.9 Summary of Thesis 10

Chapter 2: Review of Human Emotion Models 12

2.1 Introduction 12

2.2 Emotion in Human Communication 12

2.2.1 Factors affecting emotion in communication 14

2.3 Theories of Emotion 17

2.4 Models of Emotion 19

2.4.1 Basic emotion model 19

2.4.2 Dimensional model 22

2.4.3 Componential appraisal model 23

2.5 Emotion Classification 24
2.6 Chapter Summary 30
Chapter 3: Processing of Human Speech Features 32
3.1 Introduction 32
3.2 Speech Production 33
3.2.1 Respiratory system 33
3.2.2 Phonation system 34
3.2.3 Resonance system 36
3.3 Expressive speech analysis 36
3.4 Speech features 38
3.4.1 Prosodic features 38
3.4.2 Spectral features 40
3.4.3 Spectrograms 50
Chapter 4: Speech Emotion Recognition Models 52
4.1 Introduction 52
4.2 Prosodic Features 52
4.3 Spectral Features 55
4.4 Spectrogram Features 58
4.5 Hybrid Features 61
4.6 Cross-Language Acoustic Emotional Valence 64
4.7 Chapter Summary 66
Chapter 5: Research Methodology 68
5.1 Introduction 68
5.2 Emotion Model 69
5.3 Databases 70
5.3.1 Berlin database of emotional speech (EMO-DB) 71
5.3.2 Ryerson audio-visual database of emotional speech and song (RAVDESS) 72
5.3.3 Surrey audio-visual expressed emotion (SAVEE) 73

5.3.4 Italian emotional speech dataset - EMOVO 74
5.3.5 Crowd-sourced emotional multimodal actor dataset (CREMA-D) 75
5.3.6 Combined Speech emotion corpus 76
5.4 Spectrogram Generation 77
5.5 Feature Extraction 79
5.6 Feature Selection 80
5.7 Ensemble Classifiers 88
5.7.1 AdaBoost classifier (ABC) 88
5.7.2 Random decision forest (RDF) 88
5.7.3 Extra trees classifier (ETC) 90
5.7.4 Bagging classifier (BC) 91
5.7.5 Gradient boosting machines (GBM) 92
5.7.6 RDF, AdaBoost, logistic regression, and gradient boosting machine (RALOG) 93
5.7.7 XGBoost (XGB) 95
5.8 Base Inducers 96
5.8.1 Logistic regression classifier (LRC) 96
5.8.2 Classification and regression trees (CART) 97
5.9 Deep Learning Classifiers 98
5.9.1 Deep radial basis function neural network 98
5.9.2 Deep multilayer perceptron neural network 100
5.9.3 2D-CNN neural network 101
5.10 Performance Metrics 103
5.10.1 Accuracy 104
5.10.2 Precision 104
5.10.3 Recall 105
5.10.4 F1-score 105

5.10.5 Receiver operating characteristics (ROC) curve and area under curve (AUC) 105
5.10.6 Processing time 106
5.11 Chapter Summary 106
Chapter 6: Experimental Results 107
6.1 Introduction 107
6.2 Experiment 1 Results and Discussion 108
6.3 Experiment 2 Results and Discussion 125
6.4 Experiment 3 End-to-End Deep Learning Models 138
6.5 Experiment 4: Multilingual Emotion Recognition 153
6.6 Chapter Summary 164
Chapter 7: Thesis Summary, Conclusion and Future work 165
7.1 Thesis Summary 165
7.2 Conclusion 167
7.3 Limitations and Future Work 168
References 169

List of figures

Figure 1:Illustration of the vocal folds system (Titze, 1994) 14
Figure 2: Plutchik's wheel of emotions (Plutchik, 2013). 20
Figure 3:Pictorial depiction of the expression of Basic Emotions (Livingstone and Russo, 2018) 21
Figure 4: Distinctive and dimensional emotion model (Hamann et al., 2002). 24
Figure 5: Plutchnik three-dimensional model of emotion concepts (Plutchik, 2013) 26
Figure 6: Emotions mapped in Activation-Evaluation space with FEELTRACE (Cowie et al., 2000) 29
Figure 7: The Hierarchical Structure of Emotions(Shaver, Murdaya and Fraley, 2001) 29
Figure 8: The respiratory system (Ackermann, 2008) 34
Figure 9:The phonation system (Guenther, 2006) 34
Figure 10: MFCC Algorithm block diagram (Yu, Li and Fang, 2012) 43
Figure 11: Source filter model of LPCC speech production (Chia Ai et al., 2012) 45
Figure 12: Sample waveform (Ariav and Cohen, 2019) 51
Figure 13:Sample Spectrogram (Ariav and Cohen, 2019) 51
Figure 14: The architecture of the proposed research methods 69
Figure 15: Berlin Database of Emotional Speech corpus 71
Figure 16: Ryerson Audio-Visual Database of Emotional Speech (RAVDESS) corpus 72
Figure 17: Surrey Audio-Visual Expressed Emotion (SAVEE) corpus 73
Figure 18: EMOVO corpus 74
Figure 19: CREMA-D corpus 75
Figure 20: Combined speech emotion corpus 76
Figure 21: EMO-DB spectrograms showing four emotion states. 77
Figure 22: RAVDESS spectrograms 77
Figure 23: SAVEE spectrograms 78
Figure 24: jaudio feature extraction 80
Figure 25:HAF selection process flow diagram 82
Figure 26: Flowchart for the creation of cross-corpus features 84
Figure 27: Workflow of Random Forest Recursive Feature Elimination (Zvarevashe and Olugbara, 2018b) 85
Figure 28: Summarised RF-RFE algorithm 86

Figure 29: RF-RFE ranked Features 86
Figure 30: Summarised Random Forest Algorithm (Nguyen, Huang and Nguyen, 2015) 90
Figure 31:Summarised GBMs algorithm (Friedman, 2002) 93
Figure 32: RALOG Architecture 94
Figure 33: RALOG ensemble implementation 95
Figure 34: RBF neural network architecture (Wu et al., 2012) 99
Figure 35: Basic architecture of the custom 2D - CNN in processing spectrogram images 103
Figure 36: Graphical representation of Percentage Accuracy of Classifiers Trained using MFCC1, MFCC2 and HAF on Ravdess 112
Figure 37: Graphical representation of Percentage Accuracy of Classifiers Trained using MFCC1, MFCC2 and HAF on Savee 113
Figure 38: DRBFNN validation accuracy on EMODB. 140
Figure 39: DRBFNN validation accuracy on RAVDESS. 140
Figure 40: DRBFNN validation accuracy on SAVEE. 141
Figure 41: DMLP validation accuracy on EMODB. 141
Figure 42: DMLP validation accuracy on RAVDESS. 142
Figure 43: DMLP validation accuracy on SAVEE. 142
Figure 44: 2D-CNN validation accuracy on EMODB. 143
Figure 45: 2D-CNN validation accuracy on RAVDESS. 143
Figure 46: 2D-CNN validation accuracy on SAVEE. 144
Figure 47: RBFNN test loss on EMODB. 144
Figure 48: DRBFNN test loss on RAVDESS. 145
Figure 49: DRBFNN test loss on SAVEE. 145
Figure 50: DMLP test loss on EMODB. 146
Figure 51: DMLP test loss on RAVDESS. 146
Figure 52:DMLP test loss on SAVEE. 147
Figure 53: 2D-CNN test loss on EMODB. 147
Figure 54: 2D-CNN test loss on RAVDESS. 148
Figure 55: 2D-CNN test loss on SAVEE. 148
Figure 56: ROC Curve before RFE. 160
Figure 57: ROC Curve after RFE. 160

List of tables

Table 1: Plutchik's Emotion states.....22
Table 2: Turner and Ortony's Classification of basic emotions (Ortony et al., 1990).....25
Table 3:MindReader categories of emotion (Jauk, Bonafonte and Pascual, 2016).....30
Table 4: Mappings of emotions into Positive and Negative Valence.....76
Table 5: A summary of HAF features.....83
Table 6: Description of the top 20 acoustic features.....87
Table 7: ETC algorithm (Geurts, Ernst and Wehenkel, 2006).....91
Table 8: The detailed structure of the proposed custom 2D-CNN.....102
Table 9: Processing time of classifiers on MFCC1, MFCC2 and HAF.....108
Table 10: Average accuracy, recall, F1-score and precision with confidence intervals of classifiers trained with MFCC1, MFCC2 and HAF.....109
Table 11: Accuracy, Recall and Score and Precision on RAVDESS MFCC1.....113
Table 12: Accuracy, Recall and Score and Precision on RAVDESS MFCC2.....115
Table 13: Average Accuracy, Recall and Score and Precision on RAVDESS HAF.....116
Table 14: Average Accuracy, Recall and Score and Precision on SAVEE MFCC1.....118
Table 15: Accuracy, F1-Score, Recall and Precision on SAVEE MFCC2.....120
Table 16: Average Accuracy, Recall and Score and Precision on SAVEE HAF.....121
Table 17: Comparison of the proposed approach with potential works from the literature.....123
Table 18: Processing time of classifiers trained with Deep Radial Basis Neural Network Auto-generated features (DRBFNN-AGF), Deep Multilayer Perceptron Auto-generated features (DMLP-AGF) and a 2D Convolution Neural Network Auto-generated features (2D-CNN-AGF).....126
Table 19: Average accuracy, recall, F1-score and precision with confidence intervals of classifiers trained with DRBFNN-AGF, DMLP-AGF and 2D-CNN-AGF.....127
Table 20: Average Accuracy, Recall and F1-Score and Precision on RAVDESS DRBFNN-AGF.....128
Table 21: Average Accuracy, Recall and F1-Score and Precision on RAVDESS DMLP-AGF.....130
Table 22: Average Accuracy, Recall, F1-Score and Precision on RAVDESS 2D-CNN-AGF.....131
Table 23: Average Accuracy, Recall, F1-Score and Precision on SAVEE DRBFNN-AGF.....133

Table 24: Average Accuracy, Recall, F1-Score and Precision on SAVEE DMLP-AGF 134
Table 25: Average Accuracy, Recall, F1-Score and Precision on SAVEE 2D-CNN-AGF 136
Table 26: Training Time. 139
Table 27: Accuracy analysis on speech emotion spectrograms. 139
Table 28: Validation loss analysis on speech emotion spectrograms. 139
Table 29: F1-Score, Recall and Precision Analysis on Ravdess Spectrograms. 149
Table 30: F1-Score, Recall and Precision Analysis on Savee Spectrograms. 150
Table 31: F1-Score, Recall and Precision Analysis on Emodb Spectrograms. 151
Table 32: Comparison of the proposed approach with related work from the literature. 151
Table 33:Mappings of Adult Speech Emotions into Positive and Negative Valence Emotions. 155
Table 34: Processing Time. 156
Table 35*:* Average accuracy, recall, F1-score and precision with confidence intervals of 158
Table 36: Percentage precision, recall and F1-score and accuracy at each experimental stage of feature scaling with the RF-RFE algorithm. 159
Table 37: Comparison of the proposed approach with potential works from the literature. 160
Table 38**:** Percentage Precision, Recall and F1-score and Accuracy of RALOG on Individual Corpus before and after Feature Scaling with the RF-RFE Algorithm. 163
Table 39: Training Time (ms) of RALOG Individual Corpora. 163

List of Abbreviations

2D-CNN	Two-dimensional Convolution Neural Network
ABC	Airplane Behaviours Corpus
ABC	AdaBoost Classifier
A-DAE	Adaptive Denoising Autoencoders
AGF	Auto-generated Features
AUC	Area Under Curve
BC	Bagging Classifier
BEL	Brain Emotional Learning
BGRUs	Bidirectional Gated Recurrent Units
CART	Classification and Regression Tree
CASIA	Chinese emotional database
CNN	Convolution Neural Network
CREMA-D	Crowd-sourced Emotional Multimodal Acted Dataset
CWT	Continuous Wavelet Transform
DA	Domain Adaptation
DALSR	Domain-adaptive least squares regression
DBM	Deep Belief Networks
DSCNN	Deep Stride Convolutional Neural Network
ECC	Energy Cepstral Coefficient
EDA	Estimation of Distribution Algorithm
EDFLM	Emotion-discriminative and Domain-invariant Feature Learning Method
EESDB	Chinese elderly emotional database
eGeMAPS	Geneva Minimalistic Acoustic Parameter Set

ELM	Extreme Learning Machine
EMODB	Berlin Database of Emotional Speech
EPST	Emotional. Prosody Speech and Transcripts
ET	Extra Trees Classifier
FAU-Aibo	Friedrich-Alexander-Universität – Artificial intelligence bot
FDA	Functional Data Analysis
FECC	Frequency weighted Energy Cepstral Coefficient
FS	Forward Selection
GBM	Gradient Boosting Machines
GCZCMT Operator	Glottal Compensation to Zero Crossings with Maximal Teager Energy
GCZCMT	Glottal Compensation to Zero Crossings with Maximal Teager
GMM	Gaussian mixture model
HAF	Hybrid Acoustic Features
HTK	Hidden Markov Model Toolkit
HuWSF	Hu moments based Weighted Spectral Features
IEMOCAP	Interactive Emotional Dyadic Motion Capture
KNN	K-nearest neighbor
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
LLD	Low-Level Descriptors
LLD	Low-Level Descriptors
LPCC	Linear Prediction Cepstral Coefficients
LR	Logistic Regression
LSTM	Long Short-Term Memory

MCB	Multimodal Compact Bilinear Pooling
MCFS	Multi-Cluster Feature Selection
MFCC	Mel Frequency Cepstral Coefficients
MFF	Modulation Frequency Feature
MLP	Multi-Layer Perceptron
MS	Modulation Spectral
MS	Modulation Spectral
MSER	Maximally Stable Extremal Region
OFS	Orthogonal Forward Selection
PCA	Principal Component Analysis
PFS	Promising First Selection
PLPC	Perceptual Linear Prediction Coefficients
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech
RBF	Radial Basis Function
RDF	Random Forest Classifier
RECOLA	Remote Collaborative and Affective Interactions
ResNet	Residual Network
RF-RFE	Random Forest Recursive Feature Elimination
RNN	Recurrent Neural Network
SAVEE	Surrey Audio-Visual Expressed Emotion
SBC	Sub-band based Cepstral Parameter
SER	Speech Emotion Recognition
SES	Sahand Emotional Speech
SMFCC	Signal based on Mel Frequency Cepstral Coefficient

SMH	Somatic-Marker Hypothesis
SOM	Self Organising maps
STFT	Short-Time Fourier Transform
SUSAS	Speech Under Simulated and Actual Stress
SVM	Support Vector Machines
UAR	Unweighted Average Accuracy
UAR	Unweighted Average Recall
XGB	Extra Gradient Boosting Machines
ZCR	Zero Crossing Rates

Abstract

The use of customer call centres has increased exponentially in the modern business world and is the heart of marketing in the customer services industry. Previous studies have shown that the quality of services that customers receive from the call centres paint a picture of how they view the company. Reliance on the use of suggestion boxes to crowdsource customer views on call centre services is not adequate and at times, may not give a correct record about the services in question. Therefore, speech emotion recognition has been applied in customer call centres as a tool for evaluating customer service perception, emotion, and sentiment. This approach presents several advantages, for instance, the performance of call centre agents can adequately be scrutinised because their emotions can be automatically classified based on machine learning methods for emotion recognition. In recent times, various techniques and methods have been used to develop robust speech emotion recognition systems for customer call centres, but the primary problem associated with these novel applications is that most of them do not perform well in multilingual environments. In addition, most of the proposed models do not properly recognise the fear archetype of emotion. The effectiveness of a speech emotion recognition system depends largely on the strength of the features used. Consequently, the purpose of this research was to discover the most efficacious features in recognising speech emotion in call centre conversations. Therefore, this thesis reports on the development of hybrid acoustic features based on spectral and prosodic descriptors. The set of hybrid features proposed in this study comprises the logarithm of energy, fundamental frequency, zero-crossing rate, spectral roll-off point, spectral flux, spectral centroid, spectral compactness, spectral variability, fast Fourier transform, Mel frequency cepstral coefficients, and linear prediction cepstral coefficients. Furthermore, this thesis reports on the development of a novel stacked ensemble machine learning algorithm based on a combination of inducers and ensemble classifiers. The discovery of effective speech emotion features and the development of an efficient machine learning algorithm are essential stages of effective speech emotion recognition in call centre conversations. The verification and validation of the proposed speech emotion recognition methods based on feature extraction and feature classification for applications in call centre conversions were done using a series of experiments. This was accomplished by testing the crafted hybrid acoustic features on five distinct speech emotion databases. The acoustic features were evaluated against deep learning auto-generated features and a hybrid of popular acoustic features. In addition, a set of four ensemble algorithms were evaluated against the newly invented stacked ensemble algorithm. The performance of the developed stacked ensemble algorithm in this study was analysed based on the widely used statistical evaluation metrics of accuracy, precision, F-score, area under the receiver operating characteristic curve and computation time. The results have indeed demonstrated that the newly developed stacked ensemble algorithm coupled with the crafted hybrid acoustic features have consistently performed better than many other state-of-the-art algorithms and speech features across various standard speech corpora.

Chapter 1: Introduction

1.1 Preamble

Call centres across the globe are growing exponentially (Asadullah *et al.* 2019) and have become a fundamental part of the marketing and customer service strategies for most companies (Mahesh and Kasturi 2006). They function as a valuable foundation of service recovery; value addition; market intelligence and strategy benefit (Hudson, González and Rychalski 2017). Analysing customer call conversations in a call centre environment is of paramount importance in the business world of today and is a challenge (Yu *et al.* 2017). Research has revealed that the quality of services that customers get from call centre agents, determines the way they perceive a company (Greasley and Smith 2017). Some companies review the quality of their services by analysing customer sentiments using textual data (Zvarevashe and Olugbara 2018a). However, further steps can be taken by analysing the actual conversations. Companies need to review their recorded calls to understand affective behaviour of their customers and assess the way agents conduct themselves during the calls which at the end of the day will facilitate seamless decision making (Sahadev, Purani and Panda 2017).

One of the most popular, but orthodox approaches to do this, is to employ human experts who will listen to the recorded calls, analyse them and then generate reports of the corresponding behavioural descriptions. Pandharipande and Kopparapu proposed an approach to identify problematic call centre conversations to properly manage call centres (Pandharipande and Kopparapu 2012). However, this is a challenging and tiresome task when the volume of recorded calls is high as it poses a huge burden on human experts (Chakraborty, Pandharipande and Kopparapu 2016). Therefore, it is against this background that this research focused on the development of a speech emotion recognition (SER) model for customer call centres as a way of resolving the hiatus between the caller and expert responding to the calls.

1.2 Speech Emotion Recognition

Speech emotion recognition (SER) is the study of the development and variation of the emotional state of a speaker from vocal utterances (Prasada , Chandra and Hemanth 2019). Over the past decades, SER has been used to understand human emotional behaviour. One of the critical research issues in SER is the extraction of discriminative, affect-salient features from speech signals (Abo, Deriche and Mohandes 2018). The desire to create emotionally intelligent human-computer interfaces has inspired several developments in SER. It is imperative to have a deeper understanding of messages, but equally important to fully appreciate how they are conveyed (Ariav and Cohen 2019).

Kim and Park (Kim and Park 2016) suggested that for listeners to get a deeper understanding of a conversation, they need to appreciate the state of emotion of the speakers involved in the conversation. Emotions can be conveyed in several ways during a conversation and are not only crucial to human reasoning but are central to social regulation that controls the flow of dialogues (Tashev, Wang and Godin 2017). Apart from spoken words, humans express emotions in two different ways, which are modulation of facial expression (Chelali and Djeradi 2015) and modulation of the tone of a voice (Rathor and Jadon 2017). In a nutshell, human beings can express emotions through verbal, non-verbal and written communications.

During a conversation in verbal communication, words, stress, tone, phrases and enunciation play a pivotal role in shaping the meaning of spoken words and in determining how information is perceived (Badshah *et al.* 2019). This form of communication can be done in person or through various forms of communication media such as telephones, television, radio and many more. Messages conveyed in this type of communication are also referred to as explicit messages. Some elements complement verbal communication that includes general demeanour, gestures, facial expressions, clothing, and body language. These elements are the embodiment of what constitutes non-verbal communication. Messages conveyed in this type of communication are also referred to as implicit. Written communication encompasses all forms of textual communication, including blogs, text messages, emails, and the like.

1.3 Challenges in Speech Emotion Recognition

SER as an affective computing discipline has got many intrinsic challenges that have inspired lots of research over the past decades. The task of SER is generally very challenging because of several reasons, some of which are the following:

i. There are lots of speech features that are critical for distinguishing between emotions, but information regarding the most useful set of features is still vague. The presence of diverse speaking styles, speaking rates, and sentences induces acoustic inconsistencies that create inherent problems because these properties affect most of the standard features extracted (Guidi *et al.* 2019; Kwon 2020).

ii. The way an individual expresses his/her emotion is heavily dependent on the speaker, environment and culture (Alshamsi *et al.* 2019). The way emotions are expressed and interpreted varies significantly with culture, which makes it particularly difficult to develop universal SER models (Hechanova 2013). There are significant differences in the way male and female persons express emotions. This requires the development of gender-sensitive SER, which is a challenging task (Goldshmidt and Weller 2000; Kotti and Kotropoulos 2008; Skuk and Schweinberger 2013).

iii. An individual may undergo a particular emotional state for some time, which may be days, weeks or even months. This situation may suppress other emotions that may last for a couple of minutes. As a result, it may not be clear to understand the exact emotion that may be identified through SER systems (Hifny and Al 2019; Akçay and Oğuz 2020).

iv. SER depends upon the situation in which states of emotions are expressed. The various available speech emotion corpora were developed using speech from different contextual environments. Therefore, there are several types of corpora used in SER and these are induced, actor-based and natural speech corpora. Speech files in the induced emotion corpora may contain artificial emotional contents, especially if the speakers knew that they were being recorded obtrusively (Freksa *et al.* 2008). Actor-based corpora normally constitute audio files from an acted speech of amateur and professional actors (Freksa *et al.* 2008). Induced corpora are developed using audio files from emotion-induced speech (Oliver 2013). Emotion can be induced or elicited using various psychological methods

such as classical and physical induction to stimulate the required emotions (Koolagudi and Rao 2012; Oliver 2013). The actor-based emotion corpora are subjective because some actors tend to exaggerate certain emotional expressions (Cong *et al.* 2017). In that same vein, emotions that are not original are not a true reflection of everyday emotion because they are episodic (Koolagudi and Rao 2012). Furthermore, actor-based emotion corpora are usually developed using a television show material, which makes it difficult to get the extracts because production companies guard these jealously (Oliver 2013).

v. Natural speech corpora comprise audio speech files recorded from real conversations without any elicitation (Zhang *et al.* 2011; Narendra *et al.* 2019). The most realistic emotions are expressed naturally, and these may be considered ideal for SER. However, natural speech corpora are difficult to annotate, which makes it subjective to use them in SER research (Koolagudi and Rao 2012). In addition, natural speech corpora are difficult to access because of proprietary issues (Koolagudi and Rao 2012). For instance, a natural audio corpus developed by Roach was never made accessible outside Roach's group because of broadcasting ownership issues (Zhang *et al.* 2011; Narendra *et al.* 2019). The other problem associated with natural speech corpora is that sometimes it may be difficult to get a targeted emotion from the speakers. The audio speech files in natural corpora are difficult to model because they are pervasive (Koolagudi and Rao 2012; Oliver 2013). The overlapping of utterances in natural speech corpora makes it difficult to process the files for SER (Neustein 2013; Jauk, Bonafonte and Pascual 2016).

1.4 Research Problem

The development of intelligent systems to recognise emotion has become extremely important in the current business environment (Cong *et al.* 2017). The use of appropriate features in recognising emotion through speech has been a challenge for a long time. Five dimensions have been proposed as solutions to address this problem, and these are the linear predication cepstral coefficient (LPCC) approach (Chamoli, Semwal and Saikia 2017; Langari, Marvi and Zahedi 2020), prosodic approach, Mel frequency cepstral coefficient (MFCC) approach (Kerkeni *et al.* 2018; Aouani and Ayed 2020), hybrid approach (Lee *et al.* 2018; Noh *et al*. 2021; Byun and Lee 2021) and deep

learning approach (Abbaschian, Sierra-Sosa and Elmaghraby 2021; Farooq *et al*. 2020) c). Several suggestions have been proposed in the literature and evaluated using these paradigms. Several researchers have proposed to develop robust SER systems using a variety of hybrid approaches. These approaches include the use of the multimodal paradigm, which entails combining speech and text features, combining speech and facial features, combining text, speech and facial features (Hossain and Muhammad 2019). The other archetype of the hybrid approach can be implemented using the acoustic feature set paradigm, and this entails the fusion of LPCC and MFCC features (Zhao, Ye and Wang 2018).

The main problem with the hybrid angle of approach is that it has not been able to report high precision and accuracy rates (Dahake, Shaw and Malathi 2016; Abbaschian, Sierra-Sosa, and Elmaghraby 2021; Mustaqeem and Kwon 2020; Lee, Han and Ko 2020), especially for fear emotion. This problem makes it difficult to apply such features when developing SER systems that would fit environments such as customer call centres. For example, it would be disastrous if a model that records low accuracy rates is used to develop an SER for police customer call centres. Moreover, the proposed hybrid approaches such as the one that uses text are still not suitable for customer call centre environments because these environments usually receive calls from customers who speak in numerous languages such that it will be difficult to transcribe every language used (Jauk, Bonafonte and Pascual 2016). Another problem associated with using a multimodal paradigm that makes use of facial images is that most call centres do not use visual features, and this would require computationally expensive resources.

Research has shown that most SER systems perform poorly when tested on corpora that are different from the ones used for training (Kim *et al.* 2017). The prime reason being that most speech emotion audio files are recorded under different recording environments such as noisy and calm environments (Schuller *et al.* 2015). Moreover, the quality of equipment (Gideon, Provost and McInnis 2016) used in recording the vocal utterances may be different and the gender and age group of speakers may compound the matter (Anagnostopoulos, Iliou and Giannoukos 2012). In addition, there are several forms of emotion corpora, which are acted, elicited and natural that present several characteristics (Akçay and Oğuz 2020), hence the elicitation techniques used in creating SER corpora affects the processes involved (Kim *et al.* 2017). These are the issues that

significantly influence the performance of a speech recognition model when evaluated on a different corpus that poses a huge problem for SER.

The solution to the problems highlighted above boils down to the extraction of the most appropriate features because the power of a recognition model lies in the strength of the features used (Guidi *et al.* 2019). Therefore, developing efficient models for SER lies in the features used to recognise emotions. This understanding is crucial in the development of computationally inexpensive SER models that are both time-sensitive and accurate. However, there is a dearth of literature studies on the investigations of computationally inexpensive and time-sensitive speech recognition models with regards to the hybridisation of spectral and prosodic features. The moment deep learning is introduced in such discussions, either the processing time jumps through the roof or presents the need for more computational power arises (Zataraín *et al.* 2018). The problems discussed above have inspired the formulation of the research question presented in the next section.

1.5 Research Questions

SER has been for many years challenging and is still a challenging task in speech processing (Galanis *et al.* 2013; Kwon 2020; Lee *et al*. 2020; Zacarias *et al*. 2021). This study seeks to provide appropriate ripostes to the following research questions arising from the research problems.

i. *What set of effective speech features can be extracted from a customer to enhance the recognition of c*ross-language *emotional conversations?*
ii. *What efficient machine learning algorithm can be developed utilising the extracted speech features to improve the performance of a recognition system for c*ross-language *emotional conversations?*

1.6 Research Objectives

The overarching aim of this study was to develop a speech emotion model that can recognise customer emotions using highly discriminative features with the ultimate goal of predicting emotions with high accuracy. The researcher postulates that it should be possible to develop the desired model using the most discriminative features to demonstrate that developing a

computationally inexpensive SER system with high performance is achievable. Therefore, to realise the aim of this research work, the following research objectives have been set:

i. To discover a set of acoustic features that can help to improve the recognition performance of a speech emotion system for cross-language emotional conversations.
ii. To investigate whether the discovered acoustic features can give an improved performance in a cross-language emotional conversation system when compared to the auto-generated features by the deep learning method.
iii. To develop an efficient algorithm that will give an improved recognition performance for the discovered acoustic features for cross-language emotional conversations.

1.7 Methodology

The primary material for this study is five speech emotion corpora that were used to conduct a series of experiments to test the performance of the developed SER model. These corpora are the Berlin Database of Emotional Speech (EMO-DB) (Burkhardt *et al.* 2005), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018), Surrey Audio-Visual Expressed Emotion (Haq and Jackson 2009), Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) (Sierra *et al.* 2015) and EMOVO (Costantini *et al.* 2014). The primary reason for choosing these emotion corpora is that they constitute speech files from different languages, cultures, and accents, which are appropriate for solving the research question of this study.

The methods used to fulfil the objectives mentioned above include feature extraction, feature selection, spectrogram generation and emotion recognition. Firstly, a set of spectral and prosodic features were extracted from the emotion corpora using jAudio, which is a widely used speech feature extraction tool. These features were extracted into comma-separated values (CSV) files together with the corresponding emotional classes. Secondly, a set of highly discriminating features was selected from the pool of extracted features using the knowledge from the discoveries found in similar studies (Luengo, Navas and Hernaez 2010; Ram, Palo and Mohanty 2013; Khan and Roy 2017; Semwal, Kumar and Narayanan 2017; Alshamsi *et al.* 2019; Kerkeni *et al*. 2019).

Therefore, this study developed a set of hybrid acoustic features (HAF) was developed (Zvarevashe and Olugbara 2020a) using a carefully crafted combination of prosodic and spectral features. Thirdly, emotions from the selected corpora were classified using ensemble and HAF features (Zvarevashe and Olugbara 2020a). In this thesis, further experiments have been performed using the classifiers such as classification and regression trees, gradient boosting machines and random forest. Ensembles were used in this experimental study because they have been successful in other areas such as product image classification (Oyewole and Olugbara 2018) and lung cancer prediction (Adetiba and Olugbara 2015b). The audio files from the chosen corpora were then converted to spectrograms to extract auto-generated features using deep classifiers and Librosa. Librosa is a python application package that performs audio processing tasks like spectrogram generation. The classification of emotions based on machine learning methods was then performed using these auto-generated features with ensemble classifiers. This method has made it easy to compare the strength of HAF features against deep learning autogenerated features (DLAGF) in recognising emotions.

1.8 Summary of contributions

This thesis has successfully investigated the most appropriate speech features in recognising speech emotion and as a result, the following contributions were made during the process. A lot has been done in the quest to identify speech features that have the most discriminative power but the reported matrix scores can still be improved. Therefore, the contributions of this work to emotion recognition research are succinctly summarized as follows.

a) From the literature review, it was noted that most researchers perform an extra step in the form of feature selection. The problem with this structure is that it increases the processing time of the designed model. This thesis argues that this extra step can be avoided through the crafting of hybrid features. Using this design philosophy, the hybrid features noted can be extracted directly and these can be used in the SER model without performing feature selection which increases the processing time. Therefore, this thesis has identified a set of highly discriminating acoustic features with the most predictive power for recognizing

human speech emotion (Zvarevashe and Olugbara 2020a). The widely used approaches in literature concentrated on the application of MFCC features. However, combining all prosodic with MFCC features can increase processing power and reduce recognition performance because of the curse of high dimensionality. The method based on the novel hybrid acoustic features of this study agglutinates the features of speech energy, pitch, times, cepstral, shape, amplitude, moments, signature, envelope, audio, and frequency to improve the performance of the emotion recognition system when compared to the orthodox speech features previously reported in the literature. The statistical analysis of the capability of various feature parameters to recognise different emotional states was conducted in this study. The traditional speech features were compared with the proposed acoustic features by a statistical approach to demonstrate the effectiveness of the proposed speech features, which is an important contribution. Another significant contribution in this regard is that these Hybrid features are consistent in different languages. The Hybrid features were evaluated in utterances spoken in German, British English and Northern American English and the performance was constantly high across various performance benchmarks such as precision, recall, accuracy, f1-score, and computation time.

b) The identification of an efficient machine learning method based on ensemble learning of hybrid acoustic features proposed that can recognise emotion valence in a simulated multilingual environment is an important contribution of this study (Zvarevashe and Olugbara 2020b). Besides the language factor, the ensemble learning algorithm combined with HAF performed well on a combined corpus that had a variety of audio speech files from different environments, accents and was recorded using different equipment. This method can be successfully implemented in a call centre environment to recognize human emotion, valence and sentiment.

c) The development of a custom two-dimensional convolutional neural network (2-CNN) to improve the performance of cross-language emotion recognition (Zvarevashe and Olugbara 2020c) is an important contribution of this study. This is a two-dimensional convolution neural network that was designed to develop an SER model that can recognise even the most difficult emotions such as the fear emotion. Evidence in the literature shows

that as classes increase, the recognition accuracy decreases. The 2D-CNN proposed uses spectrograms as input to extract features and classify emotion. Approaches in the literature have low accuracy due to the different optimisers, the number of convolution layers, and many more. The proposed 2D-CNN reduced the number of layers and optimised the process using optimal Adam features. This model showed that with adequate computational resources, it could be the ideal model for SER if the data is enormous because this experimental study observed that it learns continuously with each run.

1.9 Summary of Thesis

The summary of this thesis is presented to increase the comprehension of the content. SER is a field that cuts across various disciplines which are computer science and psychology and is comprehensively described in chapter 1. Therefore, this chapter provides a background of emotion from a psychological point of view. Emotion is defined here as well as the factors that affect the expression of emotion across various societies and cultures. The chapter also presents various models of emotion and the present classification theories of emotions from a psychological point of view. This chapter concludes by nominating the most appropriate emotion models that can be applied in the development of a customer call centre SER model, giving reasons why.

Chapter three covers three main areas which are speech production, speech features, and expressive speech analysis. The chapter is heralded by an in-depth description of the production of speech from a physiological point of view. The purpose of the speech production section is to establish the link between speech production and speech features. This is then followed by a review of the features used in developing SER models followed by expressive speech analysis. The review done in this chapter inspired the design that was used to develop the SER model.

An analysis of the techniques that have been used to recognise speech emotion is presented in chapter four. The chapter consists of four main sections which are prosodic feature review, spectral feature review, hybrid feature review, spectrograms review and cross-language acoustic feature review. The findings in this chapter informed the research design and methods used in chapter five.

In chapter five, the thesis moves from discussing the various techniques used in the review to the actual design and methods used in the development of a customer call centre SER model. In this chapter, various intriguing components of the proposed model are discussed. The design choice presented in this chapter provides a profound understanding of the tools appropriate for the design of a computationally inexpensive customer call centre SER model.

Chapter six reports on an extensive experimental validation of the proposed SER model. The crucial purpose of this chapter is to determine and validate optimal features suitable to develop a computationally inexpensive customer call centre SER model. The results obtained in this chapter are discussed and interpreted in this chapter. A critical analysis of the results is also done in this chapter to discover some of the useful hidden facts which are essential in developing SER models.

Chapter seven is the final chapter, and it gives a concise conclusive description of the main contributions drawn from this research endeavour. It also discusses some efficient methods of implementation and limitations of the study. The chapter concludes by suggesting future directions for this and related research.

Chapter 2: Review of Human Emotion Models

2.1 Introduction

Speech emotion recognition is interdisciplinary research that requires a good understanding of related emotional psychology and computer science issues (Yoon *et al.* 2019; Petridis *et al.* 2018; Prasada, Chandra and Hemanth 2019). Therefore, to demystify the related emotional psychological issues, this chapter presents a background of what emotion is including the theories and models of emotion from a psychological point of view. Factors affecting emotions are also discussed in this chapter to paint a clear picture of how emotions are generated, expressed, and interpreted across various cultures. In addition, this chapter presents various emotion models of classification theories. These theories are presented in this chapter because they form the basis of emotional classes that are used in developing the proposed SER model.

As highlighted in the previous chapter, SER-based research is challenging because of the lack of a universal definition of emotion and cultural differences (Lee and Narayanan 2005). The definition of emotion acts as a foundation of the classification of emotions which informs the recognition used in developing SER systems. Moreover, research has shown that the way emotions are expressed and interpreted across the globe is different (Barrett, Lewis and Haviland 2016). These are some of the issues that inspired the crafting of this chapter in this body of work.

In section 2.2, various definitions of emotion are given together with the multiple factors affecting emotion in communication. Section 2.3 presents a description of the theories of emotion while the models of emotion are described in section 2.4. The classifications of emotions are described in 2.5, and the chapter is concluded with a summary in section 2.6.

2.2 Emotion in Human Communication

Emotion plays a crucial role in our day to day lives. However, it is challenging to define emotion because it has no single clear-cut definition. 28 meanings of emotion were examined in (Plutchik

2013) to understand and describe what is emotion. It was reported that there was little consistency among the definitions and that many of them were not precise in describing emotion. Furthermore, in (Richins 1997) a clarification was made on how a few authors have endeavoured to enhance their comprehension of emotion by indicating its characteristics. Emotion is sometimes defined as a subjective, short-duration state of mind that is loosely described as complicated feelings that cause physical and psychological changes that influence thought and behaviour (Russell 1980; Lisetti 1998). From a psychological perspective, emotion is defined as a complex condition of feelings that speak to the physical and psychological changes that can influence individuals' thoughts and conduct (Ekman 1992). Typically, an emotion paints a picture of the affective state that is usually a response to something that would have happened. For example, we may feel sad or angry when we are betrayed, fear when we are in danger and be surprised when an unexpected event occurs. According to research, emotion is expressed with intent and is consciously experienced (Pappas, Androutsopoulos and Papageorgiou 2016). Typically, emotion can be expressed through speech, facial expressions and hand gestures (Muthusamy, Polat and Yaacob 2015).

The act of expressing emotion is a crucial part of human communication and the ability to recognise emotions is an indispensable ingredient for successful interactions. As human beings, we have inbuilt traits that help us to be emotionally geared up during conversations. These traits can be seen in newly born babies as they imitate facial expressions (Urwin 2008). According to Ekman (Ekman 1992), some basic emotions are inborn; for example, blind babies who have no idea how the human face looks like can still put on smiles on their faces. Emotions are expressed and perceived differently in different people from one culture to another (Scollon *et al.* 2004). Some research findings have given suggestions that having too little emotion is more likely to impair decision making (Spence 1995; Damasio *et al.* 2006). In these findings, Damasio developed a Somatic-Marker Hypothesis (SMH) which suggests that emotional processes guide behaviour towards the making of decisions (Damasio *et al.* 2006). We as human beings, do not have a window to look into the inner emotions of our fellow human beings (Dew *et al.* 2014) (Dolcos, LaBarr and Cabeza 2005).

2.2.1 Factors affecting emotion in communication

Research has shown that the way human beings express emotions across the globe is different (Barrett, Lewis and Haviland 2016). This disparity is caused by several factors which include physiological, contextual, personality, emotional management methods, perceptual abilities, cognitive abilities, gender, age, social and cultural perceptions.

The human physiological built describes the physical features that human beings have and these have a huge impact on the way emotions are expressed (Bahrick *et al.* 2019). For example, individuals with stronger cheek muscles or wider smiles appear to smile more than individuals with different features (Johnson, Waugh and Fredrickson 2010). Variations in the sizes of vocal folds are also a factor that presents differences between individuals of both the same and different genders (Banse and Scherer 1996). Adult males usually have larger folds while females have much shorter vocal folds as shown in Figure 1. This is the reason why most adult men's voices are lower-pitched while most adult women's voices have a high pitch. Furthermore, people of the same sex may also have different voices because of genetics (Bailey, Nowicki and Wickline 2009). Variations in the vocal tract also contribute to the difference in voices (Bachorowski and Bachorowski 2010). Adult men have a much wider vocal tract and that is the reason why their voices have lower sounding timbres. These features have an impact in the way emotions are expressed. Changes in chemical balances and neurological functioning also affect the way emotions are expressed (Grossberg and Gutowski 1987). For instance, diseases such as Alzheimer's disease tend to cause muscular weaknesses in the vocal tract which may affect the way people express emotions (Bucks and Radford 2004).

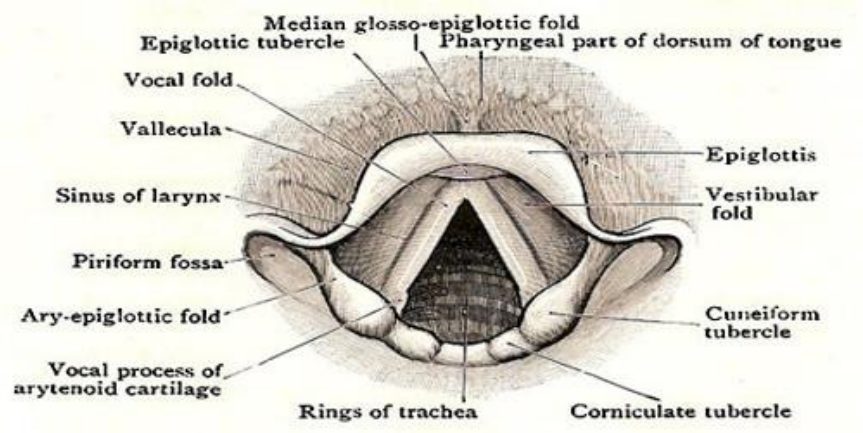


Figure 1:Illustration of the vocal folds system (Titze, 1994)

Generally, it is much easier to interpret the emotions of people we know (O'Connor 2008). For example, it is easy for a wife to know the emotional state of her husband over a telephone call. This usually happens because the wife will know the usual usage of her husband's expressions (Birkley and Eckhardt 2015). Some people always put a smile on their faces and sometimes these people need to put in extra effort when expressing their emotions in different contexts (Crivelli *et al.* 2016). For example, their smiles will have to be wider to make them significant compared to those who always wear sad faces. Mood swings also affect the way emotions are perceived and expressed. These are examples of temporary changes that affect. Contextual differences have a huge effect (Vrana and Rollock 2002). They can change the way emotional expressions are interpreted. For example, an expression can be described as a joke by one individual and at the same time, that same expression can be perceived as grief by another individual (Chaplin 2015).

Every individual has different personality traits (Ong *et al.* 2018). This explains why some people have different susceptibility to several types of affect (Perry *et al.* 2018). Research has uncovered the relationship between affect and personal traits (Bagozzi, Baumgartner and Pieters 1998). For instance, a link was discovered to exist between positive affect and extraversion. Moreover, the same kind of link was also identified in negativity and neuroticism (Bernbau, Fujita and Pfennig 2005). Furthermore, Jokinen also reported that people are motivated by different circumstances which in turn inspires them to behave differently (Jokinen 2015).

How people manage emotions differ depending on the environmental and professional setting (Taylor *et al.* 2004). People who directly communicate with customers in the service industry are trained to create a happy and welcoming environment to make the customers feel comfortable. For example, customer call centre agents are expected to be more accommodative, patient and cheerful when responding to customers regardless of their true inner feelings. This is also known as 'emotional labour' (Holland 2007). According to Izard (Izard 1992), emotion expression and awareness are essential ability that is essential in social and personal lives. After reporting that management of emotion is of paramount importance, Goleman (Goleman 2003) became a proponent of the assessment of an emotional quotient (EQ). Emotion intelligence is a special ability to recognise and manage one's emotional state. Furthermore, it describes the capacity to recognise the emotions of others (Goleman 2003).

Some people may have limited perceptual or cognitive capabilities because of blindness or deafness (Kiger 1997). This sometimes reduces their capacity to recognise multimodal expressions of emotions (Rieffe 2012). For instance, it is difficult for a blind person to see the facial expressions people make when they talk. In the same vein, it is also challenging for people with autistic spectrum disorders to perceive, express and interpret emotions according to the detects of cultural and social values (Dapretto *et al.* 2006). There are some cognitive disorders such as prosopagnosia which sometimes impairs the ability to recognise emotion in people who suffer from it (Dalrymple *et al.* 2014). Prosopagnosia is a cognitive disorder in which individuals suffering from it end up having severe face recognition difficulties (Duchaine, Parker and Nakayama 2003).

The generation and expression of emotion are highly influenced by gender (Karazi, Sasson and Lomsky 2018). Research has shown that there are some similarities and differences in the way both men and women express emotions (Ubando 2016). This has been found in the neural, behavioural and physiological nature of both sexes (MacGeorge *et al.* 2003). However, Lindquist et al (Lindquist *et al.* 2015) found it difficult to establish whether the challenges were in perception and feeling or regulation and response. The differences in the way both sexes express and generate emotions usually vary with the type of emotion (Chaplin 2016). For example, violent behaviour was discovered to be more common in men because of greater neurological responses (Ille *et al.* 2016) compared to women.

According to Goldshmidt and Weller, men and women have different approaches to communicating and expressing emotions (Goldshmidt and Weller 2000). For instance, women talk three times more than men (Brizendine 2006) and have an inborn tendency of paying attention to facial expressions (Baron-Cohen *et al.* 2010). Furthermore, it was noted that baby girls make more eye contact when communicating (Doherty *et al.* 1995) even though these abilities can be learnt over time. It has was observed that women are more expressive for most of the emotions except anger (Fischer *et al.* 2004).

Age is also a factor that influences the acknowledgement and expression of emotion (Williams *et al.* 2009). The ability to express emotions appears to be characteristic with newly born babies, thus it is considered to be an inborn trait (Mauss and Robinson 2009). Infants have a natural ability to smile in their sleep, and they can also express disgust to bitter-tasting food such as lemons (Sullivan and Lewis 2003). Other researchers such as Caron et al. (Caron, Caron and Myers 1985)

support the notion that the expression of emotions is an inborn trait and is not a result of visual learning. Inborn responses such as physical distress, interest and disgust lead to the ability to express basic emotions such as sadness, surprise, anger and joy (Urwin 2008). These are usually noticeable at around four months while fear is developed between the age of five to seven months (Bahrick *et al.* 2019).

Adults and young people use different approaches when generating and expressing emotions. For example, adults can hide their emotions when handling their day to day business while young people cannot. In support of this notion, socio-cognitive researchers suggested that adults can regulate emotions because of their life experiences (Mroczek 2001). Furthermore, it has been reported that adults are flexible enough to circumvent negative emotions and can manage their emotions to make sound decisions (Hay and Diehl 2011).

Societies across the globe vary in cultural norms and values (Von-Scheve and Ismer 2013). These have a significant influence on the generation and expression of emotions. According to (Shott 2018), emotions are shaped by events that occur within a given society such as the aftermath of the Chernobyl nuclear accident (Drottz and Sjoberg 1990). Individuals use their emotions to manage their responses to such calamities. Within a society, some emotions can be learned. Aggression is an example of an emotional state that an individual can learn within a troubled society (Amudavalli 2010).

Culture is another factor that influences the way emotions are expressed and portrayed (Swain *et al.* 2015). Cultural norms and values vary across different societies. In some cultures, (oriental cultures), it is considered rude to stare at someone while it is recommended in other cultures such as Western cultures (Oishi and Schimmack 2010). Social rules together with cultural norms and values govern the emotional values of every society (Bryant and Barrett 2008). Differences in cultures were noted between the East and West and this has been heavily linked with the corresponding emotional behaviour of the people from these cultures (Scollon *et al.* 2004).

2.3 Theories of Emotion

Various psychology researchers have postulated many theories. These emotions have been developed to describe what emotions are and also to fully understand how they operate. However, describing emotions is a difficult task because they can be analysed from different perspectives (Long *et al.* 2017). Furthermore, emotion expression can be informed by factors such as social environment and culture, as mentioned in the prevision section (Guidi *et al.* 2019). These different features have inspired the development of various theories. The theories of emotion can be categorised into three groups which are evolutionary, social and internal theories.

Evolutionary theories attempt to answer the question of why humans today have the emotions that they do (Nesse 1990). This category of theories goes through the historical analysis of emotions. Evolutionary theories are based on the premise that there is a massive possibility that emotions were present in a common ancestor since all humans have emotions (Nesse and Ellsworth 2009). There are three different ways in which evolutionary theories are described and these are natural selection in early hominids, adaptations shared by all animals, and history. The natural selection school of thought is centred on the basis that emotions have been greatly influenced by the natural selection that occurred in early hominids. The second group claims that emotions were gradual adaptations. However, it proposes that the selection transpired much earlier (Plutchik 2013). The last position suggests that emotions are historical. However, the proponents of this theory argue that emotions are not adaptations (Barrett 2012).

Social and cultural theories are premised on the notion that emotions are a result of social constructions (Levine and Safer 2002). Proponents of these theories argue that individuals learn emotions through experience (Malti and Latzko 2012). They also suggested that societies and cultures influence emotions. Anthropological studies have revealed that there are inconsistencies among the emotional words used in different societies (Lutz 1986). For example, there are several societies and cultures in which anger and sadness are regarded as similar emotions (Conradson and Mckay 2007). In some societies, emotions are regulated through prescribed norms and values (Suh, Oishi and Triandis 2017). These norms and values dictate the events that should make a person happy, angry, bored and many more other states of emotion (Keltner and Haidt 1999).

Internal theories of the emotion process are centred on the description of the emotion process (Scherer 1982). These theories can be subdivided into two groups which are cognitive and non-cognitive theories. Cognitive theories suggest that the early part of the emotion process should be

regarded as a cognitive process because they claim that it involves the manipulation of information. For example, an individual may at different times express different emotions as a response to the same event (Boucher and Brandt 1987). On the contrary, non-cognitive theories claim that appraisals are not part and parcel of the emotion process. Proponents of the non-cognitive theories describe the early stage of the emotion process as a reflex action that does not require any form of prior judgment (Ekman 1992). These theories of emotion have inspired the development of various models of emotion which are described in the next section.

2.4 Models of Emotion

There are many challenges experienced in recognising emotion. For example, it is difficult to nominate the appropriate emotions to consider in a particular scenario (Munezero *et al.* 2014). Moreover, it is difficult to consider the source to draw emotions since the data about emotion can be found in vocal utterances, body movements and facial expressions (Alshamsi *et al.* 2019). Individuals are specialists in emotions since we utilise emotions all the time in our everyday lives. We can likewise name the emotion we communicated to others and describe the emotional states of the people we would have interacted with (Cong *et al.* 2017).

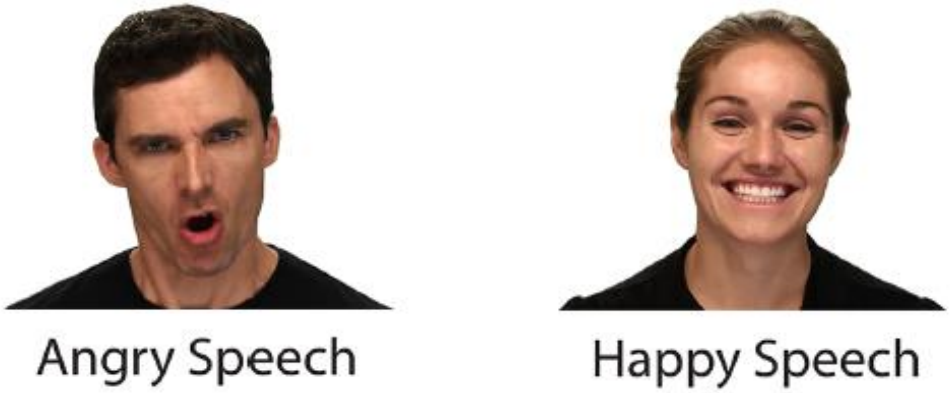
Nevertheless, it is hard to depict emotions through computational means (Chamoli, Semwal and Saikia 2017). Also, classifying emotions through the use of a few rules is significantly harder (Hossain and Muhammad 2017). Therefore, the models of emotions are critical. Several researchers have conducted extensive research to inquire about emotions, and they have proposed different models or hypotheses for the portrayal of emotions (Ozseven 2018). The research shows that emotions can be categorised into three groups which are the basic emotion model, the dimensional demonstrate and the componential appraisal model.

2.4.1 Basic emotion model

This model is anchored on the findings of Charles Darwin (Hess and Thibault 2009). Darwin postulated that emotion can be expressed in various ways such as facial expressions and physical responses which involve changes in speech as well as physiological changes. Many psychologists have expanded Darwin's research findings. Ekman proposed abstract submissions of the basic

emotion model (Ekman 1992). From his work, he reported that six basic emotions could be perceived. These basic emotions are happiness, sadness, surprise, fear, anger, and disgust as illustrated in Figure 2. He argued that more elevated emotions could be consolidated from the six fundamental emotions. Figure 3 shows the facial appearances of the six essential emotions.

Figure 2: Plutchik's wheel of emotions (Plutchik, 2013).



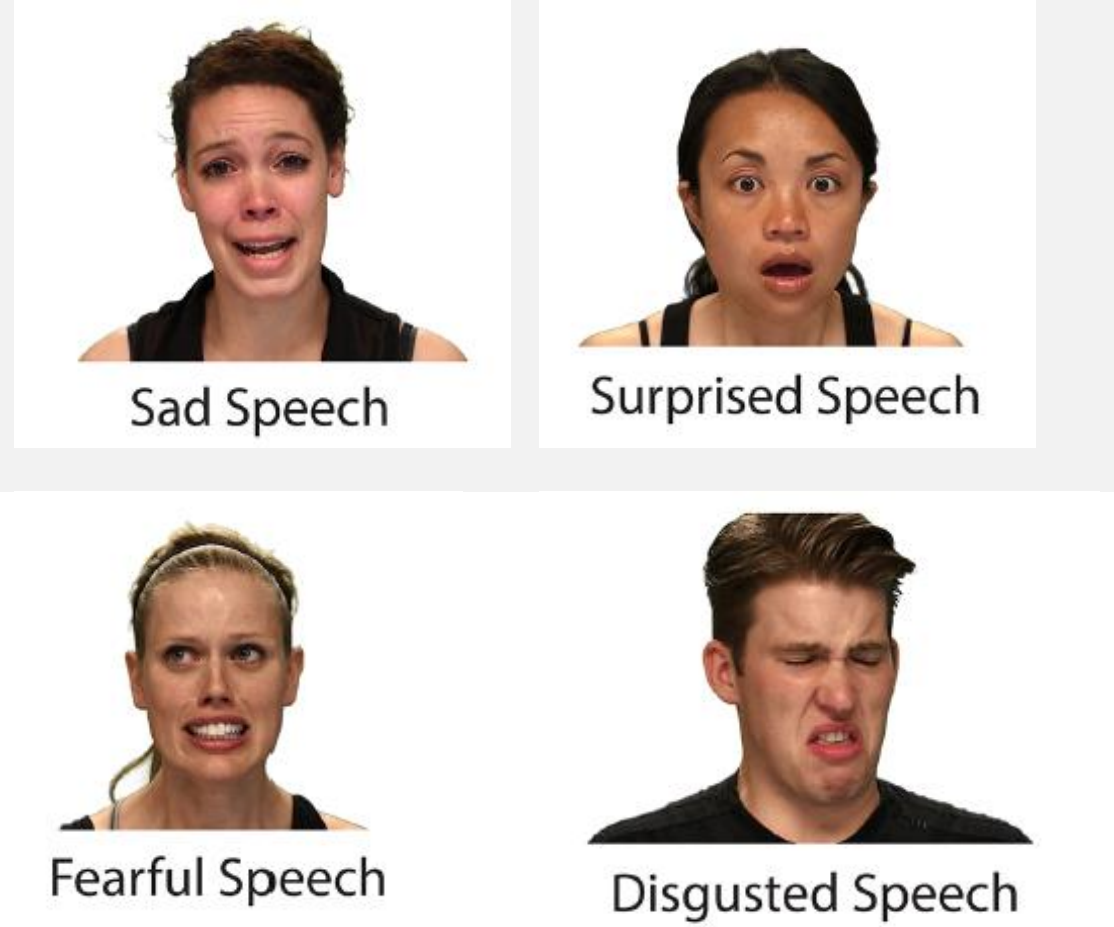


Figure 3:Pictorial depiction of the expression of Basic Emotions (Livingstone and Russo, 2018)

Dr Robert Plutchik, an American analyst, suggested that eight basic emotions serve as the foundation for all the other emotions (Plutchik 2013). He described the eight emotions as the "wheel of emotion" (Pollak *et al.* 2009). Plutchick's wheel comprises eight emotions and these are organised in opposite pairs as shown in Table 1. As shown in Figure 3, the intensity of the emotions is portrayed in separate colours. He also indicated that human beings can't experience opposite emotions at the same time. Complex emotions, which could emerge from a social setting or relationship with fundamental emotion, can be shaped by merely altering some fundamental emotions. Several researchers have proposed some basic emotions and these are in the range from 2 to 18 (Ortony *et al.* 1990). However, Ekman's theory of the six fundamental general emotions is the most commonly used in emotion recognition (Ekman 1992).

Table 1: Plutchik's Emotion states

Basic Emotion	Basic Opposite
Joy	Sadness
Sadness	Disgust
Trust	Anger
Disgust	Anticipation
Fear	Joy
Anger	Trust
Surprise	Fear
Anticipation	Surprise

Ortony (Ortony *et al.* 1990) opposed the notion that basic emotions (for instance, anger, fear, happiness and et cetera) can explain other emotions. They pointed out that it is not good enough to claim that the emotions of daily life are shaped by a combination of basic emotions. Ekman suggested that some of the emotions from the six basic emotions are difficult to classify thus creating confusion. He reported that anger and disgust are usually confused when identifying emotion (Ekman 1992).

2.4.2 Dimensional model

A dimensional framework can be used to represent emotions where the emotions can be mapped by a few factors (Faith and Thayer 2001). In this approach, the emotional states are not autonomous from each other. Valence and arousal are the variables used in a 2-dimensional space. According to Charland (Charland 2002), "valence is one of the most significant scientific concepts that lie at the core of the emotional experience". In other words, arousal describes the physical activation and valence of the pleasantness or hedonic value. The 2-dimensional space is also known as the "valence–arousal–space" (Russell 1980). For third dimensional spaces, energy is generally added. The valence dimension speaks typically to the positive or negative level of the emotion. The arousal dimension speaks to how energised the emotion is, and it ranges from low to high. The energy or control dimension speaks to the level of the energy or control over the emotion in question. These factors present a more exact portrayal of emotions since different parts of emotions are utilised concurrently in a continuous range of values.

Basic emotions can still be characterised in a dimensional model as a point or an area. Due to the differences in valence and arousal, emotion can be represented in the 2-dimensional space as illustrated in Figure 4. As shown in Figure 4, happiness can be evoked by a beautiful sunset or a smiling face in the arousal and valence dimensions. Quite a lot of research has been done using dimensional models for emotion recognition (Wöllmer *et al.* 2008). Generally, just two-dimensional models (valence-arousal demonstrate) are the most used models in emotion recognition studies. The most significant benefit of using the dimensional model is that it is exceptionally instinctive in representing emotions on some continuous scale. However, some researchers argue that the use of fewer emotions results in the loss of valuable information (Ekman 1992).

In addition, some of the fundamental emotions proposed by Ekman are very difficult to perceive in the dimensional models, for example, happiness or sadness. Nonetheless, some different emotions such as anger and disgust are difficult to recognise. Moreover, it is almost impossible to describe some of the emotions using the two-dimensional model. For instance, the surprise emotion state does not even appear in the two-dimensional model.

2.4.3 Componential appraisal model

The principle precept of the componential appraisal model is to clarify how and why particular emotions develop and reasons why people may not have the same emotions in light of a given circumstance (Harley *et al.* 2015). This model proposes that the elicitation and depiction of emotions depend on cognitive appraisals (Shuman, Sander and Scherer 2013). The main idea behind the component appraisal model is that people are always checking their surroundings to understand their world and to get ready for suitable actions if need be (Zatarain *et al.* 2018). Furthermore, the model is based on identifying emotions using events that trigger emotions. In the componential appraisal model, emotion is regarded as an episode that involves a process of continuous changes in components such as feelings, cognition, physiological reactions, motivation, and motor expressions (Wranik and Scherer 2010). This model can be viewed as an expansion of the dimensional model (Scherer 2003).

Moreover, in the componential appraisal model, there is no restriction on the number and the dimensional space of emotions. In the componential appraisal model, emotions are characterised as complex, multi-componential and dynamic processes. The model is mainly focused on the changing of emotion states and it provides multiple types of appraisal patterns. These models of emotion inspired the classification of emotions which is presented in the next session.

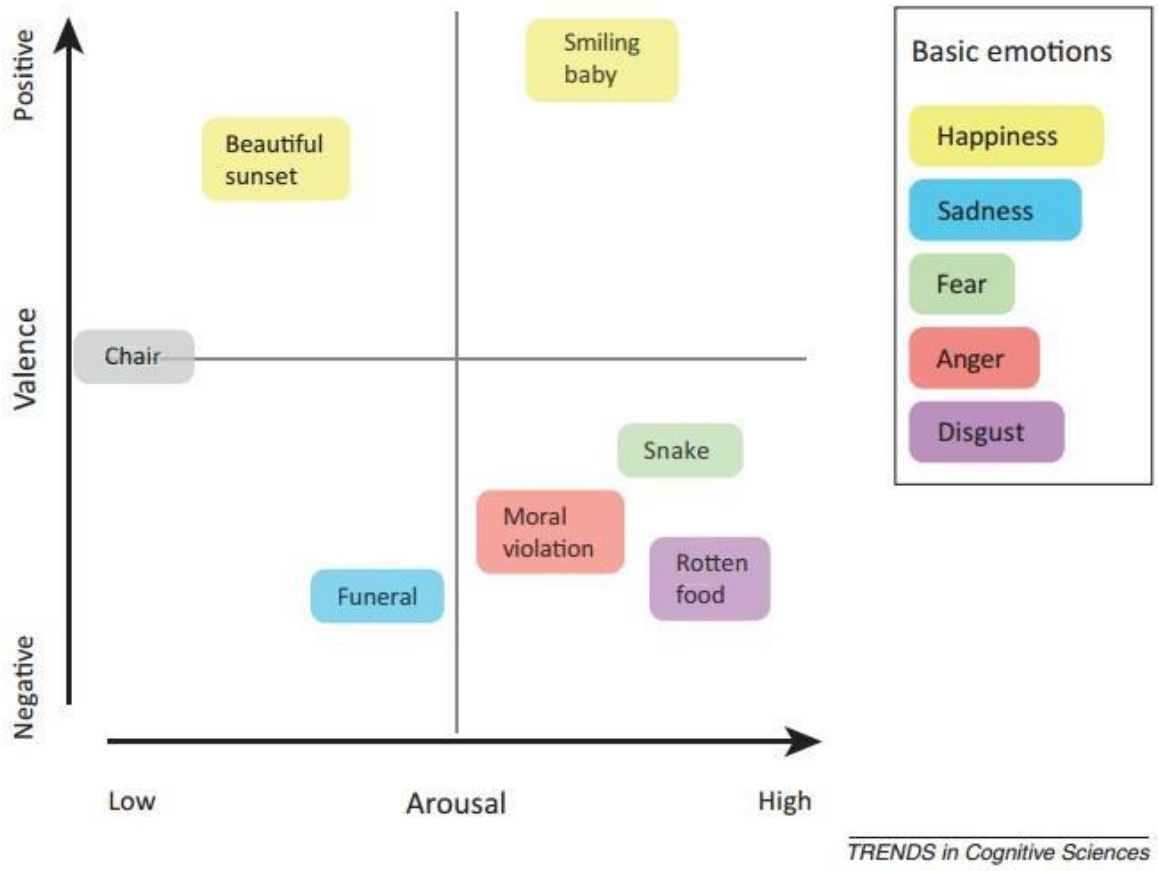


Figure 4: Distinctive and dimensional emotion model (Hamann *et al*., 2002).

2.5 Emotion Classification

Several cognitive psychologists have made some proposals to model human emotions. The classifications suggested by the researchers are different because of people's psychological perspectives and diverse experiences (Ekman 1992). Scholars are yet to come up with a uniform and standard model for classifying emotions because of differences in people's psychological

perspectives and diverse experiences. Notwithstanding the differences in perspectives, there are two forms of emotion classifications (Izard 1992) that have been widely accepted by the affective computing community. These are (1) the classification of basic human emotions and (2) the description of emotion in dimensions. The advocates of the basic emotion theory argued that every emotional state has its innate basic and contrasting characteristics in human experiences and mental stimulation patterns. Turner and Ortony (Ortony *et al.* 1990) made a summary of the taxonomies of basic emotions suggested by various scholars in this field which are shown in Table 2. This section presents emotion classifications proposed by Turner and Ortony, Plutchik, Mayasuma, Ekman, Douglas-Cowie, Shaver and Baron-Cohen.

Table 2: Turner and Ortony's Classification of basic emotions (Ortony *et al.*, 1990)

Scholars	**Basic emotion Classifications**
Arnold (Arnold 1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, and sadness
Ekman, Friesen, Ellsworth (Ekman 1992)	Anger, disgust, fear, joy, sadness, and surprise
Frijda (Frijda 2007)	Desire, happiness, interest, surprise, wonder, and sorrow
Gray (Gray 1990)	Rage and terror, anxiety, and joy
Izard (Izard 1992)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise
James (James 1884)	Fear, grief, love, and rage
McDougall (Mcdougall 1991)	Anger, disgust, elation, fear, subjection, tender-emotion, and wonder
Mowrer (Mowrer 2009)	Pain, pleasure
Oatley, Johnson-laird (Oatley and Johnson-Laird 1987)	Anger, disgust, anxiety, happiness, and sadness
Panksepp (Panksepp and Watt 2011)	Expectancy, fear, rage, and panic

Plutchik (Plutchik 1990)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, and surprise
Tomkins (Kunz 1997)	Anger, interest, contempt, disgust, distress, fear, joy, shame, and surprise
Watson (Watson and Morgan 2006)	Fear, love, and rage
Weiner ,Graham (Graham and Weiner 2011)	Happiness, sadness

In classifying, emotion numerous scientists allude to principal emotions and general emotions. Through his experiments, Robert Plutchik (Plutchik 2013) suggested that eight basic emotions were focal and central to most human encounters. His list of major emotions comprised anger, fear, sadness, joy, disgust, curiosity/interest, surprise, and acceptance. These sets of emotions ended up being the most used words in characterising key emotions. Both Plutchik and Izard (Izard 1992) view these eight emotions as being established in a transformative will to survive, which is ever-present in every single individual. In support of his concept of emotion, Plutchik (Plutchik 2013) developed a three-dimensional model as shown in Figure 4.

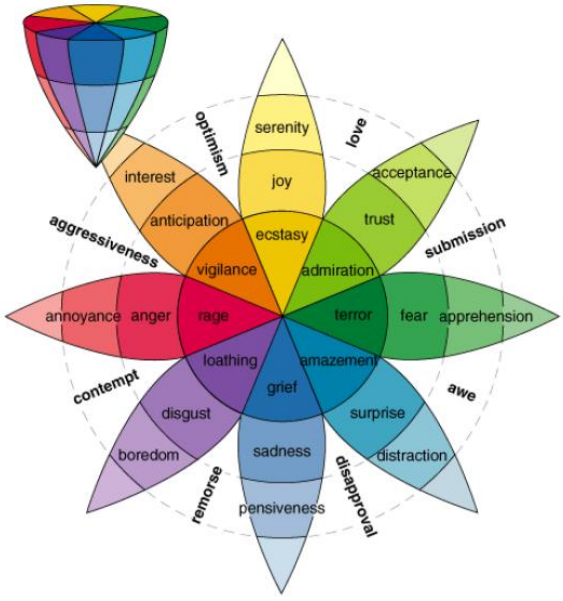


Figure 5: Plutchnik three-dimensional model of emotion concepts (Plutchik, 2013)

The proposed three-dimensional model is a fusion of basic-complex categories and dimensional theories. In this model, emotions are carefully ordered in concentric circles. Here the inner circles are more basic and outer circles more complex. Additionally, the outer circles are developed by merging the inner circle emotions and the model has eight emotional dimensions. Using this we can extricate that emotions are unfathomably perplexing and that even though they can be grouped into basic emotions, observing basic emotions alone cannot reveal the inconspicuous varieties in emotion that could affect an individual differently.

Masuyama (Masuyama *et al.* 2017) conducted similar experiments in his quest to classify emotions. In his studies, he analysed the contributing factors of 32 distinct emotions. The experiment involved a group of 19 subjects who had the responsibility of rating eight dramas. The outcome of Masuyama's analysis uncovered his rundown of basic emotions as being aversion, joy, concession, expectation, surprise anger, and sorrow. Initially, his unique rundown of 32 basic emotions included emotions such as apprehensiveness, happiness, and disappointment. It is vital to take note of how Masuyama's original and new list of emotions in terms of the number of emotions. Moreover, some of the other emotions from the original list would still be equally important if used in the right context. Therefore, it is imperative not to disparage the active part of the emotion and not to be entirely bound by our research methods in classifying emotion.

Ekman is one of the most popular specialists in the field of emotional psychology. Most researchers in this field agree that emotions can either be basic or can be surrogates of other basic emotions. According to Ekman (Ekman 1992), basic emotions are classified into the following emotion states: surprise, happiness, disgust, fear, sadness, and anger.

Just like Plutchnik's three-dimensional circumplex model, Ekman's basic emotions can be combined to form distinct variations of emotion such as rage, boredom, embarrassment, or compassion. Ekman's basic emotions are limited to only six emotions. However, each of the six emotions can have several variations of intensities depending on the context in which the emotion is triggered. To illustrate this, the anger emotion state can be used. Anger is the emotion that communicates abhorrence toward a man or circumstance which is causes revulsion. According to psychologists, anger can be described using three distinct intensity categories listed below.

i. Anger is described as that emotional state which is triggered when one is hurt.

ii. It is also described as a response to the view of being purposefully hurt or abused by others.

iii. Finally, anger is described as an individual's natural character or trait.

On the other hand, happiness is the emotion that communicates different degrees of positive emotions extending from fulfilment to extraordinary satisfaction.

Ekman passes on the developmental idea of emotions all through his extensive analysis in the area. This idea of emotion as a transformative characteristic is a subject that has been examined strongly by Charles Darwin. Like Darwin and Plutchik, Ekman holds the conviction that human emotions evolved for human survival and reproduction. From his investigations of a clan in Papua New Guinea, Ekman reasoned that these fundamental emotions are both all-inclusive and inalienable in both proficient and preliterate societies. Such research can demonstrate that emotions do influence and shape human lives. It was these culturally diverse analyses and emotion theory distributed in the mid-'80s which saw the introduction of much research in affect and affect recognition. Ekman's six fundamental emotions have been the focal point of numerous emotion research because of their general nature and the way they act as roots to every other emotion. These investigations range from facial recognition of affect, affect recognition from vocal utterances and textual information.

Douglas-Cowie et al. (Douglas-Cowie *et al.* 2007) grouped 48 emotion states into 10 groups which include negative forceful, negative/positive thoughts, caring, positive, lively, reactive, agitation, negative not in control, negative passive, and positive quiet. Additionally, they developed a tool (FEELTRACE) (Cowie *et al.* 2000) to label these emotions by mapping them on a special chart, as shown in Figure 6.

Shaver et al. (Shaver, Murdaya and Fraley 2001) developed two classification categories in their quest to classify emotions, and these categories are pleasant and non-pleasant. As shown in Figure 7, three groups of emotion states were created in each of the two main categories. Love, joy and surprise were classified in the pleasant category, while anger, sadness and fear were classified in the unpleasant category. They then grouped similar emotion states under a single emotion. For example, affection and compassion were grouped under the love emotion state while distress and worry were placed in the fear emotion state.

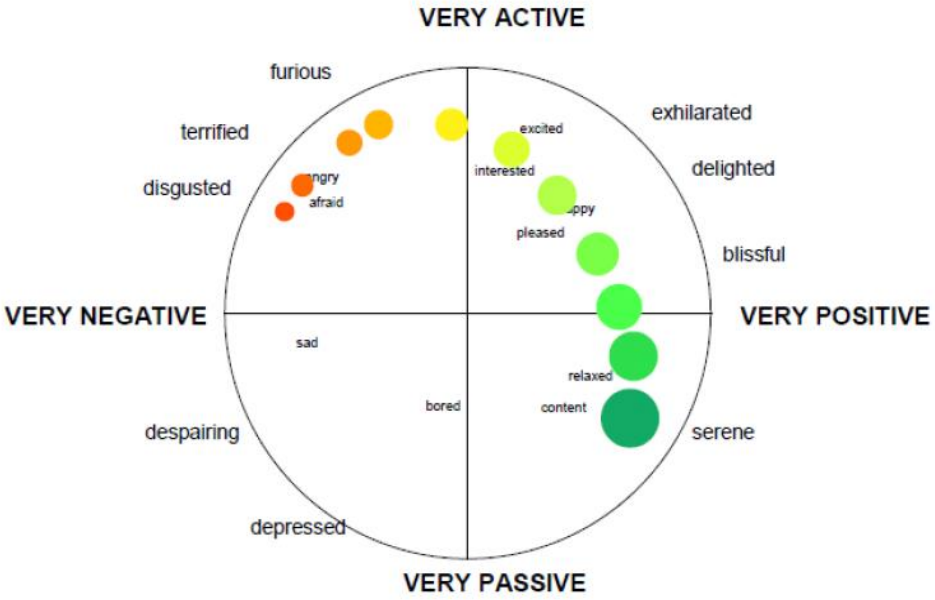


Figure 6: Emotions mapped in Activation-Evaluation space with FEELTRACE (Cowie *et al*., 2000)

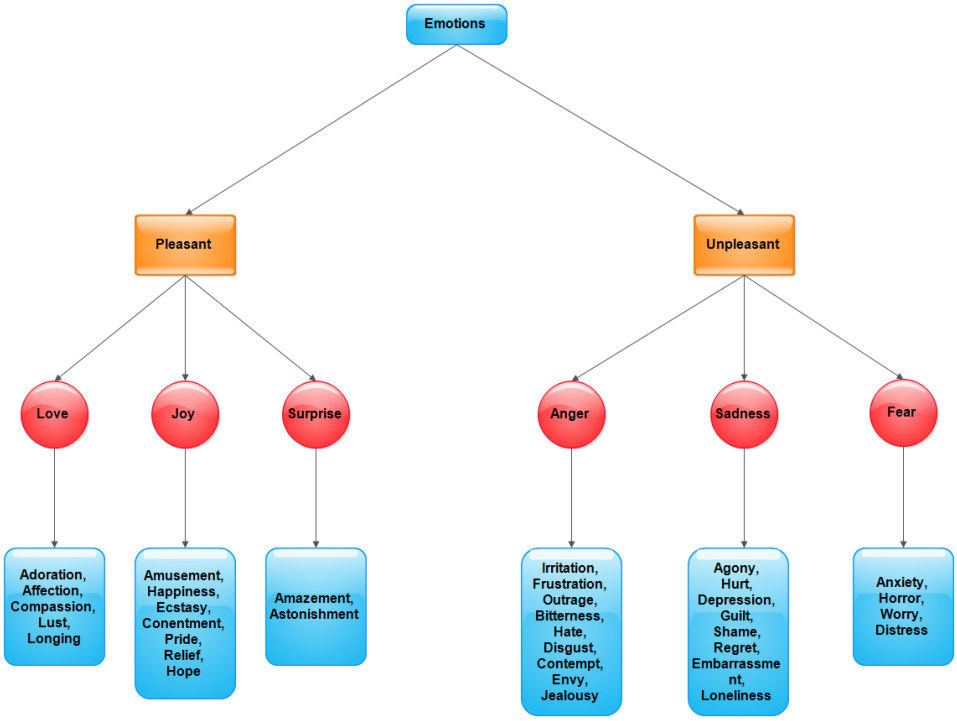


Figure 7: The Hierarchical Structure of Emotions(Shaver, Murdaya and Fraley, 2001)

In an attempt to classify emotions, Baron-Cohen (Baron-Cohen *et al.* 2010) utilised a lexicographic examination of 1150 words (412 idea words and 738 equivalent words) from the Microsoft

Thesaurus and placed them into 24 independent groups, that inspired the Mind Reading application. Mind Reading is an intelligent software program that is used to depict emotions from a multimodal stand. In addition, it uses professional actors to exhibit a list of emotions illustrated in Table 3.

Emotion words have numerous unpretentious shades of significance in various settings. These do not continuously result in clear and precise interpretations between languages. This inspired Cohen et al. to compare his classifications to the colour spectrum (Baron-Cohen *et al.* 2006).

Table 3:MindReader categories of emotion (Jauk, Bonafonte and Pascual, 2016)

Afraid	Angry	Bored	Bothered	Disbelieving	Disgusted
Excited	Fond	Happy	Hurt	Interested	Kind
Liked	Romantic	Sad	Sneaky	Sorry	Sure
Surprised	Thinking	Touched	Unfriendly	Unsure	Wanting

They argued that emotion labelling is a complicated task because the same emotion can be perceived differently by different people. In that same vein, anger might be perceived as sadness in other social circles (Brown 2014). These distinctions in the use of words may be as a result of physiological, cultural and social reasons (Lutz 1986; Von-Scheve and Ismer 2013).

2.6 Chapter Summary

From this chapter, it can be observed that emotion is a difficult word to define because of social, cultural perspectives, environmental and physiological variables. This challenge has inspired the development of several emotion theories and models, which as highlighted, are worlds apart in terms of similarity. The basic emotion model described in this chapter provides a firm representation of the emotions that would most likely be useful in assessing call centre conversation. This is also supported by Ekman's six basic emotions. These have also been used in most SER studies especially in facial emotions (Chelali and Djeradi 2015). This is because a higher

number of dimensions provided by other models are not reliable to estimate. Therefore, the basic emotion model was used to develop the design of the SER system in this thesis.

Chapter 3: Processing of Human Speech Features

3.1 Introduction

Speech features are an essential ingredient for developing SER systems because they influence the emotion class under which vocal cues fall. The recognition of emotions described in the previous chapter is highly dependent on human speech features such as prosodic and spectral features. Therefore, this chapter addresses factors involved in the production of speech features, expressive speech analysis and human speech features. The description of speech production is critical because it determines the nature of features in vocal utterances. The physiological features in a human body react in a certain way when a particular emotion is triggered and this produces a pattern of speech features that can be used to describe that emotion (Gievska, Koroveshovski, and Chavdarova 2015; Jiang *et al.* 2017). It is important to understand the physiological structure of the human body when developing SER systems because it helps in fully understanding the speech features that can be used to recognise emotion (Banse and Scherer 1996; Zaballos *et al.* 2016; Istening and Imbre 2018). SER systems should be able to process vocal utterances in precisely the same way human beings process them. Therefore, expressive analysis is presented in this chapter because it describes the link between the human ear, SER and the resulting features such as pitch (Reddy *et al.* 2011; Kumari and Ali 2015; Hübscher, Borràs and Prieto 2017; Susan and Kaur 2018). The last section of this chapter describes the features used in classifying emotion in speech. These features include prosodic, spectral and spectrogram features. Human speech features are presented in this chapter because they are the life blood of SER. Consequently, research has shown many times that a SER model is as good as the features used to develop it (Pierre 2003; Huang *et al.*2014; Alías, Socoró and Sevillano 2016).

In section 3.2, the production of speech in humans is described. Expressive speech analysis is elaborated in section 3.3. Section 3.4 presents a description of prosodic, spectral and spectrogram features.

3.2 Speech Production

The human voice carries a lot of information that can be used to recognise gender, age, and emotion (Ke *et al.* 2018). Therefore, it is essential to distinguish between two aspects of human speech which are short term segmental aspects and longer-term suprasegmental aspects. Short term segmental aspects are the ones that contain linguistic information, while the longer-term suprasegmental aspects consist of both paralinguistic and non-linguistic information (Yoon *et al.* 2019).

Indicators of a speaker's gender, age, and emotional state are believed to be carried in non-linguistic information. According to research, the most influential type of information when it comes to SER is the suprasegmental characteristics of speech (Irastorza and Torres 2017). Here the changes in the physiological mechanisms in speech production influence the emotional state of individuals.

The production of speech is done through a well-coordinated series of actions (Ying and Xue 2018) through the human physiological systems. These are the resonance system, the phonation system, and the respiratory system.

3.2.1 Respiratory system

The respiratory system is the system that enables human beings to breathe, and it comprises the thoracic cage, lungs, diaphragm, and trachea (Pützer *et al.* 2019). The main task of the respiratory system in the production of speech is to control the phonation system by regulating air pressure (Istening and Imbre 2018). The respiratory system is illustrated in Figure 8 according to the visualisations done by Ackermann.

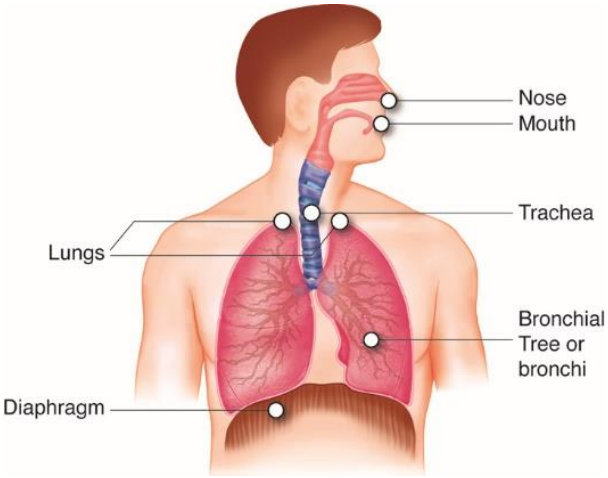


Figure 8: The respiratory system (Ackermann, 2008)

3.2.2 Phonation system

The phonation system is also known as the vocal system (Grozdić *et al.* 2017). As shown in Figure 9, the phonation system comprises the larynx, the vocal folds, and the glottis. The glottis refers to the opening between the vocal overlays through which air streams from the trachea to the pharynx (Ackermann 2008). The space between vocal folds widens when a person is calm and quiet (Gordon and Ladefoged 2001). On the contrary, the vocal folds are brought closer together by the coordinated activity of many laryngeal muscles).

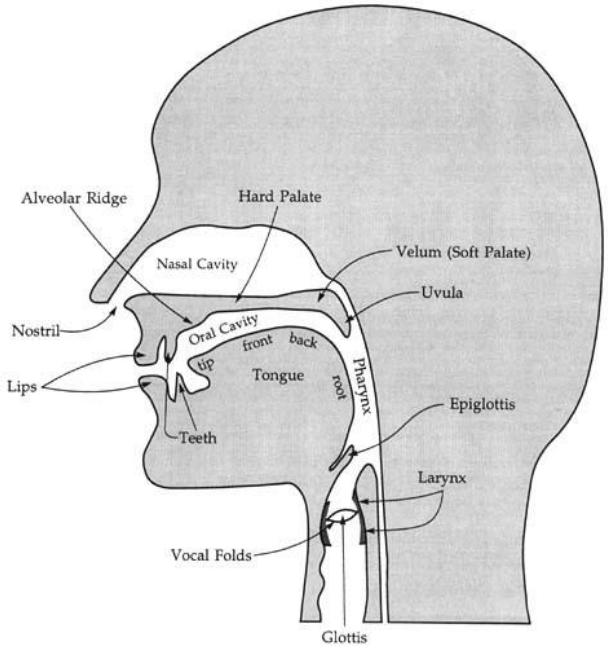


Figure 9:The phonation system (Guenther, 2006)

The air is consequently blocked, causing a pneumatic force to develop beneath the vocal folds and separating them in the process. As air travels through the glottis, the vaporous pressure in the vocal folds is reduced because of the Bernoulli Effect. This at that point makes the vocal folds close and the cycle is rehashed again and again (Titze 1994). This will then lead to a seasonal variation in the super laryngeal air pressure. This variation will be corresponding to a sound with a base frequency. This concept is referred to as the "fundamental frequency" (f_0) (Lu *et al.* 2017).

Any slight change in air pressure in the area underneath the larynx will influence the behaviour of the vocal folds, consequently creating disparities in the intensity, f_0 and the harmonic energy distribution of the sound (Titze 1994). For example, when the vocal folds are under raised strain, and subglottal pressure is high a direct result of significant expiratory effort, the vocal folds will close more rapidly, inciting an increase in basic f_0 as well as in harmonic energy distribution of the sound (Meilán *et al.* 2014). Such a vocal design may be normal for certain high excitement feelings, for example, anger (Jauk, Bonafonte and Pascual 2016). An example of speech features emanating from this approach is the fundamental frequency.

The research in this area entails the use of instruments such as Magnetic Resonance Imaging (MRI) scans, ultrasounds and X-ray snap shots of the vocal tract taken during the production of speech (Najnin and Banerjee 2019). Furthermore, the research also incorporates the development of speech features such as Linear Predictive Coefficients (LPC) (Mellahi and Hamdi 2015). This system supports the view that prosodic features are the ones that paint a clear picture of the existence of emotion within a speech signal. Additionally, the approach also argues that the expression of an emotion depends on the production of phonemes (Kannadaguli and Bhat 2018). The production approach is influenced by subtle changes in the breathing muscles and the vocal fold (Lu *et al.* 2017). It is also influenced by other factors such as vocal-tract shaping factors, which is associated with the movements of the upper articulators. The upper articulators comprise the lips, teeth, tongue, jaw, and velum (soft palate).

The location of the speech production mechanisms is illustrated in Figure 9. As mentioned earlier, prosodic features are a textbook example of features emanating from the phonation system. Prosodic features are those aspects of speech that go beyond phonemes and deal with the auditory qualities of sound (Jauk, Bonafonte and Pascual 2016). These are also referred to as suprasegmental phonology. Examples of prosodic features include spectral patterns, intensity,

duration, and fundamental frequency. The rate at which vocal folds vibrate is known as fundamental frequency (Ramdinmawii and Mittal 2016). The fundamental frequency is determined by the size and tension of the vocal fold at any given point in time. The frequency value is affected by factors such as stress, emotions, and intonations.

3.2.3 Resonance system

The resonance system is the section that makes up the remainder of the vocal tract (Hernández *et al.* 2019). It extends from the glottis, through the pharynx to the oral and nasal pits. The shape and length of the resonance system rely on the design of the tongue, velum, teeth, and lips. The latter is also referred to as articulators. Furthermore, the shape and length of the resonance system influence how specific harmonics are amplified (Watanabe *et al.* 2019). A relatively small number of formants relate to the various vowels and vocalised consonants. Formants are defined as the specific patterns of amplified and attenuated harmonics. The resonance system is said to be under greater voluntary control compared to the other systems. However, it is prone to involuntary perturbation (Ciobanu *et al.* 2014).

The bandwidths and amplitudes of the formants are affected by the amount of saliva in the mouth, the shape of the resonance tract and the tautness of articulatory muscles (Chandran, Pravena and Govind 2017). For example, many interviewees develop dry mouths when anxiously responding to interview questions, which triggers changes in the formant amplitudes and bandwidths (Shuman, Sander and Scherer 2013).

3.3 Expressive speech analysis

Expressive speech analysis refers to the investigation of non-verbal vocal communication cues and these are also referred to as paralinguistic cues (Arias, Busso and Yoma 2014). Examples of these include voice intensity, tone of voice, and many more. It can be applied in disciplines that include speech processing and linguistics, signal processing, music and psycho-acoustic, hearing, and neurology (Yogesh *et al.* 2017a). These determine the features eligible for use in SER. The most common approach in expressive speech analysis is the perception approach (Gahlawat, Malik and Bansal 2016).

A perception approach is an approach that analyses how a signal is interpreted by the human physiological system (Crivelli *et al* 2016). Moreover, this approach is used to analyse both psychological and neurological hearing mechanisms to treat hearing problems. An example of speech features emanating from this approach is pitch. The perception approach comprises of sub approaches which are the psycho-acoustic point of view, hearing point of view, linguistic point of view and music perception point of view (Lorenzo *et al.* 2018).

Some of the research in the perception approach was done using psycho-acoustic tests. These are experiments in which individuals are solicited to evaluate sound features from music and speech signals. According to experiments carried out by Mozziconacci *et al.* (Mozziconacci and Hermes 1997), the most relevant features for recognising emotions are spectral, energy and properties of the pitch contour. Other researchers attempted to develop hybrid features by combining the features that relate to different emotions but their results were sometimes contradictory (Dahake, Shaw and Malathi 2016).

It is imperative to understand the structure of the ear and how it operates to understand more about the speech features that can be used to recognise emotion (Zaballos *et al.* 2016). For instance, there is a strong link between the spectral content of speech and the ear's cochlea structure. It is believed that the human ear can process sounds that are in the range of approximately 20 Hz - 20 kHz (Alías, Socoró and Sevillano 2016).

The linguistic point of view incorporates significant portrayals of intonational phrases to interrogate stress and accent in the elocution of words and sentences (Malarkodi and Devi 2019). Moreover, it is used for prosody analysis in distinguishing between questions and sentences. Sentences usually have a falling edge towards the end of the sentence while questions have a rising edge towards the end of the sentence (Bahrick *et al.* 2019). Differences in pitch height of syllables in a vocal utterance constitute a significant factor that influences the perception of prosody. Research has shown that variations in tone and tone shapes determine emotion states within spoken utterances (Nirmal *et al.* 2017). All these different schools of thought paint a clear picture of the nature of features that can be used to describe and recognise emotion through speech. These are the features discussed in the next section.

3.4 Speech features

It is quite interesting to see how humans perceive emotional states in spoken utterances with surgical precision. Humans have an innate ability to distinguish between different human voices instantly; they can assess the tempo and mood of an utterance and also compare and contrast vocal utterances (Hess and Thibault 2009). However, this has been a problem for researchers in their effort towards developing systems that automatically recognise emotion in speech (Gong *et al.* 2015). This brought about the process of Audio Signal Classification (ASC). ASC is the process in which pertinent features are extracted from a piece of audio and are used to identify the class to which the audio file belongs (Schuller, Reiter, and Rigoll 2006).

Audio data in its raw form can be extensive and can contain massive redundancies (Fayek, Lech and Cavedon 2017). Therefore, in audio recognition, pre-processing is done first to mitigate this problem. Pre-processing involves feature extraction and feature selection. Feature extraction consists of the extraction of raw data while feature selection involves the process of choosing the most important features to use for the ultimate recognition. There is a vast amount of literature that documents the essential features (Yerigeri, Ragha and Ieee 2017) necessary for audio and speech recognition (Vivek *et al.* 2018). Speech features can be categorised into two main categories which are prosodic and spectral. These categories are described in the subsequent sections.

3.4.1 Prosodic features

Research from the phonation system discussed in section 3.2.2 of chapter 3 has shown how prosodic features are formed and these are essentially used in speech emotion recognition (Kerkeni *et al.* 2018). Speech signals carry a plethora of information that can be used to determine the emotions of humans. Some of the distinct information include prosody. Prosody is a term that is used to describe the vocal arrangement of rhythm, intonation, and stress in a spoken utterance (Ayadi, Kamel and Karray 2011). Some of the features that are used to profile the human voice from a prosodic point of view include rhythm, intensity, and pitch.

Additionally, prosody is also referred to as the suprasegmental properties of speech because prosody plays a pivotal role in structuring the flow of speech (Pérez, Reyes and Villaseñor 2012). The primary function of prosody in human speech involves the communication of linguistic, pragmatic, and affective traits of vocal utterances. Prosodic features are highly rich in paralinguistic information and for this reason, they are used in SER(Pérez, Reyes and Villaseñor 2012; Hellbernd and Sammler 2016).

Some features from different backgrounds may be referred to as prosodic features if they contain suprasegmental speech features. This is the reason why there is a thin technical line between prosodic and acoustic features. Furthermore, prosody describes those speech features that are related to larger units such as words, sentences, and phrases. Prosody represents both the physical and perceptual characteristics of speech. Human beings subconsciously identify emotion in speech using prosodic cues (Lalitha *et al.* 2014).

Prosodic features also resemble important speech features such as intensity patterns, intonation, and duration (Bahrick *et al.* 2019). Even though human beings have a natural gift of identifying emotion in speech, some still find it challenging to identify some emotion states. This is because of the nature of prosodic cues (Kahou *et al.* 2016). The emotion-specific information carried by these features is believed to be non-overlapping. Therefore, the process brings confusion in distinguishing some of the emotion states that share similar prosodic and acoustical properties. In such circumstances, individuals will then rely on linguistic and other modalities such as facial expressions in their interpretation of emotions. Prosodic features have become popular over the years because they are not affected by environmental and channel features in telephone conversations (Thyme and Hutchins 1996). The prosodic features used in this research include energy and pitch.

Energy is defined as a measure of the volume of a signal at any point in time and can, therefore, be used to identify silence and loudness in a signal. It is sometimes referred to as intensity. The structure of the amplitude of a speech signal changes with time. Normally, the amplitude of voiced signals is higher than that of unvoiced signals (Koteswara, Swarna and Hima 2016). These amplitude variations are represented by energy. The energy can be computed using the equation below.

$$E = \frac{1}{N}\sum_{i=0}^{N-1}|X(n)|^2 \quad \textbf{(3.1)}$$

where X(n) represents a speech signal in a frame while N is the total number of samples in a frame. Energy is generally high in voiced speech signals and low in unvoiced speech signals. Voiced speech signals can be identified using ZCR and other features can then be extracted from the identified voiced speech signals.

Pitch is the first perceptual feature outlined by Schuller et al. (Schuller, Reiter and Rigoll 2006). According to Sem et al. "Pitch is the fundamental frequency of the glottal pulse." (Sen, Dutta and Dey 2019). Pitch has a symbiotic relationship with prosody. Consequently, prosody is a characteristic of speech that relates to changes in pitch and phoneme duration. It assists in accentuating certain words in a phrase, hence signifying deeper meaning. A primitive example in this situation can be the rise in pitch at the end of an utterance which most certainly signifies a question. Prosody gives a lot of information concerning any given spoken utterance and is closely associated with the fundamental frequency, which is primarily perceived as pitch. Fundamental Frequency (F_0) is only informative for periodic or pseudo-periodic signals(Arias, Busso and Yoma, 2014). As discussed earlier, F_0 is directly proportional to the vibration of the vocal cords; therefore, it varies from person to person depending on gender, age. Schuller et al. (Schuller, Reiter and Rigoll 2006) clarify how the F_0 of a signal over time can be used to identify word boundaries in utterances. This is because significant deviations in F_0 are not likely to occur in the middle of a word and therefore, can be useful in defining the beginning and end of words in an utterance.

3.4.2 Spectral features

Spectral features are the features extracted from the vocal tract system (Thirunavukkarasu, Abdi and Mohajer 2016). These are some of the most explored features in speech emotion recognition. Furthermore, they are also referred to as segmental features. The most used spectral features in speech processing are MFCCs (Mel Frequency Cepstral Coefficients), LPCCs (Linear Prediction Cepstral Coefficients), and PLPCs (Perceptual Linear Prediction Coefficients). Spectral features have gained popularity in speech processing because they have an accurate representation of the vocal tract system. The spectral features discussed in this section are Zero crossing rate, spectral

centroid, spectral roll-off, spectral flux, MFCCs (Mel Frequency Cepstral Coefficients) and LPCCs (Linear Prediction Cepstral Coefficients).

The *Zero-Crossing Rate (ZCR)* is an essential element in almost every feature vector in sound recognition literature. The ZCR is a proportion of how regularly the amplitude of a speech signal crosses zero for every unit time, and it is viewed as an exceptionally instructive feature (Jayasankar, Vinothkumar and Vijayaselvi 2017). The calculation for ZCR is shown in Equation 2.

$$Z(i) = \frac{1}{2N}\sum_{n=0}^{N-1} |sgn[xi(n)] - sgn[xi(n-1)]| \qquad \textbf{(3.2)}$$

where

$$sgn[xi(n)] = \begin{cases} 1, & xi(n) \geq 0 \\ -1, & xi(n) < 0 \end{cases}$$

where $x_{-1}(N)$ is a provisional array developed to store the preceding frame values. Equation 2 shows the mathematical formula for calculating feature values using ZCR. The purpose of sgn in the equation is to assign the normalised value [−1, 0, 1]. This will be based on the range of input variable values. There is usually a strong correlation between energy distribution and ZCR with frequency because high frequencies are directly proportional to ZCRs (Ververidis and Kotropoulos 2006). It is generally believed that low ZCR shows that the speech signal is voiced while higher ZCRs depict that the speech signal is unvoiced. The following ZCR features were extracted for this experimental study:

1. Zero Crossings Overall Standard Deviation,
2. Derivative of Zero Crossings Overall Standard Deviation,
3. Running Mean of Zero Crossings Overall Standard Deviation,
4. Standard Deviation of Zero Crossings Overall Standard Deviation,
5. Derivative of Running Mean of Zero Crossings Overall Standard Deviation,
6. Derivative of Standard Deviation of Zero Crossings Overall Standard Deviation,
7. Strongest Frequency Via Zero Crossings Overall Standard Deviation,
8. Derivative of Strongest Frequency Via Zero Crossings Overall Standard Deviation,
9. Running Mean of Strongest Frequency Via Zero Crossings Overall Standard Deviation,
10. Standard Deviation of Strongest Frequency Via Zero Crossings Overall Standard Deviation,
11. Derivative of Running Mean of Strongest Frequency Via Zero Crossings Overall Standard Deviation,
12. Derivative of Standard Deviation of Strongest Frequency Via Zero Crossings Overall Standard Deviation
13. Zero Crossings Overall Average,
14. Derivative of Zero Crossings Overall Average,
15. Running Mean of Zero Crossings Overall Average,
16. Standard Deviation of Zero Crossings Overall Average,
17. Derivative of Running Mean of Zero Crossings Overall Average,
18. Derivative of Standard Deviation of Zero Crossings Overall Average,
19. Strongest Frequency Via Zero Crossings Overall Average,
20. Derivative of Strongest Frequency Via Zero Crossings Overall Average,
21. Running Mean of Strongest Frequency Via Zero Crossings Overall Average,
22. Standard Deviation of Strongest Frequency Via Zero Crossings Overall Average,
23. Derivative of Running Mean of Strongest Frequency Via Zero Crossings Overall Average,
24. Derivative of Standard Deviation of Strongest Frequency Via Zero Crossings Overall Average

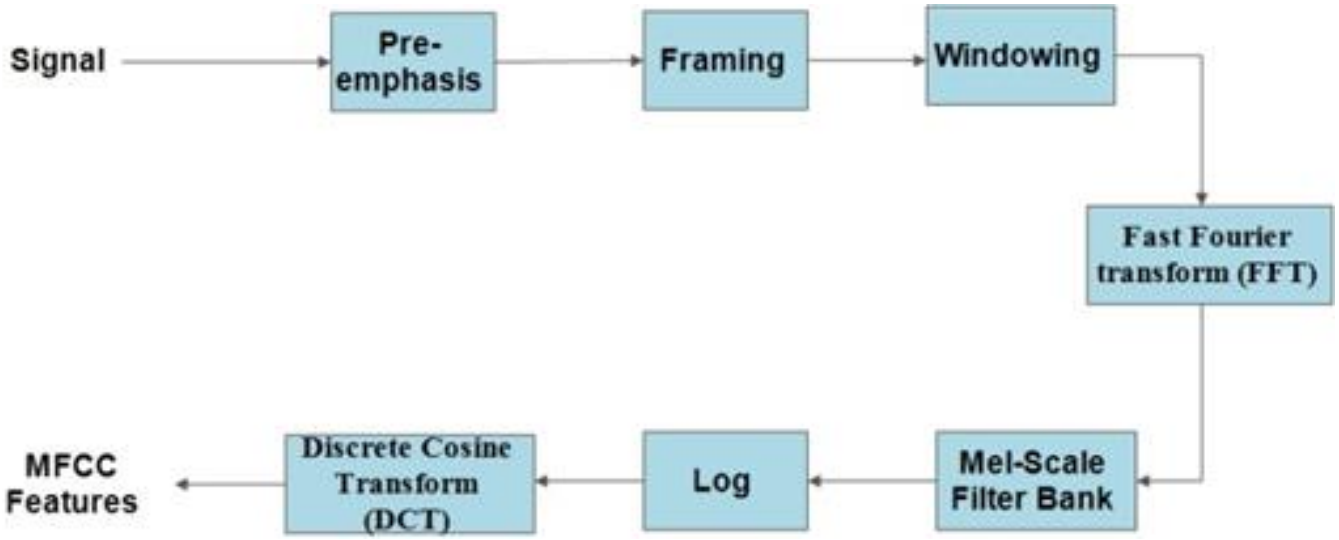


Figure 10: MFCC Algorithm block diagram (Yu, Li and Fang, 2012)

Mel Frequency Cepstral Coefficients (MFCC) are perceptually motivated features that are also based on the STFT (Majuran and Ramanan 2018). MFCCs are described as a compact representation of the spectrum of an audio signal that takes into account the non-linear human perception of pitch, as described by the Mel scale. In MFCC feature extraction, signals are first divided into small frames that contain a random number of samples. To preserve the smallest units of sound which are phonemes, the frames are subjected to the process of overlapping. The windowing technique can be used in carefully choosing the desired frames of each signal. This is usually done before computing the Fourier transform of the signals to minimize spectral leakages (Selvaraj, Bhuvana and Padmaja 2016). To implement this, the Hamming window is used. The filter coefficient of the hamming window is computed using the window function $w(n)$ listed below,

$$w(n) = \begin{cases} 0.054 - 0.46\cos\left(\frac{2n\pi}{N-1}\right), & 0 \le n \le N-1 \\ 0, \; else \end{cases} \qquad \textbf{(3.3)}$$

where N is the total number of samples and $w(n)$ is the actual current sample. Fast Fourier Transform (FFT) of each frame is calculated to speed up the process. The logarithmic Mel scale is then applied to the FFT frame, which is linear up to 1 kHz and logarithmic at greater frequencies.

The relationship between the frequency f of speech in the Hertz and Mel scale $M(f)$ has been established as follows.

$$M(f) = 2595\log\left(\frac{f+700}{700}\right) \quad \textbf{(3.4)}$$

Finally, the discrete cosine transforms which de-correlates the features and arranges them in descending order is calculated. The following MFCC features were used in this experimental study:

1. Ten MFCC Overall Standard Deviation,
2. Ten Derivative of MFCC Overall Standard Deviation,
3. Ten Running Mean of MFCC Overall Standard Deviation,
4. Ten Standard Deviation of MFCC Overall Standard Deviation,
5. Ten Derivative of Running Mean of MFCC Overall Standard Deviation,
6. Ten Derivative of Standard Deviation of MFCC Overall Standard Deviation,
7. Thirteen Area Methods of Moments of MFCCs Overall Standard Deviation,
8. Ten MFCC Overall Average,
9. Ten Derivative of MFCC Overall Average,
10. Ten Running Mean of MFCC Overall Average,
11. Ten Standard Deviation of MFCC Overall Average,
12. Ten Derivative of Running Mean of MFCC Overall Average,
13. Ten Derivative of Standard Deviation of MFCC Overall Average

Linear prediction analysis is one of the most widely used methods in speech coding, speech synthesis, speech recognition, speaker recognition, and verification (Chamoli, Semwal and Saikia 2017). LPCC methods provide incredibly accurate estimates of speech parameters, and they do it with supreme efficiency. The theory of linear prediction (LP) is highly associated with the modelling of the vocal tract system. LPCC entails the use of linear predictive analysis in speech parameter computation.

The LPCC analysis estimates the resonance of the vocal tract using a signal's waveform. Certain calculations are performed to minimise the prediction error. The basic idea of Linear Prediction is that the current speech sample can be closely approximated as a linear combination of past samples, that is,

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \quad \textbf{(3.5)}$$

Here it estimates s(n) from the $\rho(\rho << N_\rho)$ most recent values of (n) by linearly predicting the value. For Linear Prediction, the predictor coefficients (the α_k 's) are determined (computed) by minimising the sum of squared differences over a finite interval between the actual speech samples and the linearly predicted ones. Linear Prediction is based on speech production and synthesis models. Speech can be modelled as the output of a linear time-varying system, excited by either quasi-periodic pulses or noise. The summarised diagram of the LPCC model is shown in Figure 11.

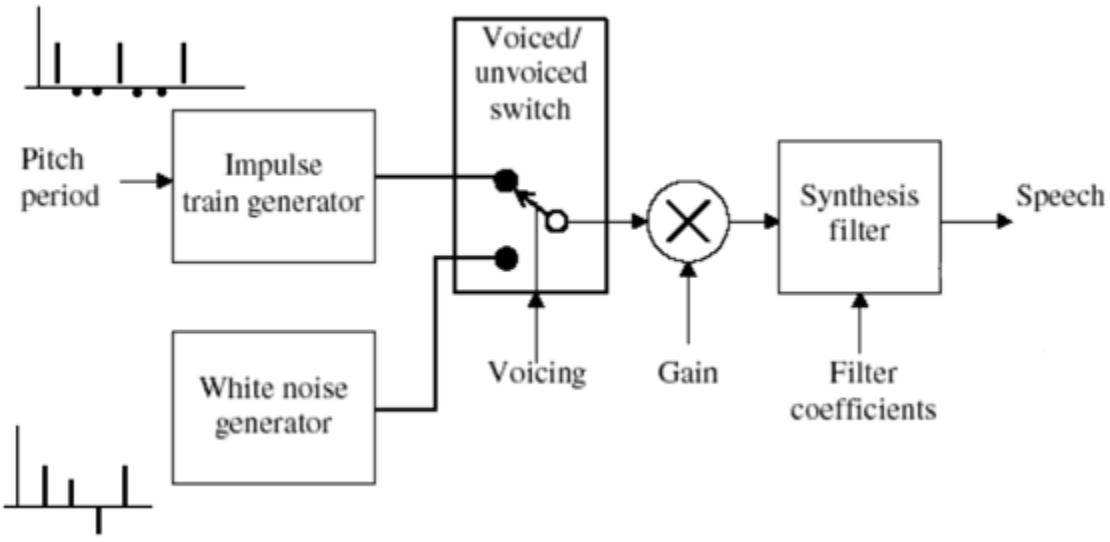


Figure 11: Source filter model of LPCC speech production (Chia Ai *et al*., 2012)

According to this illustration, voicing refers to voiced or unvoiced speech frames, gain refers to the energy level of the frame, filter coefficients define the synthesis filter response and pitch period refers to the time duration between consecutive excitation pulses (voiced). In this model, a combination of the spectral contributions of the glottal flow, the vocal tract, and the radiation of the lips is represented by a time-varying digital filter with a steady-state system function given by

$$H(z) = \frac{S(z)}{X(z)} \quad \textbf{(3.6)}$$

In the above equation (3.6) both poles and zeros exist in the transfer function. However, if the order of the denominator is high enough, H(z) can be computed by the equation below:

$$H(z) = \frac{G}{A(z)} \quad \textbf{(3.7)}$$

where

$$A(z) = 1 - \sum_{j=1}^{\rho} \alpha_j z^{-j} \quad \textbf{(3.8)}$$

Equation 3.6 will then be transformed a sample time-domain given below:

$$s(n) = Gx(n) + \sum_{j=1}^{\rho} \alpha_j s(n - j) \quad \textbf{(3.9)}$$

Equation (3.9) above is the popular LPCC equation which states that the value of the present output s(n) may be computed by adding the present input Gx(n), and a weighted sum of the past output samples. Hence, in LPCC analysis the problem can be stated as follows: given the measurements of the signal s(n), determine the parameters a j, j =1…, p which minimises Gx(n). The resulting parameters are then assumed to be the parameters of the proposed model system transfer function H(z)

If α_j represents the estimates of a_j, the error or residual is given by:

$$e(n) = s(n) - \sum_{j=1}^{\rho} \alpha_j s(n - j) \quad \textbf{(3.10)}$$

Once the above equation is computed, it is now possible to determine the estimates by minimising the mean squared error, i.e.

$$E\{e^2(n)\} = E\{[s(n) - \sum_{j=1}^{\rho} \alpha_j s(n - j)]^2\} \quad \textbf{(3.11)}$$

If the partial derivatives of the above equation for a_j is set to zero for j=1…., p the result in Equation 3.12 is obtained.

$$E\{[s(n) - \sum_{j=1}^{\rho} \alpha_j s(n - j)]\ s(n-1)\} = 0, \text{ for } i = 1,\ldots, p \quad \textbf{(3.12)}$$

In deriving the equation 3.12 above, it is assumed that the signal of the model is stationary. Nonetheless, assuming that it is indeed stationary for short segments of speech is quite reasonable. LPCC analysis can be applied in a lot of application areas such as speaker identification (Daqrouq *et al.* 2015), gender detection (Přibil, Přibilová and Matoušek 2016) and many more. The following LPCC features were extracted for the experimental study:

1. LPCC Overall Standard Deviation,
2. Ten Derivatives of LPCC Overall Standard Deviation
3. Ten Running Mean features of LPCC Overall Standard Deviation
4. Standard Deviation of LPCC Overall Standard Deviation,
5. Ten Derivatives of Running Mean of LPCC Overall Standard Deviation,
6. Ten Derivatives of Standard Deviation of LPCC Overall Standard Deviation,
7. LPC Overall Average
8. Ten Derivatives of LPCC Overall Average
9. Eight Running Mean features of LPCC Overall Average
10. Standard Deviation of LPCC Overall Average,
11. Ten Derivatives of Running Mean of LPCC Overall Average
12. Ten Derivatives of Standard Deviation of LPCC Overall Average

The *Spectral Centroid* of a signal is described as the centre of gravity of the magnitude spectrum of the Short-Time Fourier Transform. This helps in modelling the sound's sharpness (Desai, Dhameliya and Bhatt, 2015). For a given speech signal, it is defined as:

$$SC = \frac{\sum_{k=0}^{N_{FT}/2} f(k)P_s(k)}{\sum_{k=0}^{N_{FT}/2} P_s(k)} \qquad \textbf{(3.13)}$$

where P_s represent the estimated power spectrum of a speech segment while f(k) is the frequency of the kth bin and N_{FT} stands for the size of the DFT. Furthermore, the spectral centroid is referred to as a spectral moment that can be used in modelling the sharpness or brightness of sound (Kim, Moreau and Sikora 2006). The following spectral centroid features were extracted:

1. Spectral Centroid Overall Standard Deviation,
2. Derivative of Spectral Centroid Overall Standard Deviation,
3. Running Mean of Spectral Centroid Overall Standard Deviation,
4. Standard Deviation of Spectral Centroid Overall Standard Deviation,
5. Derivative of Running Mean of Spectral Centroid Overall Standard Deviation,
6. Derivative of Standard Deviation of Spectral Centroid Overall Standard Deviation,
7. Derivative of Strongest Frequency Via Spectral Centroid Overall Standard Deviation,
8. Running Mean of Strongest Frequency Via Spectral Centroid Overall Standard Deviation,
9. Three Standard Deviation of Strongest Frequency Via Spectral Centroid Overall Standard Deviation,
10. Five Derivative of Running Mean of Strongest Frequency Via Spectral Centroid Overall Standard Deviation,
11. Five Derivative of Standard Deviation of Strongest Frequency Via Spectral Centroid Overall Standard Deviation

Spectral Roll-off is defined as the frequency below which 85% of the magnitude distribution of the spectrum is concentrated (Giannakopoulos and Pikrakis 2014). This feature provides a measure of spectral shape (Desai, Dhameliya and Bhatt 2015) for speech signals and can be used to identify voiced and unvoiced signals(Kim, Moreau and Sikora 2006). Spectral Roll-off can be computed using the equation below.

$$\sum_{k=0}^{K_{roll}} |S(k)| = 0.85 \sum_{k=0}^{N_{FT}/2} |S(k)| \qquad \textbf{(3.14)}$$

where K_{roll} represents the frequency bin which corresponds to the estimated roll-off frequency. The following spectral roll-off features were extracted in the experimental study:

1. Spectral Rolloff Point Overall Average,
2. Derivative of Spectral Rolloff Point Overall Average,
3. Running Mean of Spectral Rolloff Point Overall Average,

4. Standard Deviation of Spectral Rolloff Point Overall Average,
5. Derivative of Running Mean of Spectral Rolloff Point Overall Average,
6. Derivative of Standard Deviation of Spectral Rolloff Point Overall Average
7. Spectral Rolloff Point Overall Standard Deviation,
8. Derivative of Spectral Rolloff Point Overall Standard Deviation,
9. Running Mean of Spectral Rolloff Point Overall Standard Deviation,
10. Standard Deviation of Spectral Rolloff Point Overall Standard Deviation,
11. Derivative of Running Mean of Spectral Rolloff Point Overall Standard Deviation,
12. Derivative of Standard Deviation of Spectral Rolloff Point Overall Standard Deviation

Spectral Flux describes the squared difference between the normalised magnitudes of successive spectral distributions in short-term windows (Giannakopoulos and Pikrakis 2014). It can also be described as the spectral rate of change (Desai, Dhameliya and Bhatt 2015) in a speech signal. Spectral flux can be computed using the equation below.

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} \left(EN_i(k) - EN_{i-1}(k)\right)^2 \quad \textbf{(3.15)}$$

where

$$EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} X_i(l)} \quad \textbf{(3.16)}$$

The following spectral flux features were extracted for this experimental study:

1. Spectral Flux Overall Standard Deviation,
2. Derivative of Spectral Flux Overall Standard Deviation,
3. Running Mean of Spectral Flux Overall Standard Deviation,
4. Standard Deviation of Spectral Flux Overall Standard Deviation,
5. Derivative of Running Mean of Spectral Flux Overall Standard Deviation,
6. Derivative of Standard Deviation of Spectral Flux Overall Standard Deviation
7. Spectral Flux Overall Average,

8. Derivative of Spectral Flux Overall Average,
9. Running Mean of Spectral Flux Overall Average,
10. Standard Deviation of Spectral Flux Overall Average,
11. Derivative of Running Mean of Spectral Flux Overall Average,
12. Derivative of Standard Deviation of Spectral Flux Overall Average

Fast Fourier Transform (FFT) defines a class of algorithms that provide an efficient way of calculating Discrete Fourier Transform (DFT) (Schupp 2003). FFTs became popular in the 1960s (Tukey and Cooley 1965) because of their ability to reduce the computational complexity of DFT in various application areas such as speech processing (Jameel, Siyal and Ahmed 2005; Takaki and Yamagishi 2016), spectral analysis (Jameel, Siyal and Ahmed 2005; Takamichi 2018) and many more. Given that $p(n) = x(n)$ for $n = 0,1,\ldots, N-1$ the discrete-time Fourier series of the following periodic signal is obtained.

$$a_k = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}kn} \tag{3.17}$$

Where N denotes the period of the signal and $p(n)$ is periodic. The time-domain signal can be obtained as follows:

$$x(n) = \sum_{k=0}^{N-1} a_k e^{j\frac{2\pi}{N}kn} \tag{3.18}$$

3.4.3 Spectrograms

The invention of high-performance computing resources has led to the rise of the application of deep learning models in speech processing (Petridis *et al.* 2018). Most traditional algorithms such as Support vector machines (SVM) and other deep learning models such as Recurrent Neural networks cannot handle raw audio data (Zhao, Jin and Hu 2017). Therefore, the raw audio files

can be converted to spectrograms so that various extraction methods can be applied to make the recognition process easier (Papakostas *et al.* 2017).

Spectrograms are images that represent the spectrum of frequencies of audio signals as they vary with time(Zhao, Mao and Chen 2018). In audio signal processing, spectrograms are also referred to as voicegrams, voiceprints, or sonographs (Li, Yang and Dai 2014). Spectrograms are also defined as the visual approach used in representing the signal strength (Ma *et al.* 2019), or "loudness" (Satt, Rozenberg and Hoory 2017), of a signal over time at various frequencies present in a waveform. In the typical spectrogram, time is represented by the horizontal axis, while the vertical axis resembles frequency. Lastly, the frequency of the signal at any given time and the corresponding amplitude is shown as a sort of heat map or scale of colour saturation. (Tao *et al.* 2016).

As illustrated in Figure 12, the spectrogram differs from waveforms in that, the latter cannot tell us much about the pitch, frequency, or harmonic content of a recording. Unlike, the waveform (which also displays time and amplitude), spectrograms (Figure 13) are rich in information that can be used to recognise emotion.

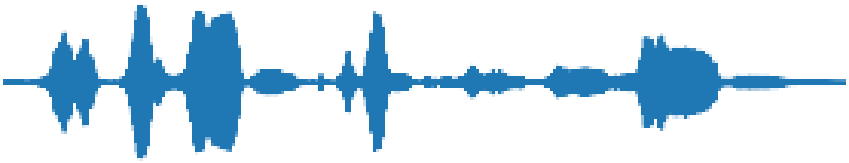


Figure 12: Sample waveform (Ariav and Cohen, 2019)

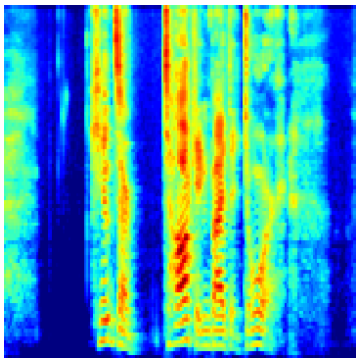


Figure 13:Sample Spectrogram (Ariav and Cohen, 2019)

Chapter 4: Speech Emotion Recognition Models

4.1 Introduction

This chapter discusses SER models built on the intrinsic features described in the previous chapter. The chapter presents a review of the use of prosodic, spectral, hybrid and spectrogram features in developing SER models. The purpose of this chapter is to present an in-depth analysis of the performance of human speech features and classifiers in developing SER models. Therefore, a review of research articles from 2010 to date is discussed. The SER models discussed in this section were developed using various speech emotion corpora which include SEMAINE, Berlin Emotion, Polish emotional database, SAVEE, RAVDESS, CASIA and many more. The classifiers reviewed in this section include Support Vector Machines (SVM), Deep Neural Networks (DNNs), SOM (self-organizing maps). The implications of each group of features combined with specific classifiers are also discussed in this chapter. The analysis of speech features and classifiers is important because it provides the strength and weaknesses of each technique as well as clues on ways that can be used to improve the performance of SER models. Research gaps were identified in this chapter and these were exploited to develop improved models presented in chapter six.

In section 4.2, the use of prosodic features in SER is described and analysed. The analysis of spectral features in SER is elaborated in section 4.3. Section 4.4 presents a review of spectrograms in SER while the use of hybrid features is analysed in section 4.5. A review of cross-language acoustic emotion valence is presented in section 4.6. The chapter is concluded with a summary in section 4.7 where the research gap is highlighted.

4.2 Prosodic Features

Human speech is coloured with a lot of prosodic features that can be used to recognise emotion (Selvaraj, Bhuvana and Padmaja 2016). Here, long term speech features such as intonation, speech energy information fundamental frequency, variations in duration and formant tracks can be used to determine the emotional states of a given vocal utterance (Jauk, Bonafonte and Pascual 2016).

This is the reason that has inspired some researchers to develop SER models using prosodic features (Getahun and Kebede 2017). Furthermore, prosodic features have been used to measure politeness given a set of vocal utterances (Hübscher, Borràs and Prieto 2017). Prosodic features have also been used to develop systems that can inject emotional content in general statements for human-computer interaction purposes (Najnin and Banerjee 2019).

Evaluation of feature selection frameworks using four feature selection algorithms and three different multiclass classifiers to detect speech emotion was proposed by Altun and Polat (Altun and Polat 2009). They reported that among all feature groups, the prosodic and sub-band energy features were the most frequently selected ones by all the algorithms in each framework. The highest accuracy recorded in their paper was 83.7% using SVM.

The Orthogonal Forward Selection (OFS) algorithm also yields good results when used to select prosodic features in speech emotion detection (Swain *et al.* 2017). Principal component analysis and linear discriminant analysis can be used jointly to construct the speaker-sensitive feature space and reduce the dimensionality of data to make meaningful predictions (Kursa and Rudnicki, 2010).

Stuhlsatz et al (Stuhlsatz *et al.* 2011) introduced the Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs) to recognise speech emotions using acoustic features. They used 6 552 dimensions of acoustic features making use of 39 functionals of 56 acoustic Low-Level Descriptors (LLDs) including first and second order delta regression coefficients. It was reported that the author has achieved an accuracy of 81.9% and 53.6% for the Berlin Emotion dataset and SUSAS (Speech Under Simulated and Actual Stress) actual stresses respectively.

Arias et al presented a novel approach based on functional data analysis (FDA) to detect local emotional prominence in the fundamental frequency on the SEMAINE database (Arias, Busso and Yoma 2014). The main aim of their study was to capture the intrinsic variability of F_0 contours. Their proposed system achieved accuracies as high as 75.8% in binary emotion recognition, which is 6.2% higher than the accuracy achieved by a benchmark system trained with global F_0 statistics.

In (Zhang *et al.* 2015) twenty-four nonlinear dynamic features were used to recognise three emotional states (angry, neutral and fear). The features were extracted from three databases i.e. the Berlin Emotion dataset, Polish emotional database, and linguistic data consortium (LCD). 69 randomly selected audio files (anger and neutral states) were chosen from the Berlin Emotion

dataset. The authors extracted a total of 24 features from each signal. 30% of the data was used for testing the proposed model while 70% was reserved for training. Their proposed system achieved recognition accuracy of 72.28%, 75.4%, and 80.75% on the Berlin Emotion dataset, Polish emotional database, and linguistic data consortium (LCD) respectively.

Pell et al (Pell *et al.* 2015) used a set of prosodic features such as fundamental frequency and intensity to develop a SER model. Their model classified three emotion states which are anger, sadness, and happiness achieving an accuracy rate of 87.3%. Ekpenyong et al (Ekpenyong, Inyang and Udoh 2018) used prosodic features to recognise emotion in an Ibibio language speech corpus. They used SOM (self-organizing maps), to classify the tone of emotion-laden statements. The set of prosodic features used in their study includes tone pattern, fundamental frequency: F0, phoneme pattern (vowels only), speech duration.

Mao et al (Mao *et al.* 2017) proposed a Domain Adaptation (DA) based method called Emotion-discriminative and Domain-invariant Feature Learning Method (EDFLM) for SER, in which both the domain divergence and emotion discrimination were considered to learn emotion-discriminative and domain-invariant features by using emotion label constraint and domain label constraint. Experimental results on the Berlin dataset produced an accuracy of 61.63 on unlabelled data.

In (Ying and Xue 2018) glottal features were used to develop an effective SER system. The authors proposed the model where the glottis was to be used for compensation of glottal features. Therefore, they extracted the feature of Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator (GCZCMT). Their experimental results obtained an accuracy of 84.45% using the Berlin Speech Database.

Narendra and Alku (Narendra and Alku 2019) proposed a new dysarthric speech recognition method from coded telephone speech using glottal features. Their method used glottal features, which were efficiently estimated from coded telephone speech using a deep neural net-based glottal inverse filtering method. Two sets of glottal features were considered: (1) time- and frequency-domain parameters and (2) parameters based on principal component analysis (PCA). Their proposed method achieved an overall accuracy of 96.38.

The work presented in the literature shows that indeed prosodic features are a true representation of an individual's emotional state. However, using these features alone really does not yield appropriate accuracies. This means developing SER models that solely rely on prosodic features can result in an unhealthy number of negative positives. This implies that there is a possibility that if applied to customer call centres, an angry customer can be perceived to be happy. Such misrecognition can lead to misrepresentation of facts hence misinform management with information that could have stimulated positive strategies.

Other prosodic features such as glottal features have been achieving excellent results. However, the main problem with these features is that the extraction process is time-consuming (Silva *et al.* 2017; Yogesh *et al.* 2017b). One of the aims of this study is to develop an SER model that is time-sensitive or that can process emotions quickly. The use of glottal features is also resource-intensive which will also defeat the aim of this research work.

4.3 Spectral Features

A variety of spectral features have been used in the field of speech processing. MFCC and LPC features have been commonly used in SER over the years (Yang *et al.* 2012; Ramdinmawii and Mittal 2016; Mansour and Lachiri 2017). Particularly, MFCC has gained so much popularity because MFCC features are believed to model the exact way in which people express their emotions (Zhu *et al.* 2017). Spectral features are also used in recognising gender, speakers and age from a set of vocal utterances (Desai, Dhameliya and Bhatt 2015; Li, Xu and Yang 2017; Ozaydin 2017). Some of the recently reported works in recognising emotion using various spectral features are presented in this section.

Pérez (Pérez, Reyes and Villaseñor 2012) et al analysed 6920 acoustic features from the IEMOCAP corpus. Their results showed that the most important feature groups for valence are MEL while MFCC (Nirmal *et al.* 2017) features were highly rated in activation and dominance. Badshah et al (Badshah *et al.* 2016) used MFCC features to develop a two-layer SER model. The first layer was responsible for detecting the existence of emotion in a set of spoken utterances. The last layer was responsible for recognising the exact emotions. They used the SAVEE emotion

corpus and the Random forest classifier. Their model obtained an overall accuracy of 82.21% across 7 emotions which are anger, boredom, disgust, fear, happy, neutral and sad.

Chia et al (Chia *et al.* 2012) conducted experiments to discover the efficiency of both MFCC and LPC features in recognising emotion in speech. They conducted these experiments using their own stuttered speech corpus. They achieved their best results using the K-nearest neighbour (KNN) classifier achieving an overall accuracy of 94.51% using LPC features. In their study, MFCC features were outperformed by LPC features.

Using the SAVEE speech corpus, Kishore and Satish compared an MFCC inspired SER model against Sub-band based Cepstral Parameter (SBC) method (Krishna Kishore and Krishna Satish, 2013). They used the Gaussian Mixture Model (GMM) to recognise emotions. The SBC method performed better than the MFCC method. The MFCC method achieved an overall accuracy of 51% while 71% was achieved using the SBC method.

A telephone complaint SER model was developed by Gong et al (Gong *et al.* 2015) to recognise emotion. They used MFCC features using the SVM classifier. The model yielded an overall accuracy of 76.39% across three emotion states which are calmness, discontent, and anger. They used the CASIA Chinese speech emotional database and the records from an undisclosed customer complaint service centre of a telecom and Internet service company.

Lalitha et al (Lalitha *et al.* 2015) developed a SER model using MFCC features. They developed a simple neural network classifier which they used to recognise emotions in the Berlin Emotional database. Their model obtained an overall recognition accuracy of 85.7% across the seven emotion states from the corpus.

Sahoo and Routray developed an improved version of an MFCC feature inspired SER model (Sahoo and Routray 2017). In their work, they optimised the analysis frequency range to achieve a higher recognition accuracy. They used the Berlin Emo-DB corpus and achieved an overall accuracy of 80% across seven emotion states which are anger, boredom, disgust, fear, happy, neutral, and sad. They also repeated the same experiment on the Assamese corpus and achieved an overall accuracy of 81%. The corpus had the same emotional content as the one contained in the Berlin Emo-DB corpus.

Likitha et al (Likitha *et al.* 2017) developed a SER model based on MFCC features. They used their custom corpus which had a total of 60 people. They evaluated their model across three emotion states which are happy, sad and angry. The model obtained an efficiency score of 80%.

In their work, Aouani and Ayed (Aouani and Ayed 2018) developed a speech recognition model using MFCC features. They performed a comparative analysis between Deep Support Vector Machine (DSVM) and the stacked auto-encoder classifier. Their results showed that the stacked auto-encoder classifier performs much better than DSVM classifier. Using the Berlin Emotional corpus, their model achieved an overall accuracy of 73.01%.

In their quest to find the best variation of SVM in recognising emotion through speech Sonawane et al (Sonawane, Inamdar and Bhangale 2018). They extracted MFCC features which they tested on linear and non-linear SVM classifiers. They extracted these features from BBC sample news. The RBF kernel outperformed the linear kernel version of the SVM classifier, and it yielded an overall accuracy of 90.82%. The corpus they used had six emotional states which are anger, disgust, happiness, sadness, surprise, and neutral.

Kerkeni et al (Kerkeni *et al.* 2018) proposed an SER that uses Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the Berlin Speech Database and Spanish dataset. Their study showed that for the Berlin database all classifiers achieve an accuracy of 83% when a speaker normalization (SN) and a feature selection were applied to the features. An overall recognition of accuracy of 94 % was achieved by RNN (recurrent neural network) classifier without SN and with feature selection. They tested their model across seven emotions which are anger, disgust, fear, joy, sadness, surprise and neutral. They also noted that disgust and fear were quite difficult to detect.

A hierarchical SVM model was presented by Ke et al (Ke *et al.* 2018) to recognise speech emotion. They extracted MFCC features from the CASIA Chinese emotion corpus. They used the Principal component analysis (PCA) algorithm to reduce the dimensionality of the data. Their optimised model achieved an overall accuracy of 76.7% across 5 emotion states which are a surprise, fear, happy, sad, and angry.

Li and Akagi (Li and Akagi 2019) proposed a multilingual SER model using four datasets which are: Fujitsu, Berlin Emo-DB, CASIA, and SAVEE. They achieved an average precision score of

91.04% after applying normalization and feature selection techniques. MFCC related features and their 1st order delta coefficients were used (Liu *et al.* 2018). The authors presented an SER model which was an improvement of the brain emotional learning (BEL) model. The model was based on the emotional operations of the human brain. Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and an ensemble of both were used to reduce the dimension of the features. Their proposed system achieved average recognition accuracy of 90.28% (CASIA), 76.40% (SAVEE), and 71.05% (FAU Aibo) for speaker-dependent (SD) speech emotion recognition.

From the review presented in section 4.3, it is observed that LPCC is hardly used in developing SER models especially since 2012. In most cases, LPCC features are used in validating the performance of new SER models and in most cases the LPCC features are outperformed. MFCC features have become the de facto spectral features in speech emotion recognition. From the literature, it can be observed that MFCC features can perform better when combined with a proper dimensionality reduction method (Li and Akagi 2019).

However, for the customer call centre environment, higher accuracies can still be achieved because most MFCC inspired-SER models achieve lower accuracy scores in recognising the fear emotion state across many languages (Selvaraj, Bhuvana and Padmaja 2016). Some MFCC inspired speech recognition models to achieve overall recognition accuracies that are above 90% but if we look at the accuracy of each emotion, we can observe that MFCC features are not very efficient in recognising some of the emotions. This is part of the research gap that this thesis intends to explore. The SER model resulting from this thesis should be able to efficiently recognise all the emotions within an identified corpus yielding recognition accuracies that are more than 90%.

4.4 Spectrogram Features

Spectrograms have become popular in speech processing over the years because of their natural ability to represent salient features with precision (Papakostas *et al.* 2017). The analysis of speech using spectrograms has become a game-changer in emotion recognition. SER has become a computer vision problem because image processing techniques are now being used to process the spectrograms (Zhao, Mao and Chen 2018). The process of analysing the patterns of emotion by

merely looking at the spectrograms is extremely difficult (Stolar *et al.* 2018). Spectrograms are also popular with deep learning models such as Conventional Neural networks because they have a superior ability to process images (Stolar *et al.* 2018; Zhao, Mao and Chen 2018). The use of spectrograms in SER is relatively new and most of the reported work is approximately five years old or less. This section presents some of the recently reported work in recognising emotion through spectrograms.

A sparse autoencoder-based feature transfer learning method for emotion recognition from the speech was proposed in (Interaction *et al.* 2013). They used quite a several databases in their experimental study. Some of the databases they used include EMO-DB and eNTERFACE database (Martin *et al.* 2006). The proposed model achieved an accuracy of 59.1% using spectrograms generated from the eNTERFACE database.

Fayek et al. (Fayek, Lech and Cavedon 2015) proposed a Deep Neural Networks (DNNs) inspired speech emotion recognition model. In their approach, they extracted features from raw speech spectrograms using DNNs. They evaluated their method on the eNTERFACE and SAVEE database achieving a recognition accuracy of 60.53% and 59.7% respectively.

Poria et al (Poria, Chaturvedi and Cambria 2016) discovered that the convolutional neural network (CNN) sometimes performs poorly when subjected to a sequence of images. It is against this background that they developed a deep learning ensemble classifier using a recurrent neural network (RNN) with CNN. They combined their new method with feature selection achieving a recognition accuracy score of 96.55%.

Zhang et al. (Zhang *et al.* 2016) developed a CNN classifier that recognises emotion classes using a combination of video and audio files. They used two separate CNNs to develop their model. The extracted features from audio spectrograms and the other CNN version were used to extract features directly from video files. The extracted features were then combined and these are the ones that were used to classify emotions. To evaluate the efficacy of their proposed model they used the RML (Ryerson Multimedia Research Lab) database achieving a recognition accuracy of 74.32%.

Fayek et al. (Fayek, Lech and Cavedon 2017) proposed a deep learning framework using convolutional neural networks (CNNs). They used CNN to extract features from spectrograms

generated from audio files. These audio files were taken from the IEMOCAP (Interactive emotional dyadic motion capture) (Busso *et al.* 2008) database. They used their proposed CNN to classify the emotions achieving a recognition accuracy of 64.78%.

Zhu et al. (Zhu *et al.* 2017) developed an ensemble of DBN (Deep Belief Networks) and SVM (support vector machine). They generated spectrograms from vocal utterances taken from the Chinese Academy of Sciences database. DBN was then used as a tool for extracting features from the spectrograms. They later used their proposed ensemble to classify the emotions achieving a recognition accuracy score of 95.8%.

Bonfiglio (Bonfiglio 2017) developed a deep learning inspired technique to recognise emotions from speech. They did this using CNN and LSTM. CNN was used to extract features from the generated spectrograms while LSTM was used to perform the ultimate classification. To validate the effectiveness of their proposed model they used the RECOLA (Remote Collaborative and Affective Interactions) database achieving a recognition accuracy of 74.1%.

Harár et al (Harár, Burget and Dutta 2017) developed a Stochastic Gradient Descent optimised Deep Neural Network for recognising emotion through speech. The converted vocal utterances to spectrograms and these audio files were taken from the EMO-DB speech emotion database. They then used the DNN to recognise the emotions achieving a recognition accuracy of 96.67%. The efficacy of the proposed model was evaluated across only three emotion states which are angry, neutral and sad.

Petridis et al (Petridis *et al.* 2018) presented BGRUs (Bidirectional Gated Recurrent Units) as an end-to-end model for recognising emotion from both audio and visual files. They converted audio files into spectrograms and extracted features from the spectrogram images using BGRUs. They also extracted features from video files using BGRUs. They then used BGRUs to recognise emotion using the fused features achieving a recognition accuracy of 98%.

Kim et al (Kim *et al.* 2018) presented a 3-dimensional convolutional network (3D CNN) deep learning algorithm as an end-to-end model for recognising speech emotion. They compared their proposed model with other classifiers and achieved a recognition accuracy score of 51.2%. They evaluated their method on several speech emotion databases which include RECOLA, EMO-DB, etc.

Yoon et al (Yoon *et al.* 2019) developed two Bi-directional Long Short-term memory (BLSTM) deep learning algorithms for extracting salient speech emotion features from spectrograms generated from audio files. They used BLSTM together with the extracted features to recognise the emotions. Using their proposed method, they achieved recognition accuracy of 76.5%

The literature in section 4.4 shows some intriguing results achieved from the use of spectrograms in recognising emotion from speech. It can be observed that applying deep classifiers in spectrograms yields some excellent results in recognising emotions. However, the main problem with this method is that spectrograms take time to generate (Wang *et al.* 2015; Zatarain *et al.* 2018; Yoon *et al.* 2019). Furthermore, to successfully apply deep learning techniques such as CNN is quite a resource-intensive exercise that requires state of art computing resources such as GPU processors which are very expensive to procure (Kahou *et al.* 2016). One of the aims of this research work is to develop a computationally inexpensive SER model that is time-sensitive which is ideal for customer call centres in developing countries such as those in Africa.

4.5 Hybrid Features

The extraction of appropriate features is the backbone of all SER systems (Avisado *et al.* 2012). It is believed that the performance of recognition algorithms in recognising emotion is hugely influenced by the features used (Mannepalli, Sastry and Suman 2018). Therefore, most researchers have attempted to fuse various speech features to come up with super features with high discriminative power (Dahake, Shaw and Malathi 2016). The systematic fusion of various features such as prosodic and spectral features is believed to boost the efficiency of SER models. Several studies on the fusion of speech features have been seen to improve the performance of speech recognition models. Some have even tried to combine text features with speech features to achieve higher overall accuracies (Williams and Mahmoud 2017). The section presents the work that has been done in recognising emotions using a combination of several speech features. The literature presented in this section dates as far as 2007 because very little research has been done on hybrid features over the past decade.

Rong et al. proposed a well-designed feature selector based on a combination of a decision tree and a random forest algorithm (Rong *et al.* 2007). They also pointed out that the well-known feature selection methods: Promising First Selection (PFS), Forward Selection (FS) and Principal Component Analysis (PCA) have a problem in that they need substantial training data sets. An automatic feature selector was developed based on an RF2TREE algorithm and the traditional C4.5 algorithm (Rong *et al.* 2007). They used these techniques to select a set of prosodic and spectral features to boost the performance of their speech recognition model and they achieved an overall accuracy of 80.58% using the random forest classifier.

An evaluation of feature selection frameworks using four feature selection algorithms and three different multiclass classifiers to detect speech emotion was proposed (Altun and Polat 2009). They reported that among all feature groups, the prosodic and sub-band energy features were the most frequently selected ones by all the algorithms in each framework. The highest accuracy recorded in their paper was 83.7% using SVM. In (Luengo, Navas and Hernaez 2010), a blend of 383 spectral, voice quality and prosodic features were extracted from the SAVEE audio files. Redundant features were removed using the minimal redundancy maximal relevance (mRMR) method achieving a recognition accuracy of 78%.

Han et al. proposed a novel approach (Han, Yu and Tashev 2014) where they used segment-level features such as Mel-frequency cepstral coefficients (MFCC), pitch period, and harmonic to noise ratio, and utterance-level features to recognise speech emotions. They used Deep neural networks (DNNs) to develop emotion probabilities in all the speech segments. The developed probabilities were then used to create the utterance-level features, which were fed to the ELM (Extreme Machine Learning) based recognition algorithm. They applied these techniques on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database (Busso *et al.* 2008) achieving an accuracy of 54.3%.

Sarker and Alam used a majority voting technique to recognise emotion from human speech using feature selection (Sarker and Alam 2014). The majority voting technique was applied over nine recognition algorithms which are Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). They used a combination of 16 low-level descriptors which include energy, Mel-Frequency cepstral coefficients (MFCC) (1-12), zero-crossing rate of time signal (frame-based), voicing probability computed from the ACF, and

fundamental frequency (F_0). The proposed system achieved an overall recognition of 84.19% on the Berlin Speech Database over four emotions such as angry, happy, neutral, sad.

Alonso et al combined pitch, spectral energy and prosodic features to recognise two classes i.e. high and low activation emotions using the SVM classifier (Alonso *et al.* 2015). They only considered four emotions and their method achieved 94.9% accuracy using the Berlin Emotion dataset. Bhaskar et al (Bhaskar, Sruthi and Nedungadi 2015) developed a novel approach for emotion recognition of audio conversation based on both speech (MFCC features) and text. They obtained 90% accuracy using the SemEval-2007 Database and SVM.

Muthusamy et al (Muthusamy, Polat and Yaacob 2015) proposed a SER model using entropy, glottal and wavelet packet energy features from three datasets i.e. Emo-DB, SAVEE and Sahand Emotional Speech (SES). These features were then enhanced through the use of the Gaussian mixture model (GMM). They obtained a maximum of 97.24% using Extreme Learning Machines ELM.

Wang et al (Wang *et al.* 2015) developed a SER model using ZCR, F_0, MFCC, Fourier parameter (FP) from three different datasets i.e. the Emo-DB, Chinese elderly emotional database (EESDB) and Chinese Emotional Speech Corpus (CASIA). They analysed six emotions which include: happiness, boredom, neutral, sad, angry, and anxiety achieving the accuracy of 92.92%, 71.48%, 87.46%, 91.21%, 98.29%, and 91.92% respectively using SVM. The average recognition accuracy has been reported as 88.88% using the Berlin Emotion dataset.

Kerkeni et al (Kerkeni *et al.* 2018) proposed an SER that uses Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the Berlin Speech Database and Spanish dataset. Their study showed that for the Berlin database all classifiers achieve an accuracy of 83% when a speaker normalization (SN) and a feature selection were applied to the features.

An attention-based CNNLSTM-DNN model was presented by Hifny and Ali (Hifny and Ali 2019) using the IEMOCAP dataset. Their experiment was conducted across five emotions namely neutral, happiness, sadness, surprise, questioning achieving a recognition accuracy score of 87.2%. In that same vein, (Ariav and Cohen 2019) used a WaveNet encoder to extract speech features from vocal utterances. Furthermore, they extracted emotion features from video files using residual

network (ResNet). They combined both the audio and video features using multimodal compact bilinear pooling (MCB). Their proposed model yielded a recognition accuracy of 91.52%.

A fusion of speech and facial features were used to develop a multimodal emotion recognition system (Prasada, Chandra and Hemanth 2019). MFCC features were combined with Maximally Stable Extremal Regions and this improved the recognition rate by 2 to 3% on the Indian Face Database and Berlin Speech Database. They obtained an overall accuracy of 85%.

Hybrid features have presented a good alternative to improve the performance of SER models. The literature in section 4.5 has shown that some of the attempts to fuse several speech features have yielded some interesting results. However, most of the models have failed to yield higher accuracies in recognising some emotional states such as fear (Nogueira *et al.* 2013; Li *et al.* 2016). This means that some of these features have carried along with them certain recognition weaknesses. The research gap in this area is to find a statistical solution to combine speech features without inheriting their weaknesses such as their poor ability to recognise some emotional states. The solutions to this research gap are expected to answer the research question of this thesis which calls for the identification of the most discriminative features across various types of speech features. Some of the hybrid features have yielded excellent results but the combined features have increased the dimensionality of the features. This often prolongs the processing time which defeats part of this thesis's research agenda which is to develop a model that processes emotion on time.

4.6 Cross-Language Acoustic Emotional Valence

Developing cross-corpus speech emotion recognition has recently become a growing area of interest in speech processing. Researchers in this area have proposed several solutions to improve cross-corpus speech emotion recognition and these include feature normalization (Zhang *et al.* 2011) and decision fusion (Schuller *et al.* 2011). However, most of the research done in speech emotion recognition has been primarily centered on single languages. Relatively very few studies have been done in cross-corpus emotion recognition. Polzehl et al (Polzehl, Schmitt and Metze 2010) proposed a set of optimal features to recognize anger using the cross-corpus approach.

In developing their proposed cross-corpus speech emotion recognition model, Latif et al used four speech emotion corpora (Latif *et al.* 2019). The corpora used in their experimental study are

EMOVO, Urdu, SAVEE (Surrey Audio-Visual Expressed Emotion), EMO-DB (Berlin Database of Emotional Speech) and these constituted of utterances spoken in English, Italian, German and Urdu. They classified the resulting emotions into two classes which are positive valence and negative valence. These two classes of emotions were classified using eGeMAPS (Geneva Minimalistic Acoustic Parameter Set) and SVM (Support Vector Machines) achieving an overall accuracy of 70.98%. The same authors refined their original model using eGeMAPS, DBN (Deep Belief Network) and the Leave-One-Out technique. Using the Leave-One-Out, they trained their model using IEMOCAP (Interactive Emotional Dyadic Motion Capture), EMO-DB, EMOVO and SAVEE. They then used the FAU-AIBO database to evaluate their model achieving an overall accuracy of 80% (Latif *et al.* 2018).

Using clean spectrograms and the DSCNN (Deep stride convolutional neural network) classifier, Mustaqeem et al (Mustaqeem and Kwon 2020) achieved an overall accuracy of 56.5%. They used converted audio files from RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and IEMOCAP speech emotion databases into spectrograms. DSCNN was then used to extract speech emotion features from the spectrograms and the same algorithm was used to make the ultimate classifications. The classification was done across four emotion states which include anger, happy, neutral, and sad. The DALSR (Domain-adaptive least squares regression) and SVM classifier were presented in (Liu *et al.* 2018). This was done using two emotion corpora which are eNTERFACE and Emo-DB. In their experimental study, six emotion states of surprise, disgust, sadness, happiness, fear, and angry were used achieving an unweighted average accuracy (UAR) of 52.27%. Li et al developed a cross-corpus speech emotion recognition model using a combination of combined MSF (Modulation spectral features) and IS16 features (Li and Akagi 2019). In a bid to solve the cross-language issues experienced in speech emotion processing, they applied the LMT (logistic model trees) classifier on four emotion states which include neutral, happy, angry, and sad. To evaluate the efficacy of their proposed model, they used Fujitsu, EMO-DB and CASIA (Chinese emotional database). Their model yielded an F1- score of 82.63%.

In (Parry *et al.* 2019), six speech emotion corpora were used to create a robust cross-language speech emotion recognition model. The model was validated on six emotion corpora which include RAVDESS, IEMOCAP, EMO-DB, EMOVO, SAVEE and EPST (Emotional. Prosody Speech and Transcripts). They mapped the emotion classes in these corpora into three distinct groups which

are neutral, negative, and positive. Deng et al used A-DAE (adaptive denoising autoencoders) on INTERSPEECH 2009 Emotion Challenge baseline feature set to develop their cross-language speech emotion recognition model (Deng *et al.* 2014). They evaluated the efficacy of their model on two German speech emotion databases which are FAU AEC and ABC (Airplane behaviours Corpus). Consequently, they achieved an unweighted average recall of 64.18%.

4.7 Chapter Summary

This chapter presented reviews of the application of prosodic, spectral, spectrograms and hybrid features in SER models paying attention to the requirements of customer call centres. It has been noted that SER models designed specifically for customer call centres should be able to recognise all the required emotion states with higher recognition accuracies. From the literature, it can be observed that using MFCC features alone is not good enough to achieve higher recognition accuracies for the fear emotion state. It was also observed that prosodic features yield low recognition accuracies which are not ideal for customer call centre environments. In the same vein, the review showed that most Hybrid features are relatively poor in recognising the fear emotion. Semwal et al. (Semwal, Kumar and Narayanan 2017) combined MFCC, energy, ZCR and F_0 features to develop a SER model. Their proposal recognised fear with 77% recognition accuracy. Likewise, Sun et al. (Sun *et al.* 2019) used deep learning autogenerated to develop an improved SER model. Their proposal recognised fear with 62.5% recognition accuracy.

Furthermore, the literature has shown that little research has been done on cross-language speech emotion recognition even though the research is now gaining traction. In this body of work, the researcher observed that most of the work reported in this area is mainly centered on the amalgamation of three languages and the submissions presented need further improvement. The results can be improved since the highest benchmarked performance reported in the literature is a recognition accuracy of 80.00% (Latif *et al.* 2018).

The chapter also revealed that the use of spectrograms together with deep classifiers achieves excellent results in speech emotion recognition. However, it was observed that the generation and analysis of spectrograms are more taxing in terms of time. Furthermore, deep learning with images

requires more processing power which is very expensive if the model is to be developed for developing countries. It was also observed that most experimental studies in SER do not develop SER models for multilingual environments (cross-language speech emotion recognition). There is very little work done in developing SER for multilingual environments and from the little that has been done during the past six years we observed that very low recognition rates were scored. This problem stimulated the research question for this body of work as mentioned in chapter one because features are the main contributing factors in classifying speech (Ayadi, Kamel and Karray 2011; Yan *et al.* 2013; Song *et al.* 2017). The presentation on hybrid features showed that some of them tend to increase in dimension which results in longer processing time which will also stimulate the need for expensive computing resources that have more processing power. Therefore, the next chapter presents the design used to cover the research gap unearthed by this chapter.

Chapter 5: Research Methodology

5.1 Introduction

This chapter presents the methodological steps followed in this experimental study to answer the research question and fulfill the objectives highlighted in chapter 1. The description of the essential components used in this experimental study is divided into six sections. The reasons for choosing the proposed emotion model are explained in section 5.2. Speech emotion corpora are described in section 5.3. Spectrogram generation is elaborated in section 5.4 while features extraction is described in section 5.5. The feature selection process is described in section 5.6. Ensemble and inducer classifiers are presented in sections 5.7 and 5.8 respectively. Deep learning classifiers are presented in section 5.9 while the performance benchmarks used in this body of work are described in section 5.10. The chapter is concluded with a summary in section 5.11.

Eight specific emotions are chosen to conduct this experimental study, and these are inspired by Ekman's model (Ekman 1992) described in chapter 2. These emotions include anger, calm, happiness, sadness, surprise, disgust, fear, and neutrality. Three speech emotion corpora are used to simulate customer call centre audio recordings, and these include EMODB, RAVDESS, EMOVO, CREMA-D, and SAVEE. These emotion corpora are chosen for this study because they constitute all the emotion classes in their respective speech files. A pre-processing exercise is done to covert the audio files into spectrograms. Spectrograms were generated to extract deep learning auto-generated features. Therefore, a custom 2D-CNN deep classifier is used to extract the deep learning auto-generated features from the spectrograms. In addition, Deep Radial Basis Neural Network and Deep Multilayer Perceptron algorithms are also used to extract the autogenerated features for comparison's sake. Prosodic and spectral features are also extracted from the raw audio files using jAudio, an audio feature extraction software. Since the audio extracted features are numerous, a feature selection exercise is done. This is done using knowledge of the performance of these features from other works to select the most prominent features in recognising emotion. This chapter also presents the inducers, ensemble classifiers, and deep classifiers used to classify the emotions. The recognition is done two-fold to fully validate the strength of the proposed feature

set. Therefore, deep learning auto-generated features are compared with combined prosodic and spectral features. Moreover, this experimental setup is done to answer the research question posed in chapter 1. Additionally, a set of performance benchmarks chosen for the valuation of the proposed model is presented in this chapter. The flow diagram of the methodology is illustrated in Figure 14.

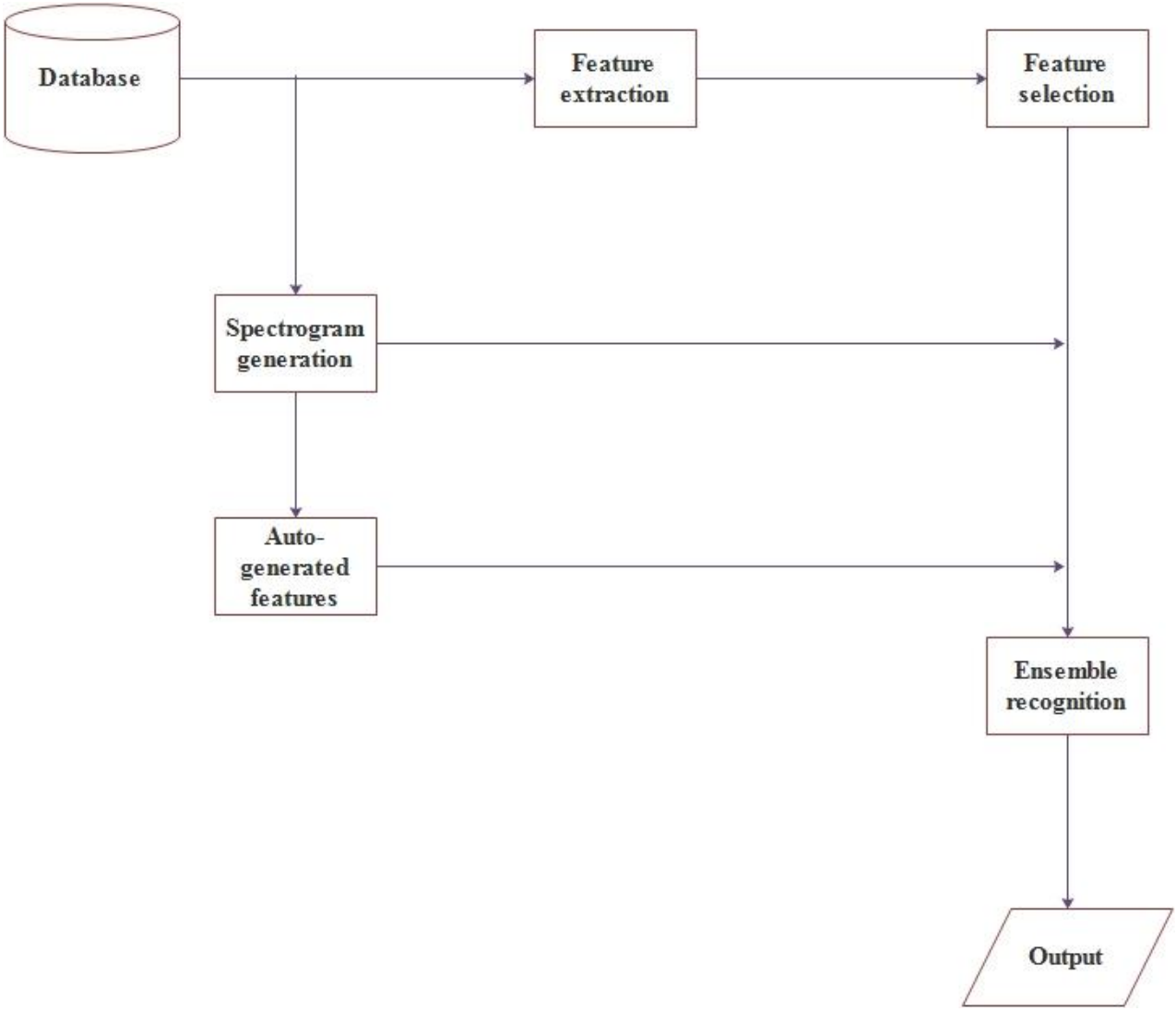


Figure 14: The architecture of the proposed research methods

5.2 Emotion Model

The literature review done in chapter 2 inspired the use of basic emotions as envisaged by Ekman (Ekman 1992). In this research work, Ekman's basic emotions were used together with universal emotions. The basic emotions have been identified as anger, disgust, fear, happiness, sadness &

surprise. These basic emotions were considered for this research because it has been adequately proven that these are universally recognised across all human cultures (Ekman 1992; Tracy and Randles 2011).

Since the proposed SER model is primarily designed for customer call centres, two additional universal emotions were added, and these are calm and neutral. The calm emotion state was chosen because, in a call centre environment, some customers express their concerns in a calm manner, and also agents are expected to be as calm as possible (Lee and Narayanan 2005; Pappas, Androutsopoulos and Papageorgiou 2016; Cong *et al.* 2017; Yu *et al.* 2017). Neutral was also identified as a very important emotional state for customer call centres because there are some circumstances where customers and agents exhibit emotions that are neither positive nor negative (Feinberg *et al.* 2002; Alam, Danieli and Riccardi 2016). Therefore, this thesis argues that any SER model designed for customer call centres must address the six basic emotions and universal emotions (calm and neutral). The emotion models used to recognise valence in this body of work were inspired by Shavers et al. and their model is described in section 2.4.5 of chapter 2 (Shaver, Murdaya and Fraley 2001).

5.3 Databases

To fully validate the robustness and consistency of the proposed SER model there was a need to use standard emotion corpora. As mentioned in the previous section, five emotion corpora were used in this thesis and these include the Berlin Database of Emotional Speech (EMO-DB) (Burkhardt *et al.* 2005), Ryerson Audio-Visual Database of Emotional Speech (Livingstone and Russo, 2018), and Song (RAVDESS), Surrey Audio-Visual Expressed Emotion (Haq and Jackson 2009), EMOVO Italian language database (Costantini *et al*. 2014) and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) (Sierra *et al.* 2015). These emotion corpora were carefully chosen because they comprise vocal utterances that cover most of the emotion states selected for this research work. Additionally, the emotional corpora were chosen because they comprise vocal utterances from three different countries which are Germany, England, and the United States of America.

5.3.1 Berlin database of emotional speech (EMO-DB)

The Berlin Database of Emotional Speech (EMO-DB) (Burkhardt *et al.* 2005) is an emotional speech corpus that comprises 535 German vocal utterances as shown in Figure 15. It is an acted emotion speech corpus that consists of 7 various emotions which are anger, joy, sadness, neutral, boredom, disgust, and fear. The corpus comprises 5 male professional native German-speaking actors and 5 female professional native German-speaking actors. An anechoic chamber was used to record these simulations using state-of-the-art recording equipment. The resulting recordings were produced at a sampling rate of 16 kHz with a 16-bit resolution and mono channel. The version of the EMO-DB corpus used in this study comprises 340 audio files. Each of the audio files is approximately 3 seconds long.

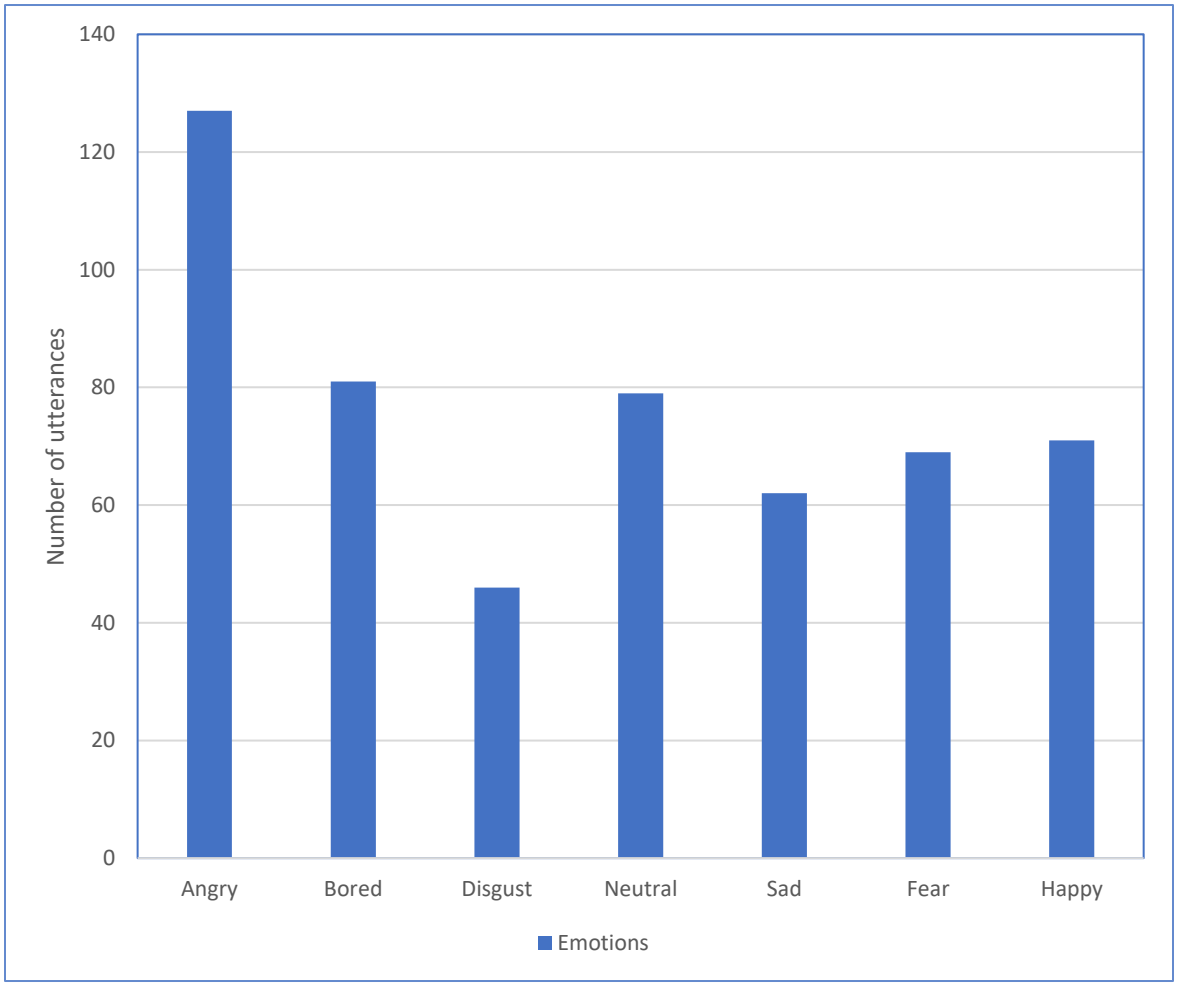


Figure 15: Berlin Database of Emotional Speech corpus

5.3.2 Ryerson audio-visual database of emotional speech and song (RAVDESS)

The RAVDESS corpus (Livingstone and Russo 2018) is an emotion corpus that consists of emotional speeches and songs. The corpus comprises 12 professional male actors and 12 female professional male actors speaking in a North American accent. The RAVDESS emotion corpus comprises eight emotion classes. These classes include angry, happy, neutral, calm, sad, surprised, fearful, and disgust expressions as depicted by Figure 16. The recorded files consist of songs (in audio and video form) and speech (in audio and video form). Only the audio speech files were used in this research since the purpose of the study was to develop a SER model appropriate for customer call centres. The RAVDESS emotion corpus is relatively new in comparison with other emotion corpora.

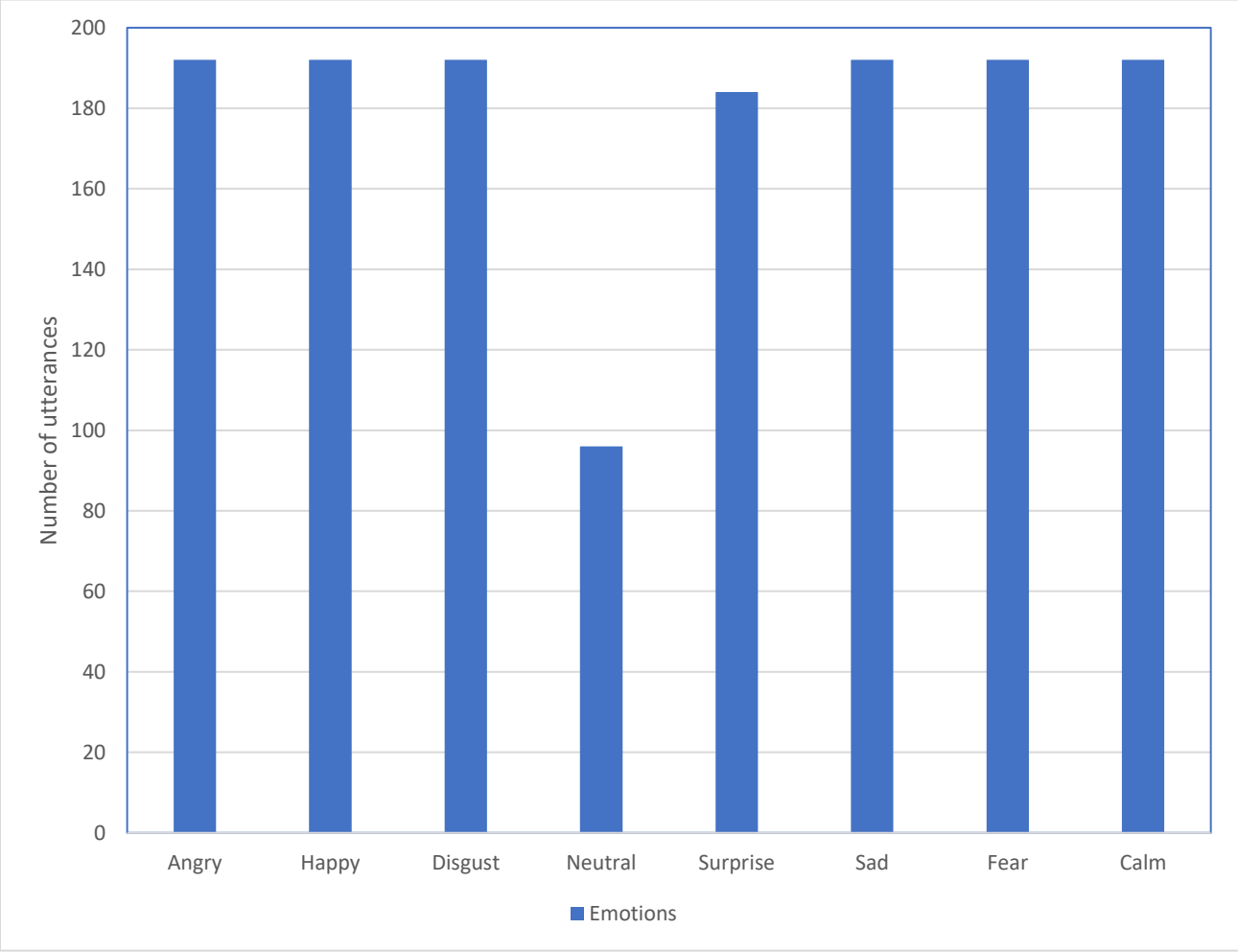


Figure 16: Ryerson Audio-Visual Database of Emotional Speech (RAVDESS) corpus

5.3.3 Surrey audio-visual expressed emotion (SAVEE)

The SAVEE corpus (Haq and Jackson 2009) is an emotion corpus that comprises 480 speech recordings from 4 male professional actors. These recordings were done across seven different emotional states (Wang *et al.* 2010). Unlike the RAVDESS emotion corpus, the speech SAVEE emotion corpus comprises 480 professionally recorded British English vocal utterances. The utterances were recorded using high-quality equipment in a state-of-the-art media lab. The corpus consists of 7 emotion classes which are angry, happy, disgust, neutral, sad, fear, and surprised emotional expressions as shown in Figure 17. A high number of recordings were done in the neutral class (120 audio files) while angry, happy, disgust, sad, fear and surprised emotion expressions had 60 audio files each.

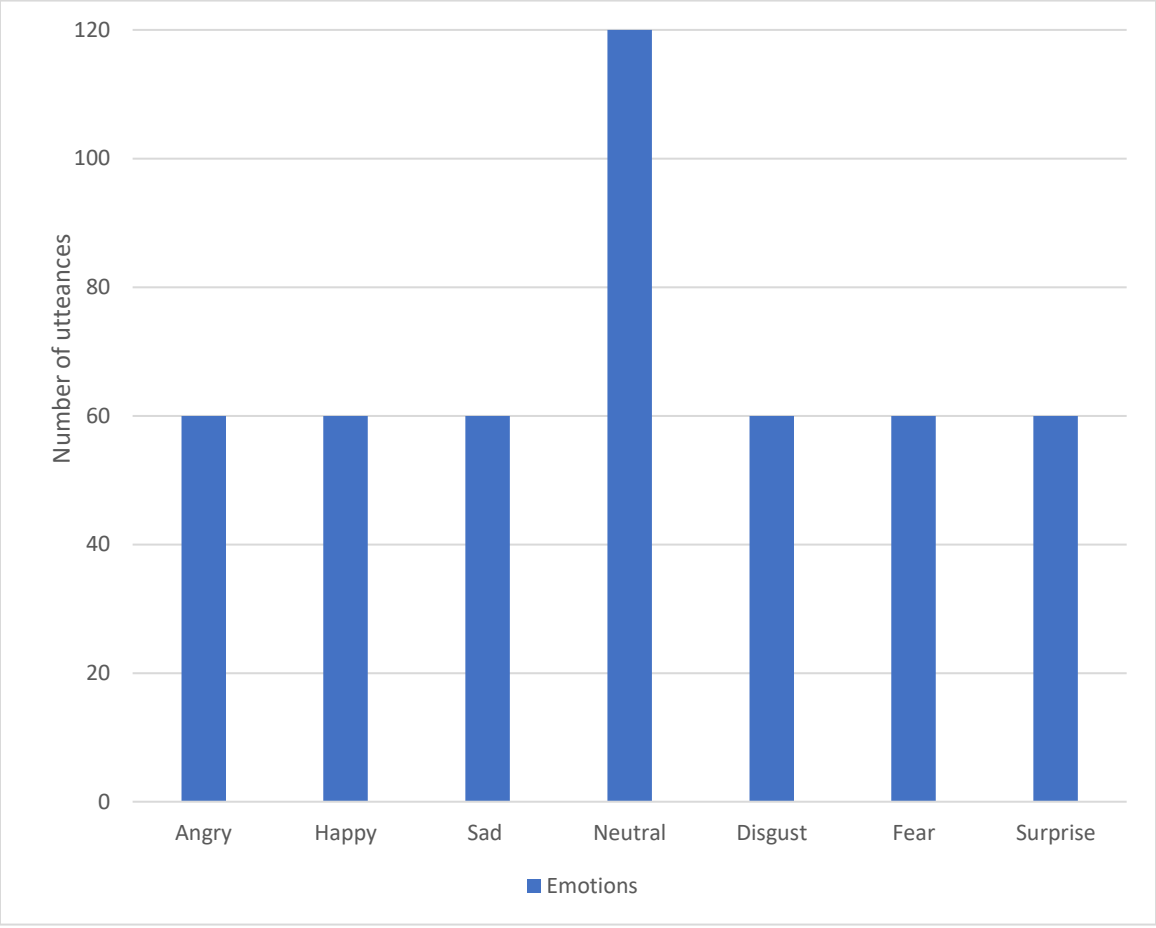


Figure 17: Surrey Audio-Visual Expressed Emotion (SAVEE) corpus

5.3.4 Italian emotional speech dataset - EMOVO

EMOVO is a speech emotion database that comprises vocal utterances from the Italian language (Costantini *et al.* 2014). In developing this database, a group of six professional actors was recorded to simulate seven emotion states which are (disgust, fear, anger, joy, surprise, sadness, and neutral) in 14 simple sentences. This is a gender-balanced speech database because the utterances were recorded from three female actors and three male actors. The vocal utterances were recorded using professional equipment in the Fondazione Ugo Bordoni laboratories in Rome. EMOVO is the first speech emotion database to be developed using the Italian language (Costantini *et al.* 2014). The experiment consisted of 12 subject recognisers who had a mammoth task of recognising the correct emotions from the vocal utterances. The resulting recordings were produced at a sampling rate of 48 kHz with a 16-bit resolution and mono channel. Therefore, a total of 588 wav files were recorded from the experiment. As shown in Figure 18, EMOVO consists of 84 speech files for each of the emotions.

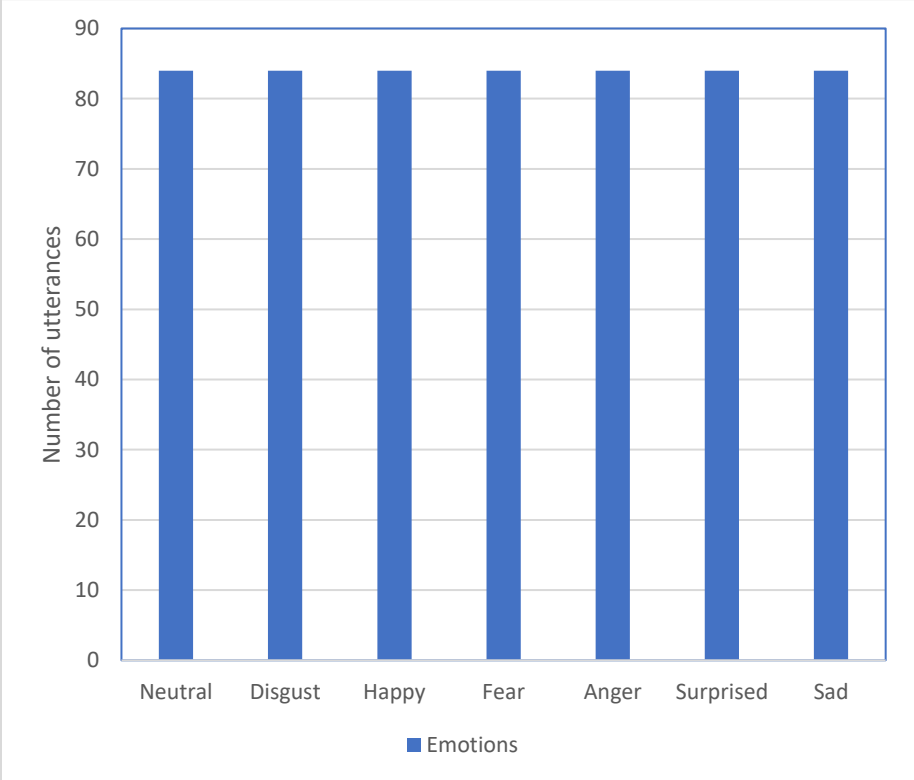


Figure 18: EMOVO corpus

5.3.5 Crowd-sourced emotional multimodal actor dataset (CREMA-D)

The CREMA-D corpus is an English multi-accent speech emotion corpus. The speech emotion database was developed using the utterances of 91 professional actors. The professional actors were carefully chosen from an assortment of various races and ethnic backgrounds (African American, Asian, Caucasian, Hispanic, and Unspecified) (Sierra *et al.* 2015). The group of actors comprised 43 females and 48 males. Therefore, the database comprises 7442 audio-visual clips. The professional actors involved in the development of the corpus were between the ages of 20 and 74. The utterances were recorded from 12 sentences and only the audio files were used in this experimental study because the study aimed to recognise emotion from speech. The database consists of six emotion states (neutral, disgust, happy, fear, anger, and sad as shown in Figure 19.

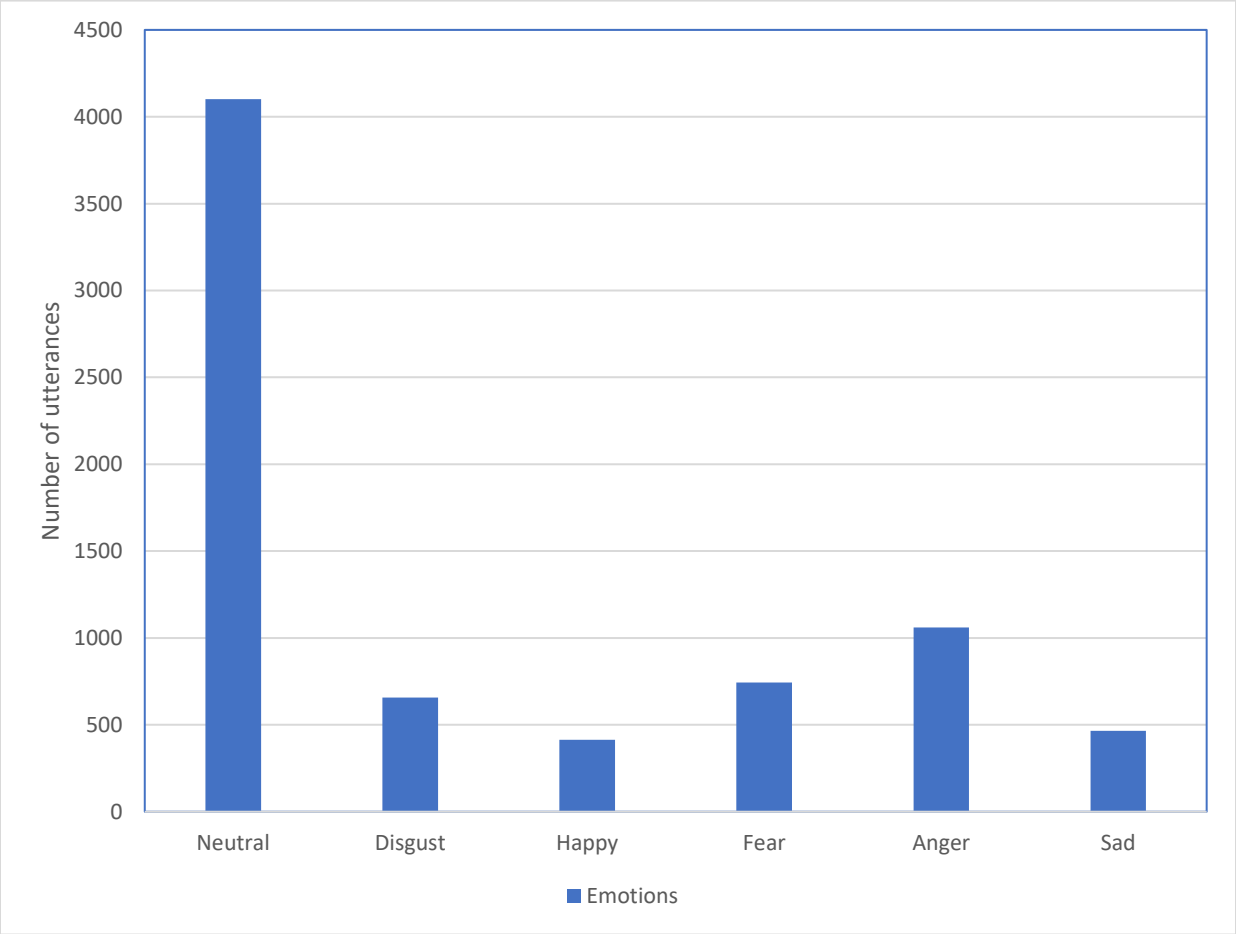


Figure 19: CREMA-D corpus

5.3.6 Combined Speech emotion corpus

The EMODB, RAVDESS, SAVEE, CREMA-D, and EMOVO emotion corpora were fused to develop a single emotion corpus which we refer to as the combined speech emotion corpus in this research. Each distinct class in the selected corpora was classified into two classes that is positive valence and negative valence. The systematic recognition was done using approaches proposed in (Latif *et al.* 2018, 2019; Parry *et al.* 2019) as illustrated in Table 4.

Therefore, anger, sadness, fear, disgust, and boredom audio speech files were grouped in the negative valence category to resemble dissatisfaction. On the other hand, the neutral, happiness, surprise, and calm speech audio files were then grouped in the positive valence category to resemble satisfaction. Consequently, the combined multilingual corpus had 6641 positive instances and 3453 negative instances shown in Figure 20.

Table 4: Mappings of emotions into Positive and Negative Valence

Corpus	Language	Age	Gender	Negative Valence	Positive Valence
Emo-Db	German	Adults	Males and Females	Anger, Sadness, Fear, Disgust, Boredom	Neutral, Happiness
SAVEE	British English	Adults	Males	Anger, Sadness, Fear, Disgust	Neutral, Happiness, Surprise
RAVDESS	North American English	Adults	Males and Females	Anger, Sadness, Fear, Disgust	Neutral, Happiness, Surprise, Calm
CREMA-D	North American English	Adults	Males and Females	Anger, Sadness, Fear, Disgust	Neutral, Happiness
EMOVO	Italian	Adult	Males and Females	Anger, Sadness, Fear, Disgust	Neutral, Happiness, Surprise

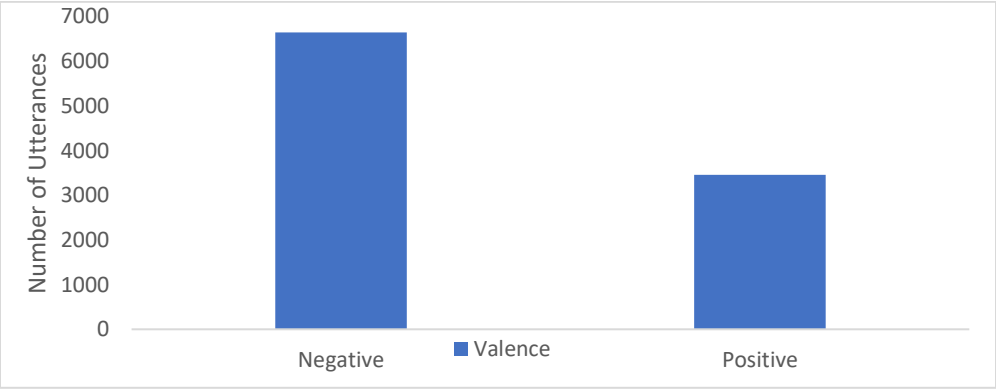
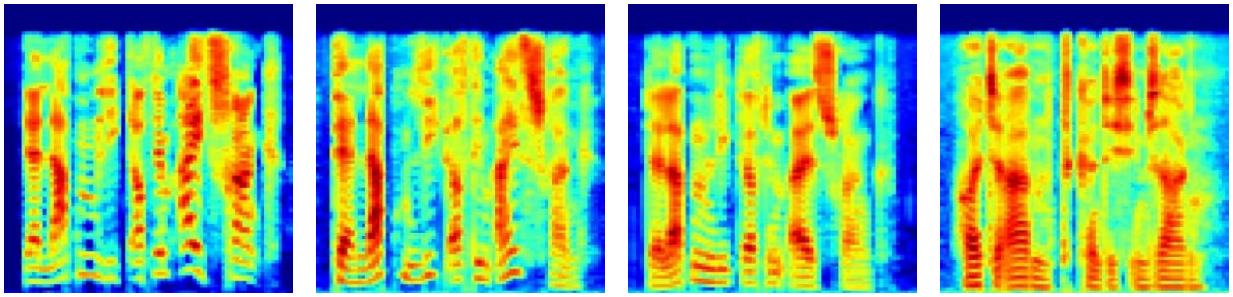


Figure 20: Combined speech emotion corpus

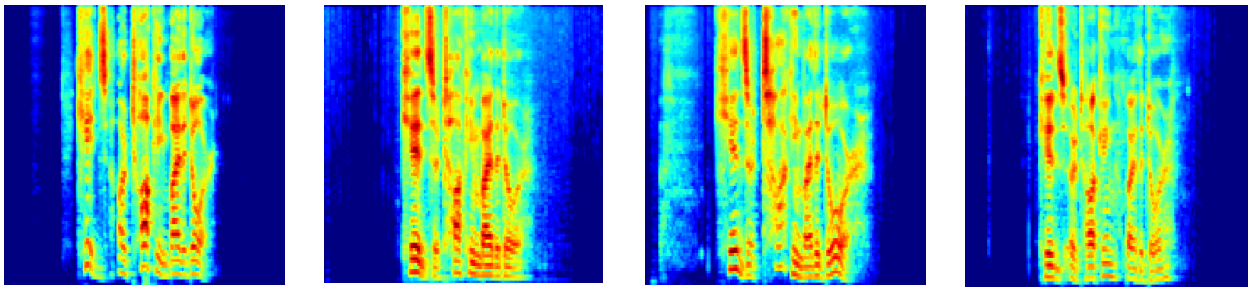
5.4 Spectrogram Generation

Spectrograms are very important in speech processing because they contain valuable speech content such as energy and formant (Cairong *et al.* 2016; Stolar *et al.* 2018; Weißkirchen, Böck and Wendemuth 2018; Z. Liu *et al.* 2018). Figures 18,19 and 20 illustrate samples of the converted spectrograms from the three databases across all emotions.

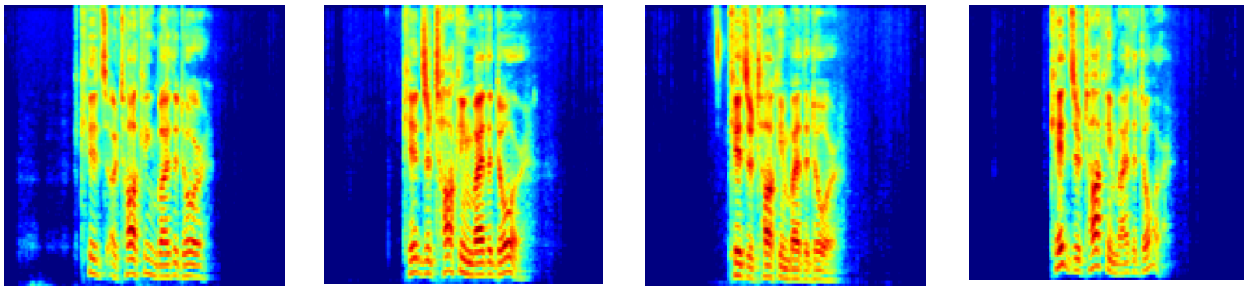


(a) (b) (c) (d)

Figure 21: EMO-DB spectrograms showing four emotion states.

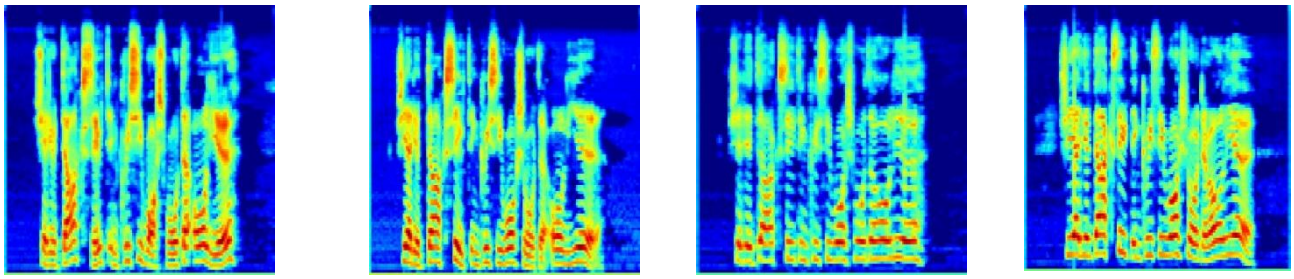


(a)Angry (b)Calm (c)Disgust (d)Sad

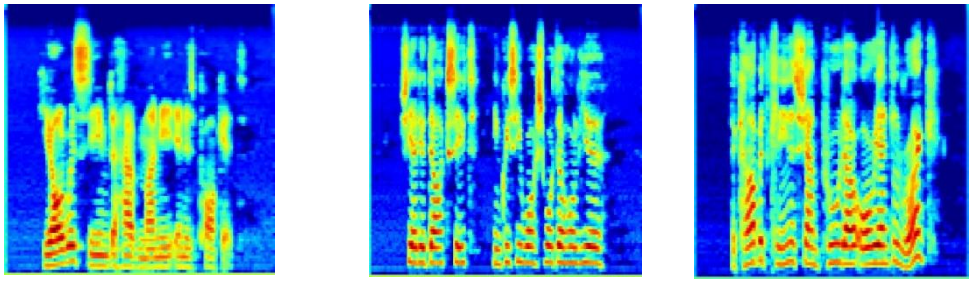


(e)Fearful (f)Happy (g)Neutral (h)Surprised

Figure 22: RAVDESS spectrograms



(b) Angry (b) Disgust (c) Fearful (d) Happy



(e) Neutral (f) Sad (g) Surprised

Figure 23: SAVEE spectrograms

The spectrogram is generated using Librosa (McFee *et al.* 2015). Librosa is a speech processing python package that can be used to generate spectrograms from raw audio files. The formula for calculating spectrograms is shown in equation 5.1.

$$L = |Y| = \left| \sum_{n=0}^{N-1} s(n)\, \omega(n)\, e^{-j(2\pi/N)kn} \right|, \quad k \in [0, N] \tag{5.1}$$

where $s(n)$ is the input signal, $\omega(n)$ resembles the humming window function while N stands for the length of the window.

Deep learning models are applied to extract affect-salient features from images, and this is done automatically without much human intervention. This technique has become popular since it reduces the burden involved in selecting the best features to perform a particular task. Auto-generated features are extracted from spectrograms using deep classifiers. 2D-CNN was used to extract these features so that they can be fed as input to ensemble classifiers for recognition. These features are referred to as deep learning auto-generated features in this body of work.

5.5 Feature Extraction

When humans communicate verbally, speech signals are produced. These speech signals carry an enormous volume of valuable information that can be used to recognise emotion, speaker identity, age, and gender. This information is extracted using various future extraction techniques. Therefore, feature extraction is an essential technique used to extract valuable insights from speech signals. It is one of the most important steps in speech emotion recognition. SER models should be fed with appropriate speech features as input since a classifier is deemed to be as good as the features used (Yogesh *et al.* 2017a).

Various tools can be used to extract trustworthy speech features from audio files. The list of many feature extraction toolkits includes the Hidden Markov Model Toolkit (HTK), the PRAAT Software, MATLAB, Python packages (PyAudioAnalysis and Librosa), and the OpenSMILE (Munich Open Speech and Music Interpretation by Large Space Extraction) toolkit. All these above-mentioned toolkits are open source which brings a lot of advantages for the research community such as rapid model development. From the evaluation that was done on the feature extraction toolkits, the jaudio toolkit was proposed in this thesis. Jaudio was proposed not just because it is open source but because it is a fast real-time and efficient open-source audio feature extractor (McEnnis *et al.* 2005). Furthermore, jaudio was chosen because it unites feature extraction algorithms from the speech processing and Music Information Retrieval communities and supports the rapid real-time extraction of relevant LLDs (Low-Level Descriptors) and functionals.

JAudio is a speech processing software that comprises advanced audio feature extraction algorithms designed. The software automatically computes the emotion features from audio files with minimal effort. It is one of the best feature extraction toolkits because of its unique capacity for handling multidimensional features.

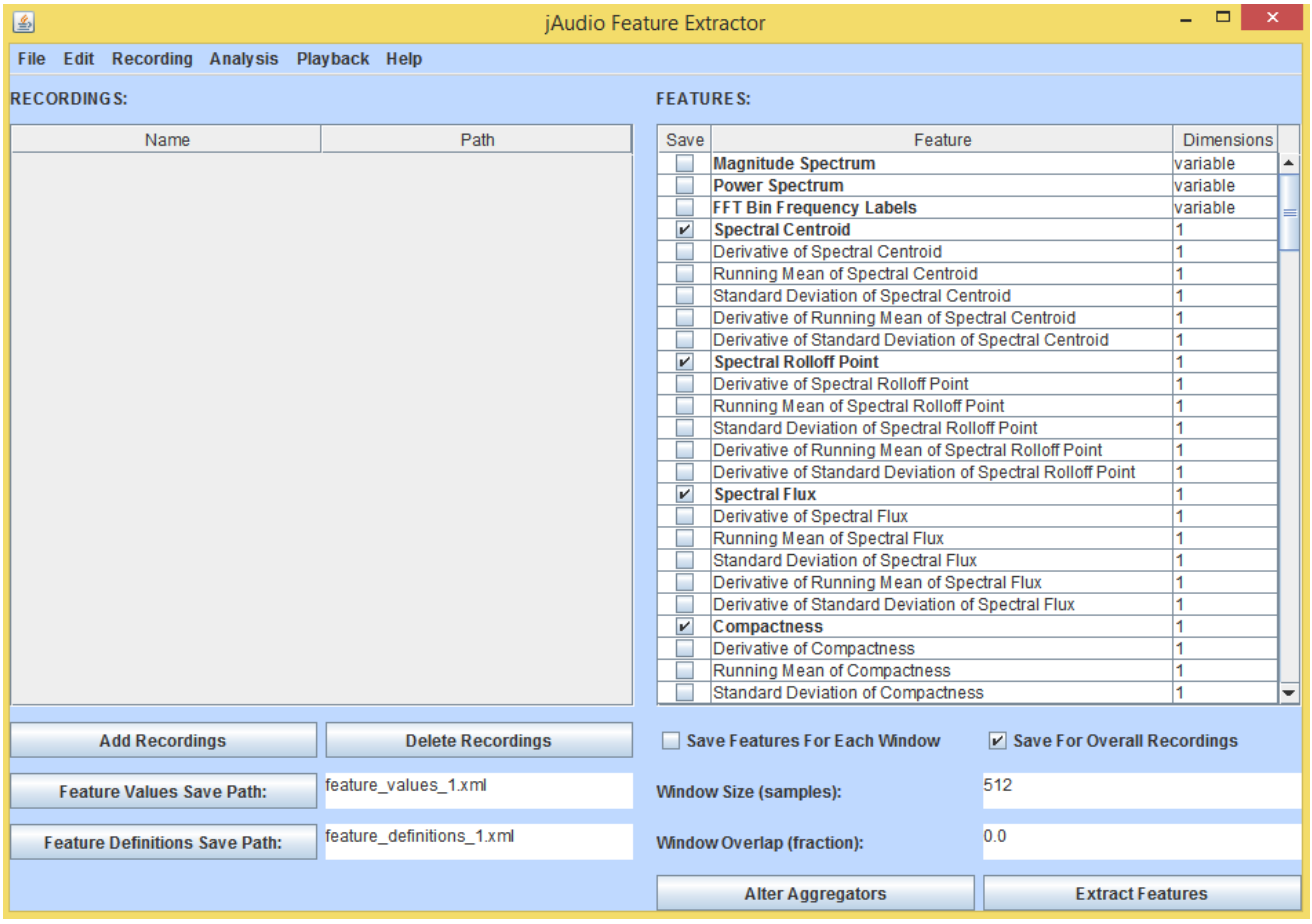


Figure 24: jaudio feature extraction

Therefore, prosodic and spectral features were extracted as shown in Figure 24. These features were extracted together with their statistical attributes such as mean, standard deviation, and the like. The features were extracted in all the three emotion corpora are EMO-DB (Burkhardt *et al.* 2005), RAVDESS (Livingstone and Russo 2018), and SAVEE (Haq and Jackson 2009).

5.6 Feature Selection

Feature selection is the process of selecting highly discriminative features from a group of features (Yan *et al.* 2013; Liu *et al.* 2015). It is sometimes referred to as a process that aids the dimensionality reduction process (Ayadi, Kamel, and Karray 2011). The reduction of dimensionality of features has been successfully applied in various tasks (Fewzee and Karray 2012; Arowolo *et al.* 2021). Feature selection is one of the main concepts of machine learning

which hugely influences the performance of recognition models (Basu *et al.* 2017). The process of feature selection was done in this research work because irrelevant or partially relevant features can negatively impact model performance. Some of the reasons that inspired the use of feature selection as a pre-processing tool in this research work include the following (Selvaraj, Bhuvana and Padmaja 2016; Basu *et al.* 2017; Badshah *et al.* 2019):

- It reduces overfitting.
- It improves accuracy.
- It reduces training time.

In line with the reasons mentioned above, feature selection was specifically chosen to fulfill the aim of this research work which is to develop a computationally inexpensive speech recognition model that is fast and accurate. An analysis of the most discriminative features from various research articles was carried out.

In machine learning, classifiers are believed to be as good as recognition features. Therefore, the main goal of this experiment was to develop an optimal dataset using the most relevant features in recognising speech emotions. The application of statistical feature selection algorithms such as PCA and LDA has proven to be an effective technique in reducing dimensionality as well as boosting the performance of recognition models. However, dependency on such techniques adds some computational costs because running such algorithms takes time.

Furthermore, a lot of researchers in the literature have proposed various combinations of features to develop robust SER models. As far as overall accuracy is concerned, some of the proposed models achieved excellent results. However, the models could not recognise the fear emotion state with higher accuracy and precision scores. Therefore, in this experimental study, a set of features were judiciously chosen using brute force and historical knowledge of the performance of such features in other application domains such as recognition of music (Agostini, Longari, and Pollastri 2003; Burger *et al.* 2013).

In this body of work, these features were chosen from the pool of acoustic features described in chapter 3. The exercise resulted in the development of 404 hybrid features (Zvarevashe and Olugbara 2020a) illustrated in Table 4. Therefore, a brute force exercise was done to come up with

the most discriminating features. The flow diagram showing the processes used to develop the HAF features is illustrated if Figure 25.

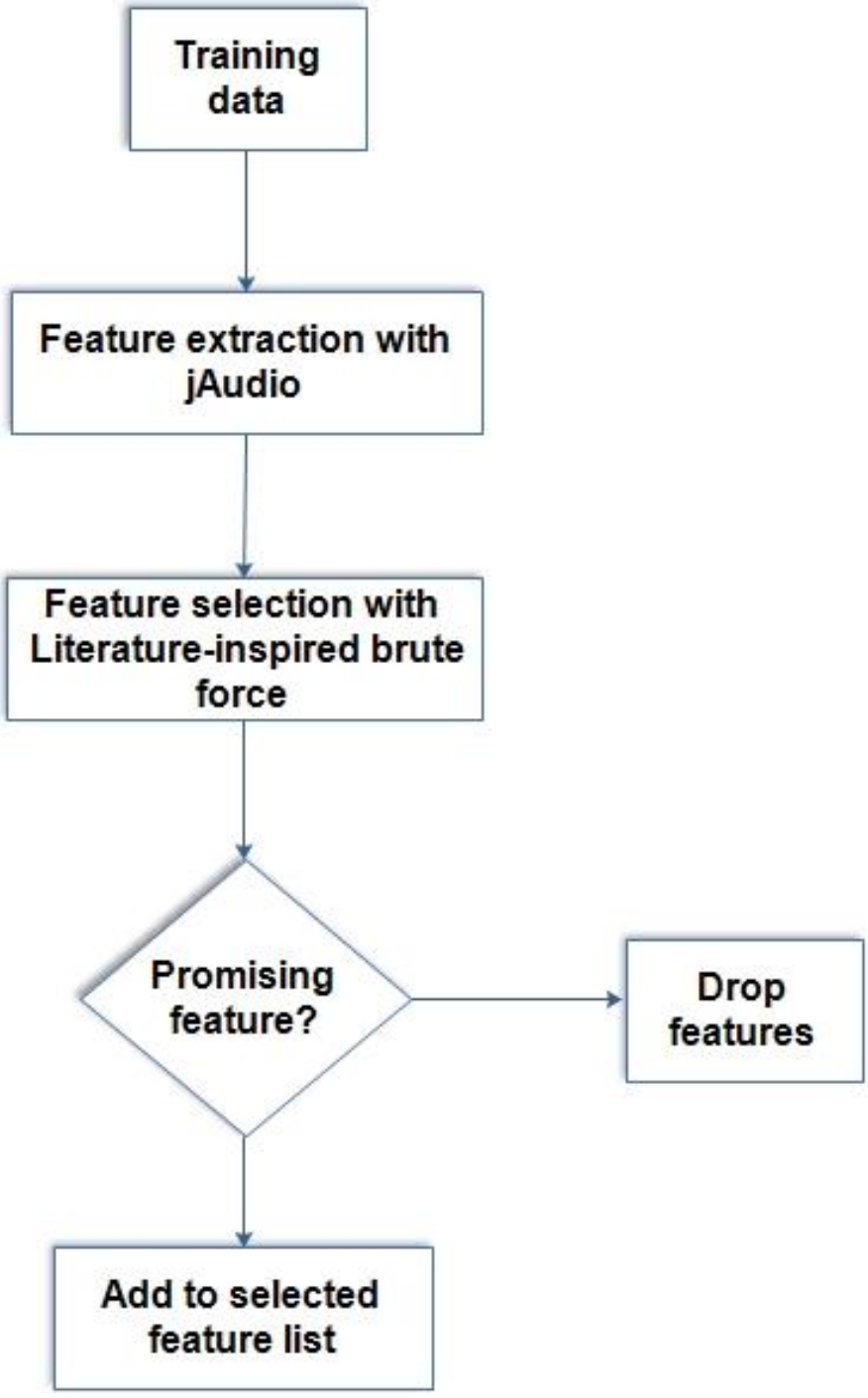


Figure 25:HAF selection process flow diagram

Table 5: A summary of HAF features

Group	Type	Number
Prosodic		
Energy	Logarithm of Energy	10
Pitch	Fundamental Frequency	70
Times	Zero Crossing Rate	24
Spectral		
Cepstral	MFCC	133
Shape	Spectral Roll-off Point	12
Amplitude	Spectral Flux	12
Moment	Spectral Centroid	22
Audio	Spectral Compactness	10
Frequency	Fast Fourier Transform	9
Signature	Spectral Variability	21
Envelope	LPCC	81
Total		**404**

The researcher observed that some of the work reported in literature achieved lower accuracy scores and had high processing times. Such model designs may be problematic if the models were to be deployed commercially (Nguyen, Huang, and Nguyen 2015; So*ng et al*. 2017). Moreover, these problems would persist more in multilingual environments because people express their satisfaction and dissatisfaction in so many different ways across various languages, cultures, and accents. Consequently, the Random Forest Recursive Feature Elimination (RF-RFE) algorithm was used to filter redundant features thus developing cross-corpus features. RF-RFE was chosen for this task because it was successful in the work presented in (Zvarevashe and Olugbara 2018b). The flow diagram of the processes involved in the creation of the cross-corpus features is illustrated in Figure 26. HAF features were extracted from the combined speech emotion corpus using jAudio. Therefore, RF-RFE was used to select the most discriminating features from the 404 HAF features.

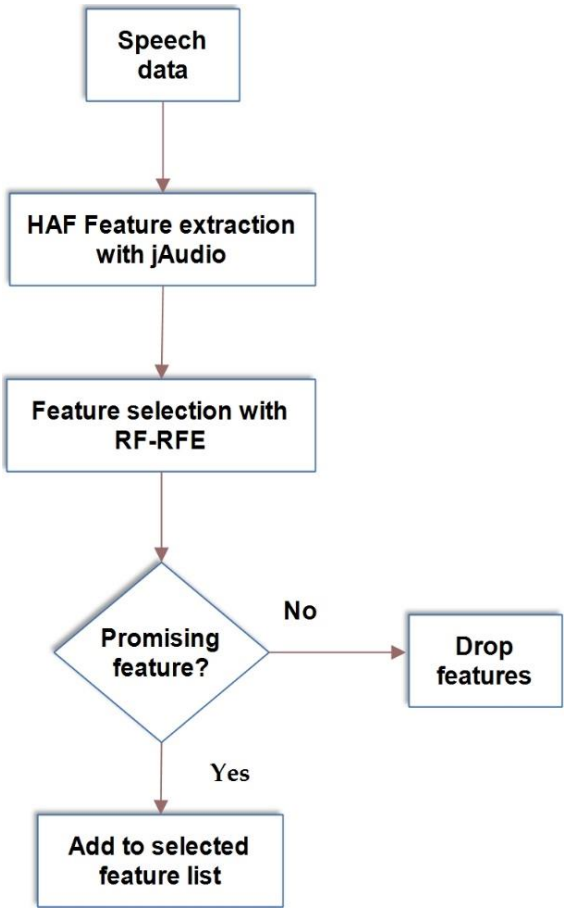


Figure 26: Flowchart for the creation of cross-corpus features

The concept of random feature elimination was proposed to improve the performance of SVM models (Guyon *et al.* 2013) and has been proposed as a solution to various recognition problems (Granitto *et al.* 2006; Xie, Hu and Yu 2006; Liu *et al.* 2015). The process involves the ranking of features and then removing the weakest features from a given dataset (Guyon *et al.* 2013; Darst, Malecki and Engelman 2018). The concept has been seen to thrive in correlated features (Gregorutti, Michel and Saint-Pierre 2017) and has been applied to Random forest algorithms to solve several problems (Granitto *et al.* 2006; Darst, Malecki and Engelman 2018; Zvarevashe and Olugbara 2018b; Bahl *et al.* 2019). The features used in this body of work are of a high dimension and this hurts the performance of a proposed model (Ayadi, Kamel and Karray 2011; Song *et al.* 2017; Song 2019). This is the reason why feature selection based on feature importance is essential before developing a recognition model (Genuer, Poggi, and Tuleau 2010). Based on this line of thinking, RF-RFE was used to rank the most discriminating features while removing the lowly ranked weak features in this body of work. The flow diagram which depicts the iterative cycle

involved in RF-RFE is illustrated in Figures 27 and 28. The feature selection exercise was done using the RF-RFE feature ranking algorithm on a combined database (EMOVO, RAVDESS, SAVEE, EMO-DB, and CREMA-D). Consequently, a group of 250 highly discriminating features was chosen. Twenty-five of the top-ranked features is presented in Figure 29.

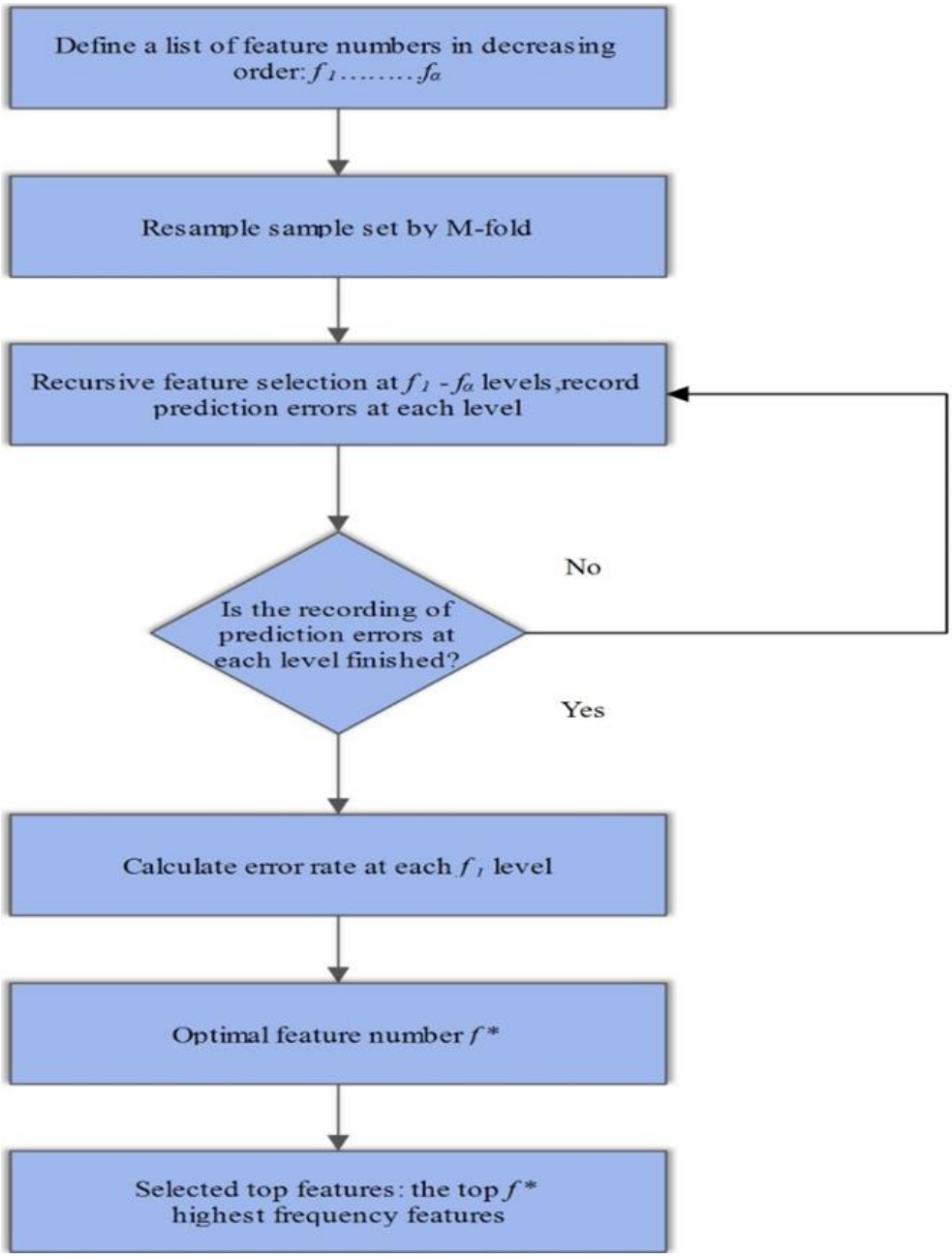


Figure 27: Workflow of Random Forest Recursive Feature Elimination (Zvarevashe and Olugbara, 2018b)

Inputs:

Training set Tr

Set of α features Fe= $\{f_1 \ldots\ldots f_\alpha\}$

Ranking method M (Tr, Fe)

Outputs:

Final ranking R

Code:

Repeat for i in {1: α}

Rank set Fe using M (Tr, Fe)

$f^* \leftarrow$ last ranked feature in Fe

R $(\alpha - i + 1) \leftarrow f^*$

Fe$\leftarrow$ Fe $- f^*$

Figure 28: Summarised RF-RFE algorithm

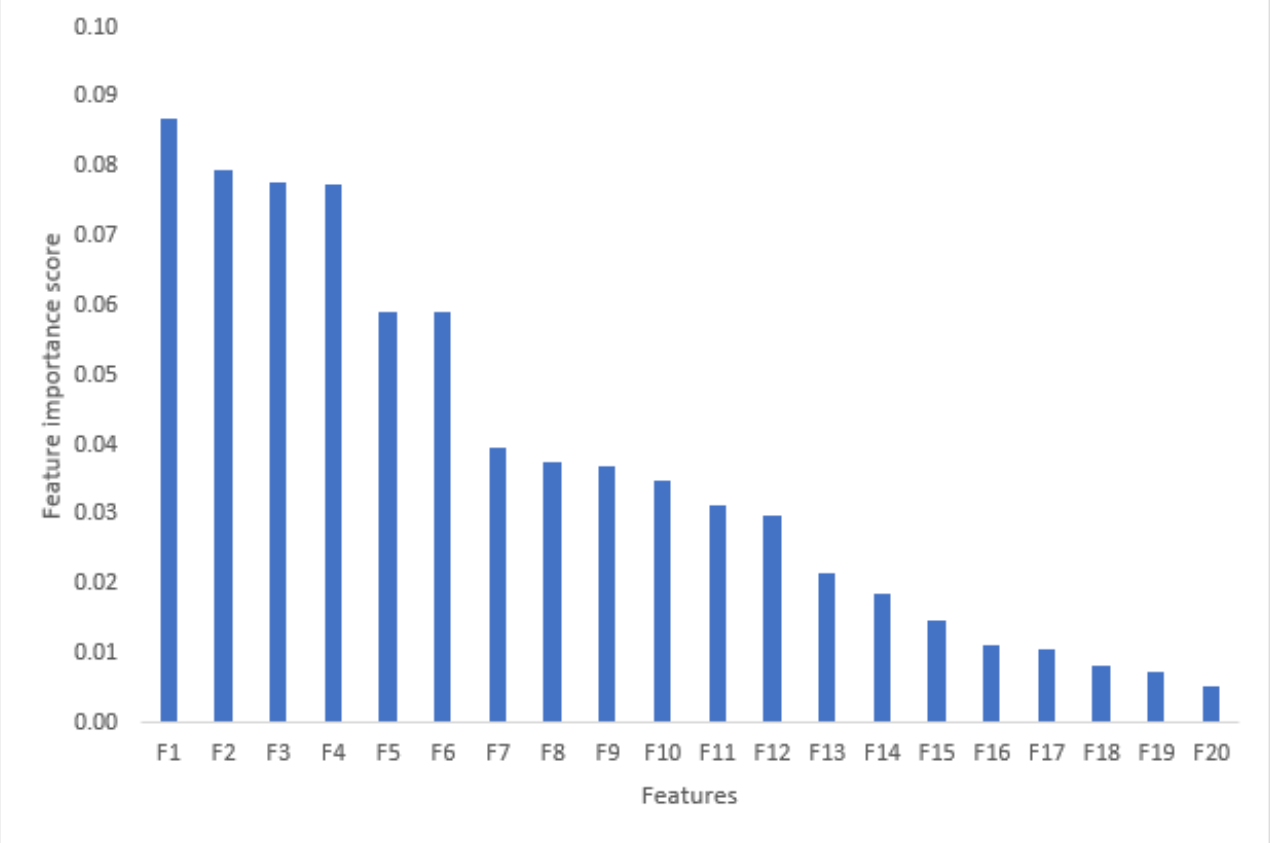


Figure 29: RF-RFE ranked Features

The 20 top-ranked discriminative acoustic features from the HAF set are presented in Table 6.

Table 6: Description of the top 20 acoustic features.

Key	Feature Description
F1	Derivative of standard deviation of area method of moments overall standard deviation (6th variant)
F2	Derivative of standard deviation of relative difference function overall standard deviation standard (1st variant)
F3	Area method of moments overall standard deviation (7th variant)
F4	Peak detection overall average (10th variant)
F5	Standard deviation of area method of moments overall standard deviation (8th variant)
F6	Derivative of area method of moments overall standard deviation (8th variant)
F7	Area method of moments of Mel frequency cepstral coefficients overall standard deviation (1st variant)
F8	Derivative of area method of moments overall standard deviation (9th variant)
F9	Derivative of standard deviation of area method of moments overall standard deviation (4th variant)
F10	Standard deviation of method of moments overall average (2nd variant)
F11	Peak detection overall average (10th variant)
F12	Peak detection overall average (9th variant)
F13	Peak detection overall average (7th variant)
F14	Derivative of standard deviation of area method of moments overall standard deviation (6th variant)
F15	Method of moments overall average (3rd variant)
F16	Peak detection overall average (3rd variant)
F17	Area method of moments overall standard deviation (8^{th} variant)
F18	Peak detection overall average (8^{th} variant)
F19	Area method of moments of MFCCs overall standard deviation (5^{th} variant)
F20	Derivative of running mean of area method of moments overall standard deviation (9^{th} variant)

5.7 Ensemble Classifiers

5.7.1 AdaBoost classifier (ABC)

ABC is a boosting ensemble classifier whose name was derived from the phrase Adaptive boosting (Freund, Schapire and Hill 1996). It was developed based on the boosting principle which is to improve the accuracy of classifiers (Cetina, Brito and Ruiz 2015). Therefore, the motivation behind the development of ABC was to transform a group of weak classifiers into strong ones. In ABC, decision trees with single splits are regarded as weak learners. ABC puts more focus on those instances that are difficult to classify while putting less emphasis on the easy-to-handle instances. The classifier can be used to solve both recognition and regression problems. ABC's recognition equation is shown below:

$$F = sign\left(\sum_{m=1}^{M} \theta_m f_m(x)\right) \quad \textbf{(5.2)}$$

where f_m resembles the m^{th} weak classifier and θ_m stands for the corresponding weight. ABC was chosen for this experimental study because it does not over-fit rapidly and it achieves higher accuracy as seen in other experimental studies (Dietterich 2000).

5.7.2 Random decision forest (RDF)

In machine learning, RDF is defined as a combination of numerous tree predictors (Breiman 2001). Here, each tree from the forest will be depending on the values of a random vector which will be sampled independently. Furthermore, the values are sampled with the same distribution for each tree in the forest. According to (Agjee et al. 2018), RDF can be described as a combination of several unpruned decision trees. These unpruned trees are then used to create an ensemble that can be used to classify complex problems (Pang *et al.* 2017). RDF was chosen in this thesis because it minimizes overfitting when applied to large corpora (Shareef, Mutlag and Mohamed 2017). Given a training dataset:

$$\mathbb{L} = \{(X_i, Y_i)^N_{i=1} \mid X_i \in R^M, Y \in \{1,2,\ldots,c\}, \quad \textbf{(5.3)}$$

where X_i stands for the features, Y is a class response feature, N is the number of training samples, and M is the number of features and a random forest model RDF described in Figure 22. Let $\hat{Y}^k$ be the prediction of tree $T_{k,}$ given an input X. The prediction of random forest with trees is K trees are:

$$\hat{Y} = \text{majority vote } \left\{\hat{Y}^k\right\}_1^K \quad \textbf{(5.4)}$$

Since each tree is grown from a bagged sample set, it is grown with only two-thirds of the samples in IL. These are also referred to as in-bag samples. Approximately, a third of the samples are excluded and these samples are referred to as OOB (out-of-bag) samples. These will then be used to compute the prediction error. The value predicted is

$$\hat{Y}^{OOB} = (1/\| ß_{\acute{\imath}} \|) \Sigma_{K \epsilon ß\acute{\imath}}, \hat{Y}^k \quad \textbf{(5.5)}$$

$$\hat{Y}^{OOB} = \ddot{Y} \quad \textbf{(5.6)}$$

Where $ß_{\acute{\imath}}$ = IL/ $ß_{\acute{\imath}}$. In this case, *i* and *í* are in-bag and out-of-bag sampled indices, /‖ $ß_{\acute{\imath}}$ ‖ is the size of OOB sub-dataset, and the OOB prediction error is

$$\hat{E}^{OOB} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \mathbb{E}(Y, \ddot{Y}), \quad \textbf{(5.7)}$$

while

$$\tilde{N} = N_{OOB} \quad \textbf{(5.8)}$$

where $\mathbb{E}$ is an error function and $\tilde{N}$ is OOB sample size.

The performance of RDF with regards to resistance to overfitting and a history of achieving higher accuracies influenced the choice to nominate it for this body of work (Abe *et al.* 2014; Hamzah *et al.* 2017; Ali and Maqsood 2018; Sagi and Rokach 2018). The other reason why RDF was nominated in this research is that it has performed very well in nanomaterial grouping (Bahl *et al.* 2019), health systems (Proniewska, Pregowska, and Malinowski 2020), damage assessment in concrete structures (Chun *et al.* 2020), remote sensing (Izquierdo and Zurita 2020) and many more. The summarized RDF algorithm is shown in Figure 30.

Input: $\mathbb{L} = \{(X_i, Y_i)_{i=1}^N \mid X_i \in R^M, Y \in \{1,2,\ldots,c\}$: the training dataset,

T: the number of trees

nsub: the number of sub spaces

Output: A Random Forest RF

for $t \leftarrow 1$ **to** *T* **do**

Draw a bagged subset of samples $\mathbb{L}_t$ from $\mathbb{L}$.

while (*stopping criteria is not met*) **do**

Select randomly *nsub* features.

for $n \leftarrow 1$ **to** | *nsub* | **do**

Calculate the decrease in the node impurity.

Select the feature with the highest decrease value and the node is split into two child nodes

Combine the *T* trees to form a random forest.

Figure 30: Summarised Random Forest Algorithm (Nguyen, Huang and Nguyen, 2015)

5.7.3 Extra trees classifier (ETC)

ETC is an ensemble supervised learning classifier that was developed based on decision trees and is also known as Extremely Randomized Trees (Geurts, Ernst and Wehenkel 2006). Just like RDF, ETC minimizes overfitting and over-learning from data by randomizing certain subsets of data and decisions. ETC is quite like RDF since it creates multiple trees splitting the nodes using randomized subsets of features. However, the key differences between ETC and RDF are that ETC does not bootstrap observations. Moreover, in ETC nodes are split using random splits unlike in RDF where this is done using best splits. This makes it less computationally expensive than RDF which is the reason why it was nominated for further exploration in this experimental study. ETC for numerical attributes is depicted in Table 7.

Table 7: ETC algorithm (Geurts, Ernst and Wehenkel, 2006)

Split_a_node(S)
Input: the local learning subset S corresponding to the node we want to split
Output: a split [a < ac] or nothing
– If **Stop_split**(S) is TRUE then return nothing.
– Otherwise select K attributes $\{a_1,..., a_K\}$ among all non-constant (in S) candidate attributes;
– Draw K splits $\{s_1,...,s_K\}$, where s_i = **Pick a random split**(S, a_i), $\forall i = 1,..., K$;
– Return a split $s*$ such that Score ($s*$, S) = $\max_{i=1,...,K}$ Score (s_i, S).

Pick_a_random split(S,a)
Inputs: a subset S and an attribute a
Output: a split
– Let a^S_{max} and a^S_{min} denote the maximal and minimal value of a in S;
– Draw a random cut-point a_c uniformly in $[a^S_{min}, a^S_{max}]$;
– Return the split $[a < a_c]$.

Stop_split(S)
Input: a subset S
Output: a boolean
– If $|S| < n_{min}$, then return TRUE;
– If all attributes are constant in S, then return TRUE;
– If the output is constant in S, then return TRUE;
– Otherwise, return FALSE.

As shown in Table 7, K stands for the randomly selected number of attributes while n_{min} resembles the minimum sample size for splitting a node. In ETC, majority voting is used to make the final prediction when solving recognition problems. ETC was used in this research because it has recorded success stories in oil formation (Seyyedattar *et al.* 2020), solar systems (Ahmad, Reynolds and Rezgui 2018) and many more.

5.7.4 Bagging classifier (BC)

BC is a bootstrap (Efron and Tibshirani 1993) ensemble meta-estimator that fits base classifiers on random subsets of training data (Breiman 1996). The classifier will then combine the individual predictions of the base classifiers to make a final prediction by either voting or by averaging the results. BC has become popular over the years because it reduces the variance of a black-box

estimator by applying randomization into its development procedure while creating an ensemble (Breiman 1999).

BC is highly efficient when applied to "unstable classifiers" such as neural networks and decision trees (Breiman 1996). BC has been successfully used in other application areas such as recognition of chicken, beef and mutton tissues (Yousaf *et al.* 2020), cardiotocography (Subasi, Kadasa and Kremic 2020), fingerprint detection (Agarwal and Chowdary 2020) and this is the reason why it was nominated for this experimental study.

5.7.5 Gradient boosting machines (GBM)

The GBMs are a group of powerful ensemble recognition algorithms that have yielded intriguing results in various applications (Natekin and Knoll 2013). In the GBM classifier, both the loss function and base learner models are selected randomly. Coming up with a solution to the required parameter estimates to obtain given some specific loss functions $\Psi(y,f)$ and $h(x, \alpha)$ can be very challenging. Therefore, a new function was developed to resolve this problem. The function $h(x, \alpha_t)$ was proposed to be the most parallel to the negative gradient $\{get\ (x_i)\}^{N}{}_{i=1}$ along with the observed data:

$$g_t(x)=\left[\frac{\partial \psi(y,f(x))}{\partial f(x)}\right]_{f(x)=f\,t-1(x)} \qquad \textbf{(5.9)}$$

A function can be chosen to improve the correlation with gt(x). This creates room for the removal of a potentially extremely hard optimization job with the classical least-squares minimization:

$$(\rho_t, \alpha_t)= \arg\min \sum_{i=0}^{n} \quad [-g_t(x) + \rho h(x_i, \alpha)]^2 \qquad \textbf{(5.10)}$$

$$\rho, \alpha$$

As proposed by Friedman (Friedman 2002), the algorithm can be summarized as shown in Figure 31. The precise state of the algorithm, including all the necessary formulae, will deeply rely on the design choices of $\Psi(y,f)$ and $h(x, \alpha)$.

GBM was nominated for this body of work because it has been successful in solving other recognition problems such as gender identification (Zvarevashe and Olugbara, 2018b), rock

permeability (Sudakov, Burnaev and Koroteev 2019), oil monitoring (Bikmukhametov and Jäschke 2019) and many more.

GRADIENT BOOSTING MACHINES ALGORITHM

Inputs:

Data $(y_i,f)^N_{i=1}$

Number of iterations M

Choice of the loss function $\Psi(y,f)$

Choice of the base learner model $h(x, \alpha)$

The Algorithm:

Initialize f_0 with a constant

For t = 1 to M do

Compute the negative gradient $g_t(x)$

Fit a new base-learner function $h(x, \alpha_t)$

Find the best gradient descent step-size ρ_t:

$\rho_t = \arg\min_\rho \sum_{k=0}^{n} \Psi[y_i, f_{t-1}(x_i) + \rho h(x_i, \alpha_t)]$

Update the function estimate:

$f_t \leftarrow f_{t-1} + \rho_t h(x, \alpha_t)$

End for

Figure 31:Summarised GBMs algorithm (Friedman, 2002)

5.7.6 RDF, AdaBoost, logistic regression, and gradient boosting machine (RALOG)

RALOG is an ensemble algorithm that was developed in this study using four classifiers which are RDF, Adaboost classifier (ABC), logistic regression classifier (LRC), and gradient boosting machine (GBM). RDF, GBM, and LR were used as the base classifiers while ABC with RDF was used as the meta-classifier. This ensemble technique was developed for this experimental study because ensemble algorithms are well known for improving the accuracy and precision of recognition models (Adetiba and Olugbara 2015b; Badshah *et al.* 2016; Malik, Farhan and Fahiem 2018). RALOG was developed using the stacking approach which has a record of clinical precision in most application domains such as malware and intrusion detection (Yan, Qi and Rao 2018; Rajagopal, Kundapur and Hareesha 2020). Stacking is an ensemble technique that combines various heterogeneous classifiers using a meta-classifier (Wolpert 1992).

The heterogeneous classifiers which are also referred to as base classifiers are initially trained on a given training set. The results processed from the base classifiers are then fed as input to the meta-classifier to conclude the processing as shown in Figure 32. LR, GBM, and RDF were therefore used as the base classifiers while AC with RDF was used as the aggregating meta-learner. The basic implementation of RALOG is illustrated in Figure 32.

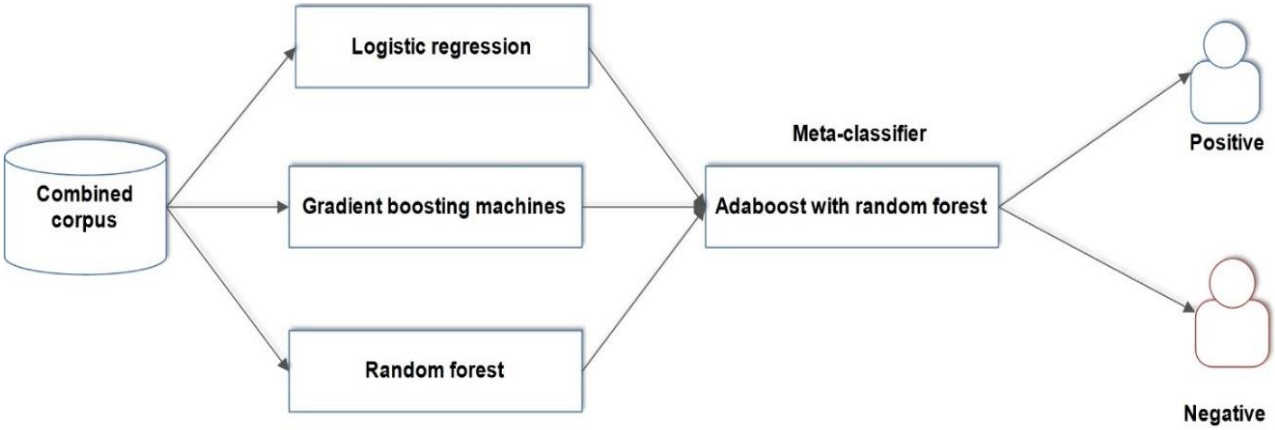


Figure 32: RALOG Architecture

Figure 33 shows the pseudocode of the RALOG stacked ensemble algorithm where $\chi = \{x_i \in R^M\}$, $Y = \{y_i \in N\}$, and a training set $D = \{(x_i, y_i)\}$. The parameters of the RALOG algorithm are that χ represents the data without the labels, x_i represents an instance in a dataset, Y represents a collection of labels from the dataset, and R^M represents a pool of instances in a dataset. S is the number of algorithms to be used to develop the stacked ensemble, x^{new}_i represents a new instance in a newly developed testing dataset. The dataset is a surrogate of D, the original dataset, a_s represents the learner-based algorithms, and a^{new} represents the new stacked ensemble algorithms developed using $(a_1(x), a_2(x), \ldots, a_S(x))$ that define a list of algorithms that are stacked together.

Input: D = $\{(x_i, y_i) | x_i \in \chi, y_i \in Y\}$
Output: An ensemble algorithm A
Step 1: Learn first-level learning algorithms
For $s \leftarrow 1$ to S do
Learn a base learning algorithm a_s based on D
Step 2: Construct a new dataset from D
For i $\leftarrow$ 1 to m do
Construct a new dataset that contains $\{x^{new}_i, y_i\}$, where
$x^{new}_i = \{a_j(x_i)$ for j = 1 to S$\}$
end
Step 3: Learn a second-level learning algorithm
Learn a new learning algorithm a^{new} based on the newly constructed dataset
Return $A(x_i) = a^{new}(a_1(x), a_2(x), \ldots, a_S(x))$

Figure 33: RALOG ensemble implementation

5.7.7 XGBoost (XGB)

XGBoost is a boosting algorithm that consists of many decision trees (Chen and Guestrin 2016; Song *et al.* 2020) and is based on the concept of gradient tree boosting (Chen and Guestrin 2016). In addition, XGB is described as an optimized version of GBM that is portable, efficient, and flexible (Leo, Luhanga, and Michael 2019; Dong *et al.* 2020). XGB has become popular because of its innate ability to avoid overfitting using shrinkage and regularisation methods (Gertz *et al.* 2020). Another benefit derived from using XGB is its ability to rank and scale features (Möller *et al.* 2016; Tamayo *et al.* 2016). According to Chen et al XGB uses the higher-order approximation concept to select the best tree structure while GBM relies on first-order Taylor expansion of loss function (Chen and Guestrin 2016).

XGB creates a strong learner by combining numerous weak learners as shown in equation 5.11.

$$\widehat{y_i} = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad \textbf{(5.11)}$$

where K represents the number of weak learners while $f_k(.)$ is the actual weak learner. XGB makes use of the Newton boosting principle which is premised on finding the optimal parameters by minimizing the loss function (Ø) as shown in equation 5.12:

$$L(\emptyset) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{5.12}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\alpha \|\omega\|^2, \tag{5.13}$$

where $\Omega(f_k)$ resembles the complexity of the kth tree model while n represents the sample size. T stands for the number of leaf nodes of the decision tree while ω resembles the weight of the leaf nodes. In addition, ω is responsible for controlling the degree of regularisation of f_k.

5.8 Base Inducers

5.8.1 Logistic regression classifier (LRC)

The LCR is a technique that can be used to solve various recognition problems (Tunç 2012). It has been successful in solving binary and multiclass recognition problems. LCR can be used to classify binary recognition problems such as cancer prediction (Zhou, Liu and Wong 2004). Furthermore, it can be used to solve multiclass recognition problems where the prediction of more than two outcomes will be expected (Liu *et al.* 2016). When using LCR, the first thing is to compute the conditional probability, and this is done as follows:

$$P\left(Y = \frac{1}{X}\right) = \pi(X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}}, \tag{5.14}$$

where $\beta X = \beta_0 + \beta_1 X_1 \cdots \beta_i X_i$, and i is the number of independent variables. This formula implies that $\pi(X)$ increases or decreases as an S-Shaped function of independent variables. The probability distribution of dependent variables is given as follows:

$$P(Y_i = y_i) = \begin{cases} \pi_i^{y_i}(1-\pi_i)^{1-y_i} & y_i = 0 \text{ or } 1 \\ 0, & \text{other wise.} \end{cases} \quad \textbf{(5.15)}$$

The likelihood function is the product of these probabilities, and the logarithm of the likelihood function is given as follows:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e(1 + \exp(\beta' X_i)). \quad \textbf{(5.16)}$$

When computing the parameters of logistic regression, the maximizing logarithmic likelihood function is used. To maximize the logarithmic likelihood, nonlinear optimization techniques are used. LCR was used in this body of work because it is not computationally expensive and a lot of success stories have been reported about it in other application areas (Zhou, Liu and Wong 2004).

5.8.2 Classification and regression trees (CART)

The CART classifier is a supervised ensemble classifier that is used to solve recognition problems with high precision and accuracy (Liu and Fan 2014). It evaluates a data set for constructing a set of rules, or questions. These rules and questions will then be used to predict classes (Majuran and Ramanan 2018). CART is also defined as a flowchart-like tree structure where an internal node represents features (or attributes). Here the branch is a representation of a decision rule while each leaf node represents the outcome. The node which sits at the very top of the decision tree is known as the root node. The decision tree learns to classify based on the attribute values. It partitions the tree using recursive partitioning. Normally, the decision tree classifier makes use of three common attribute selection techniques which are Gini index, gain ratio and information gain.

Assuming that β is the set of data samples, the attributes of class labels have η different value, and the number of different classes $ς_i$ (i = 1, 2, . . . , n) to be n. Set b_i as the number of samples in class $ς_i$. For a given sample, the expected information needed for recognition is given by the following equation:

$$I(b_1, b_2, \ldots, b_n) = \sum_{i=1}^{n} x_i \log_2(x_i) \quad \textbf{(5.17)}$$

where $x_i = b_i / b$ is the probability of any sample belonging to C_i. The decision tree classifier was chosen in this work because it is fast and it can handle high dimensional data with good accuracy(Zeng *et al.* 2014).

5.9 Deep Learning Classifiers

5.9.1 Deep radial basis function neural network

The deep radial basis function (DRBFNN) neural network is a popular artificial neural network (ANN) architecture that is used to solve both recognition and regression problems (Adetiba and Olugbara 2015a). DRBFNN was chosen in this study because of the positive results it obtained in the work done by (Adetiba and Olugbara 2015a). Furthermore, the DRBFNN can be considered as a multilayer feed-forward neural network that is used for strict interpolation in multi-dimensional space (Kumar and Yadav 2011; Memarian and Balasundram 2012). The DRBFNN neural network performs supervised recognition by measuring the input's similarity to examples from the training set (Wu *et al.* 2012). Here each DRBFNN neural network is responsible for storing a prototype. The prototype is a randomly selected example from the training set. When recognising new input data, each neuron is tasked with the duty of calculating the Euclidean distance between the input and its prototype. Therefore, casually speaking, if the input closely resembles the anger class's prototypes more than the prototypes of the sadness class, it simply means that it is classified as class anger. The basic architecture of DRBFNN is shown in Figure 34.

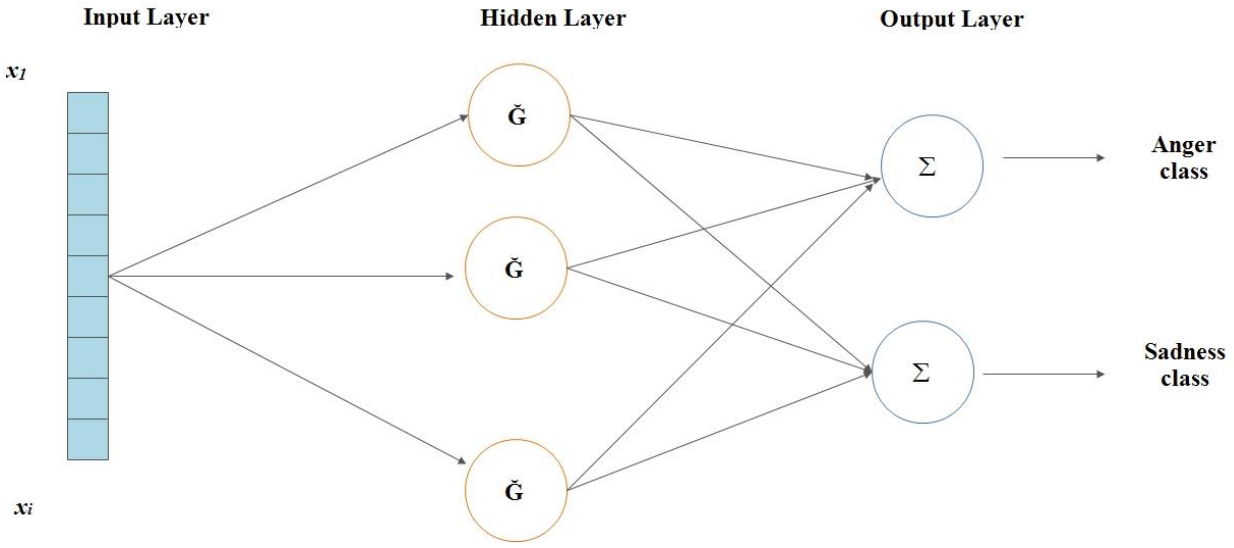


Figure 34: RBF neural network architecture (Wu *et al.*, 2012)

The DRBFNN neural network consists of three layers which are the input, hidden, and output layers. The input layer is responsible for accepting input data that will be required for recognition. The input data will be in the form of an n-dimensional vector. Conversely, the hidden layer is responsible for transforming the input data into a non-linear function. The hidden layer consists of DRBFNN neurons whereby each of the DRBFNN neurons will be responsible for storing a "prototype" vector. The output layer of the network comprises a set of nodes that will be up for recognition. Here each output node will be responsible for calculating the score for the associated class. In calculating the score for the associated class, each output node links a weight value with each of the DRBFNN neurons. It will then multiply the neuron's activation by this weight before adding it to the total response. Each output node will be having its own set of weights since each output node will be calculating the score for a different class.

The DRBFNN neuron activation function of each hidden unit in a DRBFNN neural network bears the responsibility of measuring the similarity between the input vector and the centre of that unit. The similarity distance can be computed using various similarity functions but the most popular one is based on the Gaussian function. When computing the similarity distance using the Gaussian function, each node from the hidden layer is considered as a p-multivariate Gaussian function and the equation is computed as shown in equation 5.18.

$$\check{G}\,(x, x_i) = \exp\left[\frac{-1}{2\sigma_i^2}\sum_{k=1}^{n}\ (x_k - x_{ik})\right] \quad \textbf{(5.18)}$$

where: x_i represents the mean (centre) and σ resembles the standard deviation. These functions are referred to as radial basis functions.

5.9.2 Deep multilayer perceptron neural network

Deep Multilayer perceptron (DMLP) is a popular type of feed-forward artificial neural network (Huang and Chen 2011). The DMLP uses a supervised learning method known as backpropagation for training. It is used to solve both recognition and regression problems and it was chosen in this study because of the excellent results it produced when it was used by (Getahun and Kebede 2017). Just like the RBF neural network, DMLP comprises three layers of nodes which are the input layer, hidden layer and the output layer (Hamzah *et al.* 2017). When data is fed to the MLP model through the input layer, the neurons in the first layer will propagate the weighted data and bias will be added through the hidden layers. An output response will be provided at the node using a special function known as the transfer function when the total sum is computed at the hidden nodes.

Two major features of the DMLP neural network are its non- linear processing elements and its massive interconnectivity. The non-linear processing elements usually have a non-linear activation function that is expected to be smooth. The most widely used activation function in DMLP includes the logistic function and the hyperbolic tangent (Bayram *et al.* 2016). Both these functions were used in this research work. The logistic function used is described as follows:

$$\phi(y_i) = (1 + e^{-w})^{-1} \quad \textbf{(5.19)}$$

Here, y_i represents the output of the *i*th node while w resembles the weighted sum of the input synapses. The logistic function ranges from 0 to 1. The hyperbolic tangent used is described as follows:

$$\phi(y_i) = \tanh(w_i) \quad \textbf{(5.20)}$$

Here, y_i represents the output of the *i*th node while w_i resembles the weighted sum of the input synapses. The hyperbolic tangent ranges from -1 to 1.

5.9.3 2D-CNN neural network

Convolutional neural networks (CNNs) are very popular in image and text processing. They are also used in speech and video processing. This family of deep learning algorithms was chosen for this body of work because they have been used successfully in speech processing (Farooq et al. 2020; Mustaqeem and Kwon 2020b). In speech processing, CNNs are used to extract features from raw spectrograms. These are the features that the model can use to classify the various patterns embedded within the spectrograms (Weißkirchen, Böck and Wendemuth 2018). Convolutional Neural Networks were invented in the 20th century as novel solutions to most computational problems. However, their acceptance in the research community was a bit slow because CNNs are computationally expensive. They became more popular in 2012 after AlexNet won the ImageNet challenge (Deng *et al.* 2017) and advances in processing power technology made it easier for them to gain acceptance from the research community. Furthermore, the advent of GPUs (Graphics Processing Unit) and TPUs (Tensor Processing Unit) has increased access to the required computing resources necessary to run CNNs (Chernykh and Prikhodko 2017).

CNN is a deep learning algorithm, that consists of numerous filter steps and one recognition step (Neumann and Thang 2018). These filter steps comprise 4 layers which are the pooling layer, batch normalization layer, convolutional layer, and an activation layer. The recognition stage constitutes fully connected layers and a recognition layer. Figure 35 presents the architecture of the proposed 7 layered custom 2D - CNN in processing spectrogram images. This version of CNN was developed using Keras and TensorFlow. The spectrogram images used were of size 28 x 28 = 784 pixels.

It is incredibly hard to recognize speech emotion by analysing the spectrogram images with a naked eye. Thus, CNN is applied to extract features from raw spectrogram images. The information describing the layers of the proposed custom 2 Dimension-CNN is listed in Table 8. The proposed custom 2 Dimension-CNN comprises seven layers which are four convolutional layers, only two fully connected layers, and an output layer. The filter size used to develop the proposed model is 4*4 for the first two convolutional layers. The remaining two layers had a 3 * 3 filter size. A 2 * 2 max-pooling was applied for each convolutional layer. The first fully

connected layers had 256 units while the second one had 128 units. The SoftMax classifier was used in the end to identify the emotional states.

Table 8: The detailed structure of the proposed custom 2D-CNN

Layer Type	Number of Filter	Size of Feature Map	Size of Kernel	Number of Stride	Number of Padding
Image input layer	-	28*28*1	-	-	-
Convolution Layer 1	32	28*28*32	4*4	2	2
Max Pooling Layer 1	1	14*14*1	2*2	1	0
Convolution Layer 2	64	28*28*64	4*4	2	2
Max Pooling Layer 2	1	14*14*64	2*2	1	0
Convolution Layer 3	128	28*28*128	3*3	2	2
Max Pooling Layer 3	1	14*14*128	2*2	1	0
Convolution Layer 4	128	28*28*128	3*3	2	2
Full Connection Layer 1	128	256*1	-	-	-
Full Connection Layer 2	-	128*1	-	-	-
Output Layer	-	8 * 1	-	-	-

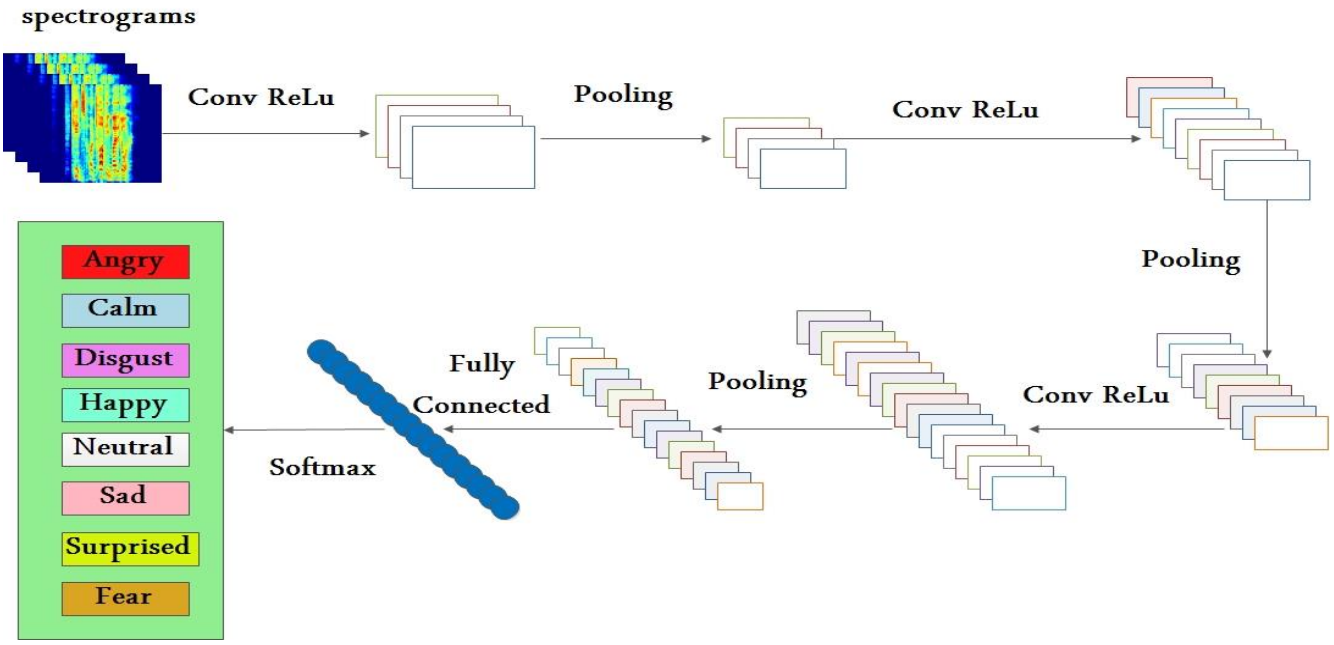


Figure 35: Basic architecture of the custom 2D - CNN in processing spectrogram images

A variety of programming tools can be used to solve recognition problems such as speech emotion recognition. These include MATLAB, R, Java, Knime, Weka and many more. Python was used to develop the proposed model in this thesis. This was mainly because python provides a high degree of flexibility to develop and optimize recognition models. For example, using python it was possible to develop customized versions of the deep learning models used in this research work (deep MLP and 2D-CNN). Moreover, Python is open source which makes it easy to access at zero cost.

5.10 Performance Metrics

The recognition of speech emotion is a perfect example of a prediction problem because the developed model is expected to predict the emotion of a given vocal utterance. Metrics are usually used as the performance benchmarks suitable for evaluating the performance of predictive models. The selection of a given metric depends on the type of application the predictive model is designed for. According to the requirements of this research work, the following metrics were carefully chosen:

5.10.1 Accuracy

Accuracy is the measure of the number of correctly classified instances divided by the total number of predictions (Li and Akagi 2019). In this case, accuracy was used to measure how the designed models fare in meeting the requirement of correctly recognising emotions. The metric is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \textbf{(5.21)}$$

Where:

- TP = True Positive
- FP = False Positive
- FN = False Negative
- TN = True Negative

5.10.2 Precision

Precision is the measure of relevant positively predicted instances. It is also defined as a measure of correctly classified instances divided by all positives including false positives. The formula to calculate this metric is given below:

$$Precision = \frac{TP}{TP+FP} \quad \textbf{(5.22)}$$

Where:

- TP = True Positive
- FP = False Positive
- FN = False Negative
- TN = True Negative

5.10.3 Recall

Recall is a performance metric that is sometimes referred to as sensitivity (Bachorowski and Bachorowski, 2010). It is a measure that shows the number of correctly classified instances divided by all the positive instances and this includes false negatives (Hernández *et al.* 2019; Ma *et al.* 2019). This metric is calculated as follows:

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(5.23)}$$

Where:

- TP = True Positive
- FP = False Positive
- FN = False Negative
- TN = True Negative

5.10.4 F1-score

This metric is a measure that shows the harmonic mean of precision and recall (Zhang *et al.* 2018). F1-score is also regarded as a measure that incorporates both false positives and negatives (Narendra and Alku 2019). The metric is calculated as follows:

$$\text{F1-score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad \textbf{(5.24)}$$

5.10.5 Receiver operating characteristics (ROC) curve and area under the curve (AUC)

The ROC curve is a graph that shows the performance of a given recognition model using a measure of the true positive rate and false-positive rate. AUC is a measure of the two-dimensional space area under the ROC curve. These two metrics are very useful when comparing models.

5.10.6 Processing time

This is a measure that estimates the average time that a model processes the recognition of emotions.

5.11 Chapter Summary

This chapter presented comprehensive details about the methods and tools used in this research work. The main reasons behind the choice of emotions used in this thesis were highlighted and this was in harmony with the review presented in chapter 2 of this thesis. The recognition algorithms such as (RDF, GBM, and CART) were discussed in this chapter and these are used for the experimental models in the subsequent chapter. RALOG, a custom-made ensemble model for coarse-grained SER was presented in this chapter. A customized deep classifier (2D-CNN) was also presented as well as the RFRFE feature selection algorithm, which was used for the experimental models.

Chapter 6: Experimental Results

6.1 Introduction

This chapter presents four experiments carried out in this body of work to fulfil the objectives highlighted in chapter 1. The chapter also presents the results of the experiments and a discussion of the results to ultimately answer the research questions posed in chapter 1. These experiments were also done to determine the most useful features and recognition algorithms in recognising speech emotion in customer call centres. All the experiments were implemented on a computer system with a Core i7 2.3 GHz processor and 8 GB of RAM. The computer system also has 1TB Hard Disk and it runs a 64bit Windows 10 operating system.

The first experiment was conducted to the evaluating the performance of HAF features (described in chapter 5) against other feature sets described in the literature. The first feature set (MFCC1) comprised of a homogeneous set of 133 MFCC spectral features while the second feature set (MFCC2) was crafted using MFCC, ZCR, energy and F_0 features (ninety features). These features were fed as input to ensemble classifiers which include AC (with CART as the base classifier), RDF, BC (with SVM as the base classifier) and GBM. The second experiment was done using deep learning auto-generated features. These features were fed as in put to the same ensemble classifiers used in the first experiment. The purpose of this second experiment was to investigate the most efficient auto-generated features. Three deep classifiers were used to extract auto-generated features from spectrograms, and these include 2D-CNN, DMLP and DRBFNN. The 10-fold cross-validation technique was applied to eliminate bias and the data was split into testing (30%) and training data (70%).

The third experiment was undertaken to discover the performance of end-to-end deep learning classifiers which are 2D-CNN, DMLP and DRBFNN. The classifiers were used together with their corresponding auto-generated features extracted in experiment 2. The fourth experiment was done to discover the most efficient features in recognising emotion in a multilingual environment using a cross emotion corpus.

6.2 Experiment 1 Results and Discussion

Experiment 1 was designed to fulfil the first research objective that is "to discover a set of acoustic features that will improve the accuracy of SER systems". Therefore, HAF, MFCC1 and MFCC2 were explored to discover the most efficient acoustic features. The features were extracted from two speech emotion databases which are RAVDESS and SAVEE. A discussion of the theoretical background of the above-mentioned feature sets has been done in Chapter 5. The training processing time result is presented in Table 9. The computational time was evaluated for each classifier using the three sets of features for the two emotion databases. It is worth mentioning that from the results shown in Table 9, RDF had the shortest processing time in comparison with the other ensemble classifiers. Amongst the features, MFCC1 and MFCC2 recorded the shortest training across all databases. However, RDF (0.43) had a slightly higher training time compared to BSVM (0.39) which took 0.39 milliseconds when SAVEE's HAF features were applied. RDF still performed better compared to the RAVDESS HAF. This inconsistency might have been achieved because SAVEE is a much smaller dataset compared to RAVDESS (1432). Furthermore, the results show that GBM (1172.3 milliseconds on MFCC2 RAVDESS) and BMLP (1319.39 milliseconds on MFCC2 SAVEE) consume a lot of processing time in comparison with the other ensemble classifiers.

Table 9: Processing time of classifiers on MFCC1, MFCC2 and HAF

	SAVEE Feature			RAVDESS Feature		
Classifier	**MFCC1**	**MFCC2**	**HAF**	**MFCC1**	**MFCC2**	**HAF**
RDF	0.03	0.38	0.43	0.052	1.62	0.06
GBM	2.30	579.40	343.10	5.90	1172.30	6.90
ABC	0.47	22.16	1.38	2.59	2.79	1.89
BSVM	0.20	49.79	0.39	0.42	211.54	0.48
BMLP	3.13	1319.39	952.46	76.00	231.16	7.18

The average overall accuracy, recall, precision, and F1-Score of the results obtained using the learning models are presented in Table 10. The overall recognition rates recorded in the two-speech

corpus differ because the databases contain utterances spoken in different accents and the way the data was collected for each corpus. The results in Table 10 show that RDF performed better than the other four ensemble classifiers regarding MFCC1 features for both the SAVEE and the RAVDESS corpus. The same performance pattern was also observed for MFCC1 for the SAVEE corpus. However, both RDF and GBM attained a similar accuracy score of 61.5%. BMLP recorded the lowest average accuracy results across all databases using all the features followed by BSVM which recorded an accuracy of 56.0% for SAVEE and 87.2% for RAVDESS.

Table 10: Average accuracy, recall, F1-score and precision with confidence intervals of classifiers trained with MFCC1, MFCC2 and HAF.

	SAVEE Feature			RAVDESS Feature		
Classifier/ Measure	**MFCC1**	**MFCC2**	**HAF**	**MFCC1**	**MFCC2**	**HAF**
			RDF			
Precision	63.4 (±0.044)	78.1 (±0.037)	99.1(±0.009)	90.1 (±0.016)	93.1 (±0.013)	99.6 (±0.003)
Recall	72.1 (±0.041)	77.4 (±0.038)	99.1(±0.009)	88.9 (±0.016)	94.0 (±0.012)	99.5 (±0.004)
F1-score	63.6 (±0.043)	76.6 (±0.039)	99.1(±0.009)	88.5 (±0.017)	92.5 (±0.014)	99.8 (±0.002)
Accuracy	61.5 (±0.044)	66.7 (±0.043)	99.3 (±0.008)	90.7 (±0.015)	93.7 (±0.013)	99.8 (±0.002)
			GBM			
Precision	64.0 (±0.043)	75.6 (±0.039)	99.4 (±0.007)	85.0 (±0.019)	86.5 (±0.018)	95.3(±0.011)
Recall	66.6 (±0.043)	74.6 (±0.039)	99.1 (±0.009)	82.9 (±0.020)	88.1 (±0.017)	94.8(±0.012)
F1-score	62.6 (±0.044)	72.7 (±0.040)	99.3 (±0.008)	82.6 (±0.020)	85.9 (±0.018)	96.8(±0.009)
Accuracy	61.5 (±0.044)	65.5 (±0.043)	99.3 (±0.008)	85.4(±0.018)	86.7(±0.018)	92.6(±0.014)
			ABC			

Precision	62.9 (±0.044)	73.4 (±0.040)	99.1 (±0.009)	81.3 (±0.020)	85.4 (±0.018)	94.3(±0.012)
Recall	64.1 (±0.044)	73.7 (±0.040)	99.4 (±0.007)	84.0 (±0.019)	87.5 (±0.017)	94.5(±0.012)
F1-score	63.1 (±0.044)	74.1 (±0.040)	97.9 (±0.013)	82.6 (±0.020)	84.5 (±0.019)	95.9(±0.010)
Accuracy	58.0 (±0.044)	62.8 (±0.043)	98.0 (±0.008)	83.0 (±0.020)	85.2 (±0.018)	92.0 (±0.014)
BSVM						
Precision	61.4 (±0.044)	71.4 (±0.041)	98.7 (±0.010)	81.0 (±0.020)	84.1 (±0.019)	91.1(±0.015)
Recall	61.0 (±0.044)	72.6 (±0.041)	99.0 (±0.009)	82.5 (±0.020)	82.3(±0.020)	92.0(±0.014)
F1-score	61.9 (±0.044)	72.0 (±0.041)	99.3 (±0.008)	80.5 (±0.021)	83.5 (±0.019)	90.3(±0.015)
Accuracy	56.0 (±0.045)	61.5 (±0.044)	96.0 (±0.013)	82.7 (±0.020)	84.5 (±0.019)	91.8 (±0.014)
BMLP						
Precision	59.4 (±0.044)	69.0 (±0.042)	97.9 (±0.013)	78.8 (±0.021)	77.5 (±0.022)	93.0 (±0.013)
Recall	60.1 (±0.044)	72.0 (±0.041)	98.4 (±0.011)	75.8 (±0.022)	80.3 (±0.021)	88.0(±0.017)
F1-score	59.6 (±0.044)	70.3 (±0.041)	98.1 (±0.012)	74.4(±0.023)	78.3 (±0.021)	89.0 (±0.016)
Accuracy	55.4 (±0.045)	60.7 (±0.044)	94.6 (±0.002)	75.6 (±0.022)	79.3 (±0.021)	91.3 (±0.015)

When the MFCC2 feature set was interrogated, an improvement was observed. This feature set was developed after drawing inspiration from the work presented by Bhaskar et al. (Bhaskar, Sruthi and Nedungadi 2015) and Sarker and Alam (Sarker and Alam 2014) where pitch, ZCR, and MFCC were combined to form hybrid features. The MFCC2, RDF outperformed all the other learning classifiers with an accuracy score of 93.7% on RAVDESS and 66.7% on SAVEE. GBM was second-best in this regard achieving an accuracy score of 86.7% on RAVDESS and 65.5% on SAVEE. BMLP was the least performing classifier, and it achieved an accuracy score of 79.3%

followed by BSVM which recorded an overall accuracy of 84.5%. The same trend was observed on the precision, recall and F1-score metrics for the classifiers and features across all the databases.

When the classifiers were applied on HAF the accuracy scores drastically rose such that all the classifiers achieved accuracy scores that were above 90%. This drastic increase was consistent across all the corpora. According to Table 10, BMLP achieved the lowest average accuracy score (91.3%) regarding HAF. All the other classifiers achieved spectacular results across the emotion corpora. Achieving an average accuracy score of 99.8 % on RAVDESS and 99.3% on SAVEE, RDF achieved the most impressive with regards to HAF. RDF was closely followed by GBM (92.6%), ABC (92.0%), BSVM (91.8%) and BMLP (91.3%). From these observations seen in these results, it can be concluded that developing SER models using ensemble classifiers and HAF feature sets is highly promising. Generally, the higher accuracy scores achieved using HAF show that the proposed hybrid features are indeed effective in SER.

From Table 10 it can be noted that there is a contrast in the percentage recognition accuracy scores. This could have been stimulated by a variety of critical factors like the type of corpus, the accent of the speakers, the data collection method, the features and classifiers used. Even though the data used in this experimental study was little, the confidence factors shown in Table 10 indicate that the estimated ranges are appropriate because the confidence intervals are in the range between 0.045 and 0.013.

As shown in Figures 36 and 37, BC achieved average recognition accuracies of 82.7% and 84.5% on the RAVDESS and SAVEE corpora which is relatively low compared to the other classifiers. In addition, Figures 36 and 37 illustrate the graphical presentation of the accuracy results on Ravdess and Savee respectively.

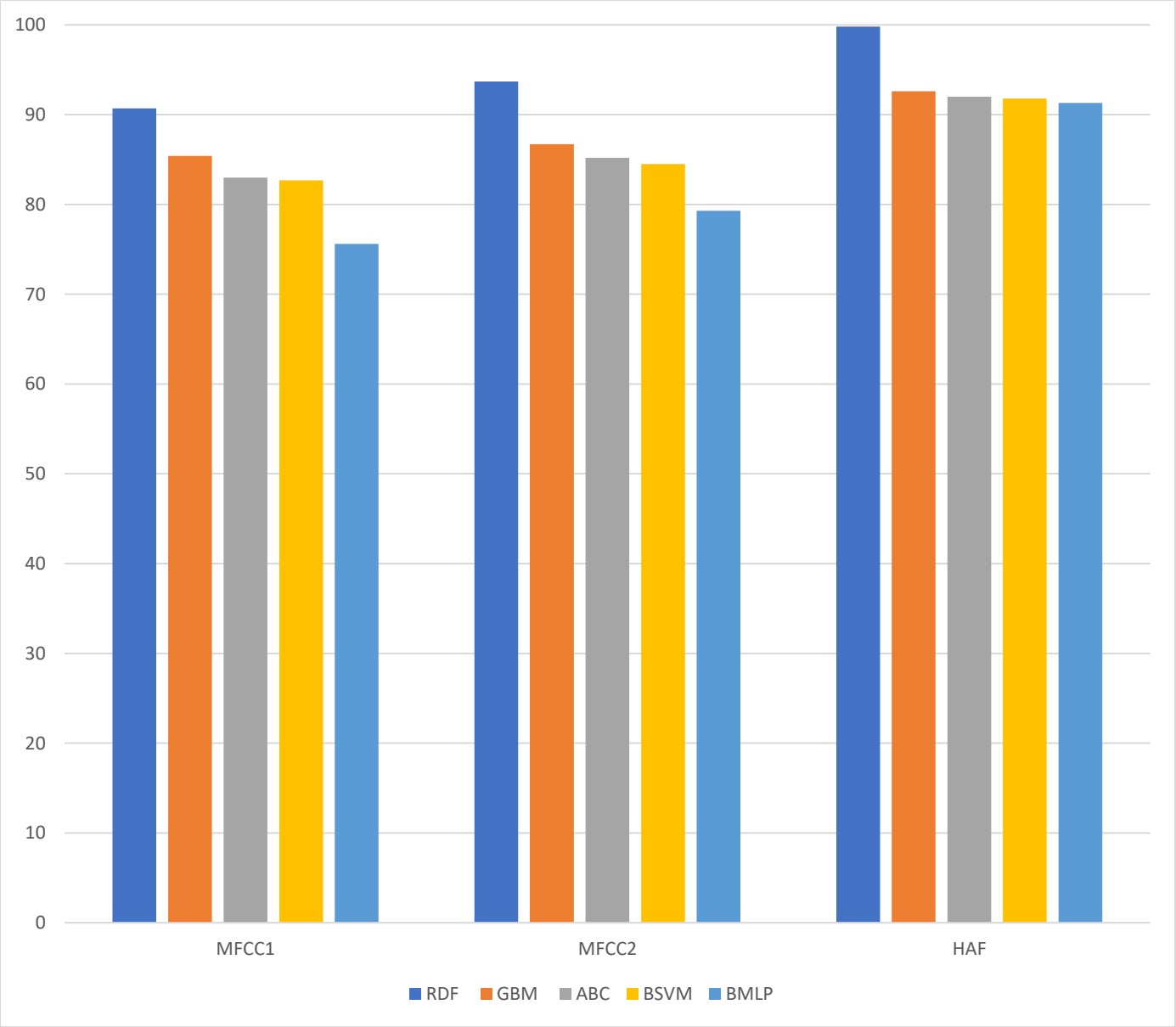


Figure 36: Graphical representation of Percentage Accuracy of Classifiers Trained using MFCC1, MFCC2 and HAF on Ravdess

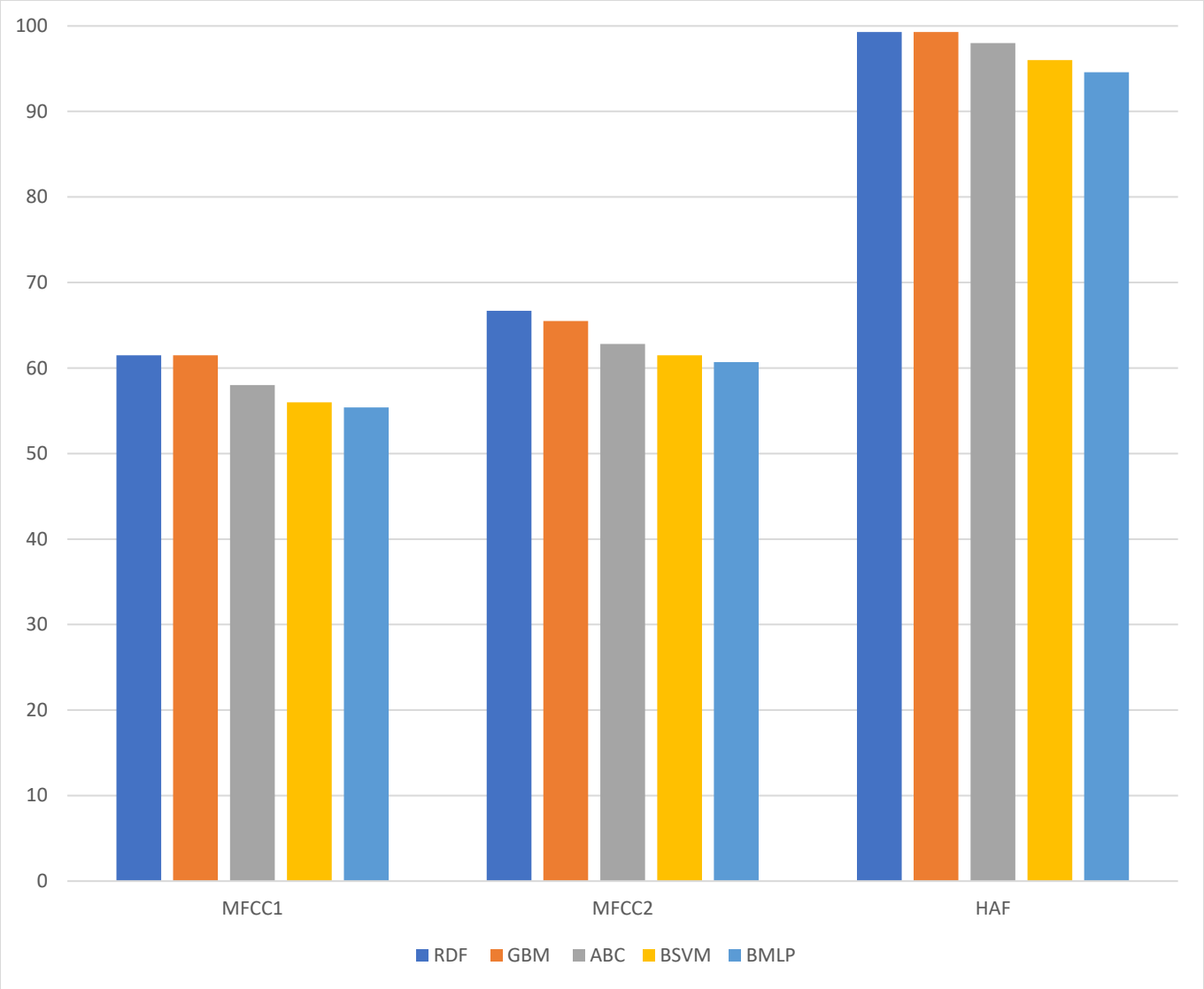


Figure 37: Graphical representation of Percentage Accuracy of Classifiers Trained using MFCC1, MFCC2 and HAF on Savee

Table 11: Accuracy, Recall and Score and Precision on RAVDESS MFCC1.

	Emotion							
Classifier/ Measure	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
				RDF				
Precision	100.0	100.0	100.0	92.0	80.0	53.0	87.0	96.0
Recall	100.0	100.0	100.0	83.0	82.0	86.0	79.0	91.0

F1-score	100.0	100.0	100.0	87.0	81.0	66.0	83.0	94.0
Accuracy	91.0	90.0	91.0	89.0	91.0	89.0	91.0	93.0
GBM								
Precision	94.0	100.0	100.0	87.0	69.0	39.0	87.0	85.0
Recall	100.0	100.0	100.0	74.0	69.0	78.0	72.0	87.0
F1-score	97.0	100.0	100.0	80.0	69.0	52.0	79.0	86.0
Accuracy	86.0	85.2	87.0	81.0	83.0	85.0	86.0	90.0
ABC								
Precision	91.0	82.0	78.0	79.0	81.0	79.0	81.0	81.0
Recall	100.0	100.0	100.0	81.0	56.0	85.0	71.0	82.0
F1-score	95.0	100.0	91.0	65.0	69.0	70.0	81.0	79.0
Accuracy	85.0	84.0	85.0	75.0	81.0	79.0	87.0	88.0
BSVM								
Precision	83.0	84.0	82.0	75.0	75.0	77.0	79.0	89.0
Recall	82.0	83.0	91.0	73.0	79.0	81.0	84.0	87.0
F1-score	81.0	82.0	79.0	79.0	81.0	79.0	82.0	84.0
Accuracy	83.0	85.0	81.0	79.0	80.6	81.0	86.0	86.0
BMLP								
Precision	72.0	81.0	78.0	68.0	70.0	70.0	75.0	81.0
Recall	78.0	84.0	82.0	70.0	76.0	79.0	78.0	83.0
F1-score	74.0	82.0	80.0	67.0	71.0	75.0	77.0	80.0
Accuracy	80.0	77.0	76.0	71.0	74.0	75.0	75.0	77.0

This section discusses the precision, recall and F1-score results for each emotion, classifier, and corpus. Disgust, clam and angry emotions can be recognised easily using ensemble classifiers with regards to RAVDESS MFCC1 features as shown in Table 11. Like what was noted in the accuracy scores, RDF achieved impressive results in precision, recall, and F1-score. However, the same

cannot be said for BMLP. In general, BMLP performed dismally compared to RDF and GBM but achieved satisfactory results in recognising sad and surprise emotions. In the same vein, ABC gave a relatively good performance in recognising sad and surprising emotions. In addition, RDF achieved a much higher precision score (92.0% in recognising fear) in comparison to the results reported in (Sun *et al.* 2019). Table 11 shows results that resonate with the discoveries found in the literature that demonstrate the ineffectiveness of MFCC features in recognising the fear emotion state. RDF achieved the highest accuracy in recognising the fear emotion but still, it is low (89%).

Table 12: Accuracy, Recall and Score and Precision on RAVDESS MFCC2.

	Emotion							
Classifier/ Measure	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
				RDF				
Precision	100.0	100.0	100.0	91.0	95.0	64.0	90.0	100.0
Recall	100.0	100.0	100.0	89.0	78.0	100.0	90.0	95.0
F1-score	100.0	100.0	100.0	93.0	86.0	78.0	90.0	98.0
Accuracy	96.0	93.0	92.0	90.0	94.0	91.0	96.0	97.6
				GBM				
Precision	100.0	100.0	100.0	87.0	79.0	55.0	80.0	86.0
Recall	100.0	100.0	100.0	87.0	75.0	86.0	67.0	90.0
F1-score	100.0	100.0	100.0	87.0	77.0	67.0	73.0	88.0
Accuracy	88.0	84.0	83.0	82.0	85.0	85.0	88.0	98.6
				ABC				
Precision	94.0	93.0	91.0	78.0	71.0	79.0	81.0	89.0
Recall	96.0	100.0	100.0	86.0	76.0	73.0	83.0	86.0
F1-score	92.0	95.0	100.0	79.0	74.0	76.0	81.0	86.0
Accuracy	86.0	82.0	83.0	81.0	87.0	82.0	88.0	92.6

				BSVM				
Precision	81.0	94.0	91.0	77.0	74.0	79.0	78.0	94.0
Recall	92.0	89.0	90.0	76.0	74.0	77.0	74.0	90.0
F1-score	94.0	91.0	91.0	77.0	74.0	78.0	76.0	92.0
Accuracy	84.0	81.0	80.0	78.0	85.0	85.0	87.0	96.0
				BMLP				
Precision	76.0	89.0	86.0	72.0	70.0	76.0	74.0	83.0
Recall	75.0	86.0	86.0	74.0	67.0	75.0	72.0	85.0
F1-score	77.0	91.0	87.0	74.0	73.0	76.0	74.0	90.0
Accuracy	85.0	81.0	78.0	73.0	78.0	77.0	78.0	84.0

The precision, recall, and F1-score analysis of RAVDESS MFCC2 features are illustrated in Table 12. The table shows that MFCC2 caused notable increases in disgust, calm and angry emotion performance scores. Nevertheless, the scores were low when the happy and neutral emotions were interrogated. This shows that the use of combined fundamental frequency, energy, ZCR, and MFCC features yields poor results in recognising neutral and happy emotions. As can be seen, MFCC2 features gave an improved performance in recognising the fear emotion compared to the results presented in (Wang *et al.* 2015), where 81.0% was realized for recognising fear. The introduction of fundamental frequency and energy improves the recognition of fear because the recognition of the fear emotion improved across all the classifiers even though the increment was marginal.

Table 13: Average Accuracy, Recall and Score and Precision on RAVDESS HAF.

					Emotion			
Classifier/ Measure	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
				RDF				
Precision	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0
Recall	100.0	100.0	100.0	100.0	100.0	96.0	100.0	100.0

F1-score	100.0	100.0	100.0	100.0	100.0	98.0	100.0	99.0
Accuracy	100.0	100.0	100.0	98.0	100.0	100.0	100.0	100.0
GBM								
Precision	93.0	100.0	100.0	94.0	100.0	95.0	99.0	93.0
Recall	92.0	100.0	100.0	93.0	97.0	100.0	76.0	100.0
F1-score	92.0	100.0	100.0	93.0	98.0	97.0	86.0	96.0
Accuracy	95.0	91.0	92.0	88.0	91.0	90.0	96.0	98.0
ABC								
Precision	92.0	100.0	96.0	94.0	100.0	95.0	93.0	97.0
Recall	91.0	100.0	100.0	95.0	93.0	76.0	100.0	100.0
F1-score	92.0	100.0	98.0	93.0	95.0	91.0	87.0	98.0
Accuracy	93.0	90.0	92.0	86.0	93.0	91.0	95.0	96.0
BSVM								
Precision	91.0	100.0	99.0	92.0	89.0	74.0	83.0	94.0
Recall	91.0	100.0	100.0	90.0	86.0	85.0	86.0	98.0
F1-score	93.0	100.0	100.0	91.0	86.0	77.0	87.0	95.0
Accuracy	93.0	93.0	92.0	86.0	90.0	91.0	93.0	96.0
BMLP								
Precision	92.0	100.0	93.0	90.0	88.0	73.0	81.0	91.0
Recall	91.0	100.0	97.0	93.0	88.0	86.0	88.0	97.0
F1-score	90.0	100.0	91.0	88.0	83.0	75.0	84.0	91.0
Accuracy	92.0	91.0	93.0	89.0	91.0	88.0	92.0	94.0

According to Table 13, the classifiers performed impressively across all the emotion corpora with regards to the RAVDESS HAF. The results show RDF's high efficiency in recognising all the eight emotions where the classifier achieved 100.0% scores on most of the emotions. In addition, RDF achieved relatively higher scores in recognising the neutral emotion. This was indeed a

superb performance since recognising the neutral emotion proved to be difficult when other sets of features were used. Even though the other classifiers were outperformed by RDF, the performance was good. These results reveal the efficiency of the proposed HAF in SER. Moreover, the results accentuate the importance of using highly discriminating features in SER. The results also show that the ensemble classifiers with HAF can recognise the fear emotion more efficiently in comparison with other techniques used (Wang *et al.* 2015). Additionally, HAF performed better than the acoustic features used to recognise six emotional states from the CASIA database (Sun *et al.* 2019) where 60.5% was realised in recognising fear.

Table 14: Average Accuracy, Recall and Score and Precision on SAVEE MFCC1.

	Emotion						
Classifier/ Measure	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
				RDF			
Precision	65.0	24.0	65.0	48.0	94.0	48.0	100.0
Recall	65.0	83.0	57.0	65.0	52.0	83.0	100.0
F1-score	65.0	37.0	60.0	55.0	67.0	61.0	100.0
Accuracy	69.0	66.0	60.0	75.0	64.0	77.0	81.0
				GBM			
Precision	65.0	29.0	75.0	26.0	91.0	62.0	100.0
Recall	62.0	67.0	48.0	55.0	62.0	72.0	100.0
F1-score	63.0	40.0	59.0	35.0	74.0	67.0	100.0
Accuracy	70.0	66.0	57.0	74.0	62.0	79.0	84.0
				ABC			
Precision	66.0	28.0	64.0	36.0	89.0	59.0	98.0
Recall	68.0	29.0	62.0	40.0	91.0	59.0	100.0
F1-score	65.0	28.0	63.0	38.0	91.0	58.0	99.0
Accuracy	60.0	57.0	53.0	66.0	62.0	79.0	83.0

BSVM							
Precision	65.0	23.0	61.0	41.0	85.0	57.0	98.0
Recall	64.0	26.0	57.0	38.0	81.0	61.0	100.0
F1-score	66.0	29.0	62.0	36.0	83.0	58.0	99.0
Accuracy	56.0	54.0	51.0	69.0	56.0	77.0	85.0
BMLP							
Precision	64.0	23.0	59.0	38.0	83.0	54.0	95.0
Recall	63.0	25.0	55.0	39.0	85.0	57.0	97.0
F1-score	64.0	25.0	59.0	37.0	81.0	55.0	96.0
Accuracy	52.0	50.0	49.0	51.0	53.0	64.0	69.0

The accuracy, F1-score, recall and precision analysis of SAVEE MFCC1 features are illustrated in Table 14. RDF outperformed all the other classifiers in recognising the neutral emotion achieving a precision score of 94.0%. This score is much higher than the score achieved using the RAVDESS corpus using the same features. RDF's average scores increased by approximately 7.9% while the other classifiers' average percentage scores increased by at least 14.7%. Furthermore, the surprise emotion was recognised easily by the ensemble algorithms achieving 100.0% recognition rates. From Table 14 it can be observed that MFCC1 features still perform poorly in recognising the fear of emotion as reported by other authors (Khan and Roy, 2017). Nonetheless, a higher percentage of precisions were achieved by RDF and GBM, which had 94.0% and 91.0% respectively. Compared to the results presented in (Kerkeni *et al.* 2018) the results achieved in the neutral emotion were relatively higher.

The analysis of the percentage F1-score, recall and precision on the SAVEE database with MFCC2 features after performing feature recognition is shown in Table 15. The results show that generally, all the classifiers performed well in recognising the surprise emotion. On the contrary, the classifiers performed dismally in recognising other emotions.

Table 15: Accuracy, F1-Score, Recall and Precision on SAVEE MFCC2.

Emotion							
Classifier/ Measure	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
RDF							
Precision	71.0	85.0	69.0	68.0	95.0	59.0	100.0
Recall	70.0	84.0	65.0	67.0	68.0	88.0	100.0
F1-score	71.0	84.0	68.0	68.0	74.0	71.0	100.0
Accuracy	77.0	73.0	63.0	70.0	69.0	82.0	100.0
GBM							
Precision	72.0	65.0	83.0	47.0	93.0	69.0	100.0
Recall	69.0	74.0	69.0	62.0	71.0	77.0	100.0
F1-score	70.0	68.0	71.0	48.0	77.0	75.0	100.0
Accuracy	74.0	71.0	59.0	73.0	69.0	86.0	92.0
ABC							
Precision	74.0	71.0	70.0	58.0	69.0	73.0	99.0
Recall	76.0	72.0	68.0	61.0	68.0	71.0	100.0
F1-score	74.0	75.0	73.0	59.0	65.0	73.0	100.0
Accuracy	71.0	67.0	56.0	69.0	69.0	71.0	99.0
BSVM							
Precision	71.0	68.0	66.0	59.0	67.0	71.0	98.0
Recall	72.0	72.0	64.0	61.0	64.0	75.0	100.0
F1-score	70.0	70.0	66.0	60.0	65.0	73.0	100.0
Accuracy	71.0	62.0	51.0	67.0	68.0	73.0	100.0
BMLP							

Precision	70.0	66.0	64.0	57.0	66.0	67.0	93.0
Recall	72.0	70.0	66.0	59.0	68.0	72.0	97.0
F1-score	71.0	68.0	64.0	59.0	65.0	69.0	96.0
Accuracy	56.0	52.0	49.0	65.0	54.0	64.0	85.0

The classifiers achieved relatively low scores across all the emotions except the surprise emotion in which perfect scores were achieved using ensemble classifiers. BMLP achieved the lowest precision score (93%) for the surprise emotion using MFCC2 features followed by BSVM (98%). Furthermore, BMLP achieved the lowest accuracy score for recognising the surprise emotion (85%). This result shows that MFCC, ZCR, energy and fundamental frequency features are not the optimum set of speech features suitable for recognising human emotions. Nevertheless, GBM achieved a relatively higher recognition percentage in comparison with the result reported in (Sun *et al.* 2019).

The analysis results of the percentage F1-score, recall and precision values achieved after performing feature recognition with SAVEE HAF are presented in Table 16. From the results, it can be observed that all the classifiers' performance improves sharply when HAF features are used to recognise all the emotions. Moreover, the combination of HAF features and ensemble learners drastically improves the recognition of the fear emotion which was seen to be difficult in the literature. GBM and ABC achieved a 99% recognition accuracy in recognising the fear emotion using HAF features while Kerkeni et al. (Kerkeni *et al.* 2018) achieved 76.16% using MFCC and MS features as shown in Table 17. The recognition F1-score, recall, precision and accuracy scores soared with the application of HAF features in recognising emotion. This shows that HAF features are effective in SER.

Table 16: Average Accuracy, Recall and Score and Precision on SAVEE HAF.

	Emotion						
Classifier/ Measure	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**

RDF							
Precision	100.0	100.0	94.0	100.0	100.0	100.0	100.0
Recall	100.0	94.0	100.0	100.0	100.0	100.0	100.0
F1-score	100.0	100.0	100.0	100.0	100.0	97.0	97.0
Accuracy	100.0	99.0	97.0	100.0	99.1	100.0	100.0
GBM							
Precision	100.0	96.0	100.0	100.0	100.0	100.0	100.0
Recall	100.0	100.0	94.0	100.0	100.0	100.0	100.0
F1-score	100.0	98.0	97.0	100.0	100.0	100.0	100.0
Accuracy	100.0	98.1	99.0	100.0	100.0	100.0	100.0
ABC							
Precision	100.0	94.0	100.0	100.0	100.0	100.0	100.0
Recall	100.0	100.0	96.0	100.0	100.0	100.0	100.0
F1-score	100.0	94.0	91.0	100.0	100.0	100.0	100.0
Accuracy	100.0	99.0	99.0	100.0	100.0	100.0	100.0
BSVM							
Precision	100.0	93.0	99.0	99.0	100.0	100.0	100.0
Recall	100.0	96.0	98.0	99.0	100.0	100.0	100.0
F1-score	100.0	96.0	99.0	100.0	100.0	100.0	100.0
Accuracy	100.0	94.0	96.0	96.0	100.0	100.0	100.0
BMLP							
Precision	100.0	90.0	98.0	97.0	100.0	100.0	100.0
Recall	100.0	93.0	97.0	99.0	100.0	100.0	100.0
F1-score	100.0	91.0	98.0	98.0	100.0	100.0	100.0
Accuracy	100.0	93.0	95.0	99.0	100.0	100.0	100.0

Table 17: Comparison of the proposed approach with potential works from the literature.

Reference	Database	Type of Recording	Number of Emotions	Feature Set	Classification Method	Result (%)
(Sarker and Alam, 2014)	Emo-DB	Acted	Angry, Happy, Neutral, Sad	Energy + MFCC + ZCR + voicing probability + fundamental frequency	FCBF + MVT	84.19
(Kerkeni *et al.*, 2018)	Spanish	Acted	Anger, Disgust, Fear, Neutral, Surprise, Sadness, Joy	MFCC + MS	RNN classifier	90.05
(Ying and Xue-Ying, 2018)	Emo-DB	Acted	Angry, Happy, Neutral	GCZCMT	SVM	84.45
(Altun and Polat, 2009)	Emo-DB	Acted	Anger, Happiness, Neutral, Sadness	Prosodic + sub-band + MFCC + LPC	SFS algorithm + SVM	83.00
(Bhaskar, Sruthi and Nedungadi, 2015)	eNTERFACE'05	Elicited	Disgust, Surprise, Happy, Anger, Sad, Fear,	Pitch, energy, formants, intensity and ZCR + text	SVM	90.00
(Liu, Wu, *et al.*, 2018)	CASIA	Acted	Neutral, Happy, Sadness, Fear, Angry, Surprise	Prosodic + quality characteristics + MFCC	correlation analysis + Fisher + ELM decision tree	89.60
(Jiang *et al.*, 2019)	IEMOCAP	Acted	Angry, Happy, Neutral, Sad	IS10 + MFCCs + eGemaps + SoundNet + VGGish	SVM	64.00
(Alonso *et al.*, 2015)	Emo-DB	Acted	Anger, Boredom, Happy, Neutral, Sadness	Prosodic features + paralinguistic features	SVM	94.90
(Z. T. Liu *et al.*, 2018)	CASIA	Acted	Surprise, Happy, Sad, Angry, Fear, Neutral	MFCC	GA-BEL + PCA + LDA	90.28

(L. Zhu *et al.*, 2017)	CASIA	Acted	Angry, Fear, Happy, Neutral, Surprise, Sad	MFCC, pitch, formant, ZCR and short-term energy	SVM + DBN	95.80
(Álvarez *et al.*, 2015)	Emo-DB	Acted	Sadness, Fear, Joy, Anger, Surprise, Disgust, Neutral	eGeMAPS	CSS stacking system +SVM	82.45
(Bhavan *et al.*, 2019)	Emo-DB	Acted	Anger, Happiness, Sadness, Boredom, Neutral, Disgust, Fear,	MFCCs	bagged ensemble of SVMs	92.45
(Shegokar and Sircar, 2016)	RAVDESS	Acted	Neutral, Surprise, Happy, Angry, Calm, Sad, Fearful, Disgust	CWT, prosodic coefficients	SVM	60.10
(Kerkeni *et al.*, 2019)	Spanish	Acted	Anger, Joy, Disgust, Neutral, Surprise, Fear, Sadness	SMFCC, ECC, MFF, MS and EFCC	RNN	91.16
Proposed model	RAVDESS/ SAVEE	Acted	Angry, Sad, Happy, Disgust, Calm, Fear, Neutral, Surprise	Prosodic + spectral	RDF ensemble	99.55

It can be noted from the results that a combination of HAF and ensemble classifiers significantly improves the recognition of emotions. The research findings confirm that ensemble classifiers perform better than inducers because of the effect of the synergy of the combined multiple inducers (Sagi and Rokach 2018; Dong *et al* 2020). This finding is consistent with the work presented by various researchers in the literature. Ensemble classifiers have a special ability to recognise human emotion through riding on the power of using several inducers in making positive decisions (Sagi and Rokach 2018). Furthermore, it can be seen across Tables 9-16 that the proposed HAF feature sets are the most effective acoustic features in SER, while MFCC1 are the worst-performing features. Moreover, the experimental results support the hypothesis that a fusion of prosodic and

spectral shows the essential characteristics of speech (Banse and Scherer 1996; Khan and Roy 2017; Liu *et al.* 2018).

The results of this experiment show that the proposed HAF features are highly effective in recognising emotion through speech. The comparative analysis done in this research shows that the HAF feature set is the most effective set of speech features as evidenced by the superior performance across all the performance benchmarks used. Furthermore, it was observed that the hybrid feature set eliminates the complexity and difficulty in recognising the fear emotion states which have been considered difficult to recognise. Hence, the results achieved in experiment 1 fulfill the first objective that is "*To discover a set of acoustic features that can help to improve the recognition performance of a speech emotion system for cross-language emotional conversations.*"

6.3 Experiment 2 Results and Discussion

Deep learning autogenerated features were extracted from raw spectrograms in this second experiment. Three deep classifiers were used to extract these affect-salient features to critically analyse performance. Each deep classifier extracted 16384 features because the spectrogram image size used was 128 (128 * 128). This performance was essential in answering this thesis' research question as well as fulfilling objective number 2 that is to investigate whether the discovered features (HAF) can give better performance when compared to auto-generated features by deep learning.

The processing time results of the training for this second experiment is presented in Table 18. The computational time was evaluated for each classifier with three types of deep learning autogenerated features across two speech emotion databases. The three auto-generated features in this experiment are Deep Radial Basis Neural Network Auto-generated features (DRBFNN-AGF), Deep Multilayer Perceptron Auto-generated features (DMLP-AGF), and 2D Convolution Neural Network Auto-generated features (2D-CNN-AGF). In general, the use of deep learning autogenerated features is time consuming compared to acoustic features. From Table 18, it was observed that RDF had the shortest processing time in comparison with other classifiers. Amongst

the features, DRBFNN-AGF and DMLP-AGF recorded the shortest training across all databases. The highest computational times were recorded when 2D-CNN-AGF were used. In addition, BMLP recorded the highest average computational time (24464 milliseconds) across all databases followed by GBM (23365 milliseconds). The results also show that the shortest processing times were recorded using the SAVEE database compared to RAVDESS. This difference was noticed because RAVDESS had more spectrograms.

Table 18: Processing time of classifiers trained with Deep Radial Basis Neural Network Auto-generated features (DRBFNN-AGF), Deep Multilayer Perceptron Auto-generated features (DMLP-AGF) and a 2D Convolution Neural Network Auto-generated features (2D-CNN-AGF)

	SAVEE Feature			**RAVDESS Feature**		
Classifier	**DRBFNN-AGF**	**DMLP-AGF**	**2D-CNN-AGF**	**DRBFNN-AGF**	**DMLP-AGF**	**2D-CNN-AGF**
RDF	15472	16789	28345	19423	21800	32600
GBM	16977	17396	29455	19988	22589	33785
ABC	16149	16946	29133	18978	22134	33155
BSVM	15954	16122	28763	18166	21878	32968
BMLP	17224	17956	32755	21386	23174	34291

From the results illustrated in Table 19, it was observed that higher average precision, F1-score, recall and accuracy scores were achieved using 2D-CNN generated features on both the RAVDESS and SAVEE database. RDF achieved the highest scores with an 85% accuracy score in recognising accuracy on the RAVDESS database. GBM had the second-highest accuracy score (83%) while BMLP achieved the lowest score (71.6%). However, the comparatively higher scores came at a cost. GBM took approximately 32600 milliseconds to process the features. DRBFNN-AGF achieved the least scores across all performance benchmarks but had the least processing time since it took 19988 milliseconds for GBM to process the features. The confidence factors shown in Table 19 indicate that the estimated ranges are appropriate because the confidence intervals are in the range between 0.042 and 0.019. In addition, the table shows that on the SAVEE database, RDF still was the best performing classifier across all the others. The same pattern of

performance was observed when the SAVEE database was used. RDF outperformed the other classifiers and BMLP was again the least performing classifier across all the features used.

Table 19: Average accuracy, recall, F1-score and precision with confidence intervals of classifiers trained with DRBFNN-AGF, DMLP-AGF and 2D-CNN-AGF

	SAVEE Feature			**RAVDESS Feature**		
Classifier/ Measure	**DRBFNN-AGF**	**DMLP-AGF**	**2D-CNN-AGF**	**DRBFNN-AGF**	**DMLP-AGF**	**2D-CNN-AGF**
			RDF			
Precision	71.3 (±0.041)	75.9 (±0.039)	79.7 (±0.036)	74 (±0.023)	80 (±0.021)	83.0 (±0.02)
Recall	71.1 (±0.041)	74.7 (±0.039)	79.6 (±0.036)	73 (±0.023)	81 (±0.02)	84.0 (±0.019)
F1-score	71.4 (±0.041)	74.3 (±0.04)	79.4 (±0.036)	73 (±0.023)	80 (±0.021)	84.0 (±0.019)
Accuracy	71.6 (±0.041)	76.7 (±0.038)	81.3 (±0.035)	74 (±0.023)	81 (±0.02)	85.0 (±0.019)
			GBM			
Precision	68.0 (±0.042)	74.1 (±0.04)	78.7 (±0.037)	70 (±0.024)	78 (±0.022)	82.0 (±0.02)
Recall	68.1 (±0.042)	73.7 (±0.04)	78.4 (±0.037)	69 (±0.024)	78 (±0.022)	81.0 (±0.02)
F1-score	68.3 (±0.042)	75.0 (±0.039)	78.1 (±0.037)	71 (±0.024)	79 (±0.021)	81.0 (±0.02)
Accuracy	69.3 (±0.042)	75.6 (±0.039)	79.3 (±0.036)	71 (±0.024)	79 (±0.021)	83.0 (±0.02)
			ABC			
Precision	64.7 (±0.043)	74.7 (±0.039)	78.3 (±0.036)	67 (±0.024)	73 (±0.023)	78.0 (±0.022)
Recall	64.3 (±0.043)	73.7 (±0.04)	76.3 (±0.037)	68 (±0.024)	72 (±0.023)	76.0 (±0.022)
F1-score	64.1 (±0.043)	74.1 (±0.04)	76.6 (±0.037)	67 (±0.024)	72 (±0.023)	77.0 (±0.022)

Accuracy	65.4 (±0.043)	74.8 (±0.039)	77.0 (±0.038)	68 (±0.024)	74 (±0.023)	79.0 (±0.021)
			BSVM			
Precision	61.0 (±0.044)	70.7 (±0.041)	74.1 (±0.04)	64 (±0.025)	69 (±0.024)	76.0 (±0.022)
Recall	63.0 (±0.044)	70.4 (±0.041)	74.0 (±0.04)	65 (±0.025)	70 (±0.024)	77.0 (±0.022)
F1-score	61.0 (±0.044)	71.1 (±0.041)	74.3 (±0.04)	65 (±0.025)	70 (±0.024)	76.0 (±0.022)
Accuracy	62.1 (±0.044)	72.3 (±0.04)	75.4 (±0.039)	66 (±0.025)	71 (±0.024)	77.0 (±0.022)
			BMLP			
Precision	59.3 (±0.044)	67.2 (±0.042)	71.4 (±0.041)	61.5 (±0.025)	65.4 (±0.025)	73.8 (±0.023)
Recall	58.7 (±0.045)	68.7 (±0.042)	72.8 (±0.04)	62.7 (±0.025)	66.2 (±0.025)	74.7 (±0.023)
F1-score	60.5 (±0.044)	67.6 (±0.042)	72.9 (±0.04)	62.3 (±0.025)	66.9 (±0.024)	73.4 (±0.023)
Accuracy	59.1 (±0.044)	68.5 (±0.042)	71.6 (±0.041)	61.9 (±0.025)	66.6 (±0.025)	74.9 (±0.023)

The results in Table 20 show that fear was difficult to recognise using DRBFNN-AGF across all ensemble classifiers. RDF outperformed the other classifiers in generally recognising emotions. RDF achieved the highest accuracy score in recognising angry emotion (79%). However, GBM achieved the highest recognition accuracy score in recognising the neutral emotion. It was noted that BSVM and BMLP are very poorly in recognising the fear emotion using DRBFNN-AGF because they achieved 59% and 58% accuracy scores respectively. The precision, F1-score, and recall scores were generally low across all the ensemble classifiers.

Table 20: Average Accuracy, Recall and F1-Score and Precision on RAVDESS DRBFNN-AGF

Classifier/	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**

Measure								
RDF								
Precision	74	74	73	74	77	74	75	71
Recall	78	71	73	69	74	74	70	75
F1-Score	76	72	66	77	77	72	68	76
Accuracy	79	75	74	72	73	72	73	76
GBM								
Precision	73	69	72	70	66	70	68	72
Recall	70	72	72	66	68	69	71	64
F1-Score	73	73	72	73	70	68	70	69
Accuracy	75	73	67	65	74	74	69	71
ABC								
Precision	67	64	68	60	71	71	66	69
Recall	72	66	71	67	63	71	67	67
F1-Score	69	64	68	60	71	71	66	58
Accuracy	73	71	73	60	71	68	64	64
BSVM								
Precision	67	60	64	59	68	66	62	66
Recall	65	68	68	66	60	63	66	64
F1-Score	66	67	66	58	64	67	68	64
Accuracy	70	68	64	59	64	68	66	69
BMLP								
Precision	61	62	59	58	59	66	64	67
Recall	60	63	60	59	60	68	66	68

F1-Score	60	61	61	58	63	64	62	67
Accuracy	60	59	59	58	61	66	66	67

An improvement was observed when DMLP-AGF was used on the RAVDESS database according to results in Table 21. This improvement was particularly noted in the recognition of the fear emotion because it improved from 72% to 77% using RDF. BMLP continued with the same trend noticed in the previous table because it was the lowest-performing classifier in recognising angry (66%), calm (65%), and fear (66%) emotions. GBM achieved the highest F1-score result in recognising the sad emotion. In the same vein, GBM achieved the highest accuracy scores in recognising the calm and fear emotions. Even though ABC was the third-best in the average recognition of the emotions, it was the joint best performing classifier in recognising the sad emotion.

Table 21: Average Accuracy, Recall and F1-Score and Precision on RAVDESS DMLP-AGF

Classifier/ Measure	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
RDF								
Precision	83	79	79	80	83	76	76	84
Recall	85	84	76	84	79	78	85	77
F1-Score	82	77	81	80	78	83	77	82
Accuracy	84	76	84	77	79	85	83	80
GBM								
Precision	82	76	79	80	75	76	76	80
Recall	80	80	76	74	80	75	80	79
F1-Score	79	80	78	78	81	79	78	79
Accuracy	81	79	78	79	79	78	78	80
ABC								

Precision	71	68	73	76	75	75	72	74
Recall	74	74	73	74	69	70	69	73
F1-Score	75	71	72	70	68	76	73	71
Accuracy	75	72	77	75	70	75	79	69
BSVM								
Precision	74	71	66	70	67	66	71	67
Recall	70	74	69	65	74	68	73	67
F1-Score	72	74	74	65	70	68	67	70
Accuracy	74	67	67	67	77	75	74	67
BMLP								
Precision	63	65	65	63	67	65	64	68
Recall	66	69	64	67	65	64	66	67
F1-Score	67	68	66	66	66	67	66	70
Accuracy	66	65	67	66	67	68	66	71

Table 22 showed that 2D-CNN-AGF is efficient in recognising human emotion compared to DRBFNN-AGF and DMLP-AGF. RDF and GBM achieved the highest accuracy scores in recognising because they both had accuracies of 87% each. An improvement was also noted in recognising the fear emotion. RDF had the biggest improvement in this regard because the accuracy score was increased from 77% to 83%. However, the fear emotion recognition accuracy did not change because it remained at 75% as shown in Tables 21 and 22.

Table 22: Average Accuracy, Recall, F1-Score and Precision on RAVDESS 2D-CNN-AGF

Classifier/ Measure	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
RDF								

Precision	87	85	83	85	83	79	81	82
Recall	85	83	84	86	87	81	81	85
F1-Score	87	82	86	85	86	79	82	85
Accuracy	87	84	86	83	84	87	86	83
GBM								
Precision	85	86	82	78	80	78	82	86
Recall	86	85	85	77	78	77	78	85
F1-Score	84	82	82	82	76	82	80	81
Accuracy	87	83	82	81	82	81	81	87
ABC								
Precision	81	81	75	81	69	80	75	82
Recall	81	81	80	79	70	70	72	75
F1-Score	82	82	71	79	74	77	81	70
Accuracy	83	81	79	75	75	84	71	83
BSVM								
Precision	82	74	77	71	77	79	81	67
Recall	82	73	68	76	79	82	77	79
F1-Score	80	78	76	71	71	78	77	76
Accuracy	83	81	79	71	79	75	72	76
BMLP								
Precision	75	74	72	72	73	72	80	74
Recall	73	74	73	74	73	77	79	77
F1-Score	74	72	71	71	72	74	76	74
Accuracy	76	74	75	75	73	75	79	76

The results shown in Table 23 show that GBM was the best performing algorithm in recognising the fear emotion achieving a recognition accuracy of 70% followed by RDF which had 69%. The worst performing algorithm in this regard was BMLP (57%) followed by BSVM which achieved an accuracy score of 58%. RDF achieved the best accuracy scores in recognising angry, happy, neutral, sad, and surprising emotions. The poor results achieved in recognising the fear emotion are consistent with the findings observed in the literature.

Table 23: Average Accuracy, Recall, F1-Score and Precision on SAVEE DRBFNN-AGF

Classifier/ Measure	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
RDF							
Precision	71	72	67	73	71	74	71
Recall	72	70	69	71	73	70	73
F1-Score	72	72	69	73	71	71	72
Accuracy	74	70	69	70	73	74	71
GBM							
Precision	70	69	67	65	67	68	70
Recall	68	69	67	69	67	69	68
F1-Score	69	67	69	69	67	70	67
Accuracy	70	71	70	68	67	69	70
ABC							
Precision	67	67	62	63	61	65	67
Recall	69	64	63	61	63	67	63
F1-Score	66	66	61	61	62	68	65

Accuracy	69	66	64	63	64	66	66
BSVM							
Precision	65	60	61	61	60	60	60
Recall	66	61	61	61	65	63	64
F1-Score	62	61	62	60	60	60	62
Accuracy	64	61	58	60	61	64	67
BMLP							
Precision	58	60	57	57	58	60	63
Recall	58	59	59	57	61	58	61
F1-Score	59	60	59	61	59	62	67
Accuracy	57	58	57	60	56	60	65

The results in Table 24 show the superiority of DMLP-AGF against other auto-generated features. All the classifiers improved their efficiency in recognising the emotions compared to the results reported in Table 23. RDF and GBM were the joint best classifiers in recognising the angry emotional state achieving recognition scores of 79%. ABC and RDF were the best classifiers in recognising the neutral emotional state because they achieved accuracy scores of 74%. GBM outperformed all the other classifiers in recognising the happy emotion state achieving an accuracy score of 76%. Although BSVM was the second-worst performing classifier using DMLP-AGF, it was second-best (76%) to RDF (78%) in recognising the surprise emotion state.

Table 24: Average Accuracy, Recall, F1-Score and Precision on SAVEE DMLP-AGF

Classifier/ Measure	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
RDF							
Precision	78	78	79	74	74	75	73

Recall	74	79	75	72	73	75	75
F1-Score	77	75	79	74	71	73	71
Accuracy	79	81	77	71	74	77	78
GBM							
Precision	77	79	74	73	73	71	72
Recall	74	78	71	70	71	73	76
F1-Score	78	78	75	73	76	72	73
Accuracy	79	77	79	76	71	73	74
ABC							
Precision	76	77	73	72	74	73	71
Recall	75	77	75	70	73	72	74
F1-Score	77	77	75	73	72	74	71
Accuracy	78	77	71	74	74	75	74
BSVM							
Precision	69	70	71	67	72	73	73
Recall	71	72	71	69	69	71	70
F1-Score	74	72	73	71	67	70	71
Accuracy	71	73	71	68	73	74	76
BMLP							
Precision	66	67	66	69	67	65	69
Recall	67	69	70	69	69	68	71
F1-Score	69	66	65	69	66	69	72
Accuracy	68	70	67	69	67	69	73

Notable improvements were noted when 2D-CNN-AGF features were used. These results are illustrated in Table 25 and they show that RDF outperformed all the other classifiers in recognising all the emotions. Even though GBM took longer to process the features, it was the second-best performing classifier recognising anger, disgust, fear, happiness, sadness, and surprise emotion state. ABC was the second-best performing classifier in recognising the neutral emotion. BMLP was generally the least performing classifier followed by BSVM. Even though BSVM was one of the worst-performing classifiers, it performed well in recognising the neutral emotion (83%) where it outperformed GBM (80%).

Table 25: Average Accuracy, Recall, F1-Score and Precision on SAVEE 2D-CNN-AGF

Classifier/ Measure	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
RDF							
Precision	83	86	75	82	89	70	73
Recall	82	82	68	82	89	77	77
F1-Score	82	84	75	78	89	73	75
Accuracy	84	83	81	76	89	79	77
GBM							
Precision	81	79	79	75	86	75	76
Recall	79	80	77	76	87	76	74
F1-Score	80	79	75	78	78	78	79
Accuracy	83	81	77	75	80	78	81
ABC							
Precision	77	87	79	79	83	77	66
Recall	72	77	70	73	84	78	80

F1-Score	74	81	74	76	84	78	69
Accuracy	77	80	73	73	86	79	71
BSVM							
Precision	76	77	72	69	79	75	71
Recall	73	75	74	65	79	78	74
F1-Score	75	78	73	67	73	78	76
Accuracy	74	79	75	71	83	77	69
BMLP							
Precision	72	66	72	70	73	70	74
Recall	70	69	76	73	73	77	73
F1-Score	71	69	76	70	76	77	72
Accuracy	68	67	68	73	76	78	74

The implications of using deep learning auto-generated features show their deficiencies in effectively recognising human emotion. These features are less efficient in recognising speech emotion compared to HAF, the proposed handcrafted features. From experiment 2, the research work revealed that AGF is not very good at recognising the fear emotion state which has been difficult to recognise according to the results recorded in literature. Furthermore, it was also observed that ADF generates a lot of features depending on the size of the spectrogram images and this hurts the processing time of the SER models proposed. For instance, the spectrogram images used in this experimental study were of size 128 therefore 16384 features were extracted by each deep learning model and the feature extraction processing time also varies depending on the number of layers found in each model. This experiment answered this thesis' research questing showing that HAF features are indeed the most efficient features in recognising human emotion.

6.4 Experiment 3 End-to-End Deep Learning Models

It was established that deep learning auto-generated features are less efficient in recognising human emotion compared to HAF features in the previous experiments. However, experiment 3 was designed to assess the strength of deep learning models on the autogenerated features to investigate the efficiency of deep learning classifiers on the very same AGF features used in the previous experiment. Therefore, end-to-end deep learning models were developed using 2D-CNN, DRBFNN, and DMLP. Another database was added to validate the findings, and this was the EMODB database described in chapter 5.

Upon completion of the experimental models, it was observed that DRBFNN had the fastest processing time across all the emotion corpora. It only took 57.4 seconds for DRBFNN to extract features and classify the corresponding emotions from the RAVDESS emotion corpus as shown in Table 26. Besides being the fastest classifier, DRBFNN also recorded the highest recognition accuracy (98.22%). Even though 2D-CNN recorded a joint top accuracy score with DRBFNN, it took too long to extract and classify the emotional states. During the last run, it took approximately 25200 seconds to process the RAVDESS spectrograms which is equivalent to approximately 7 hours. In addition, the deep learning model was run five times to boost the accuracy and this lasted for 35 hours which is approximately equivalent to a full day and 9 extra hours. Furthermore, it was observed that 2DCNN requires more epochs to achieve high accuracies as shown in Figures 44, 45, and 46. DRBFNN achieved a recognition score of 93.95% and it took approximately 66 seconds to complete the process. The performance pattern was similar when the classifiers were evaluated on the EMODB and SAVEE emotion database as shown in Tables 26, 27, and 28 as well as Figures 38 to 55.

Table 26: Training Time.

Classifier	EMODB	Number of Runs	RAVDESS	Number of Runs	SAVEE	Number of Runs
DRBFNN	7.1 seconds	1	57.4 seconds	1	9.4 seconds	1
DMLP	19.8 seconds	1	66 seconds	1	38.4 seconds	1
2D-CNN	1170 seconds	3	25200 seconds	5	14400 seconds	4

Table 27: Accuracy analysis on speech emotion spectrograms.

Classifier	EMODB	RAVDESS	SAVEE
DRBFNN	97.05	98.22	93.75
DMLP	94.11	93.95	97.91
2D-CNN	93.51	98.22	90.2

Table 28: Validation loss analysis on speech emotion spectrograms.

Classifier	EMODB	RAVDESS	SAVEE
DRBFNN	0.085	0.063	0.156
DMLP	0.112	0.263	0.124
2D-CNN	0.121	0.119	0.387

From the validation loss results illustrated in Table 28 and Figures 47 to 55, it can be noted that DRBFNN had the lowest test loss when the classifiers were evaluated on the EMODB database. DRBFNN achieved a test loss of 0.085%. Although 2D-CNN had the longest processing time, the deep learning classifier achieved the lowest test loss (0.121%) while DMLP achieved 0.112%. The same trend was observed when DRBFNN was evaluated on the RAVDESS database. It achieved an average of 0.063% test loss. On the contrary, MLP outperformed all the other classifiers since it recorded the lowest test loss score (0.124%) when it was validated on the SAVEE database. Meanwhile, CNN achieved the highest test loss on the SAVEE database.

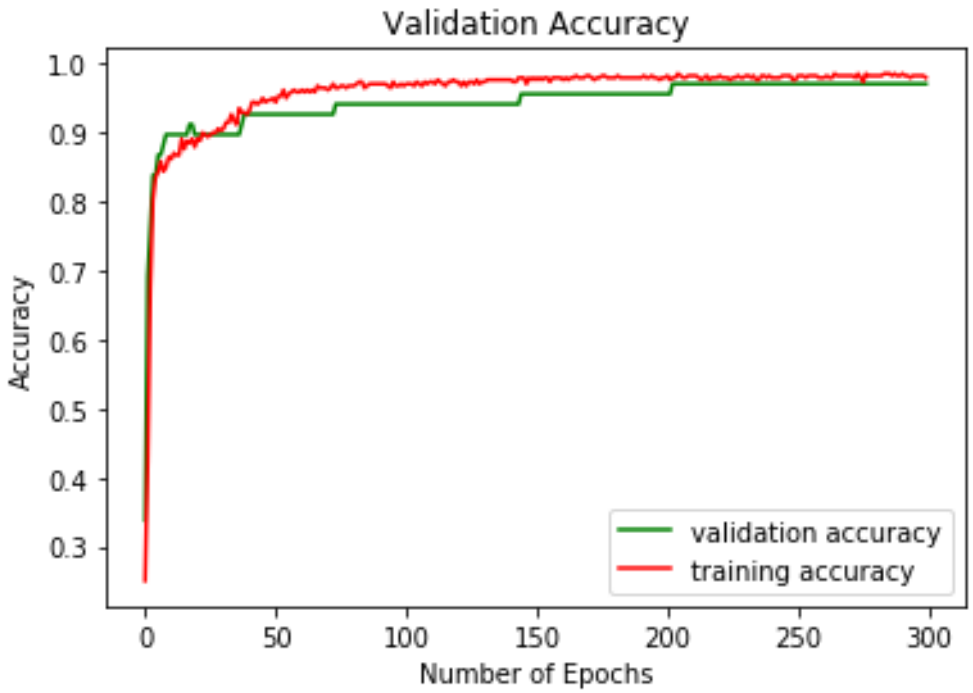


Figure 38: DRBFNN validation accuracy on EMODB.

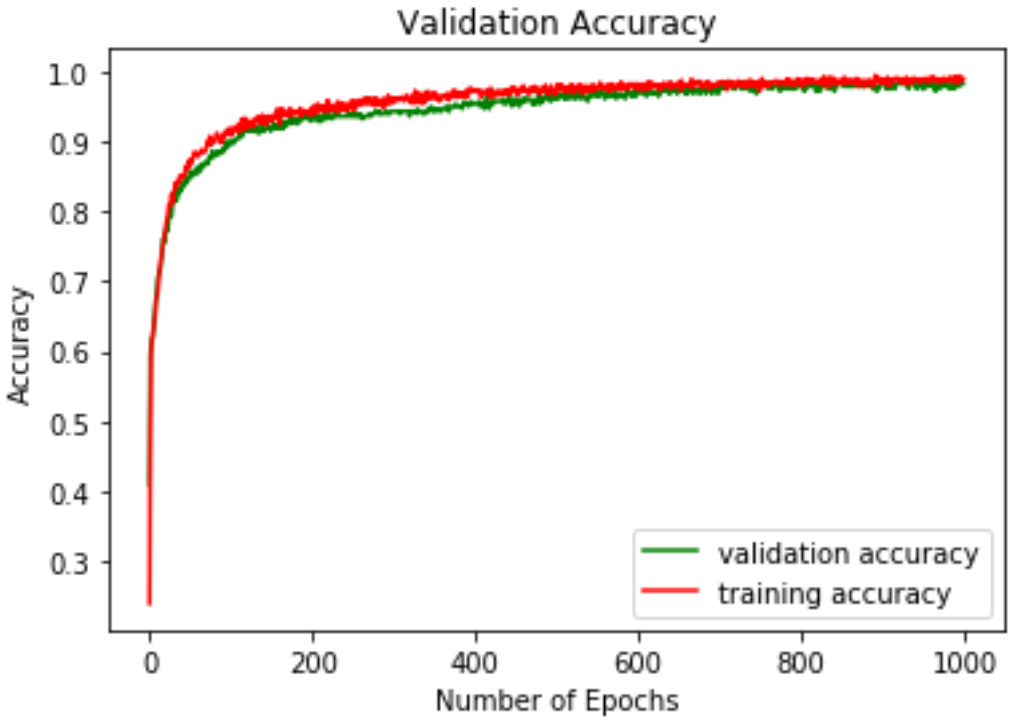


Figure 39: DRBFNN validation accuracy on RAVDESS.

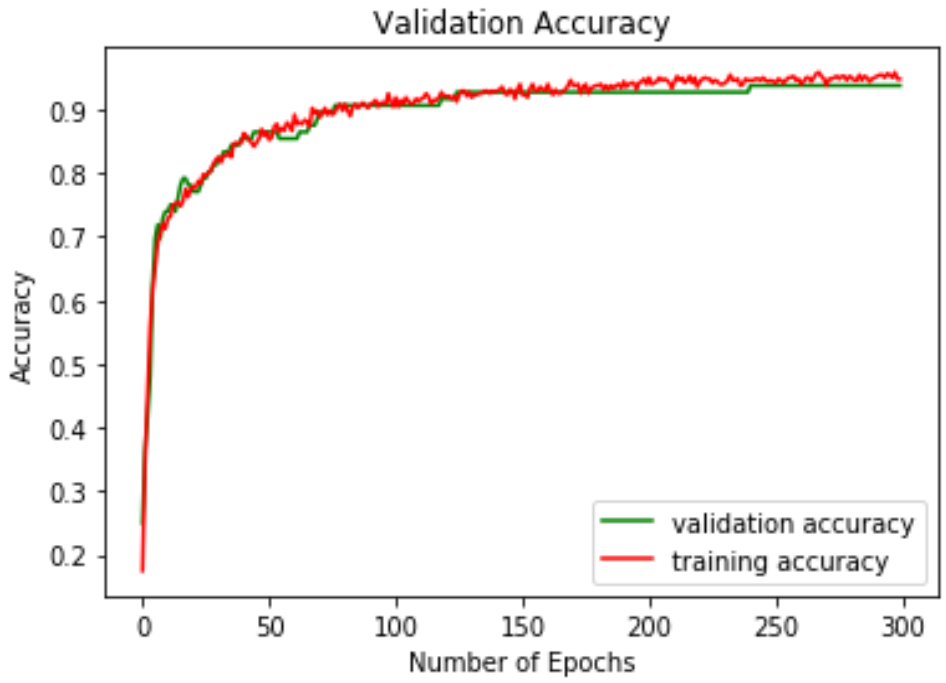


Figure 40: DRBFNN validation accuracy on SAVEE.

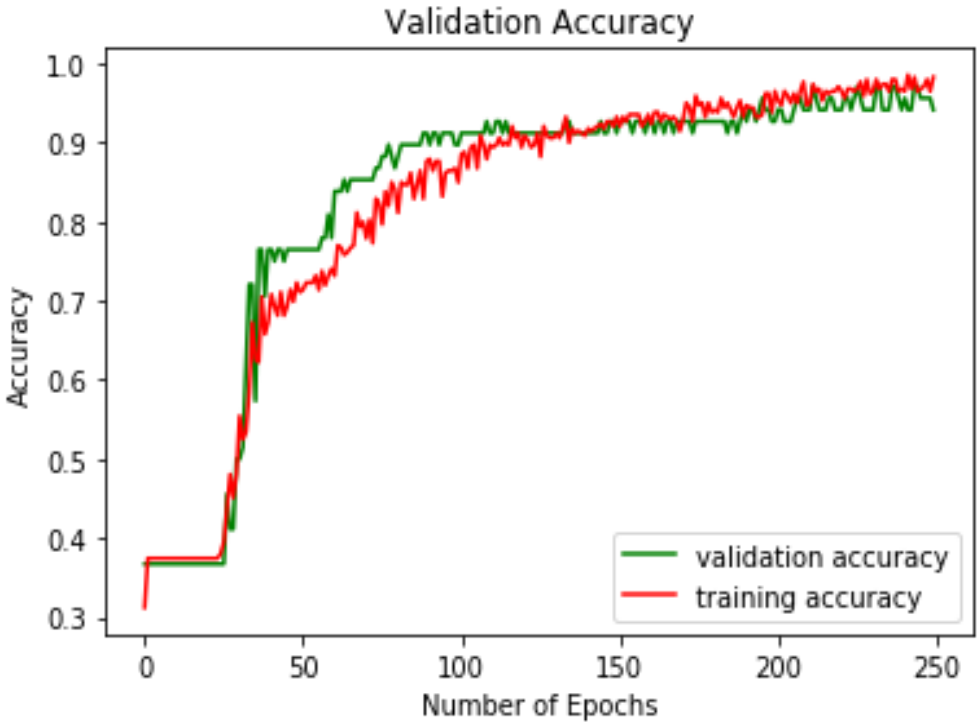


Figure 41: DMLP validation accuracy on EMODB.

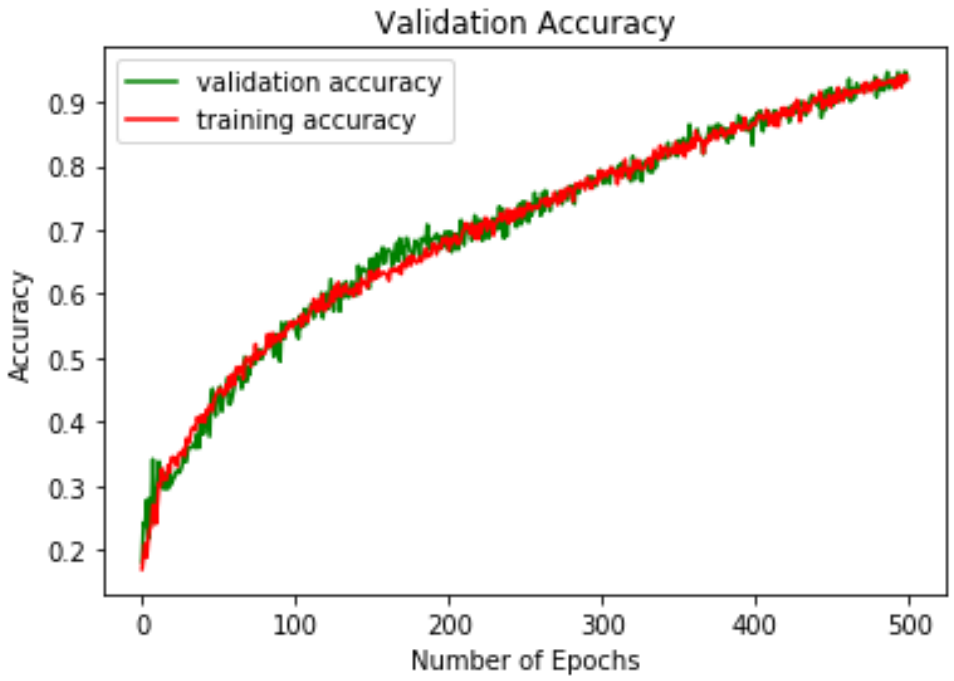


Figure 42: DMLP validation accuracy on RAVDESS.

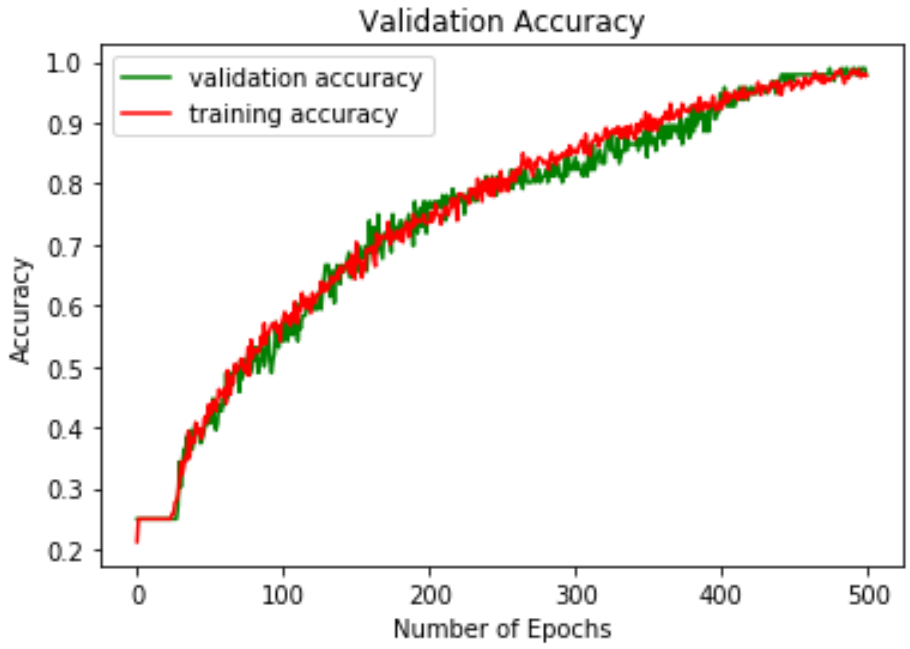


Figure 43: DMLP validation accuracy on SAVEE.

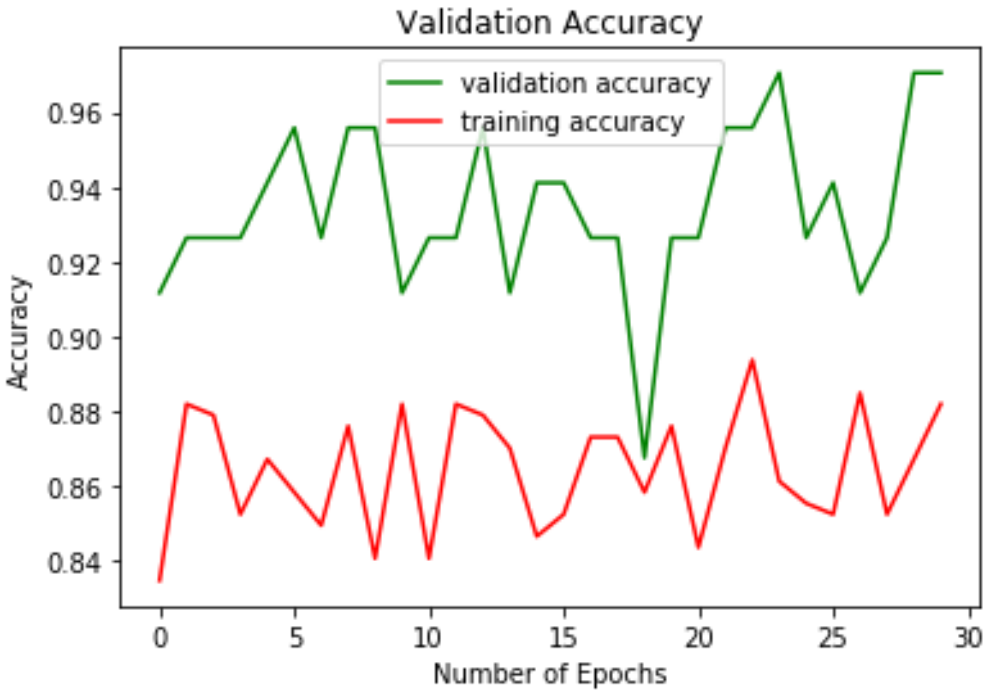


Figure 44: 2D-CNN validation accuracy on EMODB.

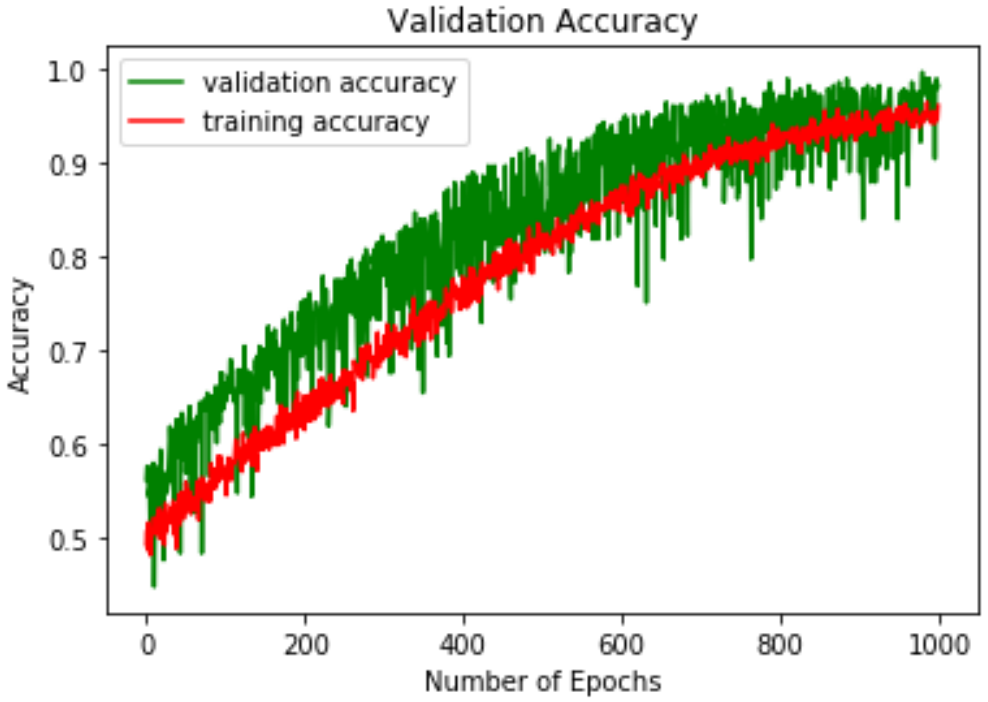


Figure 45: 2D-CNN validation accuracy on RAVDESS.

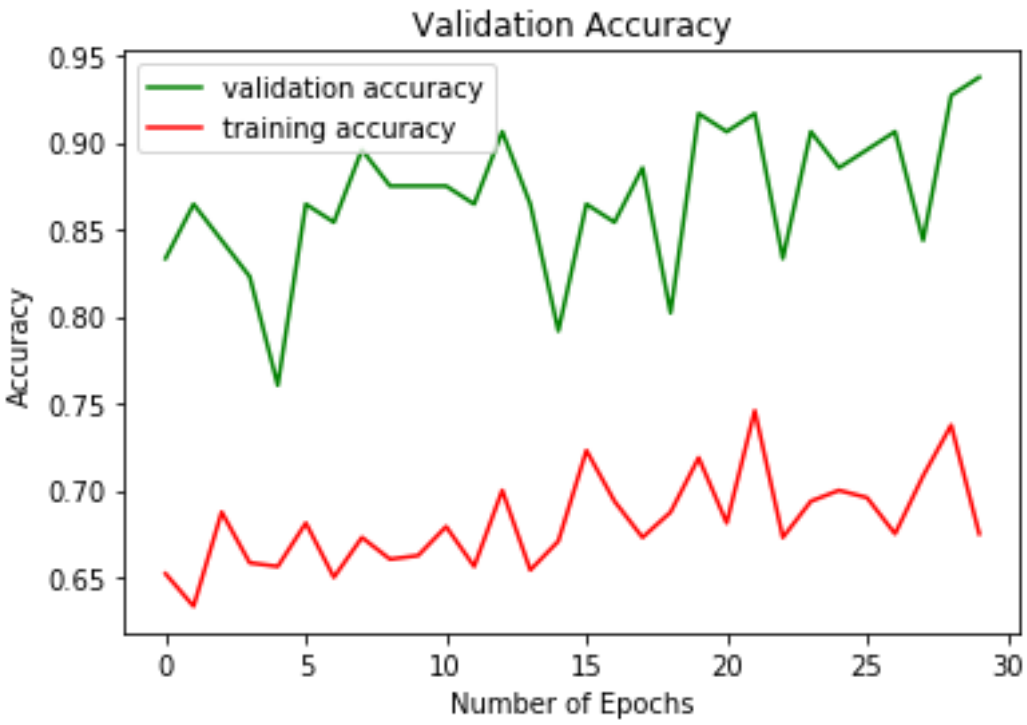


Figure 46: 2D-CNN validation accuracy on SAVEE.

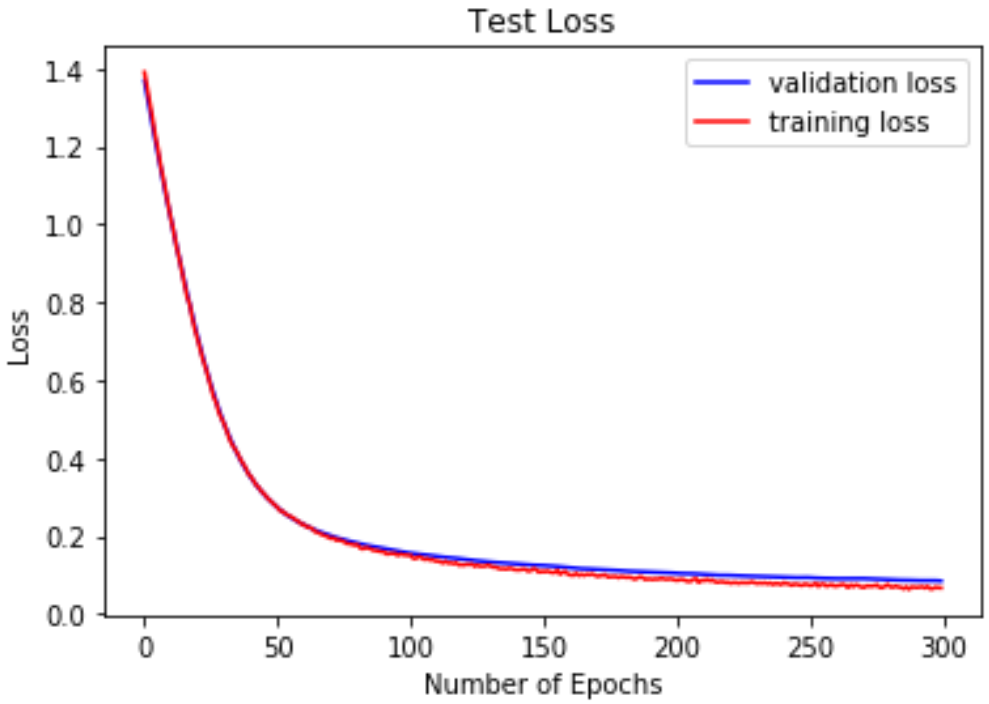


Figure 47: RBFNN test loss on EMODB.

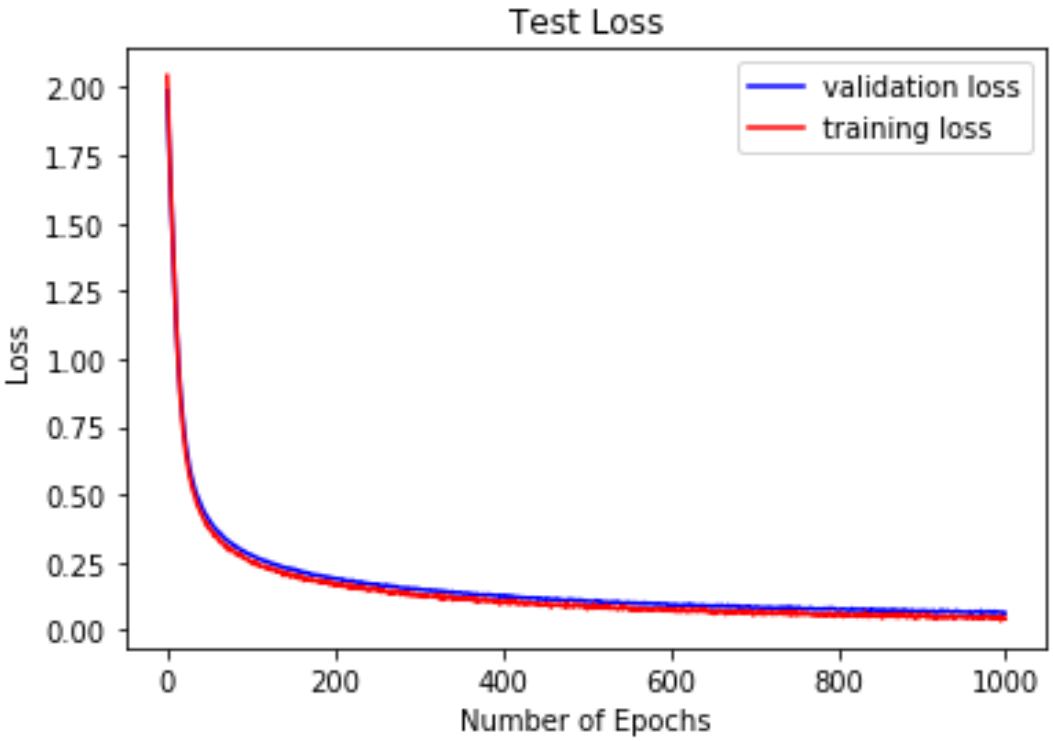


Figure 48: DRBFNN test loss on RAVDESS.

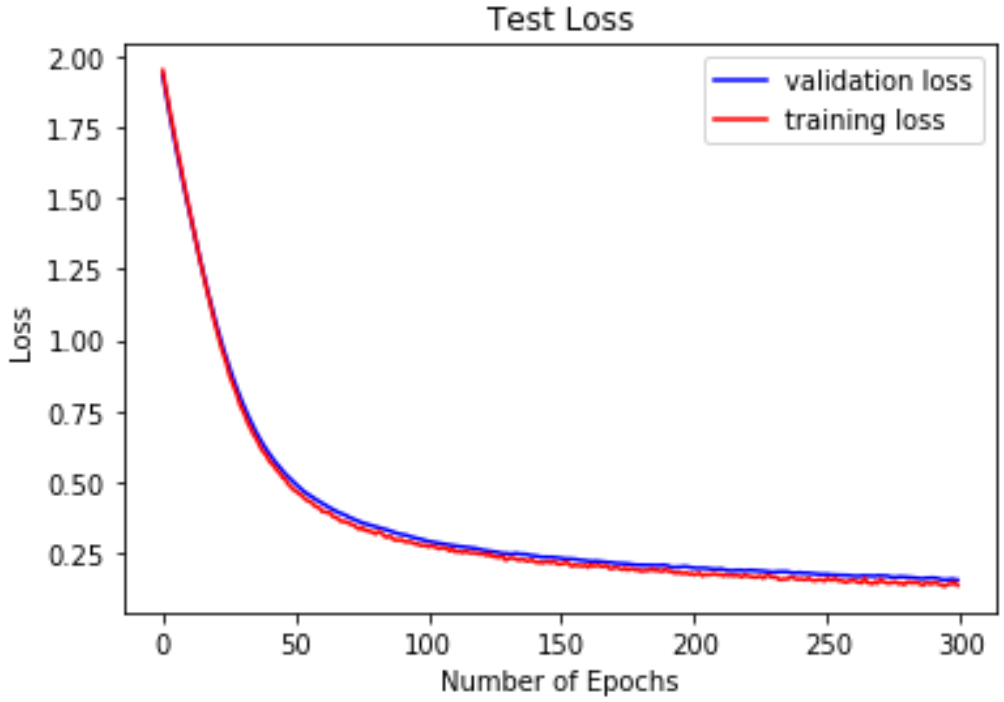


Figure 49: DRBFNN test loss on SAVEE.

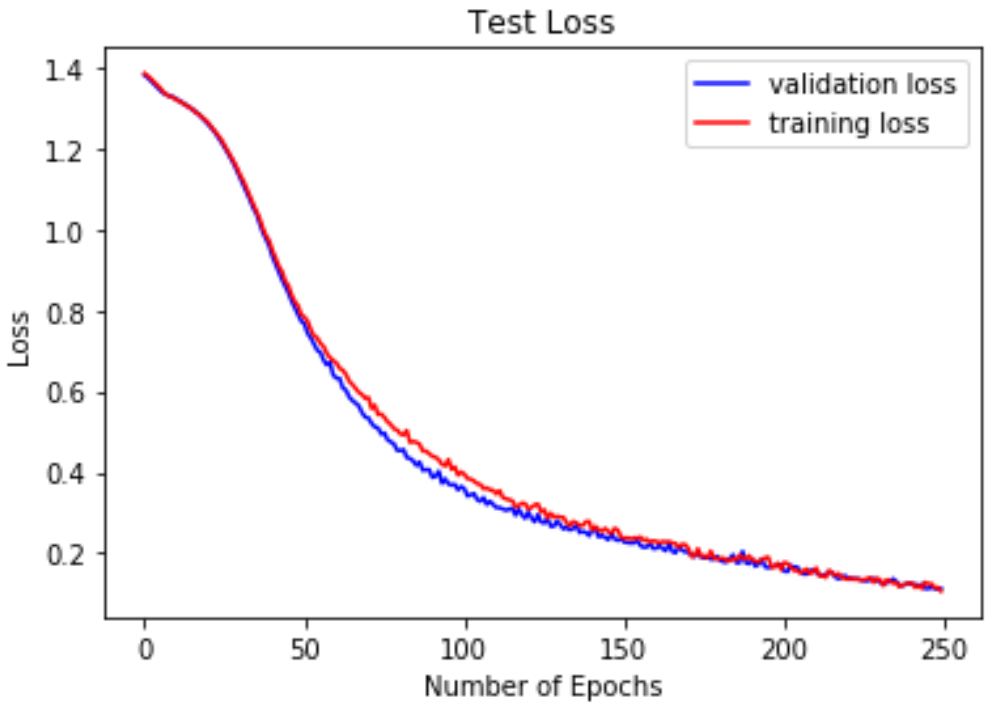


Figure 50: DMLP test loss on EMODB.

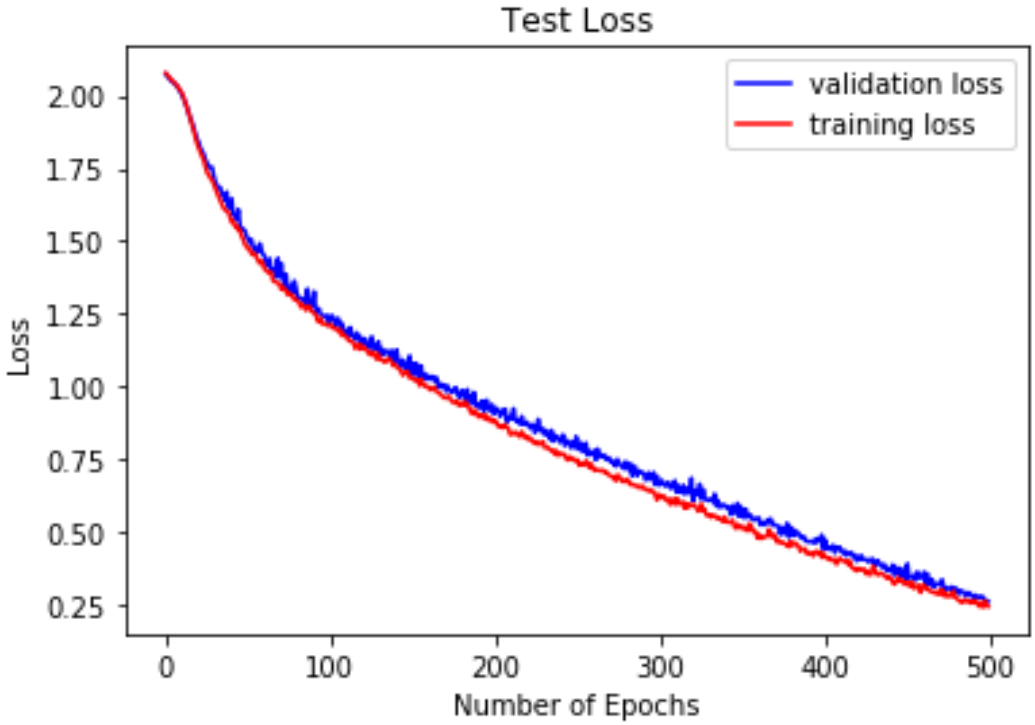


Figure 51: DMLP test loss on RAVDESS.

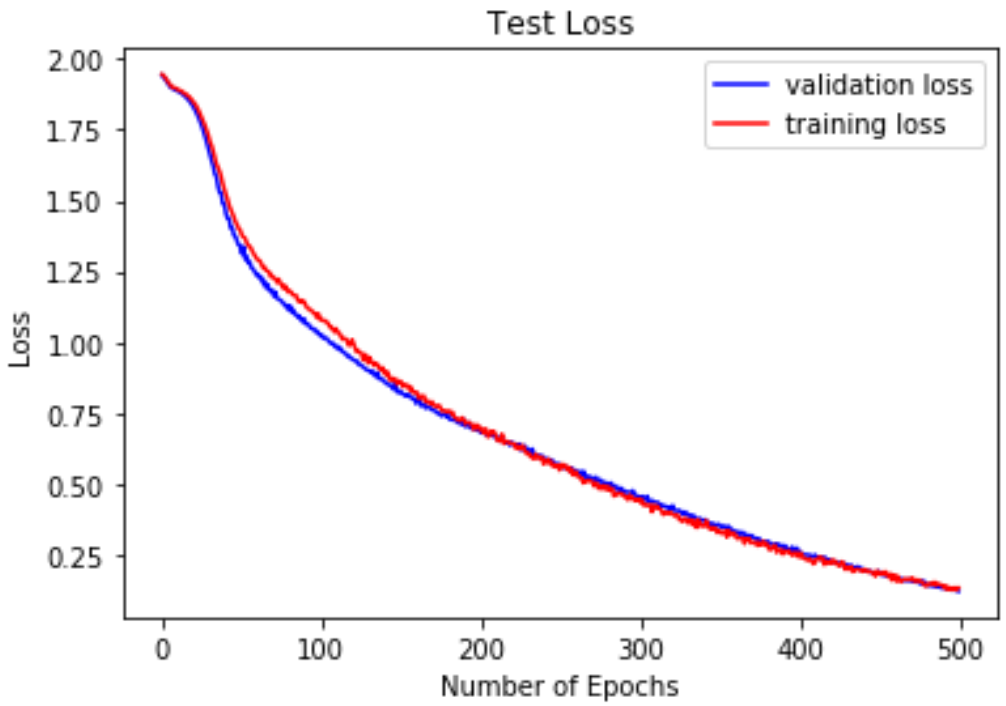


Figure 52:DMLP test loss on SAVEE.

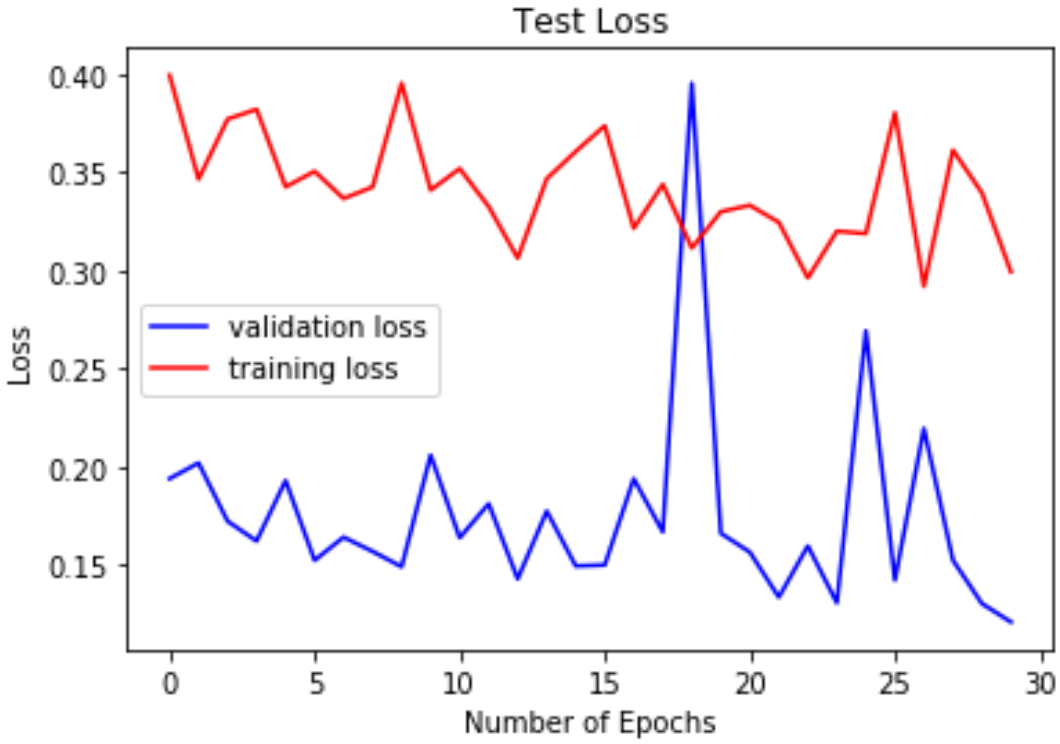


Figure 53: 2D-CNN test loss on EMODB.

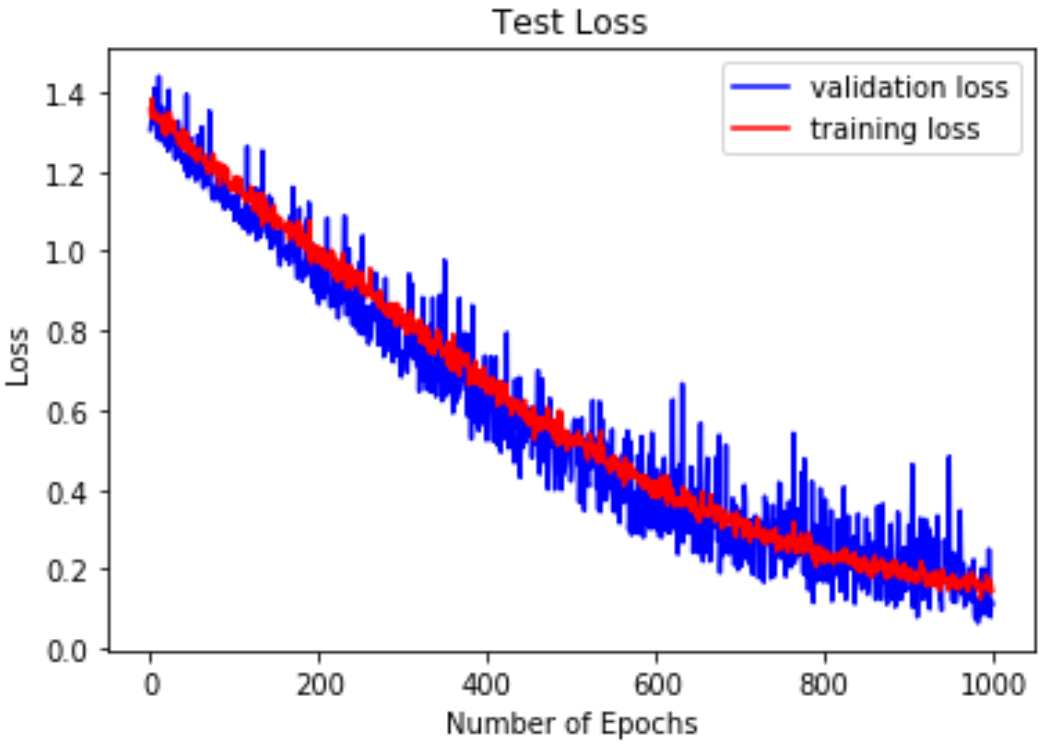


Figure 54: 2D-CNN test loss on RAVDESS.

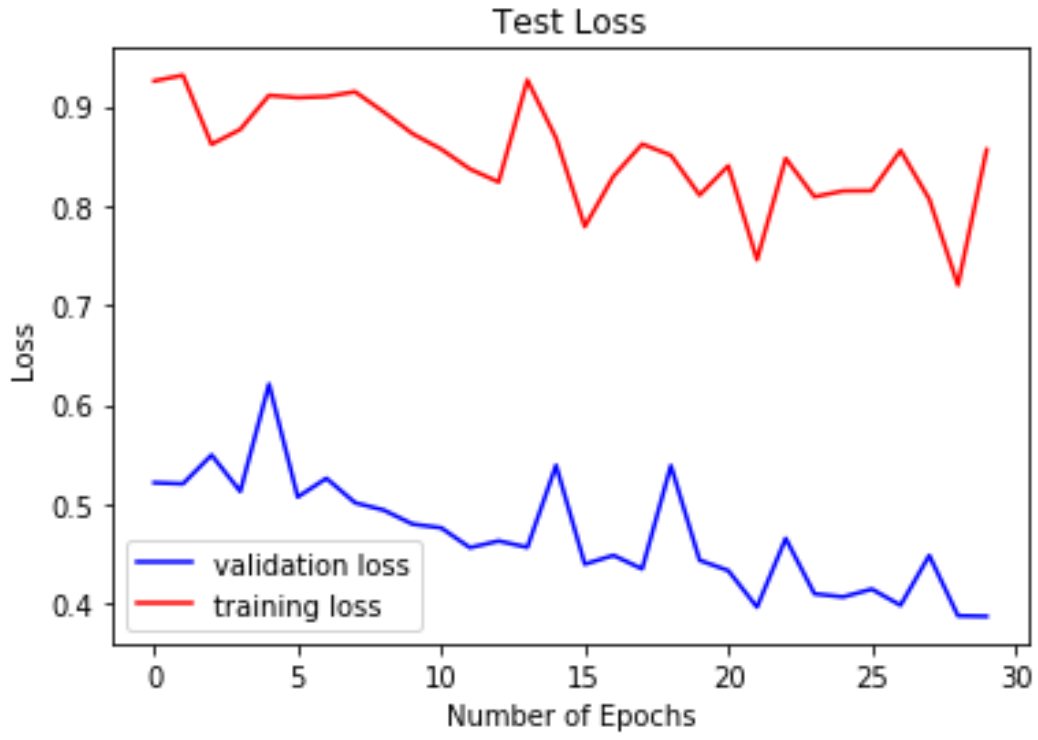


Figure 55: 2D-CNN test loss on SAVEE.

The f1-score, recall, and precision scores of the three end-to-end deep learning classifiers on the RAVDESS database are presented in Table 29. In this regard, 2D-CNN achieved the highest precision scores across all emotion classes. It achieved 100% precision scores in most emotion classes. However, it failed to make the perfect scores in the calm and fearful emotions. The model achieved the highest overall precision (98%), recall (98%), and f1-score (98%). DMLP yielded the highest scores when the classifiers were evaluated on the SAVEE database. It consistently achieved a 98% score for each benchmark as presented in Table 30. DMLP was the best performing classifier in detecting the sad, disgust, happy, neutral, and angry emotion classes since the model obtained perfect scores (100%).

Table 31 shows that DRBFNN outperformed the other classifiers in recognising the happy emotion class (96%) when the classifiers were evaluated on the EMODB database. Nevertheless, 2D-CNN achieved the best overall precision (97%), recall (97%), and f1-score (97%). The model achieved precision scores of 100% in recognising both the neutral and sad emotion classes. Additionally, DMLP outperformed all the other classifiers with regards to precision when the angry emotion was interrogated. The comparison of the results obtained in this experimental study and results presented in the literature review in chapter 3 is illustrated in Table 32. These results also show that the features mined from raw spectrograms have significant discriminative power, especially when fed to Artificial Neural Networks as presented in Table 25.

Table 29: F1-Score, Recall, and Precision Analysis on Ravdess Spectrograms.

Emotion	DRBFNN			DMLP			CNN		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Angry	76	48	59	97	97	97	100	100	100
Calm	52	77	62	87	94	90	95	100	97
Disgust	78	60	68	100	95	97	100	100	100
Fear	64	69	66	100	97	99	92	100	96
Happy	70	61	65	89	100	94	100	95	97

Neutral	55	35	43	100	86	93	100	95	97
Sad	77	64	70	87	87	87	100	97	99
Surprised	54	76	63	94	91	93	100	097	099
Mean	66	64	64	94	94	94	98	98	98

Table 30: F1-Score, Recall and Precision Analysis on Savee Spectrograms.

Emotion	**RBF**			**MLP**			**CNN**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Angry	88	78	82	100	100	100	92	100	96
Disgust	100	78	88	100	100	100	92	100	96
Fear	67	67	67	92	92	92	92	92	92
Happy	93	81	87	100	92	96	92	92	92
Neutral	14	50	22	100	100	100	100	96	98
Sad	88	62	73	100	100	100	92	100	96
Surprised	75	100	86	92	100	96	92	79	85
Mean	84	79	80	98	98	98	94	94	94

Table 31: F1-Score, Recall and Precision Analysis on Emodb Spectrograms.

Emotion	RBF			MLP			CNN		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	89	89	89	100	96	98	96	96	96
happy	100	92	96	86	100	92	93	93	93
neutral	73	85	79	100	84	91	100	100	100
sad	96	92	94	85	100	92	100	100	100
Mean	90	90	90	95	94	94	97	97	97

Table 32: Comparison of the proposed approach with related work from the literature.

Reference	Features	Classifiers	Database	Accuracy (%)
(Interaction *et al.*, 2013)	Spectrograms	Sparse autoencoder-based	eNTERFACE	59.5
(Ariav and Cohen, 2019)	Spectrograms	WaveNet encoder, ResNet, LSTM	Custom database	91.52
(Hifny and Ali, 2019)	Spectrograms	An attention-based CNNLSTM-DNN	IEMOCAP	87.2
(Zhu *et al.*, 2017)	Spectrograms	Support Vector Machines and DBN	Chinese Academy of Sciences	95.8
(Zhang *et al.*, 2016)	Spectrograms and Videos	CNN	RAVDESS	74.32
(Bonfiglio, 2017)	Spectrograms	CNN	RECOLA	74.1

(Petridis *et al.*, 2018)	Spectrograms and Videos	Audio waveforms and performs recognition using BGRUs (Bidirectional Gated Recurrent Units)	Lip Reading in the Wild (LRW) database	98.0
(Harár, Burget and Dutta, 2017)	Spectrograms	A Stochastic Gradient Descent optimised DNN	EMO-DB	96.67
(Poria, Chaturvedi and Cambria, 2016)	Spectrograms	CNN + RNN	Multimodal Sentiment Analysis	96.55
(K. Wang *et al.*, 2015)	MFCC, Fourier parameter (FP), fundamental frequency (F0), energy and zero-crossing rate (ZCR)	Support Vector Machines	Berlin Speech Database	88.88
(Sun, Wen and Wang, 2015)	Weighted spectral features based on Hu moments (HuWSF)	Support Vector Machines	SAVEE	89.32
(Ying and Xue-Ying, 2018)	Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator (GCZCMT)	Support Vector Machines	Berlin Speech Database	84.45
(Shaqra, Duwairi and Al-Ayyoub, 2019)	eGeMAPS features	MLP neural network	RAVDESS	87.95
Proposed Model		2D-CNN	RAVDESS	98.22

In this thesis, it was observed that the custom 2D-CNN is the best end-to-end model in SER compared to DRBFNN and DBMLP as well as related work shown in Table 30. This also underlined the fact that AGF is indeed effective in recognising speech emotion when used with deep learning classifiers. This experiment has shown that in SER, the power of deep learning

models lies in the algorithms rather than the features generated. 2D-CNN was seen to outperform the other end-to-end models. However, the effectiveness of the 2D-CNN model comes at a price. The model is time-consuming and is also resource-intensive because it took too long to process the spectrograms. The experimental study also revealed that the processing time increases according to the number of spectrograms. However, according to Saito et al, faster classifiers may require a huge number of training instances and their effectiveness will only be seen when applied on unseen data (Saito *et al.* 2015). This work places a strong emphasis on the trade-off between method efficiency versus method accuracy. Despite the slowness of 2D-CNN, it was noted that 2D-CNN learns continuously since the model kept on improving in accuracy with each experimental run. Conversely, DMLP and DRBFNN do not learn continuously because they keep on achieving the same scores for an arbitrary number of runs.

From Tables 29 to 31 it was observed that DMLP is more effective in recognising the angry, disgust, happy, neutral, and sad emotion states for the SAVEE emotion corpus. This clearly shows that the interpretation of speech emotion depends on language, accent and the model used since the emotion corpus used in this experiment is from three different countries. The results obtained using DRBFNN were a bit misleading because the overall accuracy was high but the scores of each emotion state were relatively low. For example, in Table 31 it was observed that even though the average accuracy scored when DRBFNN was fed with spectrograms from the EMODB corpus was high (97.05%), it recorded a low f1-score. Therefore, these results confirm that using accuracy alone as a performance benchmark is not enough (McNee, Riedl, and Konstan 2006). The proposed 2D-CNN proved to be an efficient end-to-end model for recognising emotion in speech compared to other proposals is shown in Table 32.

6.5 Experiment 4: Multilingual Emotion Recognition

Most of the research as illustrated in chapter 4, ignores the recognition of emotion and valence in a multilingual environment. This issue stimulated the idea to investigate and explore the appropriate techniques of developing a multilingual model to recognise valence through emotion using HAF features. HAF features were used in this experiment because the previous experiments showed that these are the most efficient in recognising speech emotion. Customer call centres usually constitute customers from different cultures and countries speaking a variety of languages

and accents. Many African countries have a plethora of official spoken languages. Zimbabwe has 16 official languages while South Africa and Nigeria have more than 20 official spoken languages. Developing language-specific SER models for a single organisation can be extremely expensive therefore creating a generic model that can recognises emotions from any language can minimize costs. This, therefore, presented the researchers with an opportunity to explore ways to develop a model that performs well in such environments. Therefore, this experiment was carried out to investigate the most effective models in recognising speech valence in multilingual environments.

In this experiment, five speech emotion databases were combined to develop a cross-language experimentation database and this concept was borrowed from the work presented by other researchers (Elbarougy and Akagi 2013; Song, Zhang, *et al.* 2017; Latif *et al.* 2019; Li and Akagi 2019; Parry *et al.* 2019). The decision to combine these databases was inspired by the fact that they comprise vocal utterances spoken in a variety of languages and accents which was perfect for this experiment. Eight speech emotion classes were used in the design of the cross-language corpus and these are neutral, happy, angry, surprised, sad, calm, fear and disgust. The emotion classes were grouped into two classes which are positive and negative valences as illustrated in Table 33 (Latif *et al.* 2018; Parry *et al.* 2019). This grouping of the emotions to emotional valences made it possible to combine different databases with different emotion classes to perform cross-language emotion recognition experiments. The fusion resulted in 10094 adult utterances to seamlessly conduct this body of work. 80% of the database was used for training the models while 20% was reserved for testing purposes.

After combining the five databases the new cross-language database had 6641 negative valence audio files and 3453 positive valence audio files. Although there was an imbalance in the distribution of negative and positive valences, it is of paramount importance to note that this did not introduce bias on the prediction results.

Table 33:Mappings of Adult Speech Emotions into Positive and Negative Valence Emotions.

Corpus	Language	Total Utterances		Training Utterances		Testing Utterances		Negative Valence	Positive Valence
		Male	Female	Male	Female	Male	Female		
EMO-DB	German	400	400	320	320	80	80	Anger, Sadness, Fear, Disgust, Boredom	Neutral, Happiness
SAVEE	British English	480	0	384	0	96	0	Anger, Sadness, Fear, Disgust	Neutral, Happiness, Surprise
RAVDESS	North American English	720	720	576	576	144	144	Anger, Sadness, Fear, Disgust	Neutral, Happiness, Surprise, Calm
EMOVO	Italian	294	294	235	235	59	59	Anger, Sadness,	Neutral, Happiness, Surprise
								Fear, Disgust	
CREMA-D	African, American, Asian, Caucasian, Hispanic, and Unspecified	3579	3207	2863	2566	716	641	Anger, Sadness, Fear, Disgust	Neutral, Happiness

Table 34 shows the CPU computational times of the individual ensemble classifiers used in this body of work. The GBM classifier was seen to be the slowest classifier as presented in Table 34. However, the processing time improved significantly after performing feature selection. Before scaling down the redundant features using RF-RFE GBM processed the data for 60.495 ms. A significant improvement was observed when RF-RFE was applied since it took 16.228 ms to complete the training process. The same trend was observed for all the other classifiers including RALOG. The results also show that that the proposed model had a healthy processing time since

the training process lasted for 9.05 ms. In addition, it can be observed that XGB was the fastest classifier since its training lasted for 0.00095 ms. The implications of this experimental study prove that the proposed RALOG ensemble classifier is a credible option for recognising valence especially in situations where processing time is of utmost importance (Khorram, McInnis and Mower 2019). This factor is vital since it influences the development of efficient real time systems.

Table 34: Processing Time.

	RALOG	**RDF**	**GBM**	**ETC**	**XGB**
Before RFE	43.314	5.29	60.495	0.168	0.004
After RFE	9.05	0.459	16.228	0.244	0.00095

The experimental results of the average emotional valence recognition before and after scaling features with RF-RFE obtained are shown in Table 35. The results show that RF-RFE significantly boosted the recognition accuracy. RALOG's recognition accuracy increased from 82% to 92%. In that same vein, the recognition accuracy of XGB steadily increased from 78% to 87%. All the recognition models used in this experiment experienced an increase in accuracy when RF-RFE was applied. It is important to note that the ETC classifier had the highest increment (10%) in recognition accuracy when RF-RFE was applied. However, the ETC classifier recorded the lowest accuracy score compared to the other classifiers.

From the results in Table 35, it can be observed that the RDF classifier performs fairly well in recognising speech emotions since it yielded an overall accuracy score of 85%. Additionally, the results reveal that ETC is a decent alternative for perceiving valence since it obtained a recognition accuracy score of 83% when RF-RFE was applied. The proposed classifier, RALOG proved to be efficient since it obtained the highest accuracy scores before and after applying RF-RFE. The model even outperformed other models presented in the literature. For example, the model presented by Latif et al yielded an overall recognition accuracy of 80% using eGeMAPS, DBN, and Leave-One-Out testing approach (Latif *et al.*, 2018). Consequently, it can be observed that the proposed RALOG ensemble classifier is a decent option for cross-language speech emotion recognition.

The same trend was noted when the F1-score, recall, and precision of the ensemble classifiers were analysed. As shown in Table 35, RALOG outperformed the other ensemble classifiers. The significance of performing feature selection using RF-RFE was also observed when precision scores were analysed. A surge in average precision was noticed on all the classifiers. The average precision for all the classifiers surged tremendously. The highest increment was observed when the GBM classifier was used. The precision increased from 69% to 79%. RALOG's registered a healthy increment from 82% to 90%. RALOG the proposed ensemble classifier recorded the highest average precision that is 90%. It was followed by the RDF classifier which yielded an average precision of 80%. ETC had the joint lowest precision score since it achieved 78%. This was followed by GBM which had an average precision score of 79%. From the results in Table 34, it can be inferred that RALOG is a highly efficient classifier for recognising valence in speech emotion and the results are comparable to work presented in the literature (N. Liu *et al.*, 2018).

Marginal improvements in the recall were noticed when RF-RFE was applied and this is shown in Table 35. RALOG continued to outperform the other ensemble classifier and it achieved an average recall score of 90%. The second-best performing classifier in this regard was RDF (88%) followed by XGB (85%). GBM, ETC was the lowest-performing classifiers since they recorded an average recall score of 82%. The impact of using RF-RFE was felt when GBM was used because it increased by 11% even though it had the lowest recall score. These results also make further solid proof that the combination of RALOG and RF-RFE provides a viable alternative for speech emotion recognition. The average recall percentage score obtained by RALOG is comparable to the results presented in chapter 4 (Deng *et al.* 2014; Liu *et al.* 2018). Furthermore, RALOG outperformed the other ensemble classifiers with regards to F1-Score since it achieved a score of 91%. It was followed by RDF which yielded an F1-score of 82%. The lowest F1 scores were recorded by both ETC and GBM which both achieved an average of 80%.

Table 35*:* Average accuracy, recall, F1-score, and precision with confidence intervals of classifiers before and after feature scaling with RF-RFE.

Algorithm	Stage	Precision	Recall	F1-score	Accuracy
RALOG	**Before RFE**	82(±0.007)	83(±0.007)	83(±0.007)	82(±0.007)
	After RFE	90(±0.006)	90(±0.006)	91(±0.006)	92(±0.005)
RDF	**Before RFE**	72(±0.009)	82(±0.007)	77(±0.008)	80(±0.008)
	After RFE	80(±0.008)	88(±0.006)	82(±0.007)	85(±0.007)
GBM	**Before RFE**	69(±0.009)	73(±0.009)	71(±0.009)	73(±0.009)
	After RFE	79(±0.008)	82(±0.007)	80(±0.008)	85(±0.007)
XGB	**Before RFE**	71(±0.009)	81(±0.008)	75(±0.008)	78(±0.008)
	After RFE	78(±0.008)	85(±0.007)	81(±0.006)	87(±0.007)

The average accuracy, recall, F1-score, and precision scores for the valence classes are presented in Table 36. It can be observed that ensemble classifiers achieved relatively low scores in the recognition of negative valence. The unbalanced nature of the database could be one of the culprits that contributed to this dip in performance. The performance of all the ensemble classifiers was significantly boosted when RF-RFE was applied. The best precision score in recognising negative valence was achieved by RDF (97%). It was followed by XGB which yielded a precision score of 95%. RALOG achieved a precision score of 93% in that regard and the same score was also achieved by ETC. Nevertheless, RALOG outperformed the other ensemble classifiers in recognising negative valence because it achieved a 97% accuracy score. In addition, RALOG outperformed the other classifiers by achieving a precision score of 87% in recognising positive valence. The application of RF-RFE, as well as RALOG, has significantly boosted the recognition of valence and the proposed model is comparable to the work presented in Table 37.

Table 36: Percentage precision, recall and F1-score, and accuracy at each experimental stage of feature scaling with the RF-RFE algorithm.

Algorithm	Stage	Precision		Recall		F1-score		Accuracy	
		Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
RALOG	**Before RFE**	88	76	86	79	88	77	86	78
	After RFE	93	87	94	86	94	87	97	86
RDF	**Before RFE**	83	59	84	78	82	67	82	74
	After RFE	97	86	97	62	88	87	92	72
GBM	**Before RFE**	81	56	78	64	80	62	79	71
	After RFE	92	66	89	74	90	70	91	79
XGB	**Before RFE**	84	60	85	79	83	71	84	76
	After RFE	95	61	88	82	91	70	94	80

The ROC curves of the evaluated classifiers are shown in Figures 56 and 57. Figure 56 shows the performance of the classifiers before the selection of features with RF-RFE. From the ROC curves, it can be noted that RALOG was the best performing classifier before and after (96.97%) RF-RFE was applied. This means that using RALOG with RF-RFE there is a 96.97% chance that the correct unpleasant valence will be classified. The second-best performing classifier in this regard was GBM which achieved an AUC score of 94.62%. These results show that the features selected by RF-RFE are effective in recognising valence from different languages, accents, and recording environments since the corpus used in this experimental study was an amalgam of five different emotion databases.

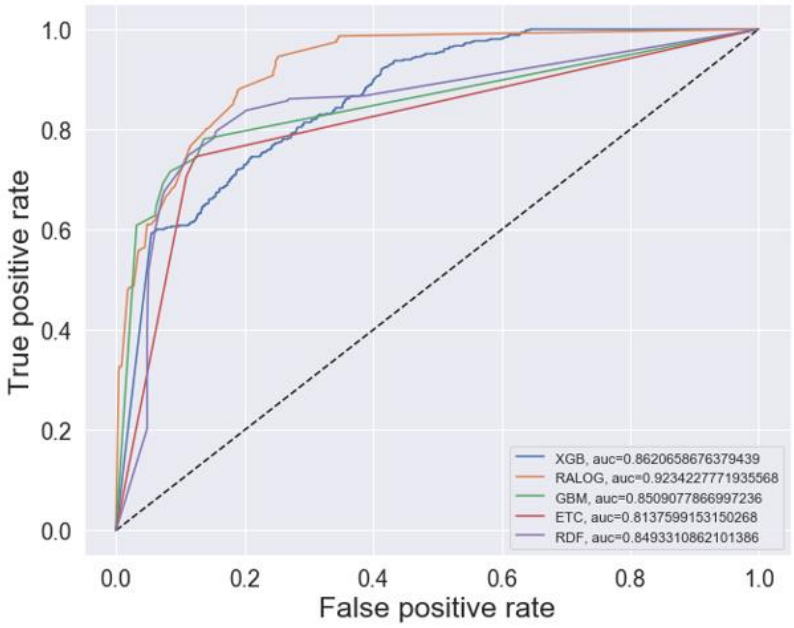


Figure 56: ROC Curve before RFE.

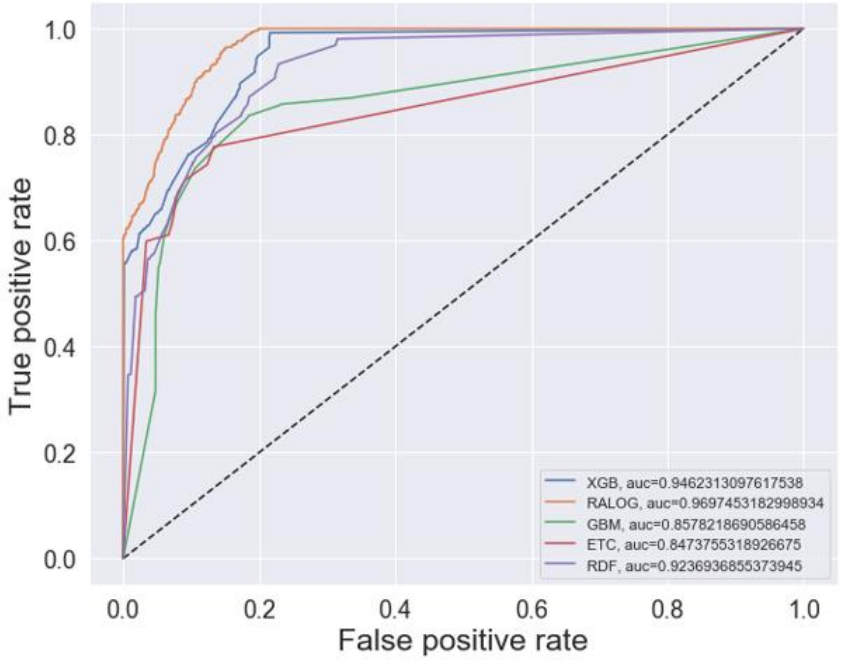


Figure 57: ROC Curve after RFE.

Table 37: Comparison of the proposed approach with potential works from the literature.

Reference	Corpora	Languages	Emotions	Recognition method	Result

Latif at al (Latif *et al.* 2019)	EMOVO, Urdu, SAVEE, EMO-DB	British English, Italian, German, Urdu	**Positive valence** (Anger, Sadness, Fear, Disgust, Boredom) **Negative valence** (Neutral, Happiness, Surprise)	eGeMAPS + SVM	70.98% (Accuracy)
Mustaqeem et al (Mustaqeem and Kwon 2020)	RAVDESS, IEMOCAP	North American English,	Anger, Happy, Neutral, Sad	Clean spectrograms DSCNN	56.5% (Accuracy)
Ocquaye et al (Ocquaye *et al.* 2020)	FAU-Aibo, IEMOCAP, EMO-DB, EMOVO, SAVEE	German, English, Italian	**Positive valence:** Surprise, motherese, joyful/happy, neutral, rest, excited, **Negative valence:** Angry, touchy, sadness, emphatic, reprimanding, boredom disgust, fear	tripple attentive asymmertric CNN model	73.11% (Accuracy)
Liu et al (N. Liu *et al.* 2018)	eNTERFACE, EmoDB	English	Anger, Happy, Neutral, Sad	DALSR + SVM	52.27% (UAR)
Latif et al (Latif *et al.* 2018)	FAU-AIBO, IEMOCAP, EMO-DB, EMOVO, SAVEE	German, English, Italian	**Positive valence:** (Motherese, Joyful/ Happy, Neutral, Rest, Excited, Surprise) **Negative valence:** (Angry, Touchy, Emphatic, Reprimanding, Sadness, Disgust, Fear, Boredom)	eGeMAPS + DBN + Leave-One-Out	80% (Accuracy)
Li et al (Li and Akagi 2019)	FUJITSU, EMO-DB, CASIA	Japanese, German, Chinese	Neutral, Happy, Angry, Sad	IS16 + MSF + LMT (logistic model trees)	82.63% (F-Measure)

Parry et al (Parry *et al.* 2019)	RAVDESS, IEMOCAP, EMO-DB, EMOVO, SAVEE, EPST	English, German, Italian	**Positive:** (elation, excitement, happiness, joy, pleasant surprise, pride, surprise **Negative:** (anger, anxiety, cold anger, contempt, despair, disgust, fear, frustration, hot anger, panic, sadness, shame **Neutral:** (boredom, calm, interest, neutral	CNN	55.11% (Average accuracy)
Deng et al (Deng *et al.* 2014)	ABC,FAU AEC	German	**Positive valence**: (Cheerful, Neut Rest, Medium stress) **Negative valence**: (aggressi intoxicated, nervous, tired, screami fear, high stress)	INTERSPEECH 2009 Emotion Challenge baseline feature set + A DAE (adaptive denoising autoencoders for)	84.18% (UAR)
Proposed Model	EMO-DB, SAVEE, RAVDESS	German, British English, North American English	**Positive valence** (Anger, Sadness, Fear, Disgust, Boredom) **Negative valence** (Neutral, Happiness, Surprise)	RF-RFE extracted features + RALOG	92% (Accuracy)

The final part of experiment 4 was performed to interrogate the performance of RALOG on the base databases that were used to create the cross-language database. The results presented in Table 38 show that the basic performance was relatively good across the databases. The highest recognition performance was achieved using RALOG which yielded an overall accuracy of 98% on the SAVEE database. This performance could have been affected by the fact that SAVEE consists of male utterances only and these are from the same country which is Britain. All the gender-balanced databases were associated with lower recognition rates in comparison with SAVEE. RALOG achieved lower recognition rates across all the performance metrics when it was applied on CREMA-D. This is attributed to the fact that CREMA-D comprises utterances from varied accents of speakers with diverse cultures. In addition, this confirms the difficulties experienced in recognizing valence emotion in a multilingual environment as reported in the literature (Yang, 2017; Latif et al., 2018). It is noteworthy to note that RALOG performs better on

individual databases in comparison with the cross-language database. However, its performance on the EMOVO database is worse than its performance on the cross-language database.

Table 38**:** Percentage Precision, Recall and F1-score and Accuracy of RALOG on Individual Corpus before and after Feature Scaling with the RF-RFE Algorithm.

Corpus	Stage	Precision	Recall	F1-score	Accuracy
EMO-DB	**Before RFE**	95	96	95	95
	After RFE	97	96	96	96
SAVEE	**Before RFE**	97	96	96	96
	After RFE	98	98	98	98
RAVDEES	**Before RFE**	94	94	94	95
	After RFE	97	97	97	97
EMOVO	**Before RFE**	92	92	93	92
	After RFE	94	94	94	95
CREMA-D	**Before RFE**	88	89	88	89
	After RFE	85	95	89	92

The training time of the proposed RALOG classifier on individual databases is shown in Table 39. The results show that RALOG processed the data faster on the SAVEE databases since it took 0.47 ms. However, it took longer to train using utterances in the CREMA-D database since it took 6.05 ms. This result was achieved because CREMA-D has more utterances (1358) while SAVEE had the least number of utterances (384).

Table 39: Training Time (ms) of RALOG Individual Corpora.

	EMO-DB	**SAVEE**	**RAVDEES**	**EMOVO**	**CREMA-D**
Before RFE	3.440	2.070	6.190	2.540	29.200
After RFE	0.780	0.470	1.380	0.580	6.050

The results of this experimental study were comparable to the results presented in the literature in Table 37. GBM proved to be a decent alternative for developing speech emotion models but its

greatest drawback is that it is computationally expensive as far as processing time is concerned. XGB, proved to be a suitable option for developing real-time valence recognition models because it is incredibly fast. RALOG may not have been the fastest classifier but it had a healthy training time which is good for developing modern-day valence recognition models. Furthermore, RALOG was the most effective classifier when all the classifiers were applied to HAF features.

6.6 Chapter Summary

This chapter presented investigative experimental studies conducted to develop a customer call centre SER model that is fast, accurate, and computationally inexpensive. This study discovered that HAF features are a viable set of features that are suitable for developing SER models. These features were evaluated against deep learning auto-generated features and the results showed that HAF features are more effective. Further experiments revealed that deep learning models are effective in recognising emotions using the AGF in an end-to-end style. However, the results showed that deep learning models are computationally expensive in terms of processing time. In addition, the last experiment showed that scaled-down HAF features are good for recognising emotional valence in multilingual environments. Multilingual environments or environments with people who speak with a variety of accents depict the real setup that is normally seen in customer call centres. Therefore, this was a viable option to set up experiment 4 in such a way. Another notable achievement was the development of the RALOG ensemble classifier for valence emotion recognition. The RALOG classifier can be used to develop models that can recognise valence in speech files with low processing time. The next and final chapter will present the summary of this thesis's contribution as well as future work.

Chapter 7: Thesis Summary, Conclusion, and Future work

7.1 Thesis Summary

In this body of work, four experiments were conducted to fulfill the objectives mentioned in chapter 1. During the first experiment, acoustic features (spectral and prosodic features) were extracted from raw audio speech files which include RAVDESS and SAVEE. Three feature sets were developed from this pool of features. The first feature set (MFCC1) constituted homogeneous MFCC features while the second one (MFCC2) comprised of a combination of MFCC, pitch (fundamental frequency), and ZCR drawing inspiration from the work presented by Bhaskar et al. (Bhaskar, Sruthi, and Nedungadi 2015) and Sarker and Alam (Sarker and Alam 2014). The third feature set (HAF) was a fusion of MFCC, fundamental frequency, energy, spectral roll-off, spectral flux, spectral centroid, spectral compactness, spectral variability, fast Fourier transform, and a set of LPCC features. These features were evaluated using five ensemble classifiers which include RDF, GBM, BSVM, ABC, and BMLP. The results of the first experiment showed that the proposed HAF features are more effective in recognising human emotions.

The second experiment was conducted to investigate and compare the performances of deep learning AGF features. Therefore, the audio speech files in the speech databases were converted to spectrograms, and three deep classifiers were used to extract AGF features from these spectrograms. The deep classifiers used to accomplish this purpose are 2D-CNN, DRBFNN, and DMLP. Consequently, this exercise resulted in three feature sets which are 2D-CNN-AGF, DRBFNN-AGF, and DMLP-AGF. The same ensemble classifiers used in the first experiment were then used to classify human emotions using the three ADF feature sets. The results showed that 2D-CNN-AGF were the better performing feature sets amongst the three feature sets. Moreover, RDF was the best performing classifier in all the feature sets.

The analysis of 2D-CNN-AGF and HAF showed that the latter was indeed the most effective set of features for recognising human emotion through speech. However, these results stimulated a need to further explore the performances of end-to-end deep learning models to evaluate the efficiency of deep learning classifiers using their feature extraction techniques. The results showed

that when AGF features are used with deep classifiers in an end-to-end fashion, they perform better. This improvement is attributed to the continuous learning behaviour of deep classifiers because the performance kept on increasing with each number of runs.

Since customer call centres operate in multilingual environments, another experiment was conducted to assess the efficiency of the proposed HAF features in cross-corpus SER. In this body of work, RAVDESS, SAVEE, EMOVO, CREMA-D, and EMODB were combined to simulate a multilingual environment. The main aim here was to perform speech valence emotion recognition therefore the classes used were positive and negative valence. A stacked ensemble classifier (RALOG) was developed to recognise emotional valence. The developed stacked ensemble together with HAF achieved results that are comparable to the work reported in the literature.

The four above-mentioned experiments played a crucial role in fulfilling the objectives of this study, which are:

i. To discover a set of acoustic features that can help to improve the recognition performance of a speech emotion system for cross-language emotional conversations.
ii. To investigate whether the discovered acoustic features can give an improved performance in a cross-language emotional conversation system when compared to the auto-generated features by the deep learning method.
iii. To develop an efficient algorithm that will give an improved recognition performance for the discovered acoustic features for cross-language emotional conversations.

The first experiment conducted in this experimental study was designed to meet the first objective since a set of acoustic features were investigated to improve the accuracy of SER. In this thesis, it was observed that the proposed HAF features are the most efficient acoustic features in SER because they achieved the highest recognition accuracy. The final experiment also demonstrated the discriminative power of scaled-down HAF features because it achieved high recognition accuracies in a simulated multilingual environment. The second experiment fulfilled the second objective because the results showed that HAF features perform way better than deep learning auto-generated features. Furthermore, it was noted that HAF features are less computationally costly compared to deep learning auto-generated features. The third objective was met in both the first and second experiments. The same ensemble algorithms were used in both experiments using

different feature sets. RDF outperformed all the other algorithms in terms of accuracy, precision and recall. Accordingly, this confirmed that RDF is indeed the best recognition in this regard. However, RALOG the proposed classifier outperformed RDF when both algorithms were used to recognise valence in experiment 4.

7.2 Conclusion

The investigations in this body of work showed that handcrafted features achieve better results in SER if the correct combination of features is used. The analysis of both the hybrid handcrafted features and deep learning auto-generated features confirmed the superiority of the former. The advantage of using deep learning AGF is that deep learning removes the burden of selecting features from the user making them a convenient choice in that regard.

However, the study also showed that deep classifiers work better when used in an end-to-end fashion. This means that deep learning models perform better when the same model is used to extract features and recognise emotions. The study showed that deep learning classifiers have a continuous learning behaviour that helps them to learn in each recognition run. However, the main drawback of this method is that it is computationally expensive.

The proposed RDF was found to be the most appropriate classifier in SER. The ensemble classifier was able to accurately classify the fear emotion that has been reported to be difficult to detect in literature. The ensemble algorithm, however, managed to classify the emotional classes with more than 95% accuracy.

The experimental study confirmed that different accents affect the performance of SER models because the performance of the various SER models in this study performed differently across two databases that had different accents. From the experiments done in the study, it was observed that some models perform better when recognising emotions from different languages. For example, DMLP performed better than the custom 2D-CNN in terms of overall accuracy when the models were applied to the SAVEE emotion corpus in the end-to-end experiment. However, when the two

models were applied to the RAVDESS and EMODB emotion corpora, 2D-CNN outperformed DMLP. According to the literature presented in chapter 2, one major challenge of recognising emotion in speech stems from the fact that emotional expression is language-specific. For example, for the same emotion, different languages can use different pitches and tones. Therefore, this thesis submits that when using spectrograms (in end-to-end setups), SER models should indeed be language-specific.

7.3 Limitations and Future Work

The primary constraint of this work is that acted data was used in the experimental study. The problem with acted speech is that at times authentic emotions may be difficult to express (Sun, Fu and Wang 2019). In addition, RALOG has a performance limitation in that it is a bit computationally expensive compared to other ensemble algorithms such as RDF. 2D-CNN has the potential of predicting various emotions with high accuracy and precision scores, but this will only be achieved after long periods of repetitive training which is computationally expensive. Another limitation is that all the experiments were done on one computer. It would have been better if the same experiments were done on different computers with different specifications such as GPU to evaluate the performance of the proposed model. The vocal utterances in the emotion corpora used in this research did not include people who stammer. Moreover, languages that have different sounds that show anger such as the Shona language that makes use of cliques to express anger were not used in the study. Therefore, the researcher would like to vigorously incorporate these factors into future work. The researcher would also like to explore other feature extraction techniques such as Histogram of oriented gradients (HOG), Speeded-up robust features (SURF), and local binary patterns (LBP) on the speech spectrogram to further interrogate the effectiveness of such features in SER. The speech corpus used in this thesis was devoid of noise. Noise can come from several sources such as cars, wind, rain, other speakers, and many more. The researcher would like to explore ways of removing such noise using various techniques such as the pixel intensity clustering algorithm (Olugbara *et al.* 2015) and perceptual colour difference saliency with morphological analysis (Olugbara *et al.* 2018) on the spectrograms.

References

Abbaschian, B.J., Sierra-Sosa, D. and Elmaghraby, A. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors.* 21: 23-37.

Abe, B.T., Olugbara, O.O and Marwala, T. 2014. Experimental comparison of support vector machines with random forests for hyperspectral image land cover classification. *Journal of Earth System Science*. 123(6). 779-790.

Abhishek, P. M. and Kopparapu, S. K. 2012. A novel approach to identify problematic call center conversations. In: Proceedings of the *9th International Joint Conference on Computer Science and Software Engineering:*1–5.

Abo, A. H., Deriche, M. and Mohandes, M. 2018. A bilingual emotion recognition system using deep learning neural networks. In: Proceedings of the *15th International Multi-Conference on Systems, Signals and Devices* (1): 1241–1245.

Ackermann, H. 2008. Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. *Trends in Neurosciences*, 31(6): 265–272.

Adetiba, E. and Olugbara, O. O. 2015a. Improved classification of lung cancer using radial basis function neural network with affine transforms of voss representation. *PLoS ONE*, 10(12): 1–25.

Adetiba, E. and Olugbara, O. O. 2015b. Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. *Scientific World Journal*, 2015.

Agarwal, S. and Chowdary, C. R. 2020. A-stacking and a-bagging: adaptive versions of ensemble learning algorithms for spoof fingerprint detection. *Expert Systems with Applications*:113-129.

Agjee, N. H., Mutanga, O., Peerbhay, K. and Ismail, K. 2018. The impact of simulated spectral noise on random forest and oblique random forest classification performance. *Journal of Spectroscopy*: 1–8.

Agostini, G., Longari, M. and Pollastri, E. 2003. Musical instrument timbres classification with

spectral features. *Eurasip Journal on Applied Signal Processing*: 5–14.

Ahmad, M. W., Reynolds, J. and Rezgui, Y. 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production*, 203: 810–821.

Akçay, M. B. and Oğuz, K. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116: 56–76.

Alam, F., Danieli, M. and Riccardi, G. 2016. Can we detect speakers empathy ? : A real - life case study. *CogInfoCom*: 59–64.

Ali, S. S. and Maqsood, J. 2018 . Net library for SMS spam detection using machine learning: A cross platform solution. In: Proceedings of the *2018 15th International Bhurban Conference on Applied Sciences and Technology*: 470–476.

Alías, F., Socoró, J. C. and Sevillano, X. 2016. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, 2016, 6(5): 143-165.

Alonso, J. B., Cabrera, J., Medina, M. and Travieso, C. M. 2015.New approach in quantification of emotional intensity from the speech signal: Emotional temperature. *Expert Systems with Applications*, 42(24): 9554–9564.

Alshamsi, Hu., Kepuska, V. and Meng, H. 2019. Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. In: Proceedings of the *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference*: 730–738.

Altun, H. and Polat, G. 2009. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. *Expert Systems with Applications*. Elsevier Ltd, 36(4): 8197–8203.

Álvarez, A., Sierra, B., Arruti, A., López-Gil, J. M., and Garay-Vitoria, N. 2015. Classifier subset selection for the stacked generalization method applied to emotion recognition in speech. *Sensors*

(Switzerland), 16(1): 1–26.

Amudavalli, A. 2010. *Theories & models of communication*. 4th ed. London: CRC Press.

Anagnostopoulos, C. N., Iliou, T. and Giannoukos, I. 2012. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2): 155–177.

Aouani, H. and Ben Ayed, Y.2018. Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder. In: Proceedings of the *2018 4th International Conference on Advanced Technologies for Signal and Image Processing*: 1–5.

Aouani, H., and Ayed, Y. B. 2020. Speech emotion recognition with deep learning. *Procedia Computer Science*, 176:251–260.

Arias, J. P., Busso, C. and Yoma, N. B. 2014. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech and Language*, 28(1): 278–294.

Ariav, I. and Cohen, I. 2019 An end-to-end multimodal voice activity detection using wavenet encoder and residual networks. In: Proceedings of the *IEEE Journal on Selected Topics in Signal Processing*:265-274.

Arnold, M. B. 1960. *Emotion and personality.* New York: Columbia University Press.

Arowolo, M.O., Adebiyi, M.O., Adebiyi, A.A. and Olugbara, O.O. 2021. Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. *Journal of Big Data*, 8(29):1-16.c

Asadullah, M. A., Peretti, J. M., Derbel, W. and Rajhi, S. 2019. Ownership-based asymmetries in training evaluation practices of call centres. *Industrial and Commercial Training*, 51(1): 13–23.

Avisado, H. G., Cocjin, J. V., Gaverza, J. A., Cabredo, R., Cu, J. and Suarez, M. 2012. Analysis of music timbre features for the construction of user-specific affect model. *Theory and Practice of Computation*, 5: 28–35.

Ayadi, M., Kamel, M. S. and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3): 572–587.

Bachorowski, A. J. and Bachorowski, J. 2010. Vocal expression and perception of emotion vocal expression of emotion. *Psychological Science*, 8(2): 53–57.

Badshah, A. M., Ahmad, J., Lee, M. Y. and Baik, S. W. 2016. Divide-and-conquer based ensemble to spot emotions in speech using mfcc and random forest. In: Proceedings of the *2nd International Integrated Conference & Concert on Convergence*: 2–9.

Badshah, A. M., Rahim, N. and Ullah, N. 2019. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78(5): 5571–5589.

Bagozzi, R. P., Baumgartner, H. and Pieters, R. 1998. Goal-directed emotions. *Cognition and Emotion*, 12(1): 1–26.

Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B. Y. and Haase, A. 2019. Recursive feature elimination in random forest classification supports nanomaterial grouping. *NanoImpact*, 15(6): 155-179.

Bahrick, L. E., McNew, M. E., Pruden, S. M., and Castellanos, I. 2019. Intersensory redundancy promotes infant detection of prosody in infant-directed speech. *Journal of Experimental Child Psychology*, 183: 295–309.

Bailey, W., Nowicki, S. and Wickline, V. B. 2009. Cultural in-group advantage: Emotion recognition in african american and european american faces and voices. *Journal of Genetic Psychology*: 170-196.

Banse, R. and Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3): 614–636.

Baron-Cohen, S., Bor, D., Billington, J., Asher, J., Wheelwright, S. and Ashwin, C. 2006. Savant memory in a man with colour form-number synaesthesia and asperger syndrome: The assumption that all human minds are wired the same. *Journal of Consciousness Studies*, 14(9): 10–237.

Baron-Cohen, S., Golan, O., Wheelwright, S., Granader, Y., and Hill, J. 2010. Emotion word comprehension from 4 to 16 years old: A developmental survey. *Frontiers in Evolutionary Neuroscience*, 2(11): 1–8.

Barrett, L. F. 2012. Emotions are real. *Emotion*, 12(3): 413–429.

Barrett, L. F., Lewis, M. and Haviland-Jones M., J. 2016. *Handbook of emotions.* 4th ed. New York: The Guilford Press.

Basu, S., Chakraborty, J., Bag, A. and Aftabuddin, A. 2017. A review on emotion recognition using speech. In: Proceedings of the *2017 International Conference on Inventive Communication and Computational Technologies*: 109–114.

Bayram, S., Ocal, M. E., Oral, E. L. and Atis, C. D. 2016. Comparison of multi layer perceptron (MLP) and radial basis function (RBF) for construction cost estimation: the case of Turkey. *Journal of Civil Engineering and Management*, 22(4): 480–490.

Bernbau, H., Fujita, F. and Pfennig, J. 2005. Consistency, specificity, and correlates of negative emotions. *Journal of Personality and Social Psychology*, 68(2): 342–352.

Bhaskar, J., Sruthi, K. and Nedungadi, P. 2015. Hybrid approach for emotion classification of audio conversation based on text and speech mining. In: Proceedings of the *Procedia Computer Science*: 635–643.

Bhavan, A., Chauhan, P. and Shah, R. R.2019. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*: 104-137.

Bikmukhametov, T. and Jäschke, J. 2019. Oil production monitoring using gradient boosting machine learning algorithm. *International federation of automatic control-PapersOnLine*, 52(1): 514–519.

Birkley, E. L. and Eckhardt, C. I. 2015. Anger, hostility, internalizing negative emotions, and intimate partner violence perpetration: A meta-analytic review. *Clinical Psychology Review*, 37: 40–56.

Bonfiglio, S. 2017. Osservatorio europeo: Il diritto alla salute dei cittadini europei e degli immigrati extracomunitari nellordinamento svizzero. *Cittadinanza Europea (La)*, (1): 121–135.

Boucher., J. D. and Brandt, M. 1987. Judgement of emotion. *Journal of corss-cultural psychology*, 12(3): 272–283.

Breiman, L. 1996. Bagging predictions. *Machine learning*, 24(2): 123–140.

Breiman, L. 1999.Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36(1): 85–103.

Breiman, L. 2001.Random forests. *Machine Learning*, 45(1): 5–32.

Brown, C. 2014. Economic insights – Trends and challenges the effects of emotional intelligence and leadership style on sales performance. *Economic Insights – Trends and Challenges*, (3): 1–14.

Bryant, G. A. and Barrett, H. C. 2008. Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, 8(1–2): 135–148.

Bucks, R. S. and Radford, S. A. 2004. Emotion processing in alzheimers disease. *Aging and Mental Health*, 8(3): 222–232.

Burger, B., Ahokas, J. R., Keipi, A. and Toiviainen, P. 2013. Relationships between spectral flux, perceived rhythmic strength, and the propensity to move. In: Proceedings of the *Proceedings of the Sound and Music Computing Conference*: 179–184.

Burkhardt, F., Paeschke, Rolfes, M. and Sendlmeier, W. F. 2005. Berlin EmoDB: A database of German emotional speech. In: Proceedings of the *Proceedings of InterSpeech*: 1517–1520.

Busso, C., Bulut, M. and Lee, C. 2008. IEMOCAP : Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*,54(2): 335–359.

Byun, S.W. and Lee, S.P. 2021. A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences*. 2021, 11, 1890.

Cairong, Z., Xinran, Z., Cheng, Z. and Li, Z. 2016. A novel DBN feature fusion model for cross-corpus speech emotion recognition. *Journal of Electrical and Computer Engineering*: 1–11.

Caron, R. F., Caron, A. J. and Myers, R. S. 1985. Do infants see emotional expressions in static faces ? *Child Development*, 56(6): 1552–1560.

Cetina, V., Brito, C. and Ruiz, H. 2015. Chagas parasite detection in blood images using adaboost. *Computational and Mathematical Methods in Medicine*, 2015.

Chakraborty, R., Pandharipande, M. and Kopparapu, S. K. 2016. Knowledge-based framework for intelligent emotion recognition in spontaneous speech. In: Proceedings of the *Procedia Computer Science*: 587–596.

Chamoli, A., Semwal, A. and Saikia, N. 2017. Detection of emotion in analysis of speech using linear predictive coding techniques. In: Proceedings of the *2nd International Conference on Inventive Systems and Control*: 1–4.

Chandran, A., Pravena, D. and Govind, D. 2017. Development of speech emotion recognition system using deep belief networks in Malayalam language. In: Proceedings of the *6th International Conference on Advances in Computing, Communications and Informatics*: 676–680.

Chaplin, T. M. 2015. Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, 7(1): 14–21.

Chaplin, T. M. 2016. Gender and Emotion Expression: A developmental contextual perspective. *Emot Rev*, 118(24): 6072–6078.

Charland, L. C. 2002. The natural kind status of emotion. *British Journal for the Philosophy of Science*, 53(4): 511–537.

Chelali, F. Z. and Djeradi, A. 2015. Face recognition using MLP and RBF neural network with gabor and discrete wavelet transform characterization: A Comparative Study. *Mathematical Problems in Engineering*, 2015: 1–16.

Chen, T. and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 785–794.

Chernykh, V. and Prikhodko, P. 2017. Emotion recognition from speech with recurrent neural networks. In: Proceedings of the *International Conference on Computational & Experimental Engineering and Sciences*: 333–336.

Chia Ai, O., M. Hariharan, Yaacob, S.B, Yaacob, S. and Chee L. S. 2012. Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*, 39(2): 2157–2165.

Chun, P., Ujike, I., Mishima, K., Kusumoto, M., and Okazaki, S. 2020. Random forest-based evaluation technique for internal damage in reinforced concrete featuring multiple nondestructive testing results. *Construction and Building Materials*:119-138.

Ciobanu, A., Luca, M., Muscă, E. and Păvăloi, I. 2014. Automatic fury recognition in audio records. In: Proceedings of the *9th International Conference on Development and Application Systems*: 176–179.

Cong, P., Wang, C., Ren, Z., Wang, H., Wang, Y. and Feng, J. 2017. Unsatisfied customer call detection with deep learning. In: Proceedings of the *10th International Symposium on Chinese Spoken Language Processing*: 1–5.

Conradson, D. and Mckay, D. 2007. Translocal subjectivities: Mobility, connection, emotion. *Mobilities*, 2(2): 167–174.

Costantini, G., Iaderola, I., Paoloni, A. and Todisco, M. 2014. EMOVO corpus: An Italian emotional speech database. In: Proceedings of the *9th International Conference on Language Resources and Evaluation*: 3501–3504.

Cowie, R., Douglas-Cowie, E., Savvidou, S. and McMahon, E. 2000. "Feeltrace": An instrument for recording perceived emotion in real time. In: Proceedings of the *ISCA Workshop on Speech & Emotion*: 19–24.

Crivelli, C., Jarillo, S., Russell, J. A., and Fernández-Dols, J. M. 2016. Reading emotions from faces in two indigenous societies. *Journal of Experimental Psychology: General*, 145(7): 830–843.

Dahake, P. P., Shaw, K. and Malathi, P. 2016. Speaker dependent speech emotion recognition using MFCC and support vector machine. In: Proceedings of the *9th International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*: 1080–1084.

Dalrymple, K. A., Fletcher, K. and Corrow, S. 2014. "A room full of strangers every day": The psychosocial impact of developmental prosopagnosia on children and their families. *Journal of Psychosomatic Research*, 77(2): 144–150.

Damasio, A. R. 2006.The somatic marker hypothesis and the possible functions of the prefrontal

cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1346): 1413–1420.

Dapretto, M., Davies, M. S. and Pfeifer, J. H. 2006. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature neuroscience*, 9(1): 28–30.

Daqrouq, K., Al-Hmouz, R., Balamash, A. S., Alotaibi, N. and Noeth, E. 2015. An investigation of wavelet average framing LPC for noisy speaker identification environment. *Mathematical problems in engineering*, 2015.

Darst, B. F., Malecki, K. C. and Engelman, C. D. 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, 19(Suppl 1): 1–6.

Deng, J., Zhang, Z, Eyben, F. and Schuller, B. 2014. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9): 1068–1072.

Deng, J., Eyben, F., Schuller, B. and Burkhardt, F. 2017.Deep neural networks for anger detection from real life speech data. In: Proceedings of the *7th Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*: 1–6.

Desai, N., Dhameliya, K. and Bhatt, N. 2015.Feature extraction and classification techniques for speaker recognition: A review. In: Proceedings of the *International Conference on Electrical, Electronics, Signals, Communication and Optimization, EESCO 2015*, 3(12): 1–5.

Dew, I. T. Z., Ritchey, M., Labar, K. S and Cabeza, R. 2014. Prior perceptual processing enhances the effect of emotional arousal on the neural correlates of memory retrieval. *Neurobiology of Learning and Memory*, 112: 104–113.

Dietterich, T. G. 2000. Ensemble methods in machine learning, In: Proceedings of the *International workshop on multiple classifier systems*: 1–15.

Doherty, R. W., Orimoto, L., Singelis, T. M., Hatfield, E. and Hebb, J. 1995. Gender and occupational differences. *Psychology*, 19: 355–371.

Dong, W., Huang, Y., Lehane, B. and Ma, G. 2020. XGBoost algorithm-based prediction of

concrete electrical resistivity for structural health monitoring. *Automation in Construction*, 114(9): 103-125.

Dong, X., Yu, Z. and Cao, W. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2): 241–258.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin J. C., Devillers, L., Abrilian, S., Batliner, A., Amir, N. and Karpouzis, K. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. *Affective Computing and Intelligent Interaction*, 4738: 488–500.

Drottz-Sjöberg, B. M. and Sjoberg, L. 1990 Risk perception and worries after the chernobyl accident. *Journal of Environmental Psychology*, 10(2): 135–149.

Duchaine, B. C., Parker, H. and Nakayama, K. 2003. Normal recognition of emotion in a prosopagnosic. *Perception*, 32(7): 827–838.

Efron, B. and Tibshirani, R. J. 1993..*An introduction to the bootstrap*. New york: Chapman and Hall/CRC.

Ekman, P. 1992. Are there basic emotions? *Psychological Review*, 99(3): 550–553.

Ekpenyong, M., Inyang, U. and Udoh, E. O. 2018.Unsupervised visualization of under-resourced speech prosody. *Speech Communication*, 101: 45–56.

Faith, M. and Thayer, J. F.2001. A dynamical systems interpretation of a dimensional model of emotion. *Scandinavian Journal of Psychology*, 42(2): 121–133.

Farooq, M.,Hussain, F., Baloch, N.K., Raja, F.R., Yu, H. and Zikria, Y.B. 2020. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors*, 20: 45-62.

Fayek, H. M., Lech, M. and Cavedon, L. 2015. Towards real-time speech emotion recognition using deep neural networks. In: Proceedings of the *9th International Conference on Signal Processing and Communication Systems:*1-9.

Fayek, H. M., Lech, M. and Cavedon, L. 2017. Evaluating deep learning architectures for speech

emotion recognition. *Neural Networks*, 92: 60–68.

Feinberg, R. A., Hokama, L., Kadam, R. and Kim, I 2002. Operational determinants of caller satisfaction in the banking/financial services call center. *International Journal of Bank Marketing*, 20(4): 174–180.

Fewzee, P. and Karray, F. 2012. Dimensionality reduction for emotional speech recognition. In: Proceedings of the *International Conference on Privacy, Security, Risk and Trust*: 532–537.

Fischer, A. H., Rodriguez, M., P. M., van Vianen, A. E. M., and Manstead, A. S. R.2 004. Gender and culture differences in emotion. *Emotion*, 4(1): 87–94.

Freksa, C., Newcombe, N. S., Gärdenfors, P. and Wölfl, S.2008. Spatial cognition learning, reasoning, and talking about space. *Lecture Notes in Artificial Intelligence*, Springer,568.

Freund, Y., Schapire, R. E. and Hill, M. 1996. Experiments with a new boosting algorithm. In: Proceedings of the *13th International Conference on International Conference on Machine Learning*: 148–156.

Friedman, J. H. 2002.Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4): 367–378.

Gahlawat, M., Malik, A. and Bansal, P. 2016. Integrating human emotions with spatial speech using optimized selection of acoustic phonetic units. *Biologically Inspired Cognitive Architectures*, 15: 51–60.

Galanis, D., Karabetsos, S., Koutsombogera, M, Papageorgiou, H., Esposito A. and Riviello, M. 2013. Classification of emotional speech units in call centre interactions. In: Proceedings of the *4th IEEE International Conference on Cognitive Infocommunications*: 403–406.

Genuer, R., Poggi, J. M. and Tuleau-Malot, C. 2010. Variable selection using random forests. *Pattern Recognition Letters*, 31(14): 2225–2236.

Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H. and Krieter, J. 2020. Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. *Computers and Electronics in*

Agriculture, 173(1):105-134.

Getahun, F. and Kebede, M. 2017. Emotion identification from spontaneous communication. In: Proceedings of the *12th International Conference on Signal Image Technology and Internet-Based Systems*: 151–158.

Geurts, P., Ernst, D. and Wehenkel, L. 2006. Extremely randomized trees. *Machine Learning*, 63(1): 3–42.

Giannakopoulos, T. and Pikrakis, A. 2014. *Introduction to audio analysis: A MATLAB Approach*, Oxford: Academic Press.

Gideon, J., Provost, E. M. and McInnis, M. 2016. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In: Proceedings of the *IEEE International Conference on Acoustics*: 2359–2363.

Gievska, S., Koroveshovski, K. and Chavdarova, T. 2015. A hybrid approach for emotion detection in support of affective interaction. In: Proceedings of the *IEEE International Conference on Data Mining Workshops*: 352–359.

Goldshmidt, O. T. and Weller, L. 2000. "Talking emotions": Gender differences in a variety of conversational contexts. *Symbolic Interaction*, 23(2): 117–134.

Gong, S., Dai, Y., Ji, J., Wang, J. and Sun, H. 2015. Emotion analysis of telephone complaints from customer based on affective computing. *Computational Intelligence and Neuroscience*, 2015.

Gordon and Ladefoged 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4): 383–406.

Graham, S. and Weiner, B. 2011. From an attributional theory of emotion to developmental psychology: A round-trip ticket? *Social Cognition*, 4(2): 152–179.

Granitto, P. M., Furlanello, C., Biasioli, F. and Gasperi, F. 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2): 83–90.

Gray, J. A. 1990. Brain systems that mediate both emotion and cognition. *Cognition and Emotion*,

4(3): 269–288.

Greasley, A. and Smith, C. M. 2017. Using activity-based costing and simulation to reduce cost at a police communications centre. *Policing*, 40(2): 426–441.

Gregorutti, B., Michel, B. and Saint-Pierre, P. 2017. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3): 659–678.

Grossberg, S. and Gutowski, W. 1987. Neural dynamics of decision making under risk: affective balance theory. *Psychological Review*, (3): 300–318.

Grozdić, D., Jovicic, S., Pavlovic, D. S., Galic, J., and Markovic, B. 2017. Comparison of cepstral normalization techniques in whispered speech recognition. *Advances in Electrical and Computer Engineering*, 17(1): 21–26.

Guenther, F. H. 2006. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5): 350–365.

Guidi, A., Gentili, C., Scilingo, E. P. Vanello, N. 2019. Analysis of speech features and personality traits. *Biomedical Signal Processing and Control*, 51: 1–7.

Guyon, I., Zhu, E. and Patel, R. 2013.Tracking cellulase behaviors. *Biotechnology and Bioengineering*, 110(1):123-148.

Hamann, S. B., Ely, T. D., Hoffman, J. M. and Kilts, C. D. 2002. Ecstasy and agony: activation of the human amygdala in positive and negative emotion, *Psychological science*, 13(2): 135–41.

Hamzah, R., Jamil, N., Samah, K. A. A. F., Mangshor, N. N. A., Sabri, N. and Roslan, R. 2017. Comparing statistical classifiers for emotion classification. In: Proceedings of the *7th IEEE International Conference on System Engineering and Technology*,(10): 183–188.

Han, K., Yu, D. and Tashev, I. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In: Proceedings of the *Proc. INTERSPEECH 2014*: 223–227.

Haq, S. and Jackson, P. J. B. 2009.Speaker-dependent audio-visual emotion recognition. In: Proceedings of the *Auditory-Visual Speech Processing*: 1–6.

Harár, P., Burget, R. and Dutta, M. K. 2017. Speech emotion recognition with deep learning. In: Proceedings of the *4th International Conference on Signal Processing and Integrated Networks (SPIN)*: 137–140.

Harley, J. M., Bouchet, F., Hussain, M. S., Azevedo, R. and Calvo, R. 2015. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, 48: 615–625.

Hay, E. L. and Diehl, M. 2011. Emotion complexity and emotion regulation across adulthood. *European Journal of Ageing*, 8(3): 157–168.

Hechanova, M. R. M. 2013. The call center as a revolving door: A Philippine perspective. *Personnel Review*, 42(3): 349–365.

Hellbernd, N. and Sammler, D. 2016. Prosody conveys speakers intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88: 70–86.

Hernández, M., Ventura, N., Costa, A., Miró-Padilla, A., and Ávila, C. 2019. Brain networks involved in accented speech processing. *Brain and Language*: 12–22.

Hess, U. and Thibault, P. 2009. Darwin and emotion expression. *American Psychologist*, 64(2): 120–128.

Hifny, Y. and Ali, A. 2019. Efficient arabic emotion recognition using deep neural networks. In: Proceedings of the *44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: 6710–6714.

Holland, J. 2007. Emotions and research. *International Journal of Social Research Methodology*, 10(3): 195–209.

Hossain, M. S. and Muhammad, G. 2017. An emotion recognition system for mobile applications. *IEEE Access*, 5: 2281–2287.

Hossain, M. S. and Muhammad, G. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49(9): 69–78.

Huang, K.Y. and Chen, K.-J. 2011. Multilayer perceptron for prediction of 2006 world cup football

game. *Advances in Artificial Neural Systems*, 2011(2): 1–8.

Huang, Z., Demrci, M. F. Yazici, A. 2014. Speech emotion recognition using CNN. In: Proceedings of the *Proceedings of the ACM International Conference on Multimedia*: 801–804.

Hübscher, I., Borràs-Comes, J. and Prieto, P. 2017. Prosodic mitigation characterizes catalan formal speech: The Frequency Code reassessed. *Journal of Phonetics*, 65: 145–159.

Hudson, S., González-Gómez, H. V. and Rychalski, A. 2017. Call centers: is there an upside to the dissatisfied customer experience? *Journal of Business Strategy*, 38(1): 39–46.

Ille, R., Wabnegger, A. and Schwingenschuh, P. 2016. Intact emotion recognition and experience but dysfunctional emotion regulation in idiopathic Parkinsons disease. *Journal of the Neurological Sciences*, 361: 72–78.

Interaction, I., Zhang, Z., Marchi, E. and Schuller, B. 2013. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: Proceedings of the *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*: 511–516.

Irastorza, J. and Torres, M. I. 201. Analyzing the expression of annoyance during phone calls to complaint services. In: Proceedings of the *7th IEEE International Conference on Cognitive Infocommunications*: 103–107.

Istening, E. M. L. and Imbre, T. 2018. Embodied listening and timbre: perceptual,acoustical, and neural correlates. *Music Perception*, 35(3): 332–363.

Izard, C. E. 1992. Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, 99(3): 561–589.

Izquierdo, E. and Zurita, R. 2020. An evaluation of guided regularized random forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 88(6):102-126.

Jameel, A., Siyal, M. Y. and Ahmed, N. 2005. FFT based analog speech scrambler using DSP. In: Proceedings of the *2005 Pakistan Section Multitopic Conference, INMIC*: 1–4.

James, W. 1884. What is an emotion? *Mind*,9(34): 188–205.

Jauk, I., Bonafonte, A. and Pascual, S. 2016. Acoustic feature prediction from semantic features for expressive speech using deep neural networks. In: Proceedings of the *European Signal Processing Conference*: 2320–2324.

Jayasankar, T., Vinothkumar, K. and Vijayaselvi, A. 2017. Automatic gender identification in speech recognition by genetic algorithm. *Applied Mathematics and Information Sciences* 913(3): 907–913.

Jiang, H., Bin, H., Zhenyu, L., Lihua, Y., Tianyang, W., Fei, L., Huanyu, K. and Xiaoyu, L. 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, 90: 39–46.

Jiang, W., Wang, Z., Jin, J. S., Han, X. and Li, C. 2019. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors (Switzerland)*, 19(12): 1–15.

Johnson, K. J., Waugh, C. E. and Fredrickson, B. L. 2010. Smile to see the forest: Facially expressed positive emotions broaden cognition. *Cognition and Emotion*, 24(2): 299–321.

Jokinen, J. P. P.2015. Emotional user experience: Traits, events, and states. *International Journal of Human Computer Studies*. Elsevier, 76: 67–77.

Kahou, S.E., Bouthillier, X., Lamblin, P., Gülçehre, Ç., Michalski, V., Konda, K.R., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., Ferrari, R.C., Mirza, M., Warde-Farley, D., Courville, A.C., Vincent, P., Memisevic, R., Pal, C.J., and Bengio, Y. 2016. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2): 99–111.

Kannadaguli, P. and Bhat, V. 2018. A comparison of bayesian and HMM based approaches in machine learning for emotion detection in native Kannada speaker. In: Proceedings of the *2018 IEEMA Engineer Infinite Conference*, (1): 1–6.

Karazi, T., Sasson-Levy, O. and Lomsky-Feder, E. 2018. Gender, emotions management, and power in organizations: The case of Israeli women junior military officers. *Sex Roles*, 78(7–8): 573–586.

Ke, X., Zhu, Y., Wen, L. and Zhang, W. 2018. Speech emotion recognition based on SVM and

ANN. *International Journal of Machine Learning and Computing*, 8(3): 198–202.

Keltner, D. and Haidt, J. 1999. Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5): 505–521.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M. A. 2018. Speech emotion recognition: methods and cases study. In: Proceedings of the *10th International Conference on Agents and Artificial Intelligence (ICAART2018)*: 175–182.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M. A. and Cléder, C. 2019. Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication*, 114(5): 22–35.

Khan, A. and Roy, U. K. 2017. Emotion recognition using prosodic and spectral features of speech and naïve bayes classifier. In: Proceedings of the *2017 International Conference on Wireless Communications*: 1017–1021.

Khorram, S., McInnis, M. and Mower Provost, E. 2019. Jointly aligning and predicting continuous emotion annotations. *IEEE Transactions on Affective Computing*: 1–16.

Kiger, G. 1997. The structure of attitudes toward persons who are deaf: Emotions, values, and stereotypes. *Journal of Psychology: Interdisciplinary and Applied*, 131(5): 554–560.

Kim, H. G., Moreau, N. and Sikora, T. 2006. *MPEG-7 Audio and beyond: Audio content indexing and retrieval*. New York, Wiley:45- 66.

Kim, J., Englebienne, G., Truong, K.P. and Evers, V. 2017. Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning. In: Proceedings of the *Annual Conference of the International Speech Communication Association*: 1113–1117.

Kim, J., Truong, K.P., Englebienne, G., and Evers, V. 2018. Learning spectro-temporal features with 3D CNNs for speech emotion recognition. In: Proceedings of the *7th International Conference on Affective Computing and Intelligent Interaction*: 383–388.

Kim, J. B. and Park, J. S. 2016. Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition. *Engineering Applications of Artificial Intelligence*,

52: 126–134.

Koolagudi, S. G. and Rao, K. S. 2012. Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15(2): 99–117.

Koteswara, R. A., Swarna, K. and Hima, D. V. 2016. Acoustic modeling for emotion recognition. Berlin, Springer International Publishing: 66-89.

Kotti, M. and Kotropoulos, C. 2008. Gender classification in two emotional speech databases. In: Proceedings of the *19th International Conference on Pattern Recognition*: 1–4.

Krishna, K. V. and Krishna, P. 2013. Emotion recognition in speech using MFCC and wavelet features. In: Proceedings of the *3rd IEEE International Advance Computing Conference, IACC 2013*: 842–847.

Kumar, M. and Yadav, N. 2011. Multilayer perceptrons and radial basis function neural network methods for the solution of differential equations: A survey. *Computers and Mathematics with Applications*, 62(10): 3796–3811.

Kumari, M. and Ali, I. 2015. An efficient algorithm for gender detection using voice samples. In: Proceedings of the *2015 Communication, Control and Intelligent Systems*: 221-226.

Kunz, R. E. 1997. Miniature integrated optical modules for chemical and biochemical sensing, Sensors and Actuators. In: Proceedings of the *Part of special issue:3rd European Conference on Optical Chemical Sensors and Biosensors*: 13-28.

Kursa, M. B. and Rudnicki, W. R. 2010. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11): 24-53.

Kwon, S. 2020. CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM Network. *Mathematics*, 8, 2133.

L.Brizendine 2006. The Female Brain, *Nature*, 443(7112): 634–634.

Lalitha, S., Lalitha, S., Madhavan, A., Bhushan, B., and Saketh, S. 2014. Speech emotion recognition. In: Proceedings of the *2014 International Conference on Advances in Electronics, Computers and Communications*: 1–5.

Lalitha, S., Geyasruti, D., Narayanan, R., and Shravani, M. 2015. Emotion detection using MFCC and cepstrum features. In: Proceedings of the *Procedia Computer Science*, 70: 29–35.

Langari, S., Marvi, H., and Zahedi, M. 2020. Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked,* 20: 13-28.

Latif, S., Rana, R., Younis, S., Qadir, J., and Epps, J. 2018. Transfer learning for improving speech emotion classification accuracy. In: Proceedings of the *Annual Conference of the International Speech Communication Association*: 257–261.

Latif, S., Qayyum, A., Usman, M., and Qadir, J. 2019. Cross lingual speech emotion recognition: Urdu vs. Western languages. In: Proceedings of the *2018 International Conference on Frontiers of Information Technology*: 88–93.

Lee, S., Han, D.K. and Ko, H. 2020. Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition. *Sensors*, 20, 6688.

Lee, C. M. and Narayanan, S. S. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2): 293–303.

Lee, C. W., Song, K.Y., Jeong, J., and Choi, W.Y. 2018. Convolutional attention networks for multimodal emotion recognition from speech and text data. In: Proceedings of the *Grand Challenge and Workshop on Human Multimodal Language*: 28–34.

Leo, J., Luhanga, E. and Michael, K. 2019. Machine learning model for imbalanced cholera dataset in Tanzania. *Scientific World Journal*, 2019.

Levine, L. J. and Safer, M. A. 2002. Sources of bias in memory for emotions. *Current Directions in Psychological Science*, 11(5): 169–173.

Li, D., Yang, Y. and Dai, W. 2014. Cost-sensitive learning for emotion robust speaker recognition. *Scientific World Journal*, 2014.

Li, S., Xu, L. and Yang, Z. 2017. Multidimensional speaker information recognition based on proposed baseline system. In: Proceedings of the *IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference*: 1776–1780.

Li, X., Pang, J., Mo, B., and Rao, Y. 2016.Hybrid neural networks for social emotion detection over short text. In: Proceedings of the *2016 International Joint Conference on Neural Networks*: 537–544.

Li, X. and Akagi, M. 2019. Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model. *Speech Communication*, 110(1): 1–12.

Likitha, M. S., Gupta, S.R., Hasitha, K., and Raju, A.U. 2017. Speech based human emotion recognition using MFCC. In: Proceedings of the *2017 International Conference on Wireless Communications, Signal Processing and Networking*: 2257–2260.

Lindquist, K. A., Wager, T.D., Kober, H., Bliss, E., and Barrett, L.F. 2015. The brain basis of emotion: A meta-analytic review. *Behavior Brain Science,* 35(3): 121–143.

Lisetti, C. L. 1998. Affective computing. *Pattern Analysis and Applications*, 1(1): 71–73.

Liu, D. S. and Fan, S. J. 2014. A modified decision tree algorithm based on genetic algorithm for mobile user classification problem. *The Scientific World Journal*, 2014.

Liu, H., Li, T., Chen, L., Zhan, S., Pan, M., Ma, Z., Li, C., and Zhang, Z. 2016. To set up a logistic regression prediction model for hepatotoxicity of chinese herbal medicines based on traditional chinese medicine theory. *Evidence-Based Complementary and Alternative Medicine*: 1–9.

Liu, N., Zong, Y., Zhang, B., Liu, L., Chen, J.J., Zhao, G., and Zhu, J. 2018. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning. In: Proceedings of the *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*: 5144–5148.

Liu, S., Meng, J., Yang, J., Zhao, X.J., He, F., Qi, H., and Ming, D. 2015. Within-stimulus emotion recognition may inflate the classification accuracies based on EEG signals. In: Proceedings of the *IEEE 7th International Conference on Awareness Science and Technology*: 115–118.

Liu, Z., Wu, Z., Li, T., Li, J., and Shen, C. 2018. GMM and CNN Hybrid method for short utterance speaker recognition. *IEEE Transactions on Industrial Informatics*, 14(7): 3244–3252.

Liu, Z. T., Xie, Q., Wu, M., Cao, W., Mei, Y., and Mao, J. 2018. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, 309: 145–156.

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).

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., and Cai, H. 2017. Detecting depression in speech: Comparison and combination between different speech types. In: Proceedings of the *2017 IEEE International Conference on Bioinformatics and Biomedicine*: 1052–1058.

Lorenzo, J., J., Henter, G.E., Takaki, S., Yamagishi, J., Morino, Y., and Ochiai, Y. 2018. Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication*, 99: 135–143.

Lu, L., Liu, L., Hussain, M.J., and Liu, Y. 2017. I sense you by breath: Speaker recognition via breath biometrics. *IEEE Transactions on Dependable and Secure Computing*, 15(2): 1–15.

Luengo, I., Navas, E. and Hernaez, I. 2010. Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6): 490–501.

Lutz, C. 1986. The anthropology of emotions. *Annual Review of Anthropology*, 15(1): 405–436.

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., and Kosir, A. 2019. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, 46: 184–192.

MacGeorge, E. L., Gillihan, S.J., Samter, W., and Clark, R.A. 2003. Skill deficit or differential motivation? *Communication Research*, 30(3): 272–303.

Mahesh, V. S. and Kasturi, A. 2006. Improving call centre agent performance: A UK-India study based on the agents point of view. *International Journal of Service Industry Management*, 17(2): 136–157.

Majuran, S. and Ramanan, A. 2018. A feature-driven hierarchical classification approach to emotions in speeches using SVMs. In: Proceedings of the *2017 IEEE International Conference on Industrial and Information Systems*: 1–5.

Malarkodi, C. S. and Devi, S. L. 2019. Generic feature selection methodology to named entity detection from Indian and European languages. *Advances in Electrical and Computer Engineering*,

19(1): 79–88.

Malik, F., Farhan, S. and Fahiem, M. A. 2018. An ensemble of classifiers based approach for prediction of alzheimers disease using fmri images based on fusion of volumetric, textural and hemodynamic features. *Advances in Electrical and Computer Engineering*, 18(1): 61–70.

Malti, T. and Latzko, B. 2012. Moral emotions. *Encyclopedia of Human Behavior*, 15(4): 644–649.

Mannepalli, K., Sastry, P. N. and Suman, M. 2018. Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University - Computer and Information Sciences*: 1319-1578.

Mansour, A. and Lachiri, Z. 2017. A comparative study in emotional speaker recognition in noisy environment. In: Proceedings of the *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications*: 980–986.

Mao, Q., Xu, G., Xue, W., Gou, J., and Zhan, Y. 2017. Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Communication*, 93: 1–10.

Martin, O., O., Kotsia, I., Macq, B.M., and Pitas, I. 2006. The eNTERFACE 05 audio-visual emotion database. In: Proceedings of the *22nd International Conference on Data Engineering Workshops*: 8-8.

Masuyama, N., Islam, M.N., Seera, M., and Loo, C.K. 2017. Application of emotion affected associative memory based on mood congruency effects for a humanoid. *Neural Computing and Applications,* 28(4): 737–752.

Mauss, I. B. and Robinson, M. D. 2009. Measures of emotion: A review. *Cognition and Emotion*, 23(2): 209–237.

Mcdougall, C. 1991. Aggression , anger control. *Personality and Individual Differences*, 12(6): 625–629.

McEnnis, D., McKay, C., Fujinaga, I., and Depalle, P. 2005. jAudio: A feature extraction library.

In: Proceedings of the *International Conference on Music Information Retrieval*, (1): 600–603.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In: Proceedings of the *14th Python in Science Conference*, (Scipy): 18–24.

McNee, S. M., Riedl, J. and Konstan, J. A. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: Proceedings of the *Conference on Human Factors in Computing Systems*: 1097–1101.

Meilán, J.J., Martínez, F., Carro, J., López, D.E., Millian, L., and Arana, J.M. 2014. Speech in alzheimers disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(6): 327–334.

Mellahi, T. and Hamdi, R. 2015. LPC-based formant enhancement method in kalman filtering for speech enhancement. *International Journal of Electronics and Communications*, 69(2): 545–554.

Memarian, H. and Balasundram, S. K. 2012.Comparison between Multi-layer perceptron and radial basis function networks for sediment load estimation in a tropical watershed. *Journal of Water Resource and Protection*, 4(10): 870–876.

Moller, A., Ruhlmann-Kleider, V., Leloup, C., Neveu, J., Palanque, N., Rich, J., Carlberg, R., Lidman, C., and Pritchet, C. 2016. Photometric classification of type Ia supernovae in the supernova legacy survey with supervised learning. *Journal of Cosmology and Astroparticle Physics*, 2016(12).

Mowrer, O. H. 2009. *Learning theory and behavior.* New York: John Wiley and Sons Inc.

Mozziconacci, S. J. L. and Hermes, D. J. 1997. A study of intonation patterns in speech expressing emotion or attitude: Production and perception. *IPO Annual Progress Report*: 154–160.

Mroczek, D. K. 2001. Age and emotion in adulthood. *Current Directions in Psychologiccal Science*, 10: 87–90.

Munezero, M., Montero, C.S., Sutinen, E., and Pajunen, J. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*,

5(2): 101–111.

Mustaqeem and Kwon, S. 2020a. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors (Switzerland)*, 20(1).

Mustaqeem and Kwon, S. 2020b CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM Network. *Mathematics*, 8: 1- 21.

Muthusamy, H., Polat, K. and Yaacob, S. 2015. Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. *Mathematical Problems in Engineering*, 2015: 1-13.

Najnin, S. and Banerjee, B. 2019. Speech recognition using cepstral articulatory features. *Speech Communication*, 107: 26–37.

Narendra, N.P., Airaksinen, M., Story, B.H., and Alku, P. 2019. Estimation of the glottal source from coded telephone speech using deep neural networks. *Speech Communication*, 106(12): 95–104.

Narendra, N. P. and Alku, P. 2019. Dysarthric speech classification from coded telephone speech using glottal features. *Speech Communication*, 110: 47–55.

Natekin, A. and Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(12): 1-21.

Nesse M, R. 1990. Evolutionary explanations of emotions. *Human Nature*, 1(3): 261–289.

Nesse, R. M. and Ellsworth, P. C. 2009. Evolution, emotions, and emotional disorders. *American Psychologist*, 64(2): 129–139.

Neumann, M. and Thang Vu, N. G. 2018. CRoss-lingual and multilingual speech emotion recognition on English and French. In: Proceedings of the *IEEE International Conference on Acoustics*: 5769–5773.

Neustein, A. 2013. Language identification using excitation source features. Berlin: SpringerBriefs in Electrical and Computer Engineering Speech Technology Series Editor.

Nguyen, T.T Huang, J and Nguyen, T2015. Unbiased feature selection in learning random forests for high-dimensional data. *Scientific World Journal*, 2015: 1–18.

Nirmal, J.H., Zaveri, M.A., Patnaik, S., and Kachare, P.H. 2017. Novel approach of MFCC based alignment and WD-residual modification for voice conversion using RBF. *Neurocomputing*, 237: 39–49.

Nogueira, P.A., Rodrigues, R.A., Oliveira, E.C., and Nacke, L.E. 2013. A hybrid approach at emotional state detection: Merging theoretical models of emotion with data-driven statistical classifiers. In: Proceedings of the *2013 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, 2(3): 253–260.

Noh, K.J., Jeong, C.Y., Lim, J., Chung, S., Kim, G., Lim, J.M. and Jeong, H. 2021. Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets. *Sensors* 2021, 21, 1579.

OConnor, K. E. 2008. "You choose to care": Teachers, emotions and professional identity. *Teaching and Teacher Education*, 24(1): 117–126.

Oatley, K. and Johnson, P. N. 1987. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1(1): 29–50.

Ocquaye, E. N. N., Mao, Q., Xue, Y., Song, H. 2020. Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *International Journal of Intelligent Systems*, 35(9): 1-19

Oishi, S. and Schimmack, U. 2010. Culture and well-being. *Perspectives on Psychological Science*, 5(4): 463–471.

Oliver, J. 2013. Emotion-oriented systems the humaine handbook. *Journal of Chemical Information and Modeling*,23(2): 23-52.

Olugbara,O.O., Adetiba, E. and Oyewole, S.A. 2015. Pixel intensity clustering algorithm for multilevel image segmentation. *Mathematical problems in engineering*, 2015: 1-19.

Olugbara, O.O., Taiwo, T.B. and Heukelman, D., 2018. Segmentation of melanoma skin lesion

using perceptual color difference saliency with morphological analysis. *Mathematical Problems in Engineering*, 2018: 1-19.

Ong, A.D., Benson, L., Zautra, A.J., and Ram, N. 2018. Emodiversity and biomarkers of inflammation. *Emotion*, 18(1): 3–14.

Oyewole, S.A and Olugbara, O.O. 2018. Product image classification using eigen colour feature with ensemble machine learning. *Egyptian Informatics Journal*, 19(2): 83-100.

Ortony, A., and Turner, T.J. 1990. What's basic about basic emotions ? *Psychological Review*, 97(3): 315–331.

Ozaydin, S. 2017. An isolated word speaker recognition system. In: Proceedings of the *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*: 1-5.

Ozseven, T. 2018. The acoustic cues of fear: Investigation of acoustic parameters of speech containing fear. *Archives of Acoustics*, 43(2): 245–251.

Pang, B., Yue, J., Zhao., G. and Xu, Z. 2017. Statistical downscaling of temperature with the random forest model. *Advances in Meteorology*, 2017: 1–11.

Panksepp, J. and Watt, D. 2011. What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion Review*, 3(4): 387–396.

Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P. and Makedon, F. 2017. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation,* 5(4):26-41.

Pappas, D., Androutsopoulos, I. and Papageorgiou, H. 2016. Anger detection in call center dialogues. In: Proceedings of the *6th IEEE Conference on Cognitive Infocommunications:*139–144.

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M. and Hofer, G. 2019. Analysis of deep learning architectures for cross-corpus speech emotion recognition. In: Proceedings of the *Annual Conference of the International Speech Communication Association*: 1656–1660.

Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S. and Rigoulot, S. 2015. Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, 111: 14–25.

Pérez, H., Reyes, C. A. and Villaseñor, L. 2012. Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model. *Biomedical Signal Processing and Control*, 7(1): 79–87.

Perry, N. B., Dollar J. M, Calkins, S. D, Keane, S. P and Shanahan, L. 2018. Childhood self-regulation as a mechanism through which early overcontrolling parenting is associated with adjustment in preadolescence. *Developmental Psychology*, 54(8): 1542–1554.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G. and Pantic, M. 2018. End-to-end audiovisual speech recognition. In: Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*: 6548–6552.

Pierre, O. 2003. The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human Computer Studies*, 59(1–2): 157–183.

Plutchik, R. 2013. Emotions and psychotherapy: a psychoevolutionary perspective, emotion, psychopathology, and psychotherapy. Amsterdam: Academic Press.

Pollak, S. D., Messner, M., Kistler, D.J and Cohn, J. F. 2009. Development of perceptual expertise in emotion recognition. *Cognition*, 110(2): 242–247.

Polzehl, T., Schmitt, A. and Metze, F. 2010. Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection. In: Proceedings of the *Proc of Speech Prosody*: 3–6.

Poria, S., Chaturvedi, I. and Cambria, E. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings of the *IEEE 16th International Conference on Data Mining*: 439–448.

Prasada, K., Chandra, M. and Hemanth , N. 2019. An integrated approach to emotion recognition and gender classification. *Journal of Visual Communication and Image Representation*, 60: 339–345.

Přibil, J., Přibilová, A. and Matoušek, J. 2016. GMM-based speaker gender and age classification after voice conversion. In: Proceedings of the *1st International Workshop on Sensing, Processing and Learning for Intelligent Machines*: 1-5.

Proniewska, K., Pregowska, A. and Malinowski, K. P. 2020. Identification of human vital functions directly relevant to the respiratory system based on the cardiac and acoustic parameters and random forest. *Innovation and Research in BioMedical engineering*, 1: 5–10.

Pützer, M., Moringlane, J. R., Sikos, L., Reith, W. and Krick, C. M. 2019. fMRI and acoustic analyses reveal neural correlates of gestural complexity and articulatory effort within bilateral inferior frontal gyrus during speech production. *Neuropsychologia*, 132: 107-129.

Raelin, J. D. 2007. The laws of emotion. *Academy of Management Review*, 32(3): 995–998.

Rajagopal, S., Kundapur, P. P. and Hareesha, K. S. 2020. A stacking ensemble for network intrusion detection using heterogeneous datasets. *Security and Communication Networks*, 2020: 1–9.

Ram, R., Palo, H. K. and Mohanty, M. N. 2013. Emotion recognition with speech for call centres using lpc and spectral analysis. *International Journal of Advanced Computer Research*, 3(3): 182–187.

Ramdinmawii, E. and Mittal, V. K. 2016.Gender identification from speech signal by examining the speech production characteristics. In: Proceedings of the *2016 International Conference on Signal Processing and Communication*: 244–249.

Rathor, S. and Jadon, R. S. 2017. Text independent speaker recognition using wavelet cepstral coefficient and butter worth filter. In: Proceedings of the *8th International Conference on Computing, Communication and Networking Technologies*: 1–5.

Reddy, S. A., Singh, A., Kumar, N. S. and Sruthi, K. S. 2011. The decisive emotion identifier? In: Proceedings of the *3rd International Conference on Electronics Computer Technology*, 2: 28–32.

Richins L. Marsha 1997. Measuring emotionsin the consumption experience. *Journal of Consumer Research*, 24(2): 127–146.

Rieffe, C. 2012. Awareness and regulation of emotions in deaf children. *British Journal of Developmental Psychology*, 30(4): 477–492.

Rong, J., Chen, Y.P., Chowdhury, M.U., and Li, G. 2007. Acoustic features extraction for emotion recognition. In: Proceedings of the *6th IEEE/ACIS International Conference on Computer and Information Science*: 419–424.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6): 1161–1178.

Sagi, O. and Rokach, L. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4): 1–18.

Sahadev, S., Purani, K. and Kumar Panda, T. 2017. Service employee adaptiveness: Exploring the impact of role-stress and managerial control approaches. *Employee Relations*, 39(1): 54–78.

Sahoo, S. and Routray, A. 2017. MFCC feature with optimized frequency range: An essential step for emotion recognition. In: Proceedings of the *2016 International Conference on Systems in Medicine and Biology*: 162–165.

Saito, P. T. M., Nakamura, R. Y. M., Amorim W. P., Papa J. P., Rezende, P. J. and Falcão A. X. 2015. Choosing the most effective pattern classification model under learning-time constraint. *PLoS ONE*, 10(6).

Sarker, K. and Alam, K. R. 2014. Emotion recognition from human speech: emphasizing on relevant feature selection and majority voting technique. in *3rd International Conference On Informatics, Electronics & Vision 2014*: 89–95.

Satt, A., Rozenberg, S. and Hoory, R. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In: Proceedings of the *Annual Conference of the International Speech Communication Association*: 1089–1093.

Scherer, K. 1982. Emotion as a process: Function, origin and regulation. *Social Science Information*, 21(4): 555–570.

Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech*

Communication, 40(2): 227–256.

Von-Scheve, C. and Ismer, S. 2013. Towards a theory of collective emotions. *Emotion Review*, 5(4): 406–413.

Schuller, B., Zhang, Z., Weninger, F. and Rigoll, G. 2011. Using multiple databases for training in emotion recognition: To unite or to vote? In: Proceedings of the *Annual Conference of the International Speech Communication Association*: 1553–1556.

Schuller, B., Vlasenko B., Eyben, F., Wöllmer, M., Stuhlsatz, A. and Wendemuth, A. 2015. Cross-corpus acoustic emotion recognition: Variances and strategies. In: Proceedings of the *2015 International Conference on Affective Computing and Intelligent Interaction*: 470–476.

Schuller, B., Reiter, S. and Rigoll, G. 2006. Evolutionary feature generation in speech emotion recognition. *Evaluation*: 5–8.

Schupp, S. 2003. Lifting a butterfly - A component-based FFT. *Scientific Programming*, 11(4): 291–307.

Scollon, C. N., Diener, E., Oishi, S. and Biswas-Diener, R. 2004. Emotions across cultures and methods. *Journal of Cross-Cultural Psychology*, 35(3): 304–326.

Selvaraj, M., Bhuvana, R. and Padmaja, S. 2016. Human speech emotion. *International Journal of Engineering and Technology* (10): 311-323.

Semwal, N., Kumar, A. and Narayanan, S. 2017. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models, In: Proceedings of the *2017 IEEE International Conference on Identity, Security and Behavior Analysis*: 1–6.

Sen, S., Dutta, A. and Dey, N. 2019. *Audio processing and speech recognition: Concepts, Techniques and Research Overviews*. Singapore: Springer.

Seyyedattar, M., Ghiasi, M.M., Zendehboudi, S., and Butt, S. 2020. Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models. *Fuel*, 269(12): 116-149.

Shaqra, F. A., Duwairi, R. and Al-Ayyoub, M. 2019. Recognizing emotion from speech based on

age and gender using hierarchical models. In: Proceedings of the *Procedia Computer Science*, 151(2018): 37–44.

Shareef, H., Mutlag, A. H. and Mohamed, A. 2017. Random forest-based approach for maximum power point tracking of photovoltaic systems operating under actual environmental conditions. *Computational Intelligence and Neuroscience*: 1–17.

Shaver, P. R., Murdaya, U. and Fraley, R. C. 2001. Structure of the Indonesian emotion lexicon. *Asian Journal of Social Psychology*: 201–224.

Shegokar, P. and Sircar, P. 2016. Continuous wavelet transform based speech emotion recognition. In: Proceedings of the *10th International Conference on Signal Processing and Communication Systems*: 1–8.

Shott, S. 2018. Emotion and Social Life : A symbolic interactionist. *American Journal of Sociology*, 84(6): 1317–1334.

Shuman, V., Sander, D. and Scherer, K. R. 2013. Levels of valence. *Frontiers in Psychology*, 4(5): 1–17.

Sierra, H., Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R.2015. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 135(2): 612–615.

Silva, M., Vellasco, M. M. B. R. and Cataldo, E. 2017. Evolving spiking neural networks for recognition of aged voices. *Journal of Voice*, 31(1): 24–33.

Singh, N., Collins, M. and Hazen, T. J. 2007. Dimensionality reduction for speech recognition using neighborhood components analysis. In: Proceedings of the *Annual Conference of the International Speech Communication Association*: 1397–1400.

Skuk, V. G. and Schweinberger, S. R. 2013. Gender differences in familiar voice identification. *Hearing Research* ,296: 131–140.

Sonawane, A., Inamdar, M. U. and Bhangale, K. B. 2018. Sound based human emotion recognition using MFCC & multiple SVM. In: Proceedings of the *IEEE International Conference on*

Information, Communication, Instrumentation and Control: 1–4.

Song, K., Yan, F., Ding, T., Gao, L. and Lu, S. 2020. A steel property optimization model based on the XGBoost algorithm and improved PSO. *Computational Materials Science*, 174(12): 109-135.

Song, P., Ou, S., Du, Z., Guo, Z., Ma, W., Liu, J. and Zheng, W. 2017. Learning corpus-invariant discriminant feature representations for speech emotion recognition. *IEICE Transactions on Information and Systems*,12(5): 1136–1139.

Song, P. 2019. Transfer linear subspace learning for cross-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, 10(2): 265–275.

Spence, S. 1995. Descartes error: Emotion, reason and the human brain. *British Medical Journal*, 310: 1213-1235.

Stolar, M. N., Lech, M., Bolia, R. S. and Skinner, M. 2018. Real time speech emotion recognition using RGB image classification and transfer learning. In: Proceedings of the *11th International Conference on Signal Processing and Communication Systems*: 1–8.

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke,T., Meier, G. and Schuller, B. 2011. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In: Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*: 5688–5691.

Subasi, A., Kadasa, B. and Kremic, E. 2020. Classification of the cardiotocogram data for anticipation of fetal risks using bagging ensemble classifier. In: Proceedings of the *Procedia Computer Science*, 168(2019): 34–39.

Sudakov, O., Burnaev, E. and Koroteev, D. 2019. Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks. *Computers and Geosciences*, 127(11): 91–98.

Suh, E., Oishi, S. and Triandis, H. C. 2017. The shifting basis of life satisfaction judgments across cultures: Emotions versus norms. *Journal of Personality & Social Psychology*, 74(2): 482–493.

Sullivan, M. W. and Lewis, M. 2003. Emotional expressions of young infants and children a

practitioners primer. *Infants and Young Children*, 16(2): 120–142.

Sun, L., Zou, B., Fu, S., Chen, J. and Wang, F. 2019. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*, 115: 29–37.

Sun, L., Fu, S. and Wang, F. 2019. Decision tree SVM model with Fisher feature selection for speech emotion recognition. *Eurasip Journal on Audio, Speech, and Music Processing*: 23-44.

Sun, Y., Wen, G. and Wang, J. 2015. Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control*, 18: 80–90.

Susan, S. and Kaur, A. 2018. Measuring the randomness of speech cues for emotion recognition. In: Proceedings of the *10th International Conference on Contemporary Computing*: 1–6.

Swain, M., Sahoo, S., Routray, A. and Kabisatpathy, P. 2015. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. *International Journal of Speech Technology*, 18(3): 387–393.

Swain, M., Routray, A., Kabisatpathy, P. and Kundu, J. N. 2017. Study of prosodic feature extraction for multidialectal Odia speech emotion recognition. In: Proceedings of the *IEEE Region 10 Annual International Conference*: 1644–1649.

Takaki, S. and Yamagishi, J. 2016. A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis. In: Proceedings of the *IEEE International Conference on Acoustics*: 5535–5539.

Takamichi, S. 2018. Modulation spectrum-based speech parameter trajectory smoothing for DNN-based speech synthesis using FFT spectra. In: Proceedings of the *9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*: 1308–1311.

Tamayo, D., Silburt, A., Valencia, D., Menou, K., Ali, M., Petrovich, C., Huang, C., Rein, H., Laerhoven, C. and Paradise, A. 2016. A machine learns to predict the stability of tightly packed planetary systems. *The Astrophysical Journal*, 832(2):22-57.

Tao, H., Liang, R., Zha, C, Zhang, X. and Zhao, L. 2016. Spectral features based on local hu moments of gabor spectrograms for speech emotion recognition. *IEICE Transactions on*

Information and Systems, 8(1): 2186–2189.

Tashev, I. J., Wang, Z. Q. and Godin, K. 2017. Speech emotion recognition based on gaussian mixture models and deep neural networks. *Information Theory and Applications Workshop*: 1–4.

Taylor, S. E., Lerner, J. S, Sage, R. M., Lehman, B. J. and Seeman, T. E . 2004. Early environment, emotions, responses to stress, and health. *Journal of Personality*, 72(6): 1365–1393.

Thirunavukkarasu, S. G., Abdi, H. and Mohajer, N. 2016. A smart HMI for driving safety using emotion prediciton of EEG signals. In: Proceedings of the *2016 IEEE International Conference on Systems, Man, and Cybernetics*: 4148–4153.

Thyme, A. E. and Hutchins, S. E. 1996. On using prosodic cues in automatic language identification. In: Proceedings of the *International Conference on Spoken Language Processing*: 1768–1771.

Titze, I. R. 1994. Mechanical stress in phonation. *Journal of Voice*, 8(2): 99–105.

Tracy, J. L. and Randles, D. 2011. Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, 3(4): 397–405.

Tukey, J. W. and Cooley, J. W. 1965. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90): 297–301.

Tunç, T. 2012. A new hybrid method logistic regression and feedforward neural network for lung cancer data. *Mathematical Problems in Engineering*, 2012: 1–10.

Ubando, M. 2016. Gender differences in intimacy, emotional expressivity, and relationship satisfaction. *Pepperdine Journal of Communication Research*, 4(13): 19–29.

Urwin, C. 2008. A review of "Imitation in infancy". *Infant Observation*, 3(2): 106–109.

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

Vivek, J., Gokilavani, A., Kavitha, S., Lakshmanan, S. and Karthik, S. 2018. A novel emotion recognition based mind and soul-relaxing system. In: Proceedings of the *4th International*

Conference on Innovations in Information, Embedded and Communication Systems: 1–5.

Vrana, S. R. and Rollock, D. 2002. The role of ethnicity, gender, emotional content, and contextual differences in physiological, expressive, and self-reported emotional responses to imagery. *Cognition and Emotion*, 16(1): 165–192.

Wang, J., Wang, D. and Bie, F. 2015. DNN-based discriminative scoring for speaker recognition based on i-vector. *Center for Speech and LanguageTechnologies Technical Report*: 1–19.

Wang, K. An, N., Li, B. N., Zhang,Y. and Li, L. 2015. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1): 69–75.

Wang, W. 2010. *Machine audition: Principles, algorithms and systems*. University of Surrey: IGI Global Press: 398–423.

Watanabe, H., Tanaka, H., Sakti, S. and Nakamura S. 2019. Synchronization between overt speech envelope and EEG oscillations during imagined speech. *Neuroscience Research*, 153:48-55.

Watson, J. B. and Morgan, J. J. B. 2006. Emotional reactions and psychological experimentation. *The American Journal of Psychology*, 28(2): 163-187.

Weißkirchen, N., Böck, R. and Wendemuth, A. 2018. Recognition of emotional speech with convolutional neural networks by means of spectral estimates. In: Proceedings of the *7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*: 50–55.

Williams, G. and Mahmoud, A. 2017. Analyzing, classifying, and interpreting emotions in software users tweets. In: Proceedings of the *IEEE/ACM 2nd International Workshop on Emotion Awareness in Software Engineering*: 2–7.

Williams, L. M., Mathersul, D., Palmer, D. and Gur, R. 2009. Explicit identification and implicit recognition of facial emotions: Age effects in males and females across 10 decades. *Journal of Clinical and Experimental Neuropsychology*, 31(3): 257–277.

Wöllmer, M., Eyeben, F., Reiter, S. and Schuller, B. 2008. Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies. In: Proceedings of

the *Annual Conference of the International Speech Communication Association:* 597–600.

Wolpert, D. 1992. Stacked generalization (stacking). *Neural Networks*, 5: 241–259.

Wranik, T. and Scherer, K. R. 2010. *International handbook of anger*. New York: Springer-Verlag: 243–266.

Wu, Y., Wang, H., Zhang, B. and Du, K. L. 2012. Using radial basis function networks for function approximation and classification. *Applied Mathematics*, 2012: 1–34.

Xie, Y., Zhu, C., Lu, Y. and Zhu, Z. 2019. Towards optimization of boosting models for formation lithology identification. *Mathematical Problems in Engineering*, 2019.

Xie, Z., Hu, Q. and Yu, D. 2006. Improved feature selection algorithm based on svm and correlation. In: Proceedings of the *Advances in Neural Networks Third International Symposium on Neural Networks*: 1373–1380.

Yan, J., Wang, X., Gu, W. and Ma, L. 2013. Speech emotion recognition based on sparse representation. *Archives of Acoustics*, 38(4): 465–470.

Yan, J., Qi, Y. and Rao, Q. 2018. Detecting malware with an ensemble method based on deep neural network. *Security and Communication Networks*, 2018.

Yang, N., Muraleedharan,R., Kohl, J., Demirkol, I., Heinzelman,W. and Sturge-Apple, M. 2012. Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion. In: Proceedings of the *2012 IEEE Spoken Language Technology Workshop*: 455–460.

Yerigeri, V. V and Ragha, L. K. 2017. Marathi speech emotion detection: A retrospective analysis. In: Proceedings of the *8th International Conference on Computing, Communication and Networking Technologies*: 1–6.

Ying, S. and Xue-Ying, Z. 2018. Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition. *Future Generation Computer Systems*, 81: 291–296.

Yogesh, C. K., Ruzelita, N., Hariharan, M. and Abdul, A. 2017. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert*

Systems with Applications, 69: 149–158.

Yoon, S., Dey, S., Byun, S. and Jung, K.2019. Speech emotion recognition using multi-hop attention mechanism. In: Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*: 2822–2826.

Yousaf, M. S., Ahmad, I., Khurshid, A. and Ikram, M. 2020. Machine assisted classification of chicken, beef and mutton tissues using optical polarimetry and Bagging model. *Photodiagnosis and Photodynamic Therapy*, 31(3): 101-129.

Yu, B., Li, H. and Fang, C. 2012. Speech emotion recognition based on optimized support vector machine. *Journal of Software*, 7(12): 2726–2733.

Yu, M., Gong, J., Tang, J. and Kong, F. 2017. Delay announcements for call centers with hyperexponential patience modeling. *Industrial Management and Data Systems*, 117(6): 1037–1057.

Zaballos, M. T. P., Plasencia,D.P, González, M.L.Z, Miguel, A.R and Macías, A.R. 2016. Air traffic controllers long-term speech-in-noise training effects: A control group study. *Noise & Health*, 18(85): 376–381.

Zacarias, N., Pancardo, P., Hernández, J.A. and Garcia, M. 2021. Attention-inspired artificial neural networks for speech processing: A systematic review. *Symmetry*, 13, 214.

Zahorian, S. A., Singh, T. and Hu, H. 2007. Dimensionality reduction of speech features using nonlinear principal components analysis. In Proceedings of the *Annual Conference of the International Speech Communication Association*, 1(5): 281–284.

Zataraín, R., Barrón-Estrada, M.L., Hernández, F.G., and Rangel, H.R. 2018. Emotion recognition using a convolutional neural network, *Lecture Notes in Computer Science*: 208–219.

Zeng, X. Yuan, S., Li,Y. and Zou, Q. 2014. Decision tree classification model for popularity forecast of chinese colleges. *Journal of Applied Mathematics*, 2014.

Zhang, M., Chen, Y., Li, L. and Wang, D. 2018. Speaker recognition with cough, laugh and "Wei". In: Proceedings of the *9th Asia-Pacific Signal and Information Processing Association Annual*

Summit and Conference: 497–501.

Zhang, S., Zhang,S., Huang,T. and Gao,W. 2016. Multimodal deep convolutional neural network for audio-visual emotion recognition. In: Proceedings of the *2016 ACM on International Conference on Multimedia Retrieval*: 281–284.

Zhang, Z., Weninger, F., Wöllmer, M. and Schuller, B. 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. In: Proceedings of the *2011 IEEE Workshop on Automatic Speech Recognition and Understanding*: 523–528.

Zhang, Z., Coutinho, E., Deng, J. and Schuller, B. 2015. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 23(1): 115–126.

Zhao, H., Ye, N. and Wang, R. 2018. A survey on automatic emotion recognition using audio big data and deep learning architectures. In: Proceedings of the *4th IEEE International Conference on Big Data Security on Cloud*: 139–142.

Zhao, J., Mao, X. and Chen, L. 2018. Learning deep features to recognise speech emotion using merged deep CNN. *IET Signal Processing*, 12(6): 713–721.

Zhao, Y., Jin, X. and Hu, X. 2017. Recurrent convolutional neural network for speech processing. In: Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*: 5300–5304.

Zhou, X., Liu, K. Y. and Wong, S. T. C. 2004. Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*, 37(4): 249–259.

Zhu, J., Zhang, J., Chen, Q. and Tu, P.. 2017. Speaker recognition based on the improved double-threshold endpoint algorithm and multistage vector quantization. In: Proceedings of the *9th IEEE International Conference on Communication Software and Networks*: 1056–1061.

Zhu, L., Chen, L., Zhao, D., Zhou, J. and Zhang, W. 2017. Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN. *Sensors (Switzerland)*, 17(7): 1127–1138.

Zvarevashe, K. and Olugbara, O. O. 2018a. A framework for sentiment analysis with opinion mining of hotel reviews. In: Proceedings of the *International Conference on Information Communications Technology and Society*: 1–4.

Zvarevashe, K. and Olugbara, O. O. 2018b. Gender voice recognition using random forest recursive feature elimination with gradient boosting machines. In: Proceedings of the *International Conference on Advances in Big Data, Computing and Data Communication Systems*:1–6.

Zvarevashe, K. and Olugbara, O. O. 2020a. Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, 70(13): 1–24.

Zvarevashe, K. and Olugbara, O. O. 2020b. Recognition of cross-language acoustic emotional valence using stacked ensemble learning. *Algorithms*, 13(10): 1–21.

Zvarevashe, K. and Olugbara, O. O. 2020c. Recognition of speech emotion using custom 2d-convolution neural network deep learning algorithm. *Intelligent Data Analysis*, 24(5): 1065–1086.