Sudan University of Science & Technology

Faculty of Computer Science and Information Technology

A Speech-Based Emotion Recognition Framework

إطار للتعرف على العاطفة مبني على الكلام

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy in Computer Science

by

Inshirah AbdElrahman Mohamed Idris

Supervisor

Dr. Mohamed Elhafiz

May, 2019

Dedication

To the memory of my mother

all I have are memories and your picture in a frame
your memory is my keepsake, with which I'll never part

To the memory of my father

who gave me the thirst for new knowledge and the potential to seek it
who provided support and encouraged me to believe in myself
who dedicated all his life to us and provided us with endless love

To the memory of my nephew AbdElrahman

my heart knows that you are in a better place where there is no pain
you are at peace

We will be together again InshaAllah.
Until then, my love will always be with you

Acknowledgements

First and foremost, I thank Allah the Most Merciful for giving me the strength and persistence to complete this research. I would also like to offer my sincere thanks to several people who in one way or another contributed to the completion of this thesis by their assistance and guidance.

I would like to convey grateful thanks to my supervisor Dr. Mohammed Al Hafiz for his guidance and support throughout this study.

I would like to sincerely thank to my teacher Prof. Alaa Sheta for guiding me at the start of this journey and setting me on the right direction.

It is with immense gratitude that I acknowledge my teacher and advisor Dr. Akod for encouragement and continuous support. I could not have imagined having a better advisor and mentor for my Ph.D study. Your guidance helped me at all times.

I cannot find words to express my immeasurable gratitude to Dr. Md Sah Hj Salam for such immense knowledge, invaluable guidance, inspiring discussions, constant support, encouragement, incredible patience, and caring. It has been an honour to be your student. Many thanks for believing in me.

I owe my deepest gratitude to the Ph.D. program members in SUST for their efforts to provide and allow us this wonderful opportunity. Therefore, I am particularly indebted to Prof. Izzeldin Mohammed Osman (Academic Consultant for Vice Chancellor Sudan University) and all the professors who taught me in the first and second semesters.

I am indebted to my many friends in Sudan and Malaysia (especially Lab two members). I thank you all for the support, laughter and for making this journey an enjoyable one. Your

friendship makes my life a wonderful experience.

I share the credit of my work with my friend and sister Dr. Taqwa Ahmed. I cannot thank you enough for massive assistance in the writing of this thesis, the encouragement you always give and for the constant support you always provide. Many thanks to your small and big families for their incredible patience.

Last but not least, my deepest gratitude also extends to all my family members who have provided assistance, continued support and endless love through the writing of this thesis and my life in general. My brothers: Mohamed AbdElrahman; AbdElhamed AbdElrahman; Mohamed Alhasan; and Amar Awad. My sisters: Intesar AbdElrahman; Izdihar AbdElrahman; Alaa Alnor; and Marwa Awad. My nephews and nieces: Jalal; Amal; Mimona; AbdElrahman; Inshirah; Salwa; Mohamed; Mustafa; Mira; Nora and Maria.

Abstract

Human-computer interaction (HCI) has become one of the most challenging areas of research in the field of artificial intelligent (AI) at the present time. Speech emotion recognition (SER) introduces a new means of communication between humans and machines. Enabling a machine to understand human emotion renders it more capable of understanding the speech process. Despite the great progress and intensive research performed in this area, there is still a lack of naturalness in identifying emotions. There is a need to fill the gap between commercial interest and current performances. The key is to find significant speech emotion features that can map emotion correctly and efficiently. The previous works of SER extracted and selected different sets of acoustic features. However, the most significant features have not yet been found. These problem is addressed in this research by proposing a speech emotion recognition framework that provide an enhancement of features extraction technique and hybrid feature selection method respectively. The voice quality prosodic spectral-based feature extraction (VQPS) is implemented using prosodic and spectral features extraction technique in addition to new and traditional voice quality features extraction technique. At the same time, the balanced hybrid filter-based feature selection (BHFFS) consists of two layers: the balancing layer; and the hybrid filter-based layer. The proposed features extraction technique and selection method was successfully experimented through the use of EMO-DB dataset. The experimental results proved that using VQPS leads to performance improvement upon previous works. In addition, it demonstrates that the voice quality features are important in developing the SER system. In the same manner, BHFFS performance outperforms the previous work performance.

المستخلص

أصبح التفاعل بين الإنسان والحاسوب (HCI) واحد من اكثر مجالات البحوث تحديا في مجال الذكاء الاصطناعي (AI) في الوقت الحاضر. وبما إن تمكين الآلة من فهم العاطفة البشرية يجعلها أكثر قدرة على فهم عملية الكلام فإن أنظمة التعرف على العاطفة من الكلام (SER) تقدم وسيلة اتصال جديدة بين البشر والآلات. على الرغم من التقدم الكبير والأبحاث المكثفة التي أجريت في هذا المجال، لا يزال هناك نقص في عملية تحديد العواطف، و لازال هناك حاجة لملء الفجوة بين الرغبة التجارية والأداء الحالي. مفتاح الحل لهذه المشكلة يكمن في العثور على ميزات عاطفة الكلام الهامة التي يمكن أن تحدد العاطفة بشكل صحيح وكفء. استخرجت واختارت الأعمال السابقة في هذا المجال مجموعات مختلفة من ميزات العاطفة الصوتية. ومع ذلك، لم يتم العثور على أهم تلك الميزات حتى الآن. يتم تناول هذه المشكلة في هذا البحث من خلال اقتراح إطار يساعد على تحسين تقنية استخلاص الميزات وطريقة اختيار الميزات على التوالي. تم تطبيق تقنية استخلاص الميزات (VQPS) باستخدام ثلاثة انواع من الميزات: الميزات الكلامية والطيفية بالإضافة إلى ميزات جودة الصوت الجديدة والتقليدية. في نفس الوقت، تتألف طريقة اختيار الميزة (BHFFS) من طبقتين: طبقة التوازن وطبقة التصفية المختلطة. تم بنجاح تجربة تقنية استخراج الميزات المقترح وطريقة الاختيار من خلال استخدام قاعدة البيانات الالمانية (EMO-DB). أثبتت النتائج التجريبية أن استخدام (VQPS) يؤدي إلى تحسين الأداء اكثر من الأعمال السابقة. بالإضافة إلى ذلك ، فإنه يوضح أن ميزات جودة الصوت مهمة في تطوير أنظمة التعرف على العاطفة من الكلام (SER). بنفس الطريقة، تفوق أداء (BHFFS) على أداء طرق اختيار الميزة المستخدمة في الاعمال السابق.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

HCI	Human-computer interaction
SER	Speech Emotion Recognition
MFCC	Mel Frequency Cepstral Coefficients
LPC	Linear Prediction Coefficients
ASR	Automatic Speech Recognition
HNR	Harmonics to Noise Ratio
kNN	K-Nearest Neighbor
SVM	Support Vector Machine
ANN	Artificial Neural Network
RF	Random Frost
HMM	hidden Markova Models
GMM	Gaussian Mixtures Model
DBN	Deep Belief Network
ZCR	Zero Crossing Rate
LPCC	Linear Prediction Cepstral Coefficients
PLP	Perceptual Linear Predictive Coefficients
LFCC	Linear Frequency Cepstral Coefficient

LFPC	Log Frequency Power Coefficients
Rasta-PLP	Relative Spectral Transform Perceptual Linear Prediction
RF	Relief Algorithm
IGR	Information Gain Ratio
GR	Gain Ratio
RS	Rough Set Theory
FDR	Fisher Discriminant Ratio
OLDA	orthogonal-Linear Discriminant Analysis
CFS	correlation-based Feature Selection
FCBF	Fast Correlation-based Filter
mRMR	minimum Redundancy Maximum Relevance
FS	Forward Feature Selection
SFS	Sequential Forward Selection
SFFS	Sequential Floating Forward Selection
LFFS	linear floating forward selection
ERFTrees	Ensemble Random Forest to Trees
SUSAS	Speech under Simulated and Actual Stress
DES	Danish Emotional Speech
LLD	Low-Level Descriptor
TEO	Teager Energy Operator-based
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
IITKGP-SEHSC	Simulated Emotion Hindi Speech Corpus

RDA	Regularized Discriminant Analysis
BHUDES	Beihang University Database of Emotional Speech
PDREC	Persian Drama Radio Emotional Speech Corpus
MEDC	Mel-Energy Spectrum Dynamic Coefficients
MESDNEI	Multilingual Emotional Speech Database of North East India
MLS	Mean of the Log-Spectrum
IEMOCAP	Interactive Emotional Dyadic Motion Capture
FBANK	Filter Bank
NHR	Noise to Harmonic Ratio
GEMEP	French Geneva Multimodal Emotional Portrayal
HammI	Hammarberg Index
eNTERFACE	Audio-Visual Emotion
FFT	Fast Fourier Transform
CASIA	Automation Chinese Academy of Sciences
EESDB	Chinese Elderly Emotion Database
VDERM	Voice Driven Emotion Recognizer Mobile Phone
LSP	Line Spectral Pairs
ESD	Emotional Speech Database
CLDC	Chinese Linguistic Data Consortium
SRC	Sparse Representation Classifier
MIC	Maximal Information Coefficient
SWLDA	Stepwise Linear Discriminant Analysis
BBO	biogeography optimization

LSBOUND	Least Squared Bound
PCC	Pearson correlation coefficient
MIM	Mutual Information maximization Criterion
EMOVO	Italian Dataset
SAFE	English Situation Analysis in a Fictional and Emotional
MLSTSVM	Multi Least Squares Twin Support Vector Machine
AMDF	Average Magnitude Difference Function
SHS	Subharmonic Summation
DCT	Discrete Cosine Transform
PSO	Particle Swarm Optimization
LIBSVM	Library for Support Vector Machines
SMO	Sequential Minimal Optimization
QP	Quadratic Programming
KKT	Karush Kuhn Tucker

List of Publications

1. Idris, I. and Salam, M.S.H., 2014, December. Emotion detection with hybrid voice quality and prosodic features using Neural Network. In 2014 4th World Congress on Information and Communication Technologies (WICT 2014) (pp. 205-210). IEE.

2. Idris, I. and Salam, M.S.H., 2015. Voice Quality Features for Speech Emotion Recognition. Journal of Information Assurance and Security, 10(4).

3. Idris, I., Salam, M.S.H. and Sunar, M.S., 2015. Speech emotion classification using SVM and MLP on prosodic and voice quality features. Jurnal Teknologi, 78(2-2).

CHAPTER 1

Introduction

1.1 Introduction

The field of Human-Computer Interaction (HCI) is a challenging one considering the major differences between the manner in which humans and machines understand speech. The difference between the human and the machine is that the human not only communicates using speech, but also by using text; facial expressions (Pao et al., 2005a; Koolagudi et al., 2018); heart rate; skin temperature; body gestures (such as hand waving and eye movements)(Swain et al., 2018) as well as emotions. The emotion of speech plays an important role in explaining the words uttered by the speaker by focusing on how the words were expressed rather than what was said. Sometimes, the same sentences expressed through different emotions have different meanings. This field of research in HCI is known as Speech Emotion Recognition (SER).

SER is an interesting subject and has attracted many researchers at the present time because it introduces a new means of communication between humans and machines. This is important in many applications, including: education (Schuller et al., 2004; Tickle et al., 2013); security systems (Saste and Jagdale, 2017); healthcare (San-Segundo et al., 2009;

Lopez-de Ipiña et al., 2015); call centers and mobile communications (Vidrascu and Devillers, 2005; Tarng et al., 2010); and robotics (Alonso-Martín et al., 2013).

SER, in general, seeks to resolve pattern recognition problems. It starts by taking speech samples and extracting a suitable feature set, prior to classifying the emotion. The speech signal has many sets of features that can be extracted. These speech features are an important factor in SER systems because they affect the classification performance (Lanjewar and Chaudhari, 2013b). Three types of acoustic emotional features are used to determine the emotional state of a speaker, namely, voice quality, as well as prosodic and spectral features(Luengo et al., 2010; Pérez-Espinosa et al., 2012; Henríquez et al., 2014; Chen et al., 2014; Valstar et al., 2016; Sudhkar and Anil, 2016). Prosodic and spectral features are used in the recognition of emotion in speech because both of these features contain emotional information (Joshi and Kaur, 2013).

Prosodic features are the most frequently-used features in SER because they provide a reliable indication of emotions state (El Ayadi et al., 2011). It is also referred to as the fundamental indication of the speakers emotion (Ingale and Chaudhari, 2012). Prosodic features such as pitch, energy, and speaking rate have been examined widely in previous works, and they are the most common type used in SER research (Wu et al., 2011; Sun et al., 2009).

Spectral features are extracted from the vocal tract system. They describe the speech signal in the frequency domain by using Fourier Transform to convert the time signal into the frequency domain. Examples of such features are mel frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) respectively. These features have been found to be successful in various speech tasks such as speaker recognition and automatic speech recognition (ASR). These features are also among the most common feature types in the SER specially MFCC feature.

Further, voice quality features are used as a complement to spectral and prosodic features, and are rarely used alone. Voice quality features are also considered to be an important feature (Hassan and Damper, 2012; Lugger and Yang, 2008) since many speakers express their emotional state by altering their voice quality (Lugger and Yang, 2008). However, its use is a significant challenge because there are many different measures of voice quality (Lugger and Yang, 2008; Batliner et al., 2011; Schuller et al., 2011), and it is mostly used for evaluating constant vowels only. Most SER researchers use three types of voice quality features, namely: pitch irregularity (jitter); amplitude irregularity (shimmer); and the harmonics-to-noise ratio (HNR).

Implementing all these feature extraction techniques results in large numbers of extracted emotional features from speech. However, not all of these extracted features are useful and important for SER. Therefore, the implementation of an accurate SER system depends on the selection of these sets of features. In order to improve SER, previous researchers have tended to be concerned with finding the best feature set. However, until now, the optimum feature set has not been found.

1.2 Problem Statement and Significance

In spite of efforts that were carried out to improve the SER system, there is still a gap among the SER current performance (recognition accuracy) and commercial needs (Schuller et al., 2007). The key element in improving SER recognition accuracy is selecting the appropriate reduced emotional feature set. However, up until now, there has been no proof of what is the most appropriate speech emotion feature set that can assist in classifying different emotions (Sezgin et al., 2012). The difficulty involved in defining the features that affect the SER makes it a challenging task (Sezgin et al., 2012).

Features extraction and selection techniques in previous works conducted several limitations. First, there are no standard or a specific combination of features to be developed and used in SER. Every researcher extracted different acoustic feature subsets (single or combination) based on their own experience and knowledge which has led to a low recognition accuracy. This because most of the researchers concentrate on using some features extraction techniques like prosodic and spectral features and neglecting some other techniques such as voice quality features extraction technique. Even more, some types of voice quality feature are extensively used while other feature are ignored.

The second problem arise from using the different combination of features extraction techniques that result in large extracted feature set that can number up to many thousands. For Example: (Esmaileyan and Marvi, 2014; Manolov et al., 2017). This number of features led in many cases to cures of dimensionality. This problem could be solved by using features selecting techniques. However, most of this technique implements a single features selection method instead of implementing a hybrid selection method in spite of its advantage. Also, the focusing is on wrapper method implementation connects features to a specific classification algorithm. Which mean that the selected subset features that perform well with specific classifier may not perform as well on other classifiers. Filter methods, on the other hand, do not depend on the classification algorithm which makes it a much proper choice. Another issue in features selection that some of the selection algorithms do not perform well in unbalanced datasets which is the case of most emotional datasets.

1.3 Research Methodology

This section introduces the research methodology adopted in this research. The details are presented in Chapter 3. Two research methodology have been conducted the constrictive and action research methodology. The constrictive research has been used to create the theoretical model of SER features that represent the relation between features and SER performance. While action research has been used to create the research framework. The proposed SER framework was developed in five phases that address the research aims and objectives.

Phase 1 investigate the performance of the existing single features extraction techniques to discover its importance in representing the emotions. In Phase 2, investigate the existing hybrid features extraction technique to identify the combination that helps in the improvement of recognition accuracy. Phase 3, introduces an enhancement to the existing features combination depending on new and brute force traditional voice quality features. Phase 4, introduce an enhanced selection method that combines data balancing to hybrid filter selection. Finally, Phase 5 compares the result achieved by the proposed framework with other researcher results.

1.4 Research Questions

The main research question is: How to extract and select a compact set of features that improve SER performance?

Based on this question, the sup questions are:

1. What is the impact of features in SER recognition accuracy?
2. What consideration should be given in designing features extraction techniques and selection methods?
3. What is the impact of the different type of features in emotion recognition?

4. What is the best feature extraction technique combination that is appropriate for emotion recognition?
5. How to enhance the feature extraction technique that obtains a better representation of emotion?
6. How to design a selection method that selects a significant set of features that improve recognition accuracy?

1.5 Research Hypothesis

The research hypothesis is:

The recognition accuracy of a speech emotion recognition system could be improved by enhancing the features extraction technique and selection method to identify the most significant features set that can be recognize emotion effectively.

1.6 Research Objectives

This study aims to identify minimum significant features set for SER that represent the emotion of the speech signal efficiently with high recognition accuracy.

The objectives of the research are as follows:

1. Investigate SER existing features extraction techniques and the capability of the extracted features.
2. Identify the best feature extraction technique combination which increases the recognition accuracy.

3. Enhance the feature extraction technique based on prosodic, spectral and a combination of new and traditional voice quality features that provide a better representation of emotion and improve the recognition accuracy.

4. Reduce the features dimensionality using enhanced hybrid filter-based feature selection method to remove irrelevant and redundancy features.

5. Identify the most significant features set.

1.7 Research Scope

The study is limited to the following:

1. Speaker-dependent system.

2. Acoustic Static features.

3. Berlin emotional speech dataset(EMO-DB). This is used to evaluate the recognition model and involves 535 records and seven emotions.

1.8 Research Structure

This thesis consists of six chapters. This chapter provides an introduction to this research. Chapter 2 provides a brief survey of a selection of previous work performed in the area of SER. It also includes the architecture of the SER system. Chapter 3 presents an overview of the methodology which has been used by the researcher to address SER problem. Chapter 4 presents the development and evaluation of proposed SER framework. Chapter 5 concludes the thesis by presenting the highlighting features of the work. It also discusses future directions for extending the research work.

CHAPTER 2

Literature Review

2.1 Introduction

This chapter starts with a general view of the speech emotion recognition (SER) system, as illustrated in Figure 2.1 below. It then presents a detailed overview of SER features extraction and selection. In addition, this chapter discusses the issues and problems facing SER features. A comprehensive review of SER features extraction techniques as well as selection methods and their limitations are also described. Finally, the possible techniques for SER are presented.

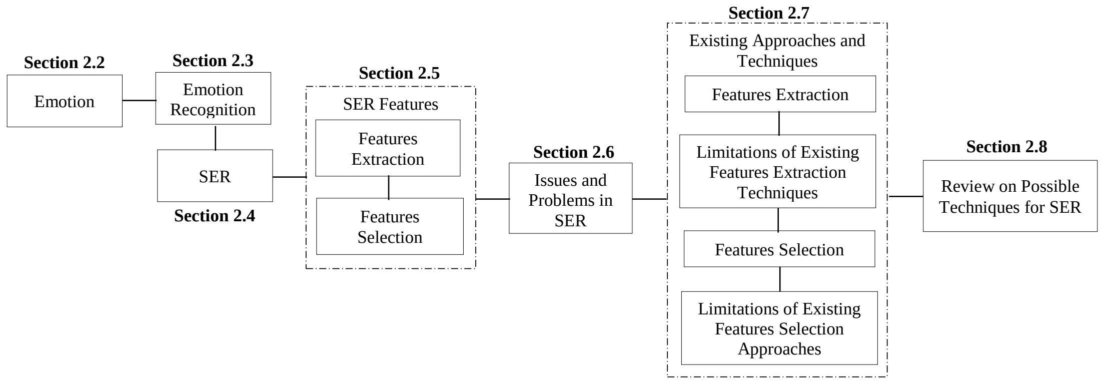

Section 2.2
Emotion
Section 2.3
Emotion Recognition
SER
Section 2.4
Section 2.5
SER Features
Features Extraction
Features Selection
Section 2.6
Issues and Problems in SER
Section 2.7
Existing Approaches and Techniques
Features Extraction
Limitations of Existing Features Extraction Techniques
Features Selection
Limitations of Existing Features Selection Approaches
Section 2.8
Review on Possible Techniques for SER

Figure 2.1: Literature Review Structure

2.2 Emotion

Emotions play an important role in human-human communication as they provide important information regarding the speakers and their responses to the outside world. As a result of this, identifying emotions that are to be recognized by machines became a critical issue in developing an SER system. However, since there are about 300 different emotional states in real life, identifying and recognizing all these emotions make the recognition process complicated (Utane and Nalbalwar, 2013; El Ayadi et al., 2011; Ingale and Chaudhari, 2012). In the last decades, all the research on emotion recognition has generally followed one or two emotion theories, namely, the discrete and continuous emotion theories (Pérez-Espinosa et al., 2012).

2.2.1 Discrete Emotion

This is also called the palette theory as it assumes that any emotion can be represented by a combination of a number of basic emotions (Lugger and Yang, 2008; Pérez-Espinosa et al., 2012). This theory is similar to the theory of the seven colours of light that can be combined to generate any colour (Ingale and Chaudhari, 2012). A variety of basic emotion numbers were adopted by previous researchers. The most commonly adopted are the set of six basic emotions: anger, disgust, fear, joy, sadness, and surprise (Razak et al., 2005; El Ayadi et al., 2011; Ingale and Chaudhari, 2012; Ram and Ponnusamy, 2014; Ooi et al., 2014) and a further set of five basic emotions: anger, disgust, fear, happiness, and sadness (Murray and Arnott, 2008). The discrete theory is more suitable in recognizing a pre-defined set of emotions e.g. for the acted dataset. Most of the existing emotional speech datasets are based on the six basic emotions (Mustafa et al., 2018). However, this approach does not cover the entirety of the human emotions (Pérez-Espinosa et al., 2012).

2.2.2 Continuous Emotion

This theory assumes that any emotion can be represented by a combination of numbers of low dimensional space (Lugger and Yang, 2008; Pérez-Espinosa et al., 2012) which usually comprise two to three dimensions (Lugger and Yang, 2008; Hassan and Damper, 2012). The most widely-used is the one representing two basic dimensions: arousal and valence (Pérez-Espinosa et al., 2012). Arousal (activation) refers to the excitement and activeness of the speaker while expressing a certain emotion (Pérez-Espinosa et al., 2012; Hassan and Damper, 2012; Mustafa et al., 2018). Therefore, angry and happy should have high arousal while sad, bored and relaxed have low arousal. On the other hand, valence represents both positive and negative emotions of the speaker (Pérez-Espinosa et al., 2012; Hassan and Damper, 2012; Mustafa et al., 2018). This dimension is used to separate angry emotions from happy ones. The two-dimensional approach is widely applied in cross-corpus emotion recognition (Mustafa et al., 2018). From the viewpoints of psychologists, dimensional space can be mapped to the three dimensions of arousal, valence and dominance (power)(Rao and Koolagudi, 2011; Koolagudi and Rao, 2012). The dimension of dominance describes the degree of control with which the individual intends to take on the situation (Pérez-Espinosa et al., 2012).

2.3 Emotion Recognition

Researchers name a different resource that can be used for emoting recognition. For instance, emotion recognition can evolve from speech signals, text and facial expressions (Pao et al., 2005a; Koolagudi et al., 2018). Gestures, heart rate and skin temperature are also conceded as being a source for emotion recognition (Swain et al., 2018). From literature studies, it can be noted that speech, text and facial expression represent the most-used emotion source in SER. Among the entire three sources, speech signals were considered the more efficient source for emotion recognition compared to text and image. Text emotion recognition is a challenge due to the lack of clarity at the level of syntactic and semantic. Facial expressions alone cannot identify all the emotions required for emotion recognition (Koolagudi et al., 2018). In addition, detecting a particular emotion from facial expressions requires a high quality camera for capturing face images; thereby rendering it difficult and expensive (Tomar et al., 2014).

Speech is considered to be the most natural, primary, and fast method in human communication (El Ayadi et al., 2011; Ingale and Chaudhari, 2012; Basharirad and Moradhaseli, 2017; Kurpukdee et al., 2017; Manolov et al., 2017; Liu et al., 2018). This has inspired many researchers to use it in human-machine interaction implementation (El Ayadi et al., 2011; Ingale and Chaudhari, 2012; Basharirad and Moradhaseli, 2017). As a result, an extreme study has been conducted to analyze emotion from speech (Basharirad and Moradhaseli, 2017).

2.4 Speech Emotion Recognition

Speech emotion recognition (SER) is a technology that aims to recognize emotions from speech signals (Ververidis and Kotropoulos, 2006; Altun and Polat, 2009; El Ayadi et al., 2011; Shen et al., 2011; Ingale and Chaudhari, 2012; Pan et al., 2012; **?**; Alonso et al., 2015; Samantaray et al., 2015; Wang et al., 2015; Yogesh et al., 2017; Kurpukdee et al., 2017; Koolagudi et al., 2018). Most pattern recognition systems generally contain three main components: feature extraction, feature selection and classification (You et al., 2006; Pao et al., 2006; Altun and Polat, 2009; Gharavian et al., 2013; Sun and Wen, 2015; Liu et al., 2018) as illustrated in Figure 2.2 below.

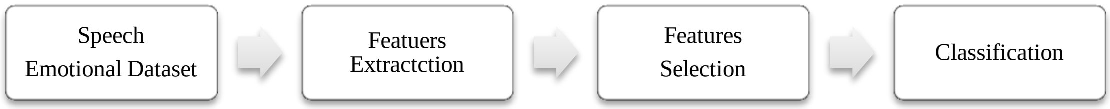


Figure 2.2: SER System General Components

The first component is features extraction, which aims to transfer the speech signal from emotional dataset to a sequence of features vector that is considered an important factor in emotion recognition performance (Ingale and Chaudhari, 2012; Waghmare et al., 2014; Muthusamy et al., 2015). Emotions represent a large number of features and any variation in emotions will result in a change in these features (Shen et al., 2011; Ingale and Chaudhari, 2012). The second component is features selection. This aims to select the significant features from the extracted features in order to improve the classifier performance in terms of accuracy and time by eliminating irrelevant and redundant features. Finally, the last component is the classification process. The goal of classifier algorithms is to classify the emotion (Kuchibhotla et al., 2014; Zhao et al., 2014) derived from the extracted speech features (Zheng et al., 2014). Popular classifiers for SER include: k-nearest neighbor (kNN); support vector machine (SVM); artificial neural network (ANN); random forst (RF); hidden Markova models (HMM); Gaussian mixtures model (GMM); and deep belief network (DBN).

Different researchers have been working on different components separately and independently (Hassan and Damper, 2012; Mustafa et al., 2018). According to Hendy and Farag (Hendy and Farag, 2013), some researchers worked on extracting and selecting suitable features from the speech signal that gave the best representation for emotions as in (Borchert and Dusterhoft, 2005; Saste and Jagdale, 2017; Renjith and Manju, 2017; Manolov et al., 2017). Others, meanwhile, worked on the classification algorithms that can recognize and identify emotions, for example (Razak et al., 2005; Lanjewar et al., 2015; Trabelsi et al., 2016; Wang et al., 2017; Kurpukdee et al., 2017; Anoop et al., 2018; Koolagudi et al., 2018). In addition, some researchers have devised a contraption for studying or creating a dataset which contains emotional speech that is used by the classification algorithm to recognize the emotion as in (Burkhardt et al., 2005).

It has been noted that the majority of modern research concentrates on the concept of feature extraction and selection to identify emotions and improve SER performance (Swain et al., 2018). Determining the most efficient features may be more critical to SER performance than the classifier itself (Hendy and Farag, 2013). Therefore our research focuses only on emotion features. This chapter reviews the relevant literature related only to emotion features.

2.5 Speech Emotion Recognition Features

In any classification system, the first processes are usually the extraction of features from the raw data. These features are then reduced in a feature selection process before being presented to a classification algorithm. Feature extraction is the imperative process for generating and producing essential emotional features (Schuller et al., 2011; Ingale and Chaudhari, 2012; Reddy and Vijayarajan, 2017). Furthermore, the feature selection process has been used commonly in literature to select significant subset features that represent emotions accurately.

In SER, two groups of features were used, namely, the acoustic and the linguistic features (Schuller et al., 2011; Batliner et al., 2011; Kamaruddin et al., 2012; Ramakrishnan and El Emary, 2013). The linguistic techniques focus on explicit linguistic messages (dialogue related features) while the acoustic techniques focus on implicit messages (Gharavian et al., 2013; Henríquez et al., 2014). Performance compression between acoustic and linguistic features shows that their impacts on SER depend on the type of dataset used. For instance, linguistic features perform well with a spontaneous dataset and are worthless with an acted dataset (Schuller et al., 2011; Batliner et al., 2011). This is in contrast to acoustic features that work well with the acted dataset with writing scripts. Furthermore, its reported form works to adopt linguistic features. SER has not yet reached the level required to work well with spontaneous datasets (Anagnostopoulos et al., 2015). As a result of this, and since the dataset used is the acted dataset, only the acoustic features extraction will be discussed in this research.

2.5.1 Features Extraction

Feature extraction aims to represent speech signals by a sequence of speech features (Lanjewar and Chaudhari, 2013a). Extraction of these speech features is an important factor in the SER system (El Ayadi et al., 2011; Lanjewar and Chaudhari, 2013a). Different speech features have been analyzed; however, until now there has been no agreement on a fixed set of features (Shirani and Nilchi, 2016). The most commonly-used features in SER are acoustic features (Vogt and André, 2005; Neiberg et al., 2006; Fu et al., 2008; San-Segundo et al., 2009; Yang and Lugger, 2010; Lee et al., 2011; Pérez-Espinosa et al., 2012; Hassan and Damper, 2012; Sun and Wen, 2015; Jassim et al., 2017; Ding et al., 2018). They are also it conceded as being the most effective features in emotion recognition derived from speech (Klaylat et al., 2018). The acoustic features can be categorized as, namely, voice quality, prosodic, and spectral. This categorized follows (Luengo et al., 2010; Pérez-Espinosa et al., 2012; Henríquez et al., 2014; Chen et al., 2014; Valstar et al., 2016; Sudhkar and Anil, 2016).

Voice Quality Features

Voice quality features (VQ) are known as such because they depend on the human voice (Borchert and Dusterhoft, 2005). It is considered as one of the most important features in emotion representation (Hassan and Damper, 2012; Lugger and Yang, 2008). Often it is called the 4th dimension of prosody because any change in voice quality results in a different emotional state (Lugger and Yang, 2008). However, its usability is a significant challenge because it has many different measures (Lugger and Yang, 2008; Batliner et al., 2011; Schuller et al., 2011). The most popular voice quality features are jitter, shimmer, and the harmonics-to-noise ratio (HNR) (Tahon et al., 2012).

Prosodic Features

Prosodic Features are known as the primary indicator of the speakers emotional state (You et al., 2006; Chen et al., 2006; Ingale and Chaudhari, 2012; Lanjewar and Chaudhari, 2013a). According to many researchers (Kuchibhotla et al., 2014; Zhao et al., 2014; Ahmad, 2016) they are conceded as being among the most widely-used features in SER research since they provide a reliable indication of an emotion (El Ayadi et al., 2011). In addition, they are easier to use (Borchert and Dusterhoft, 2005). The prosodic features can be classified into pitch, intensity, energy, loudness, zero crossing rate (ZCR), formant and duration respectively.

Spectral Features

These features are successfully used in speech and speaker recognition systems (Milton et al., 2013). Furthermore, it has also been able to successfully classify emotion due to its ability to estimate the shape of the vocal tract which is unique for different emotions (Bitouk et al., 2010; Kuchibhotla et al., 2014). As a result, spectral features have been widely used in speech emotion recognition (Zhou et al., 2009). Spectral features extracted from spectral content of the speech signal (Kuchibhotla et al., 2014) indicate the use of different extraction techniques (Koolagudi et al., 2018). The most popular techniques include: mel frequency cepstral coefficients (MFCC); linear prediction coefficients (LPC); linear prediction cepstral coefficients (LPCC); perceptual linear predictive coefficients (PLP); linear frequency cepstral coefficient (LFCC); log frequency power coefficients (LFPC); and Relative Spectral Transform Perceptual Linear Prediction (Rasta-PLP).

2.5.2 Features Selection

Selecting the significant features is an important stage in building real-time systems (Hendy and Farag, 2013; Reddy and Vijayarajan, 2017) because it simplifies the hardware implementation of the classifier in terms of processing speed and memory requirements (Hendy and Farag, 2013). In addition, it helps in saving the time needed for training and classification, which often takes quite a while when using the whole features set. Moreover, it assists in avoiding over-fitting which is caused by high dimensionality features. Features selection approaches are divided into four categories, namely: filter, wrapper, embedded and hybrid approach. Figure 2.3 below illustrates the taxonomy for art selection approaches that are used in SER.

Filter Approach

In this approach, the features are evaluated independently from the classifier based on their relation to class label. The algorithms in this approach can be categorized into two groups,

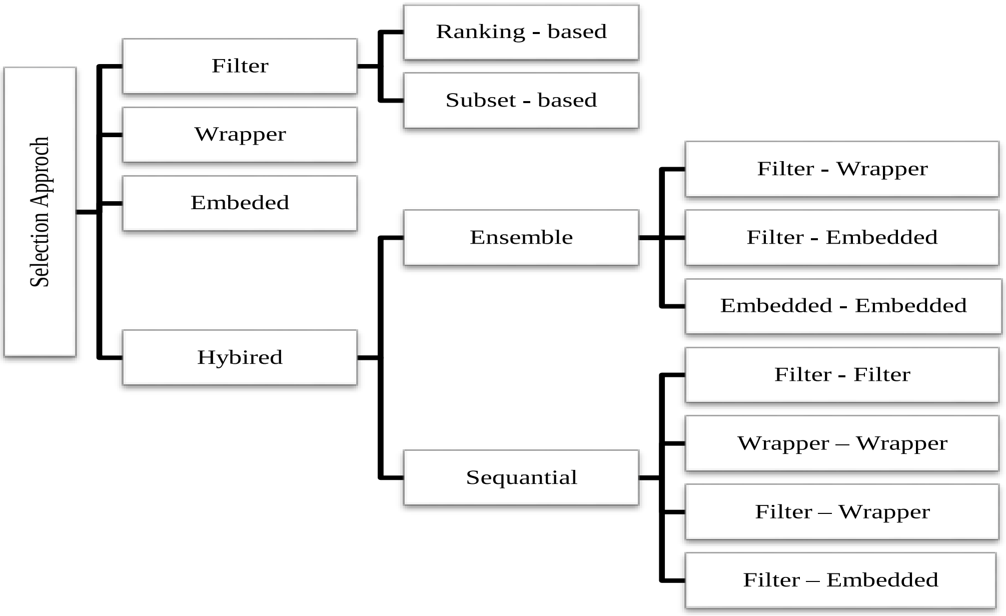


Figure 2.3: Features Selection Taxonomy

specifically:

1. **Filter Ranking-Based:** ranks the features individually based on their relevance to the class labels. Some of the algorithms that are used in SER include: relief algorithm (RF); information gain ratio (IGR); gain ratio (GR); rough set theory (RS); fisher discriminant ratio (FDR); and orthogonal-linear discriminant analysis (OLDA).

2. **Filter Subset-Based:** searches through the possible number of combinations of the feature subsets guided by a certain evaluation measure that captures the goodness of each subset. An optimal subset is selected when the search stops. Some of the algorithms that are used in SER include: correlation-based feature selection (CFS); fast correlation-based filter, (FCBF); and minimum redundancy maximum relevance (mRMR).

Ranking-based filter helps in eliminating irrelevant features. These represent the features that did not affect class identification in any way. Conversely, subset-based filter helps in eliminating redundant features which represent the features that did not add anything new to the class identification.

Wrapper Approach

The wrapper approach selects the feature subsets using search techniques then evaluates these subsets using a classification algorithm with a selected criterion. The criterion used in selection could include the classification error (Schuller et al., 2007; Sun et al., 2009) or accuracy (Lin and Wei, 2005; Schuller and Rigoll, 2006; Ververidis and Kotropoulos, 2008; Bitouk et al., 2010). The selection of features stopped when adding or removing new ones failed to increase or decrease the chosen criterion. A stopping rule also could be set by selecting the preset features number when this number of features reached the selection stops and no more features would be added or removed. Examples of wrapper features selection algorithms include: forward feature selection (FS); sequential forward selection algorithm (SFS); sequential floating forward selection algorithm (SFFS); linear floating forward selection (LFFS); and ensemble random forest to trees (ERFTrees).

Embedded Approach

Embedded approaches propose a combination of the advantages of both previous approaches. They often provide a good trade-off between performance and computational cost. The classification algorithm implements feature selection and classification simultaneously. For features selection embedded approach, it is necessary to use internal information of the classification algorithm to perform feature selection (e.g. use of the weight vector in SVM).

Hybrid Approach

The hybrid features selection approach combines two of the previous feature selection approaches. This is generally done to combine the desired characteristics of each. This approach is categorized as follows:

1. **Ensemble Features Selection:** can be described as a group of selection algorithms which implement together to obtain the final selected features in a combine. It can be categorized further to filter-wrapper, filter-embedded or embedded-embedded according to the type of combined approaches used in SER implementation.
2. **Sequential Features Selection:** can be described as a group of selection algorithms that were implemented sequentially in two or more steps to obtain the final selected features. It can be categorized further according to the type of selection approach used in SER, namely: filter-filter; wrapper-wrapper; filter-wrapper; and filter-embedded.

2.6 Research Issues and Problems in Speech Emotion Recognition

The research issues in SER related to features extraction can be categorized into four main categories, specifically: speech processing diversity; analysis unit variety; disagreement on the efficient features; and curse of dimensionality. In each category, the research problems are identified and highlighted in the following subsections.

2.6.1 Speech Processing Diversity

Despite the importance of the processing phase before and after extracting features, there is a lack of standardization of processing methods in SER research. Moreover, there is no agreement between previous works in the use of these methods. This has resulted in the adoption of different methods by various researchers. There are two major methods of speech processing that can be named according to when it performs, before or after features extraction. This is: the pre-processing and the post-processing.

Pre-processing

This refers to processes that are required to be performed on the speech signal before features have been extracted (El Ayadi et al., 2011). Its importance comes from its ability to enhance the efficiency of the extracted features (Kuchibhotla et al., 2014) thereby improving classification performance of the SER system (Hendy and Farag, 2013). These processes include: noise reduction; pre-emphasis; framing; windowing; and silence removing.

1. **Noise Reduction:** noise can be defined as undesired signals that are normally presented by recording hardware or recording environment (Kuchibhotla et al., 2014). This signal has a great impact on the quality of speech signals and, as a result, impact on SER system performance. An increase in the level of noise in speech signals causes a decrease in the accuracy of SER (Schuller et al., 2006). Consequently, a noise reduction processor is required before extraction of features can be performed (Basu et al., 2017; Reddy and Vijayarajan, 2017) since it does not change the original signal. This is considered a difficult task in SER because it can differ depending on data (Basu et al., 2017). For the dataset that is recorded in a noise condition such

as the speech under simulated and actual stress dataset (SUSAS), noise reduction is necessary (Basu et al., 2017; Reddy and Vijayarajan, 2017). However, for standard datasets that were recorded in an isolated environment with high-quality equipment such as EMO-DB dataset and Danish emotional speech dataset (DES), there is no need for this step. In fact, to study the impact of noise in SER for this dataset noise was added as described in (Schuller et al., 2006). Noise reduction can be carried out by applying filtering techniques, for instance: high pass filter (Kuchibhotla et al., 2014); or wiener filter; and spectral subtraction (Basu et al., 2017).

2. **Pre-emphasis:** a pre-emphasis filter is used to process a speech signal before the features extraction process (El Ayadi et al., 2011; Chen et al., 2012a). The filter increments the signal energy level to provide more information (Basu et al., 2017). In addition, it balances the impact of the transmission of speech signal through air (Chen et al., 2012a). Transmission of speech signals having low amplitude through the air result in more sensitivity to noise effect.

3. **Framing and Windowing:** before features can be extracted, each of the speech signals is divided into a small unit called frame (Lanjewar and Chaudhari, 2013a; Kuchibhotla et al., 2014). The feature vector is then created form each frame. This process is necessary to solve the variation of human speech length problem (Basu et al., 2017) and render it stationary (Swain et al., 2018). A process called windowing is performed after the framing process. This maintains the speech sample continuity (El Ayadi et al., 2011; Kuchibhotla et al., 2014; Basu et al., 2017) while most of the researchers agree to use a Hamming window (Kockmann et al., 2011; Koolagudi and Rao, 2012; Hendy and Farag, 2013; Kuchibhotla et al., 2014; Verma et al., 2016; Jalili et al., 2018). The controversy is related to frame size. According to Anagnostopoulos et al. (Anagnostopoulos et al., 2015) the frame size is typically 25-50 msec. A

review of previous works confirms this claim. For instance: (Kockmann et al., 2011; Koolagudi and Rao, 2012; Verma et al., 2016) chose 20 msec frame; (Waghmare et al., 2014) chose 25 msec frame; and (Hendy and Farag, 2013) chose 30 msec frame. However, there are also different sizes such as 55 msec size frame (Henríquez et al., 2014) and 60 msec frame size (Yang et al., 2012) respectively.

4. **Silence Removing:** the process of removing the non-speech signal. Usually, in speech signal processing, the silence region is removed. However, in SER there are conflicting opinions about it. The first opinion considers that the silence region should be kept because it carries important information about the emotion (El Ayadi et al., 2011). The second opinion does not believe that the silence and unvoiced frames contain any useful information (Milton et al., 2013) and hence should be removed (Hendy and Farag, 2013; Milton et al., 2013). This was detailed as in (Shami and Verhelst, 2007; Kockmann et al., 2011; Yang et al., 2012; Milton et al., 2013; Hendy and Farag, 2013; Henríquez et al., 2014; Alonso et al., 2015).

Post-processing

Post-processing refers to the process that is required to be performed after features extraction and before the features feed to the classifier (El Ayadi et al., 2011). It usually includes missing data handling and features normalization.

1. **Missing Data Handling:** some researchers have suggested that if the missing data in certain features is larger than 2% of the total data the features should be discarded (Ververidis and Kotropoulos, 2005a; Kotti and Paternò, 2012; Hendy and Farag, 2013). However, most of them do not describe what happens if the missing data is less than that. Ververidis et al. (Ververidis and Kotropoulos, 2005a) proposed that

if the missing data is small then 1% can be replaced with the sample mean.

2. **Features Normalization:** denotes the approach that is used to scale the range of the features to ensure that the classification is not made based on one feature that varies significantly more than the other features. However, in SER it not always defined. Moreover, there are different approaches that can be used to normalize features such as z-score (Yang et al., 2012; Alonso et al., 2015) and min-max (Henríquez et al., 2014) normalization. In some papers, features were normalized before selection stage as in (Ververidis and Kotropoulos, 2005a; Kamaruddin et al., 2012; Kotti and Paternò, 2012); while in many others, this step was not mentioned as in (Razak et al., 2005; Khanchandani and Hussain, 2009; Hendy and Farag, 2013).

2.6.2 Analysis Units Variety

Determining the proper analysis unit for speech signals so as to prepare it for the feature extraction stage is an important issue in SER research (El Ayadi et al., 2011; Ingale and Chaudhari, 2012; Joshi, 2013; Lanjewar and Chaudhari, 2013a). However, it has not received much attention in the SER system. Few efforts have been made to compare various types of analysis units (Vogt et al., 2008). The analysis unit of the speech signal can be categorized to frame (supra-segmental) and utterance (segmental) (El Ayadi et al., 2011; Ingale and Chaudhari, 2012; Anagnostopoulos et al., 2015; Ingale and Chaudhari, 2012). There is no agreement as to which is better; the supra-segmental or segmental features (El Ayadi et al., 2011).

Supra-segmental Features

Also called the local dynamic or short time features. This features calculate from every frame (El Ayadi et al., 2011; Anagnostopoulos et al., 2015) which results in different numbers of features for each sample in the emotional dataset (Henríquez et al., 2014). Supra-segmental features achieve a higher performance than the segmental features with the complex classifiers that need a large number of features such as HMM and SVM classifiers (El Ayadi et al., 2011). A few years ago low-level descriptor (LLD) features were presented (El Ayadi et al., 2011) which were extracted from each short-time frame. The LLD and functional are becoming the standardized features for the SER system (Mustafa et al., 2018).

Segmental Features

These features are assessed from the entire utterance length (El Ayadi et al., 2011; Anagnostopoulos et al., 2015) by calculated statistics of all speech features extracted from the whole utterance. They are also called global static or long-time features. The most common approaches on SER rely on segmental features (Shami and Verhelst, 2007).

Some researchers argue that segmental features contain rich information about emotion and suggest that these are more appropriate to identify emotion than the supra-segmental features (Koolagudi and Rao, 2012; Lupu, 2011). They claim that segmental features outperform the supra-segmental features in terms of classification accuracy (Schuller and Rigoll, 2006) and time (El Ayadi et al., 2011). In addition, they are less sensitive to linguistic information (Ververidis and Kotropoulos, 2008; Sun et al., 2015).

Since the segmental features calculate the statistical function from the whole utterance, there are much fewer numbers of features than for the local features (El Ayadi et al., 2011). However, this is also a disadvantage for the segmental features because it will render them unreliable to be used with complex classifiers that need a large number of features such as HMM and SVM classifiers. Another disadvantage of segmental features is that the temporal information present in speech signals is completely lost. It is also claimed that the segmental features cannot classify emotions with similar arousal e.g. anger versus joy (El Ayadi et al., 2011) and it can only distinguish between high-arousal emotions versus low-arousal emotions respectively.

2.6.3 Disagreement on Efficient Features

Over the years, various features have been explored in SER. However, determination of the most useful features for emotion recognition is still an open issue (Hendy and Farag, 2013). Researchers have not yet identified the best speech features for this task (El Ayadi et al., 2011). This problem has two aspects: the first one is the disagreement between researchers in determining whether the acoustic features only are enough for SER or other types of features need to be included in the feature set.

The majority of researchers believe that the acoustic features alone are enough for emotion classification; hence they make use of acoustic features only in their SER implementation. Recently different types of emotional features have been combined with acoustic features in order to improve performance (Chen et al., 2006; Zhao et al., 2014); for instance: linguistic (lexical); discourse, facial (visual); and gender information.

Linguistic information needs a language model that can describe constraints on possible word sequences in a certain language so as to recognize the word sequence of the speech.

A system that makes use of lexical features has the assumption that certain words can be correlated with emotion state (Chen et al., 2006). However, the relationship between words and emotions is ambiguous in that a single word may convey a different number of emotions. In addition, lexical information always needs manual transcription for each utterance, which is difficult to be realized automatically (Chen et al., 2006). Schuller et al. (Schuller et al., 2005) combine acoustic and linguistic features which reduce the classification error rates up to 8.0%. However, they reported that acoustics features outperform linguistic features when conducted alone.

Discourse markers are linguistic expressions that convey explicit information about the structure of the discourse. It was reported that it can improve performance when combined with acoustic features (Chen et al., 2006; Lee and Narayanan, 2005). Chen et al. (Chen et al., 2006) present an enhanced SER system based on discourse information between humans. The enhanced method makes improvements at almost all of the emotional states. Lee et al. (Lee and Narayanan, 2005) combined acoustic, lexical, and discourse information to improve the performance of the SER system. The results show that significant improvements can be made by combining this information.

Facial features were used before speech features became the favored method for emotion recognition. However, the combination of facial and speech features has had less attention [99] implementing SER systems using both facial and acoustic features. The results reveal that the combination of facial and speech features lead to performance improvements. In addition, the facial features alone outperform the performance of speech features alone. However, speech features contain emotional information that cannot be extracted from visual information.

Gender information can help to improve SER performance (Vogt and André, 2006; Verma et al., 2016). However, the gender of a speaker must be a priori given (Vogt and André, 2006). This can be done only in an offline system or in an academic experience (Vogt and André, 2006). Vogt et al. (Vogt and André, 2006) propose a solution which depends on the use of automatic gender detection before the SER system by using acted and spontaneous datasets. The results illustrated that gender-dependent emotion recognizers perform better than gender independent ones. However, the problem of finding significant features for emotion recognition will still not be solved even with gender separation. Furthermore, they reported that this method could have a negative effect upon the overall classification accuracy if the automatic gender detection was not 100% correct.

The second aspect is the absence of uniformity in acoustic features classification. Overview and studding for these features and its feature extraction techniques require a taxonomy that provides a unique and preferred distribution of features into categories. However, no such taxonomy currently exists.

Different researchers gave different classifications for acoustics features. For example, (Kuchibhotla et al., 2014; Cao et al., 2015; Ahmad, 2016; Reddy and Vijayarajan, 2017; Jalili et al., 2018) classify the acoustic features into only two categories; prosodic and spectral features; while (Koolagudi et al., 2018) add the combination of prosodic and spectral features as the third category. In (Koolagudi and Rao, 2012; Milton et al., 2013; Zhao et al., 2014; Vydana et al., 2015; Swain et al., 2018; Klaylat et al., 2018) three categories were found: prosodic; spectral (vocal tract); and excitation source features. (Luengo et al., 2010; Pérez-Espinosa et al., 2012; Henríquez et al., 2014; Chen et al., 2014; Valstar et al., 2016; Sudhkar and Anil, 2016) name three categories, namely: voice quality; prosodic; and spectral features while (El Ayadi et al., 2011; Henríquez et al., 2014; Muthusamy et al., 2015; Manolov et al., 2017) categorize them into continuous, qualitative (voice quality), spectral and teager energy operator-based features (TEO).

Even inside the category, there is no agreement on the type of each feature. For instance, (Ververidis and Kotropoulos, 2005a; Esmaileyan and Marvi, 2014; Ahmad, 2016; Koolagudi et al., 2018; Klaylat et al., 2018) classify formants as spectral features; while (Borchert and Dusterhoft, 2005; Zhao et al., 2014; Xiaoqing et al., 2017; Samantaray et al., 2015) classify it as a voice quality feature. Another example is the classification of HNR feature as a prosodic feature by (Partila et al., 2015); in the same manner as jitter (Koolagudi et al., 2018; Mariooryad and Busso, 2014), shimmer (Koolagudi et al., 2018), probability of voicing (Mariooryad and Busso, 2014) and voiced-unvoiced ratio (Altun and Polat, 2009).

In SER, it is important to identify important features that have the ability to recognize different emotions (Basharirad and Moradhaseli, 2017). However, due to the lack of unification in emotional speech features classification, evaluation of the features can be confusing.

2.6.4 Curse of Dimensionality

Emotion recognition from speech signals has a huge amount of extracted features; in some researches, it can number up to many thousands of extracted features. For example: Esmaileyan et al. (Esmaileyan and Marvi, 2014) extracted over 2000 features; Schuller et al. (Schuller et al., 2006, 2007) extracted over 4000 features; and Manolov et al. (Manolov et al., 2017) extracted over 6000 features. This number of features led in many cases to what researchers called the curse of dimensionality (Batliner et al., 2011; Kotti and Paternò, 2012; Hendy and Farag, 2013) which causes the degradation of classification performance even if the number of features increases. The curse of dimensionality problem can be avoided by minimizing the features space using features reduction technique or by searching for significant features using features selection techniques (Vogt et al., 2008; Batliner et al., 2011).

Features reduction is used to generate new features containing most of the valuable speech information (Zhao et al., 2014). This is done by finding a suitable linear or nonlinear mapping from the original feature space to another space with reduced dimensionality while preserving as much relevant classification information as possible. Common reduction techniques used in SER comprise: principal component analysis (PCA) and linear discriminant analysis (LDA) respectively. Feature reduction does not retain the original features after the transformation (Batliner et al., 2011). That is needed to determine the most significant feature set.

Features selection represents techniques that aim to find the feature subset that achieves the best possible classification between classes. Many researchers claimed that searching for and selection of the right features selection (Schuller et al., 2006; Schuller and Rigoll, 2006; Schuller et al., 2005) is a mandatory step in SER. Others, however, consider the selection of suitable features as a key problem for SER because it directly affects the performance of SER (Zhou et al., 2006). However, most of the SER research according to (Rong et al., 2009) has focused on improving SER accuracy by building a better classification model. Little effort has been made to search which feature subset will be the most effective for this classification model. Moreover, Hendy and Farag (Hendy and Farag, 2013) argued that significant efforts were spent in literature to extract more features from speech in order to enhance classification accuracy. This was done at the same time ignoring other factors that have been proven to be equally important in achieving good classification results. They prove this by showing it is possible with a reduced number of features to obtain a good classification when ANN is used.

2.7 Existing Approaches and Techniques in Features Extraction and Selection

This section concentrates on the existing SER previous works. It provides a detailed review on feature extraction and selection techniques.

2.7.1 Features Extraction

Much of the existing literature recommends the use of a combination of a variety of acoustic features rather than merely using individual features to achieve an improvement in the accuracy of emotion recognition. They justify that by stating that the use of combined features will correct the errors that occur at different points when using individual features (Pao et al., 2005a, 2006). Several studies suggest that a better performance can be obtained in emotion recognition with a combination of features rather than individual features (Pao et al., 2005a, 2006; Zhou et al., 2009; Koolagudi and Rao, 2012; Kuchibhotla et al., 2014; Koolagudi et al., 2018; Kuchibhotla et al., 2014). The following sections provide a brief review of the combined features sets that are derived from voice quality, prosodic, and spectral features for SER.

Prosodic and Spectral Feature Set

In existing SER literature, there are many different combinations of acoustic features; the most often considered is the combination of prosodic and spectral features. Since both of them are the most commonly used acoustic features in SER (Ververidis and Kotropoulos, 2006; Zhou et al., 2009; Zhang et al., 2013; Kuchibhotla et al., 2014) their combination is also widely used in SER research. It is believed that it is more effective to use prosodic and spectral features in combination rather than merely using them individually. Not only do the literature studies agree that the combination of prosodic and spectral features will improve the performance of SER (Zhou et al., 2009; Ooi et al., 2014; Koolagudi et al., 2018), but some literature also suggests that using them individually will substantially degrade the system performance (Kuchibhotla et al., 2014).

Intensive work studies can be found in literature using different combinations for prosodic and spectral features. For example, Cao et al. (Cao et al., 2015) used 988 prosodic and spectral features extracted from the German EMO-DB and the English LDC datasets bay implementing ranking SVM classifier. The results indicate different classification accuracies for the two datasets: 83.5% for EMO-DB; and 50.4% for LDC. Similarly, Padmaja and Rao (Padmaja and Rao, 2017) have proposed their work on SER using spectral and prosodic features. PCA was used to reduce the dimensionality of the features before classification. The results show that the accuracy of SER has increased significantly with the usage of PCA with Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) emotional dataset and GMM classifier.

MFCC is a well-known acoustic feature and considered as one of the best features used in SER (Pao et al., 2005a). Many researchers have explored the combination of MFCC spectral features with different prosodic feature sets. One of these combinations is to blend MFCC with pitch (F0); that is conceded to be the most important parameter in differentiating among basic emotions.

In (San-Segundo et al., 2009), a combination of F0 and MFCC related features was used to classify anger, happiness, neutral, sadness and surprise respectively. The recognition rate was 81.5% for the Spanish Emotional Speech dataset.

In a similar way, Neiberg et al. (Neiberg et al., 2006) combined standard MFCC (300-3400 Hz) and MFCC-low (20-300Hz) with pitch and its first derivation on the frame level.

This was in order to implement GMM with two Spontaneous dataset: the Swedish voice-controlled telephone services (Swedish VP) and the English meetings (ISL meeting). The results show that MFCC-low outperforms the pitch features. However, the two MFCC features have the same performance.

The combination of MFCC with energy can be found in (Ghai et al., 2017) that implements the SER system with EMO-DB using energy and MFCC, together with three classifiers, namely: SVM; Random Decision Forest; and Gradient Boosting. The results show that the highest accuracy was obtained when using the Random Decision Forest of 81.05%.

Kuchibhotla et al. (Kuchibhotla et al., 2014) proposed a combination of energy, pitch prosodic features and MFCC spectral features respectively. The prosodic and spectral features were classified individually and in combination using LDA, regularized discriminant analysis (RDA), SVM and KNN classifiers. The results are validated over EMO-DB and Spanish emotional speech dataset and show that the use of prosodic and spectral features in combination can lead to performance improvement of 20% for each classifier compared to the performance with the features individually. RDA and SVM classifiers provide the best classification accuracy.

The same features combination is found in the works of Vogt and Andr (Vogt and André, 2005) and (Vogt and André, 2006) that combine 1289 features related to pitch energy and MFCC features. In (Vogt and André, 2005) Naive Bayes classifier with EMO-DB was implemented. The recognition accuracy achieved was about 69.1%. In addition they found after selection that the pitch-related features are the most important feature for acted dataset. Conversely.

In (Vogt and André, 2006) Naive Bayes classifier was implemented with two different datasets, specifically: EMO-DB acted dataset; and SmartKom mobile spontaneous dataset.

A very small subset of relevant features was then selected with the best-first search using the classification accuracy of a Nave. They reported that using the reduced features set (69.1%) outperformed the use of the full feature set (77.4%).

The above works discuss the combination of MFCC with pitch and energy which are prosodic features within a low frequency domain. In contrast, ZCR and formants are high frequency features. Different works present a combination of such features with MFCC. Ramakrishnan and El Emary (Ramakrishnan and El Emary, 2013) compare the performances of spectral (MFCC) and prosodic (F0 and formants) features using HMM and SVM classifiers with EMO-DB and DES. They reported that both pitch an MFCC which can be used to distinguish high-arousal emotions (anger, fear and joy) from low-arousal ones (e.g. sadness). In addition, they are efficient in the classification of emotions that have similar arousal (anger versus joy). In (Shaw et al., 2016) a recognition rate of 86.87% was obtained for pitch, energy, formant and MFCC features extracted from their recorded speech dataset that has four emotions (neutral, happy, angry and sad) using ANN classifier.

Chen et al. (Chen et al., 2012a) use energy, ZCR, pitch, and formants with MFCC to identify six emotions (anger, fear, happiness, sadness, surprise, disgust) from Beihang University Database of Emotional Speech (BHUDES) using SVM classifier. The recognition accuracy was 50.3%. (Manolov et al., 2017) used EMO-DB to extract 6669 features related to energy, ZCR, pitch and MFCC. The feature set was then reduced using different feature selection algorithms before implementing ANN classifier. The result shows that an accuracy of 85% was achieved using 200 features.

Finally, a combination of duration and intensity prosodic features together with MFCC can be found in the work of Rao and Koolagudi (Rao and Koolagudi, 2011). This work identifies six emotions, namely, anger, disgust, fear, happiness, neutrality and sadness from IITKGP-SEHSC dataset. Prosodic and spectral features were extracted from speech

(MFCC, durations, pitch and energy). The recognition performance was found to be 81% for Auto associative neural network (AANN) and 78% for SVM. (Xiaoqing et al., 2017) used SVM with 45 spectral and prosodic features (including pitch, energy, duration, formant, MFCC) extracted from EMODB. The recognition accuracy was 74.26%.

SER system can be found in (Verma et al., 2016) with three different classification algorithms, namely, KNN, multi-layer perceptron (MLP) and SVM with IITKGP-SEHSC that has five basic emotions (happiness, sadness, anger, fear and neutrality). Pitch, intensity, speech rate and MFCC were extracted. SVM classifier shows the highest accuracy of 78% as compared to MLP 75% and KNN 68%.

However, there exists another set of spectral features that were previously combined with prosodic features. For example, LPC, LPCC, LFPC, PLP, Rasta-PLP and wavelet are widely known spectral features used in literature.

Esmaileyan and Marvi (Esmaileyan and Marvi, 2014) investigated the impact of prosodic (pitch, energy, ZCR and formats) and spectral (MFCC, LPC, PLP) feature sets. These were extracted from the proposed Persian Drama Radio Emotional Speech Corpus (PDREC) and EMO-DB datasets of five emotions (anger, fear, joy, sadness and neutrality). The two feature sets were tested individually and in combination using LDA classifier. The result illustrates that for females, spectral features are more effective than prosodic features in terms of performance using both EMODB and PDREC. However, the best performance is attainable when a combination of prosodic and spectral features is used.

Liqin Fu et al. (Fu et al., 2008) used a Mandarin dataset with five emotions, specifically, anger, happiness, surprise, sadness and disgust to implement speaker-independent emotion classification. A combination of prosodic (pitch, energy and formant) and spectral (LPCC and MFCC) features were extracted in frame level. HMM and SVM classifier was used for emotional classification. An accuracy rating of 76.1% was obtained using a fusion system instead of using HMM only.

Pao et al. (Pao et al., 2006) applied an experiment with different combinations of features in order to find the best features combination. This was carried out by using SVM and ANN classifiers together with a Mandarin emotion dataset. The features related to pitch, formants, LPC, LPCC, MFCC, LFPC, PLP and Rasta-PLP respectively. An accuracy rate of 84.2% was obtained with LPC, MFCC, LPCC and LFPC features using SVM classifier; while accuracy of 80.8% was achieved with LPC, LFPC, Rasta-PLP, LPCC, and MFCC feature using ANN.

Lanjewar et al. (Lanjewar et al., 2015) focuses on the detection of six emotions (happiness, anger, neutrality, surprise, fear and sadness) from the extracted speech features related to MFCC, wavelet and pitch. GMM and KNN were used as a classifier with EMO-DB dataset as the GMM classifier provides the best accuracy.

Tarng et al. (Tarng et al., 2010) use a combination of spectral (MFCC and wavelet) and prosodic (F0 and ZCR) features extracted from EMO-DB dataset. An accuracy rate of 90.9% could be obtained using SVM classifier with all emotions.

Kumar et al. (Samantaray et al., 2015) proposed a novel approach that uses a combination of 196 features related to pitch, ZCR, energy, entropy, formant, MFCC, LPCC and mel-energy spectrum dynamic coefficients (MEDC). The SVM used for classification emotion is derived from the Multilingual Emotional Speech Database of North East India (MESD-NEI). The features provide 82.26% classification accuracy for speaker independent emotion recognition.

Pan et al. (Pan et al., 2012) extracted pitch, energy, LPCC, MFCC, and MEDC. They then used SVM to recognize three emotional states: happiness, sadness and neutrality.

They also used different combinations of features from EMO-DB and their collected Chinese datasets. The results indicate that the performance of spectral features is higher than prosodic features. In addition, using both spectral and prosodic features is better in than only spectral or prosodic features are used. The combination of MFCC, MEDC and energy has the highest accuracy rate for both Chinese emotional (91.3%) and EMO-DB (95.1%) datasets.

Albornoz et al. (Albornoz et al., 2011) implemented experiments using a combination of two spectral features, specifically: the mean of the log-spectrum (MLS), MFCC; as well as two prosodic features (energy and F0) using EMO-DB and hierarchical classifier. The result shows that the combination of the two types of features improved results in the emotion recognition process.

Kurpukdee et al. (Kurpukdee et al., 2017) implemented the SER system using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset with SVM classifier. They examined different combinations of features relating to energy, MFCC, PLP, pitch and filter bank (FBANK) respectively. The best accuracy rating was achieved for SVM of 58.40%. In addition, they reported that the combination of energy with FBANK and pitch features is the most suitable feature.

Voice Quality and Prosodic Feature Set

Since it is conceded that voice quality features are important for the process of emotion recognition, combining it with prosodic features can improve SER performance (Yang and Lugger, 2010; Monzo et al., 2014). Lugger and Yang (Yang and Lugger, 2010) proposed a novel set of harmony-based voice quality features which were compared in terms of the classification rate with pitch, duration, formant, and ZCR prosodic features. The comparison was done for prosodic features only, for voice quality features only and for both of them in combination using the Bayesian classifier and EMO-DB. They reported that the combination of both feature types led to improved recognition performance.

Lalitha, et al. (Lalitha et al., 2014) proposed an approach that combines the same voice quality features set (jitter, shimmer, HNR and autocorrelation) with a different prosodic feature set (pitch, entropy, energy and ZCR). Seven emotions were investigated through this study from EMO-DB. A recognition accuracy rating of 81.13% was obtained using SVM classifier.

Joshi et al. (Joshi, 2013) proposed a hybrid classifier using HMM and SVM classifiers. Fourteen features related to prosodic (pitch, intensity, entropy and ZCR) and voice quality (jitter, shimmer, HNR, autocorrelation and noise-to-harmonic ratio (NHR)) were extracted from a collected dataset. The result shows that the proposed hybrid classifier provides an accuracy rating of 98.1%.

Marpaung and Gonzalez (Marpaung and Gonzalez, 2014) designed SER by combining five different features: jitter; shimmer; HNR; NHR; and pitch. Four emotions were investigated in this study: anger; joy; fear; and sadness. The overall recognition rate achieved was 62% by using KNN classifier with the French Geneva Multimodal Emotional Portrayal (GEMEP) dataset.

Borchert and Dusterhoft (Borchert and Dusterhoft, 2005) extracted 63 features related to: jitter; shimmer; voiced to unvoiced frames ratio; spectral energy; pitch; intensity; and formant. This was in addition to new proposed HNR (using different frequency bands). The best recognition rate of 76.06% was obtained using the the sequential minimal optimization (SMO) classifier and EMO-DB. The experiments with new HNR features show that there is little improvement if using different frequency bands for these features.

Monzo et al. (Monzo et al., 2014) used a Spanish expressive speech dataset with five emotions (neutral, happy, sensual, aggressive, and sad) to extract features related to voice quality (jitter, shimmer, HNR, Hammarberg index (HammI) and pe1000) and prosodic (F0, duration and energy) features. After classification, the result shows that the combination of prosodic and voice quality features improve the performance compared with using only prosodic or voice quality features.

Voice Quality and Spectral Feature Set

It is rare to find a combination of spectral and voice quality features. Pao et al. (Pao et al., 2005b) used LDA, K-NN and HMMs classifiers to classify five emotions namely: anger; boredom; happiness; neutrality and sadness. Jitter, shimmer, MFCC, LPC, LPCC, LFPC, and PLP were extracted from two Mandarin datasets. The results obtained show that the proposed system yielded top recognition rates of 88.3% - 88.7% for both datasets.

Voice Quality, Prosodic and Spectral Feature Set

Four voice quality features, namely, HNR, jitter, shimmer, and the probability of voicing are the features that are most combined with both prosodic and spectral features. For instance, Lee et al. (Lee et al., 2011) extracted 384 features related to HNR, F0, energy, ZCR, and MFCC from IEMOCAP and Artificial Intelligence Robot (AIBO) datasets. The features set has been reduced using binary logistic regression in standard statistics software (SPSS) used with a step-wise forward selection. The result obtained showed inspiring accuracy for their proposed hierarchical decision tree.

Vstruc et al. (Štruc et al., 2010) extracted a combination of voice quality (HNR), prosodic (pitch, energy and ZRC) and spectral (MFCC) features from the audio-visual emotion dataset (eNTERFACE) dataset together with six emotions (anger, disgust, fear, happiness, sadness and surprise). An accuracy rating of 62.9% was obtained using SVM classifier.

Research by Schuller et al. (Schuller et al., 2005) compare the performance of acoustic and linguistic features using different classifiers and both the EMO-DB and EMO-AL datasets. A set of 276 acoustic features have been extracted based on the HNR, pitch, energy, duration, formant, ZCR, MFCC and fast Fourier transform (FFT) respectively. The feature set was than reduced to only 75 features using SVM-SFFS. The results showed that acoustics features give a better performance when used alone rather than with the use of only linguistic information. However, the overall performance increases by 3.51% by combining them in one vector. The best result of acoustic features was with SVM with EMO-DB of 87.50%. In addition, using IGR to rank the top 30 selected features shows that HNR ranked as 21 features and MFCC-based features are the top-ranked.

Zhou et al. (Zhou et al., 2010) proposed a hybrid emotion recognition system that combines a GMM-based subsystem and an SVM-based one through the use of F0, loudness, harmonic and MFCC features. An average recognition rate of 91% was achieved for five emotions (anger, happiness, neutrality, sadness and surprise). In addition, the result illustrates that the proposed hybrid system improves performance for all emotions except for the emotion of anger.

Wang et al. (Wang et al., 2015) proposed new Fourier parameter (FP) (harmonically-related) features to improve SER performance. FP features are combined and compare with MFCC spectral features and F0, energy and ZCR prosodic features respectively. EMO-DB, the Institute of Automation Chinese Academy of Sciences (CASIA) and the Chinese elderly emotion (EESDB) datasets were used to validate speaker-independent emotion recognition by using SVM. The study showed that FP features achieved higher average recognition rates than MFCC and prosodic features specific to the EMODB dataset. Moreover, the prosodic features led to the worst performance while the proposed FP and FP+MFCC features improved SER performance.

Zhao et al. (Zhao et al., 2014) used Praat toolkit to extract 204 features related to voice quality (HNR, jitter, shimmer, and spectral energy), prosodic (pitch, intensity, duration and formant) and spectral (MFCC) features from EMO-DB. An accuracy rating of 78.75% was obtained using SVM classifier.

Razak et al. (Razak et al., 2005) use the Voice Driven Emotion Recognizer Mobile Phone (VDERM) dataset that has Malay and English emotions to evaluate emotion features combinations. They used a set of 18 features related to jitter, pitch, energy, duration and LPC. MLP and fuzzy model were used as a classifier. They reported that LPC and jitter are very important features in emotion recognition. In addition, jitter improved the recognition rate achieved by both classification methods.

Chen et al. (Chen et al., 2014) compared SVM and DBN classifiers using 988 features related to the probability of voicing, pitch, intensity, loudness, ZCR, and MFCC respectively. All features were extracted from CASIA. DBN obtained the best accuracy rating with 92.5% while SVM obtained an accuracy rating of 87.5%. Klaylat et al. (Klaylat et al., 2018) used the same features used by Chen with an addition of Line spectral pairs (LSP) features. These features were extracted from an Arabic speech corpus. By compression of different classification models, the SVM classifier obtained the best result with 95.52% accuracy.

Mariooryad and Busso (Mariooryad and Busso, 2014) used voice quality (jitter and probability of voicing), prosodic (F0, energy, and loudness) and spectral (RASTA-filtered auditory, MFCC and Spectral energy) features. They were extracted from the IEMOCAP dataset using only four emotions (neutral, happiness, anger, and sadness). An accuracy rating of 55.32% was obtained using the SVM classifier.

Shirani and Nilchi (Shirani and Nilchi, 2016) extracted 68 features relating to HNR, jitter, shimmer, pitch, intensity, energy, power, ZCR, duration, formants, amplitude, and MFCC respectively. This was followed by SER implements using ANN and SVM classifier with the Persian emotional speech database (Persian ESD) and EMODB dataset. For the ESD and speaker-dependent process, the use of ANN achieved a recognition rate of 98.33%; while a recognition rate of 98.89% was obtained using SVM. However, EMO-DB and speaker-dependent using ANN achieved a recognition rate of 84.70%; while a recognition rate of 86.53% was obtained using SVM.

Ahmad et al. (Ahmad, 2016) used SFS to select 3000 features out of 7956 features related to HNR, jitter, shimmer, the probability of voicing, pitch, energy, ZCR, formant as well as MFCC extracted from EMO-DB dataset. The result obtained using ANN and SVM showed that ANN outperforms SVM in accuracy. The accuracy ratings were as follows: ANN provided 91.2%, while SVM gave 86.3%. Huang et al. (Huang et al., 2013) extracted HNR, jitter, shimmer, the probability of voicing, F0, energy, ZCR, speech rate, duration, formant, and MFCC from collected dataset. An accuracy rating of 63.3% was obtained using four emotions: anger, happiness, neutrality and sadness with GMM classifier.

Oflazoglu and Yildirim (Oflazoglu and Yildirim, 2013) provided a total of 1,532 features related to: voice quality (jitter, shimmer and voicing probability); prosodic (F0 and loudness); and spectral (MFCC, log Mel Freq. Band, LSP) features. These were extracted from Turkish Emotional Speech database (TURES) with neutral, sad, happy, and angry emotions.

Ding et al. (Ding et al., 2018) extracted 1582 features related to: voice quality (HNR, Jitter, shimmer, voicing probability); prosodic (Pitch, loudness); and spectral (Log Mel freq. band, LSP and MFCC) features. OpenSMILE toolkit was used to extract a 1582 INTERSPEECH 2010 features. Only 40 audio files for each kind of emotion were selected from EMO-DB. A recognition rate of 74.29% was obtained using SVM.

The same features set were extracted by Sun and Wen (Sun and Wen, 2015) from three emotional datasets the SAVEE, EMO-DB and CASIA. In addition, they proposed a new normalization method (SSMCFS). The proposed normalization method contains two steps: firstly all the features are normalized to the mean and standard deviation; and second, all features are shifted by means of unlabeled data. They performed the experiment ten (10) times using SVM with polynomial kernel. After rebating the experiment ten (10) times, mean accuracy of 84.58% was obtained for EMO-DB, 71.83% for SAVEE and 75.42% for CASIA. After normalization mean accuracy rating of 86.63% was obtained for EMO-DB, 73.33% for SAVEE and 78.56% for CASIA datasets.

2.7.2 Limitation of Existing Features Extraction Techniques

The detailed survey offered by previous sections demonstrated several limitations of the existing features extraction works that had been conducted in past researches. These limitations will be discussed in this section.

Based on the above discussions, there are no standards or a specific combination of features to be developed and used in SER systems. Every researcher extracts a different number of feature subsets based on their own experience and knowledge. Some SER features were extensively used like spectral and prosodic features; while other features like voice quality were less used, as illustrated in Figure 2.4 below.

Even in combining a feature set it can be observed that the combination of prosodic and spectral features is the most-used combination. Little effort has been done to investigate

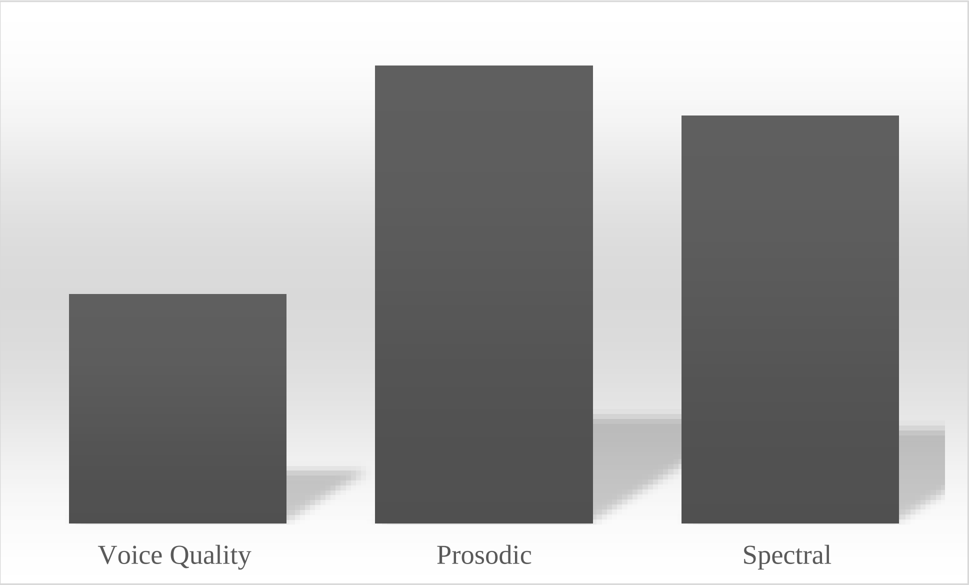


Figure 2.4: Summary of Acoustic Features in Previous Works

the combination of voice quality and spectral features. Table 2.1 following summarizes the different combinations used by different researchers

Furthermore, some voice quality features were totally ignored. The discussion thus far is illustrated in Table 2.1. As shown in Table 2.2 there are some features represented in bold that are commonly used among many authors, especially in EMO-DB.

The table shows that the most popular types of voice quality features used by researchers in SER are the HNR, harmonic, jitter, shimmer, and voice probability features. However, features such as NHR, HammI and autocorrelation are not often used. Furthermore, the glottal-to-noise excitation ratio (GNE) feature has not been used before in SER.

Table 2.1: Summary of Acoustic Feature Combinations in Previous Works

Feature Set	Reference
Prosodic and Spectral	(Cao et al., 2015; Padmaja and Rao, 2017; San-Segundo et al., 2009; Neiberg et al., 2006; Ghai et al., 2017; Kuchibhotla et al., 2014; Vogt and André, 2005, 2006; Ramakrishnan and El Emary, 2013; Shaw et al., 2016; Chen et al., 2012a; Manolov et al., 2017; Rao and Koolagudi, 2011; Xiaoqing et al., 2017; Verma et al., 2016; Esmaileyan and Marvi, 2014; Fu et al., 2008; Pao et al., 2006; Lanjewar et al., 2015; Tarng et al., 2010; Samantaray et al., 2015; Pan et al., 2012; Albornoz et al., 2011; Kurpukdee et al., 2017)
Voice Quality and Prosodic	(Yang and Lugger, 2010; Lalitha et al., 2014; Joshi, 2013; Marpaung and Gonzalez, 2014; Borchert and Dusterhoft, 2005; Monzo et al., 2014)
Voice Quality and Spectral	(Pao et al., 2005b)
Voice Quality, Prosodic and Spectral	(Lee et al., 2011; Štruc et al., 2010; Schuller et al., 2005; Zhou et al., 2010; Wang et al., 2015; Zhao et al., 2014; Razak et al., 2005; Chen et al., 2014; Klaylat et al., 2018; Mariooryad and Busso, 2014; Shirani and Nilchi, 2016; Ahmad, 2016; Huang et al., 2013; Oflazoglu and Yildirim, 2013; Ding et al., 2018; Sun and Wen, 2015)

2.7.3 Feature Selection

This section reviews works that have used features selection algorithms to reduce the features for SER.

Filter Approach

The filter approach was one of the earliest used approaches implemented using either ranking-based or subset-based filter methods. The simplest form was the ranking-based filter methods. Different researchers worked with ranking-based filter methods in their SER implementation.

Schuller et al. (Schuller et al., 2006) used IGR to reduce the feature dimensionality; the results were obtained applying SVM with DES, EMO-DB, and SUSAS datasets. For EMO-

Table 2.2: Summary of Voice Quality Features in Previous Works

Voice Quality Feature	Reference	Dataset
HNR	(Lee et al., 2011)	AIBO
	(Štruc et al., 2010)	eNTERFACE
	(Schuller et al., 2005)	EMO-DB, EMO-AL
Harmonic	(Yang and Lugger, 2010)	EMO-DB
	(Zhou et al., 2010)	CASIA
	(Wang et al., 2015)	EMO-DB, CASIA, EESDB
Jitter	(Razak et al., 2005)	VDERM
Voicing probability	(Chen et al., 2014)	CASIA
	(Klaylat et al., 2018)	Arabic datasets
Jitter, Shimmer	(Pao et al., 2005b)	Mandarin datasets
Jitter, Voicing probability	(Mariooryad and Busso, 2014)	IEMOCAP
HNR, Jitter, Shimmer	(Shirani and Nilchi, 2016)	EMO-DB
Jitter, Shimmer, Voicing probability	(Oflazoglu and Yildirim, 2013)	TURES
HNR, Jitter, Shimmer, Voicing probability	(Ahmad, 2016; Ding et al., 2018)	EMO-DB
	(Sun and Wen, 2015)	EMO-DB, SAVEE, CASIA
	(Huang et al., 2013)	Collected dataset
HNR, Jitter, Shimmer, spectral energy	(Zhao et al., 2014)	EMO-DB
HNR, Jitter, Shimmer, Autocorrelation	(Lalitha et al., 2014)	EMO-DB
HNR, Jitter, Shimmer, NHR	(Marpaung and Gonzalez, 2014)	GEMEP
HNR, Jitter, Shimmer, Autocorrelation, NHR	(Joshi, 2013)	Collected dataset
HNR, Jitter; Shimmer, Voiced-unvoiced ratio, Spectral energy	(Borchert and Dusterhoft, 2005)	EMO-DB
HNR, Jitter, Shimmer, HammI, pe1000	(Monzo et al., 2014)	Spanish datasets

DB the correct recognition rate after IGR improves from 86.7% to 86.9%. The correct recognition rate for DES increased from 68.7% to 74.5%. Finally, SUSAS achieved 77.8% correct recognition rate using the full feature set. By using IGR reduction accuracy improved to 84.9%. Another implementation of IGR was found in a study by (Borchert and Dusterhoft, 2005) Borchert and Dusterhoft implementing SVM algorithm and EMO-DB dataset with IGR to rank the top 25 prosodic (pitch, intensity, and formant) and voice quality (HNR, jitter, shimmer, ratio of voiced to unvoiced frames and spectral energy) features from 63 features for SER. The ranking parameter shows that spectral energy ranked first (1^{st}), shimmer ranked 8^{th}, HNR ranked 15^{th}, and jitter ranked 24^{th}. This indicated that

HNR is not more of an important voice quality feature than spectral energy distribution and shimmer. They conclude that voice quality features as well as prosodic features are important for emotion recognition. In addition, it shows that pitch and intensity are the most important features of emotional speech.

A novel filter approach based on RS and SVM was proposed by Zhou et al. (Zhou et al., 2006) in an attempt to reduce the calculation cost while keeping a high recognition rate. Prosodic features related to pitch, energy, formant, and speaking rate were extracted from the Chinese Linguistic Data Consortium (CLDC) dataset. A comparison of recognition performance with and without the feature selection process was done. An accuracy rating of 74.75% with selected 13 features obtained. An accuracy rating of 77.91% with the 37 full feature set obtained. These are considered a good result because the features space was reduced by 64.86%, while the performance decreased by only 3.16%.

Chen et al. (Chen et al., 2012b) proved that FDR is superior to PCA in comparative experimental processes using SVM, in addition to ANN classifiers with BHUDES Mandarin dataset. This experimental test consisted of four models: FDR + SVM, PCA + SVM, FDR + ANN, and PCA + ANN. They found that combining FDR + SVM resulted in a better performance than the three other combinations. Zheng et al. (Zheng et al., 2014) also applied FDR in combination with three classifiers (SVM, the Nearest Neighbor (NN), and sparse representation classifier (SRC)), together with two emotional speech datasets: eNTERFACE, and AIBO. The result shows that the best performance was for NN+FDR combination with an increase of 0.37% when decreasing the features set from 1582 to 1500 features.

The ranking-based filter does not meet the need for feature selection for high dimensional features very well because it selects features relevant to the class label even if they are highly correlated with each other. Accordingly, it helps in eliminating only the relevant features and does not help in the elimination of redundant features. However, redundant features also affect the speed and accuracy of classification algorithms and should be removed.

Therefore, a better selection would be the subset-based filter. This filter has the capability to remove both irrelevant and redundant features. Different researchers tend to rely on subset-based methods. For example, Vogt and Andre (Vogt and André, 2005) search for the most relevant features for EMO-DB dataset using CFS with best-first search. With Nave Bayes classifier, the original 1280 features derived from pitch, energy and MFCC reduced to about 90-160. The accuracy rating increased from 69.1% using the full feature set to 77.4% for selected features.

Shirani and Nilchi (Shirani and Nilchi, 2016) compared CFS using best-first search method with SVM features selection using ANN and SVM classifiers. They performed their evaluation on a Persian ESD, German and EMO-DB dataset. A total of 68 features related to duration, pitch, intensity, formants, amplitude, HNR, jitter, shimmer, energy, power, ZCR and MFCC were extracted. The result shows that the proposed SVM feature selection method provides better performance compared to CFS and baseline feature set. The recognition rate achieved using the proposed method was 99.44% for ESD and of 87.21% for EMO-DB for SVM with speaker-dependent classification; while for ANN with speaker dependent classification the recognition rate was 98.89% for ESD and 85.16% for EMO-DB.

Zhao et al. (Zhao et al., 2014) applied FCBF to reduce 204 acoustic features (including HNR, jitter, pitch, intensity, duration, formants, spectral energy, MFCC) using a threshold of 0.001 with six different classifiers. Firstly, FCBF was used to reduce features set to 90 features for EMO-DB and 70 for the polish dataset. The selected features were then tested to identify the best number of features that would provide the highest recognition accuracy. For the polish dataset, 64 features gave the best accuracy with 87 features for the

EMO-DB dataset. The result indicates that applying FCBF improves emotion recognition performance; hence, all used classification methods obtain better recognition performance after feature selection.

Sun and Wen (Sun and Wen, 2015) compare mRMR, PCA, and LDA with a proposed semi-supervised feature selection method (SMCFS). SMCFS method can handle unlabeled data. Three emotional datasets were used, namely, the SAVEE, EMO-DB and CASIA. The experiment was repeated ten (10) times using SVM with polynomial kernel. The mean of the ten (10) results was adopted as a final performance. The accuracy ratings obtained were, specifically: 88.85% for EMO-DB; 77.44% for SAVEE; and 78.10% for CASIA datasets.

Jassim et al. (Jassim et al., 2017) proposed emotion classification under clean and noisy environments based on a combination of traditional and proposed features. The traditional features were presented by INTERSPEECH 2010 paralinguistic emotion challenge features; while the proposed features were neural-responses based. The system was evaluated using LIBSVM (Matlab implementation) with RBF kernel for two well-known datasets: EMO-DB and eNTERFACE. The traditional features were extracted using openSMILE toolkit, utilizing the configuration file IS10-paraling:conf. After the extraction, the features vector was normalized between [0-1]. The mRMR algorithm was then employed to remove the irrelevant features and reduce the features dimensionality from 1582 to 946. An accuracy rating of 89.45% was obtained for the traditional features in a clean environment with EMO-DB using seven emotions (neutrality, anger, fear, joy, sadness, disgust, and boredom).

Huang et al. (Huang et al., 2013) proposed the Maximal information coefficient (MIC) filter based upon features selection. MIC is a new statistic tool that measures linear and nonlinear relationships between paired variables. With features set included: HNR; jitter; shimmer; voiced frames; unvoiced frames; unvoiced to voiced frame ratio; voiced to total frame ratio; pitch; energy; duration; speech rate; formant; ZCR and MFCC features that extracted from a collected dataset. The result that obtained using GMM indicated that MIC is a powerful algorithm.

Wrapper Approach

The previous section describes the related work using filter approaches. However, since the wrapper approach uses a classification performance to evaluate the features it usually provides a more improved feature than filter approaches.

Pao et al. (Pao et al., 2006) used FFS algorithm with SVM and NN classifiers to classify five (5) emotions from Mandarin emotional speech. The experimental result shows that among a selection of ten (10) best selected features LPC is considered to be the best feature and pitch is the worst one. Kotti and Paterno (Kotti and Paternò, 2012) also applied FFS algorithm followed by PCA using EMO-DB for a total number of 2327 extracted features. Several numbers of features ranging from 50 to 100 were tested, with an accuracy rating of 97.0% accuracy being obtained for the linear SVM. In (Arias et al., 2014) FFS was also applied by Arias et al. to reduce the number of features for the EMA and EMO-DB dataset to 20.

Hendy and Farag (Hendy and Farag, 2013) used FFS to reduce 175 extracted features related to pitch, duration, formants, energy, ZCR, jitter, shimmer, MFCC, and LPC features in an attempt to find a robust and fast ANN classifier suitable for use in a real-life application. An accuracy rating of 85% was obtained using 129 reduced features set.

You et al. (You et al., 2006) compared SFS with the enhanced Lipschitz embedding method. SVM was used to evaluate the performance of the SER system. A total of 64 features were extracted relating to pitch, energy and formant from the Mandarin dataset. The proposed method obtained improvements of 9%-26% in speaker-independent and 5%-20% in speaker-dependent classification.

Ververidis and Kotropoulos (Ververidis and Kotropoulos, 2005a) performed a comparison between SFFS and SFS algorithms. A set of 65 features related to pitch, energy, and formants were extracted from the DES dataset. The results showed accuracy improvement of the Bayes classifier by an increment of 3%. Moreover, they reported that the SFFS algorithm is more powerful for feature selection than the SFS with regard to emotional speech. The same features were also used in another study (Ververidis and Kotropoulos, 2005b). They reported that the simple SFS algorithm was subjected to nesting problems; accordingly, they implemented SFFS (which is considered to be an improved version of the SFS algorithm). The results show that Bayes classifier obtained a correct classification rating equal to 55% compared to the human classification score which is 67%.

Schuller et al. in (Schuller et al., 2005) used SFFS to select the best 75 feature from a 276 feature set related to voice quality, prosodic and spectral features. Using different classifiers the best result was SVM of 87.50% for the selected features. The recognition accuracy for the full features was 84.84%. In addition, they reported that the performance could be improved by acoustic features selection. Furthermore, the extraction effort could be saved.

Schuller et al. (Schuller and Rigoll, 2006) used the FSSF with SVM in order to find the optimal classifier such as: instance-based nearest neighbor (1NN and kNN); MLP; a decision tree (C4.5); Naive Bayes (NB) and Bayesian Networks (BN); as well as ensemble classification construction. The final result shows that not all classifiers show better performance with the reduced set, but that SVM gave the best result. This is an accepted result since the feature set is optimized by SVM-SFFS.

Another comparative study was carried out by Lugger and Yang in (Lugger and Yang, 2007) for SFFS and fisher transform to attempt to reduce the feature numbers from 208 to eight (8) for speaker independent emotion classification. They concluded that no significant changes could be observed for either of the algorithms. The fisher transform generally performs better than SFFS for the same number of final features.

Wang et al. (Wang et al., 2014) used SFFS to compare both linear SVM (LSVM) and radial basis function kernel SVM (RSVM) classifiers with German and Chinese datasets. When SFFS was used with RSVM an improvement of the recognition rate by 14.9% and 4.3% respectively was detected on the German and Chinese datasets. In addition, an improvement of the recognition rate by 17.4% and 10.1% were detected on German and Chinese datasets when SFFS used LSVM.

Gharavian et al. (Gharavian et al., 2013) used a collected Farsi emotional dataset to extract: F0; energy; formant; and MFCC. GMM was used to classify three emotions namely anger, happiness, and neutrality. FCBF and ANOVA feature selection methods have been used to select various combinations of the features. The results indicated that energy and pitch features are important features for SER. In addition, MFCC features are generally discarded when using FCBF and ANOVA methods.

Yogesh et al. (Yogesh et al., 2017) proposed PSOBBO (a wrapper feature selection technique) in addition to OSBSBCF higher order spectral features. PSOBBO is a new particle swarm optimization assisted by Biogeography-based algorithm. OSBSBCF (1632 features) was extracted using openSMILE toolbox; this feature is a combination of 50 BSBCFs proposed features and 1582 features used in Inter-speech 2010. Different experiments were conducted using extreme learning machine (ELM) classifier with three emotional datasets namely: EMODB, SAVEE, and SUSAS. The results showed that the proposed selection technique provides the best accuracy with the minimum number of features for the three datasets. For EMO-DB the accuracy rating was 88.36% for speaker independent and 97.54% for speaker dependent with only 177 features respectively. For SAVEE, the accuracy rating was 59.63% for speaker independent and 69.75% for speaker dependent with only 336 features. SUSAS obtained an accuracy rating of 84.12% with only 258 features.

Muthusamy, et al. (Muthusamy et al., 2015) proposed a new feature enhancement method based on GMM model. To validate the proposed methods SAVEE, EMD-DB and Sahand Emotional Speech database (SES) datasets were used by implementing KNN and ELM classifiers. The stepwise linear discriminant analysis (SWLDA) was used to reduce the enhanced wavelet packet, energy and entropy features. For speaker-dependent ELM obtained an accuracy rating of 98.98% with EMO-DB, 97.60% with SAVEE and 92.79% with SES. KNN obtained an accuracy rating of 59.14% with EMO-DB and 94.27% with SAVEE. For speaker-independent ELM achieved an accuracy rating of 97.24% with EMO-DB and 77.92% with SAVEE. With KNN obtained an accuracy rating of 49.12% together with EMO-DB, 69.17% with SAVEE and 84.58% with SES.

Ding et al. (Ding et al., 2018) proposed an optimization method based on the biogeography optimization algorithm (BBO) which attempted to solve the problem of high features dimensionality in SER. Features related to voice quality, prosodic and spectral features were extracted. Only 40 audio files for each kind of emotion ware selected from EMO-DB. After the selection, the result showed that BBO-SVM can filter a lot of redundant features. In addition, it provided a better performance compared to GA under the same parameter setting. BB-SVM resulted in 90.13% average recognition rate while GA-SVM resulted in 81.26% average recognition rate.

Embedded Approach

Embedded methods have been proposed to combine the advantages of both filter and wrapper approaches. Not much research can be found for the single embedded approach in SER. More work on embedded algorithms can be found in hybrid approaches. In (Altun and Polat, 2009), Altun and Polat compared four feature selection algorithms: SFS wrapper; Least Squared Bound (LSBOUND) filter; and two embedded algorithms, namely, the Mutual Information Based Feature Selection (MUTINF) and W2R2. Using EMO-DB dataset 58 features related to voiced-unvoiced ratio, F0, sub-band energy, MFCC, and LPC were extracted. The results showed that LSBOUND outperformed the other algorithms in reducing average CV error. With regard to the features, it was reported that for all algorithms the most frequently selected ones were the prosodic and sub-band energy features. In addition, MFCC features are more informative than LPC features.

Hybrid Approach

In addition to the previously-mentioned categories, some researchers elected to work with a hybrid method by combining two or more approaches to overcome some method limitations. In some instances, the first approach (which is regularly filtered) can be used as a pre-process step before using the main method. The hybrid features selection approach can direct categories to ensemble or sequential methods depending on the method approaches implemented. In SER literature three types of an ensemble method were noticed. The first one combines the filter-wrapper approach while the second combines a filter- embedded approach. The final approach combines two embedded approaches.

An example of filter-wrapper approach can be found in (Tickle et al., 2013) Tickle et al. This study selected a 71 hybrid features set by implementing both IGR and classify Subset

Eval algorithms. Firstly, the Subset Eval algorithm (which is WEKA wrapper algorithm with best-first search) was used to select 11 out of 998 features. However, the performance of these features alone was very poor; hence, IGR was implemented to obtain the ranking of all features. Further, the top 63 features were combined with the previous 11 features. The final 71 hybrid set of features was used with MLP classifier and EMO-DB.

The filter-embedded approach was used by Mencattini et al. (Mencattini et al., 2014) who performed a comparison between three hybrid models. This was carried out by employing the embedded stepwise regression (SWR) with ranking-based filter Relief algorithm and two subset-based filter algorithms (namely, Pearson correlation coefficient (PCC), and Mutual Information maximization Criterion (MIM)) using EMOVO Italian dataset. The results show that the Relief-SWR approach does not provide widely acceptable results. The author explained that it was probably due to the Relief not being able to deal with the redundancy features. Further, it was reported that the PCC-SWR approach outperforms MIM-SWR approach in performance.

The embedded-embedded approach was used by Rong et al. (Rong et al., 2009) to select the most effective acoustic features that improve the performance of the SER system. A total of 84 features were extracted from Chinese (Mandarin) dataset. These features are related to pitch, intensity, ZCR and MFCC. Ensemble Random Forest to Trees (ERFTrees) has been proposed as a selection method that involves two components of feature selection and voting strategy. The feature selection uses two algorithms, namely, C4.5 Decision Tree and RF. The voting strategy uses a voting-by-majority method to combine these two subsets of candidate features. A total of 16 acoustic features were selected from the original 84 feature set. The result shows that the selected 16 feature subset provides higher recognition accuracy than the original feature set.

In SER literature, four types of a sequential method were noticed. The first one combines more than one filter approach while the second method combines more than one wrapper approach. The third method combines the filter-wrapper approach; while the final one combines a filter-embedded approach.

The filter-filter approach is applied by Clavel et al. (Clavel et al., 2008) who used FDR to reduce the feature space. This was carried out by selecting the 40 most relevant features from the English Situation Analysis in a Fictional and Emotional (SAFE) dataset. The selection was performed in two steps. Firstly, 100 features were selected from voice quality, prosodic, and spectral features separately. Then a second selection step with the same algorithm was applied to the selected features. This strategy was adopted in order to avoid having strong redundancies between the selected features.

Liu et al. (Liu et al., 2018) used two filter feature selection methods, specifically: the correlation analysis and the Fisher criterion with CASIA Chinese dataset as well as two classifiers, ELM and SVM classifiers. A total of 34 spectral and prosodic features were extracted then analyzed by correlation analysis to dispose of the redundant features. The Fisher criterion was then implemented to select 20 features. The best accuracy rating for ELM was 88.25% before selection and 90.43% after selection. In the case of SVM, the best accuracy rating was 87.63% before selection and 87.73% after selection.

The wrapper-wrapper approach implemented by Mariooryad et al. (Mariooryad et al., 2014) uses SVM with a two-level wrapper feature selection approach to reduce the dimension of the 4368 feature vector of SEMAINE database. These features are related to prosodic, spectral and voice quality features. Firstly, FFS selection used them to reduce the number of features to 500. Following this, 100 features were selected from them using the same algorithm.

The filter-wrapper approach in (Schuller et al., 2007) was used to save computation time prior to using SFFS wrapper. Schuller et al. applied the GR ranking-based filter to reduce features sets extracted from DES and EMO-DB. The result was obtained using RF classifiers. It stated that the reduction helped to increase performance. For DES, the accuracy rating increased from 53.5% to 57.1%; while for EMO-DB the accuracy rating increased from 72.3% to 72.5%.

Wang et al. (Wang et al., 2017) used IGR with a threshold of 0.0032 in order to select 2535 features from 5760 Wavelet features. SFS with SVM was then used to select the best 1279 features. SVM, AdaBoost and RF classifiers were used to implement SER using Chinese elderly emotion (EESDB) dataset. The use of AdaBoost resulted in accuracy ratings of 93.9%, RF of 92.8% and SVM of 94.2% respectively.

Wu and Liang (Wu and Liang, 2011) proposed a two-stage feature selection scheme to reduce the number of features. The first stage calculates the FDR and ranks each feature individually to remove irrelevant features using a threshold of 0.15, as they noticed no improvement in performance results when increasing the threshold. In the second stage, the SFS were compared to the multi-class linear discriminant analysis (LDA) to select features with SVM classification. The result shows that using six LDA selected features delivered higher accuracy than using 50 SFS selected features.

Tomar et al. (Tomar et al., 2014) used the combination of two feature selection techniques, namely, the ranking-based filter F-score and the wrapper SFS to select 24 significant features using Multi least squares twin support vector machine (MLSTSVM) classifier. Firstly, the F-score for each feature was calculated, and the SFS was then used for obtaining 24 feature subsets or models. The features inside every model were ordered by the F-score. The result shows that model 16 gave the highest accuracy rating of 87.28% for linear MLSTSVM, 92.89% for Gaussian MLSTSVM, and 88.87% for the polynomial respectively.

The filter-embedded approach used by Esmaileyan and Marvi (Esmaileyan and Marvi, 2014) used a two-stage filter and embedded feature selection algorithm to reduce 2461 extracted prosodic and spectral features. Firstly, features are ranked by FDR and features with a low FDR score are eliminated. Then, the features which are selected by the FDR filtering are reduced in dimensions using LDA feature selection algorithm. PDREC and EMO-DB with five emotions (anger, fear, joy, sadness and neutrality) were then used. An accuracy rate of 55.74% for females and 47.28% for males was achieved using the PDREC dataset. Also, accuracy ratings of 78.64% and 73.40% were obtained for Berlin database for females and males, respectively.

2.7.4 Limitation of Existing Features Selection Approaches

The comprehensive review given by previous sections explains various limitations of existing features selection works that have been conducted in past researches. These limitations will be discussed in this section. According to the above discussions, feature selection techniques were used to reduce the enormous number of extracted features and select significant features that represent emotional accurately. However, the best significant set of emotional SER features that increase the classification accuracy has not yet been found.

In the reported works above, the features selection approaches can be split into two groups. In the first group, a single feature selection algorithm was employed to find the most informative features. Table 2.3 summarizes the single selection methods used in previous works.

However, the quality of the selected features is dependent on the ability of the used algorithm to rank or select the set of features. Consequently, each feature selection algorithm will end up with a different subset of features as the best feature set. Therefore, an obvious

Table 2.3: Summary of Single Selection Approaches in Previous Works

Selection Approach	**Reference**
Ranking based filters	(Schuller et al., 2006; Borchert and Dusterhoft, 2005; Zhou et al., 2006; Chen et al., 2012b; Zheng et al., 2014)
Subset-based filter	(Vogt and André, 2005; Shirani and Nilchi, 2016; Zhao et al., 2014; Sun and Wen, 2015; Jassim et al., 2017; Huang et al., 2013)
Wrapper	(Pao et al., 2006; Kotti and Paternò, 2012; Hendy and Farag, 2013; You et al., 2006; Ververidis and Kotropoulos, 2005a,b; Schuller et al., 2005; Schuller and Rigoll, 2006; Lugger and Yang, 2007; Wang et al., 2014; Gharavian et al., 2013; Yogesh et al., 2017; Muthusamy et al., 2015; Ding et al., 2018)
Embedded	(Altun and Polat, 2009)

need to define a framework which it is more likely to obtain a reliable subset of features.

In the second group, a hybrid features selection algorithm was employed to select the best representative features that give higher accuracy. Table 2.4 summarizes the hybrid selection methods that were used in previous works.

In the second group, a hybrid features selection algorithm was employed to select the best representative features that give a higher accuracy. Table 2.4 summarizes the hybrid selection methods that used in previous works.

Table 2.4: Summary of Hybrid Selection Approaches in Previous Works

Selection Approach	**Reference**
Ensemble	
filter-wrapper	(Tickle et al., 2013)
filter-embedded	(Mencattini et al., 2014)
embedded-embedded	(Rong et al., 2009)
Sequential	
filter-filter	(Clavel et al., 2008; **?**)
wrapper-wrapper	(Mariooryad et al., 2014)
filter-wrapper	(Schuller et al., 2007; Wang et al., 2017; Wu and Liang, 2011; Tomar et al., 2014)
filter-embedded	(Esmaileyan and Marvi, 2014)

As can be seen from Tables 2.3 and 2.4, most of the work done by single selection methods rely on wrapper algorithms. Even for the hybrid selection methods a considerable amount of work has done using wrapper algorithms. Even though wrapper methods often achieve better classification accuracy than filter methods, they tend to be much slower than filter methods because they must repeatedly call the induction algorithm. Filter methods are generally much faster than the wrapper and embedded methods and are more practical for use on data of high dimensionality.

In addition, wrapper methods depend on classification algorithms which connect the features to a specific classifier. For instance, in (Schuller and Rigoll, 2006) SFFS was used with SVM to select features in order to find the optimal classifier. However, the result indicates that not all classifiers show better performance with the reduced set, but that SVM gave the best result. On the other hand, the results achieved using filter feature selection methods are independent of classifiers and yield a much more general conclusion,

The implementation of a hybrid filter-filter approach appears to be the most suitable choice. A few number of studies have been performed using this approach. Only one research could be found that made use of a two-layer sequential ranking-based filter. The ranking-based filter approach has much less complexity and low-cost computation compared to the subset-based filter approach. However, the subset-based filter approach can eliminate irrelevant and redundant features. In contrast, the ranking-based filter approach can eliminate irrelevant features only. A hybrid approach that can combine the advantages of the two filter approaches could be promising.

2.8 Reviews on Possible Techniques for Speech Emotion Recognition

This section reviews the techniques required for further improvements.

2.8.1 Features Extraction

This section concentrates on describing different voice quality, prosodic and spectral features extraction techniques.

Loudness

A measurement of the sound level. It is closely linked to the frequency and the duration of the sound. This scale has been built from psychoacoustics measurement methods called direct measures (Stevens, 1956). It is obtained by first calculating 1000Hz voice pressure ratios under different intensities and then takes $1/10$ of the logarithm of the result.

Pitch and Autocorrelation

Pitch represents periodicity candidates as a function of time that may refer to acoustics, perception or vocal fold vibrations. There are many types of Pitch Detection Algorithms (PDA) in the literature. Time domain, frequency domain, and time-frequency domain related. The time domain method includes: the short-time average magnitude difference function (AMDF); short-term autocorrelation function (ACF); the frequency domain method which includes harmonics enhancement based on instantaneous frequency and Subharmonic Summation algorithm (SHS) methods. Finally, the time-frequency domain method includes pitch detection based on Hilbert-Huang transform. In the context of this research, only ACF and SHS functions will be discussed.

ACF-based algorithms are simpler to implement and are quite robust against noise. In addition, it is the most frequently used F0 estimators. $ACF(f)$ for a discrete time domain signal $x(t)$ is expressed in Equation 2.1 below.

$$ACF(f) = \frac{1}{S} \sum_{m=0}^{S-K-1} x(t)x(t+k) \tag{2.1}$$

Where:

$x(t)$ = is signal in time domain
S = is total number of samples in a window
f = is the lag index

SHS pitch estimation is the value of $f = 2^s$ for which $H(f)$ is maximum. $H(f)$ is the function that represents the sub-harmonic sum spectrum. This is illustrated in Equation 2.2 below.

$$H(f) = \sum_{n-1}^{N} h_n P(nf) \tag{2.2}$$

Where:

n = is the compression factor
h_n = is a decreasing sequence implying that higher harmonic contribute less to the pitch than lower harmonics do (0.84^{n-1})
N = is the number of harmonic that are taken into account (equal to 15)

Harmonic

Harmonic frequency is a signal or wave whose frequency is a multiple of fundamental frequency. Fundamental frequency itself is considered as the first harmonic. HNR provides an indication of the overall period of the voice signal by measuring the ratio between the periodic (harmonic part) and aperiodic (noise) components. NHR is a measure that quantifies the amount of additive noise in the voice signal.

This can be found from the relative height of the maximum of the autocorrelation function. The autocorrelation function (ACF) is described in equation 2.1. The function has a global maximum at the lag x = 0. The signal is said to have at least a periodic part if the highest local maximum is at lag x_{max} and its height $ACF(x_{max})$ is large enough. The harmonic strength $R_0 = ACF(x_{max})$ is a number between 0 and 1 and results from normalized autocorrelation function ACF.

$$ACF_f(x) = \frac{ACF_f(x)}{ACF_f(0)} \tag{2.3}$$

At lag x_{max}. If noise n_n is added to a periodic signal h_n of period T_0 and n_n and h_n are uncorrelated, the autocorrelation function of the resulting signal f_n at zero lag is:

$$ACF_f = ACF_h(0) + ACF_n(0) \tag{2.4}$$

If white noise is added a local maximum can be found at lag $x_{max} = T_0$ with height $ACFf(x_{max}) = ACF_h(T_0)ACF_h(0)$. The autocorrelation function at zero lag equals the power of the signal. Hence, the normalized autocorrelation at lag x_{max} represents the relative power of the periodic (harmonic) component of the signal where its complement represents the relative power of the noise component:

$$ACF_f(x_max) = \frac{ACF_h(0)}{ACF_f(0)} \tag{2.5}$$

$$1 - ACF_f(x_{max}) = \frac{ACF_n(0)}{ACF_f(0)} \tag{2.6}$$

HNR is presented in Equation 2.7 as shown below:

$$HNR(dB) = -10\frac{1}{m}\sum_{j=1}^{m} log(1 - \frac{1}{ACF_f(x_{max})}) \tag{2.7}$$

NHR is presented by Equation 2.8 as shown below:

$$NHR(/) = 100\frac{1}{m}\sum_{j=1}^{m}[1 - \frac{1}{ACF_f(x_{max})]} \tag{2.8}$$

Jitter and Shimmer

Jitter refers to the cycle to cycle variations of the fundamental frequency $(f0)$ and is calculated as shown in Equation 2.9:

$$Jitter(i) = \frac{|f0(i+1) - f0(i)|}{(f0(i)} \tag{2.9}$$

Shimmer indicates cycle to cycle variation in the energy (E) and is calculated as shown in Equation 2.10:

$$Shimmer(i) = \frac{|E(i+1) - E(i)|}{(E(i)} \tag{2.10}$$

Jitter is a measure of frequency instability, while shimmer is a measure of amplitude instability.

Period

It is time for one cycle. This is measured by calculating the length of time for a known number of cycles and then dividing it by the number of cycles. For N cycles and t length of time, the period is given as shown in Equation 2.11:

$$Period = \frac{t}{N} \tag{2.11}$$

Voice Break

The number of voice breaks denotes the number of distances between consecutive pulses that are longer than 1.25 divided by the pitch floor. Similarly, the degree of voice breaks indicates the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal. The silences at the beginning and the end of the signal are not considered as breaks.

Pulses

A model of the excitation of the vocal tract where each pulse represents a new excitation of the vocal tract. The distance between pulses corresponds to the inverse of the local pitch. For instance, if the local pitch in an interval is 100 Hz, the assigned pulses in the interval are 0.01.

Voice to Unvoiced Frame Ratio

The outline of the voice frame over the unvoiced frame is given in Equation 2.12.

$$voice - unvoiceframe - ratio = \frac{numberofvoicedframes}{numberofunvoiceframes} \tag{2.12}$$

Unvoiced to Total Frame Ratio

The outline of the voice frame over the total frame is given in Equation 2.13.

$$unvoiced - totalframe - ratio = \frac{numberofunvoicedframes}{numberoftotalframes} \tag{2.13}$$

Voice to Total Frame Ratio

The outline of the voice frame over total frame ratio is given in Equation 2.14.

$$voice - totalframe - ratio = \frac{numberofvoicedframes}{numberoftotalframes} \tag{2.14}$$

Probability of Voicing

This is conceded to be one of the judging emotion classes, since it uses short-time average magnitude to detect voice. This is shown in Equation 2.15.

$$M_n = \sum_{\infty}^{+\infty} |x(m)|w(n-m) \tag{2.15}$$

Where:

M_n = is short-time average magnitude
$x(n)$ = is speech signal
$w(n)$ = is window function

Hammarberg index

HammI is the difference between the maximum energy in the 0-2000 Hz and 2000-5000 Hz frequency bands. Equation 2.16 describes HammI.

$$Hamml = \frac{max(E_{0-2000Hz})}{max(E_{2000Hz-5000Hz})} \tag{2.16}$$

Where:

$E_{0-2000Hz}$ = the energy between 0–2000 Hz frequency bands
$E_{2000Hz-5000Hz}$ = the energy between 2000–5000 Hz frequency bands

Drop-off energy above 1000Hz

Drop-1000 denotes the relative amount of energy above 1000 Hz versus the low frequency range. See Equation 2.17.

$$Drop_{1000} = 10log\frac{\sum_{f-1000Hz}^{\frac{f_s}{2}} E_f}{\sum_{f=0}^{1000Hz} E_f} \tag{2.17}$$

Where:

fs = the sampling rate
E_f = the energy in the frequency band

Glottal to Noise Excitation Ratio

GNE calculates the glottal to noise excitation ratio which was first introduced by Michael et al. (**?**) in 1997. It is based on the correlation for Hibert which envelopes different frequency channels. GNE determines whether the voice signal was caused by a produced noise in the vocal tract or from vibrations of the vocal folds. Figure 2.5 following presents the steps for extraction of mean and standard deviation of GNE features.

For each speech signal, the following steps should be performed:

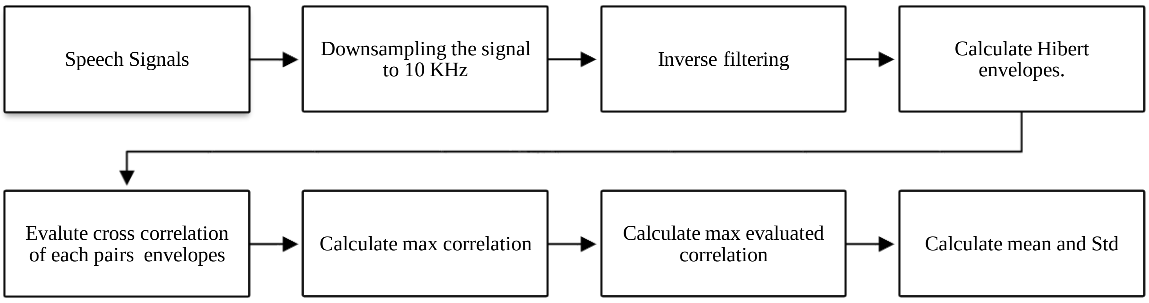

Speech Signals
Downsampling the signal to 10 KHz
Inverse filtering
Calculate Hibert envelopes.
Evalute cross correlation of each pairs envelopes
Calculate max correlation
Calculate max evaluated correlation
Calculate mean and Std

Figure 2.5: GNE Procedure

1. Down- sampling the signal to 10 KHz.
2. Inverse filtering of the speech signal to detect glottal pulses.
3. Calculate the Hibert envelope bands.
4. Evaluate the cross correlation function between such envelopes where the central frequencies of the band are greater than half of the bandwidth.

5. Pick the maximum value of each correlation between pairs of the frequency bands.
6. Pick the maximum from step 5 (which is the GNE for the time window).
7. Compute the mean and the standard deviation of resulting vector.

Mel-frequency Cepstral Coefficients

MFCC is based on the human auditory perception system that does not follow a linear scale of frequency. MFCC is one of the most widely-used features in speech recognition because of its superior performance over other features (Caballero-Morales, 2013). The MFCCs are robust, contain much information about the vocal tract configuration regardless of the source of excitation, and can be used to represent all classes of speech sounds (Pao et al., 2005a).

Extracting features using MFCC techniques are presented through six processes, namely: Pre-emphasizing; Framing and Windowing; Fast Fourier Transform; Mel-Frequency Filter Bank; Logarithm; and Discrete Cosine Transform. The whole processes involved in the MFCC technique are shown in Figure 2.6 following.

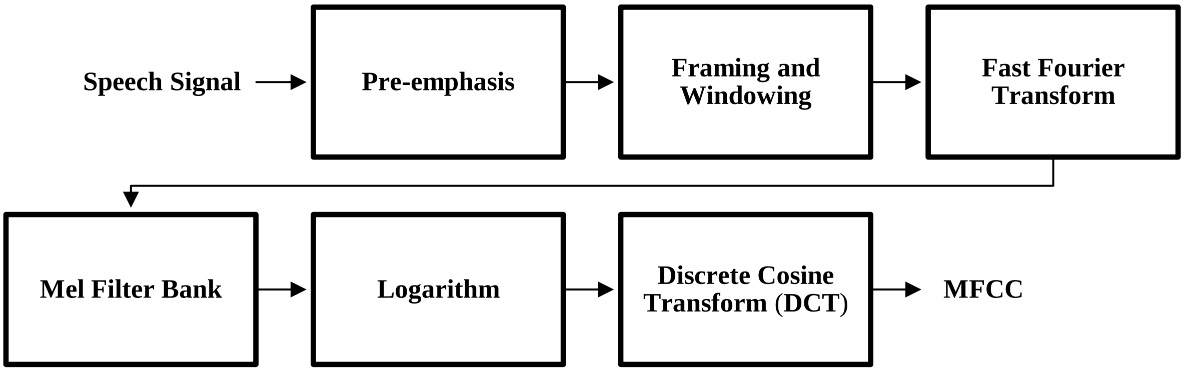


Figure 2.6: MFCC Feature Extraction Technique

In the first process, the speech signal is pre-emphasized using a high-pass finite impulse response (FIR) filter of order 1. In the framing and windowing process, the length of the frame was set to 25 ms and the frame was shifted by 10 ms. For windowing, Hamming window function is used. Following that, Fast Fourier Transform converts each frame of the input speech signal from time domain into frequency domain. The result after this process is often referred to as spectrum or period gram and to obtain a good frequency resolution, a 512 point Fast Fourier Transform (FFT) is used. In Mel Frequency Filter Bank process, a filter bank is created by calculating a number of peaks, uniformly spaced in the Mel-scale. It is then transformed back to normal frequency scale. These are used as peaks for the filter banks. Next, in the Logarithm process, the logs of the powers at each of the Mel frequencies are calculated. Following this, Discrete Cosine Transform (DCT) process is used to achieve the Mel- cepstral coefficients.

Log Mel Frequency Band

Mel frequency is a description of the short-term power of a sound by taking the logs of the powers at each of the Mel frequencies. A frequency in Mel is a logarithmic function of the frequency in Hertz. This is described in Equation 2.18.

$$mels = 2595log(1 + \frac{hertz}{700} = 1127ln(1 + \frac{hertz}{700}) \tag{2.18}$$

Line Spectral Pairs Frequency

Line spectral pairs (LSP) (Kabal and Ramachandran, 1986) are derived from linear prediction coefficients (LPC) for transmission over a channel. Since LSP has less sensitivity to quantization noise, that makes it superior to LPC. LSP is obtained by decomposition of the

LPC coefficient polynomial into a symmetrical and asymmetrical part. Figure 2.7 below shows the decomposition procedure.

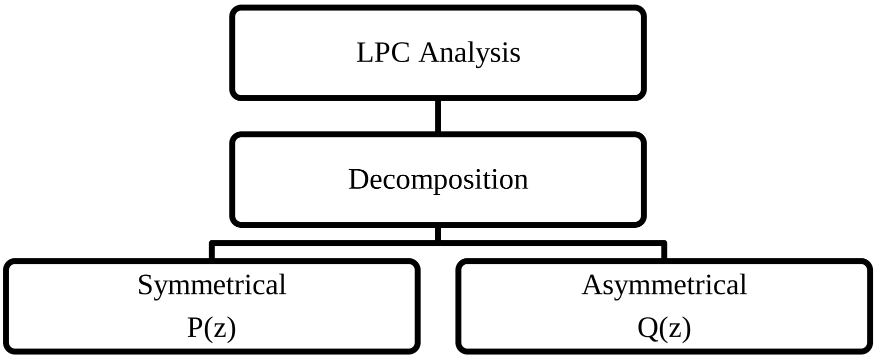


Figure 2.7: LSP Decomposition Procedure

In the z-domain $H(z)$ the two polynomials $P(z)$ and $Q(z)$ are illustrated by the following Equations:

$$P(z) = H(z) + z^{-(p+1)}H(z^{-1}) \tag{2.19}$$

$$Q(z) = H(z) - z^{-(p+1)}H(z^{-1}) \tag{2.20}$$

Where $P(z)$ and $Q(z)$ represent the vocal tract system with the glottis closed and opened, respectively.

2.8.2 Feature Selection

The IGR and CFS-PSO features selection algorithms are described in the following sections.

Information Gain Ratio

Information gain ratio(IGR) measures its importance and relevance to the class label. Computing the information gain for a feature involves computing the entropy of the class label for the entire dataset and subtracting the conditional entropies for each possible value of that feature. The entropy calculation requires a frequency count of the class label by feature value. All instances are selected with some feature value *e* and then the number of occurrences of each class within those instances can be counted. Following this, the entropy for *e* is computed. This step is repeated for each possible value e of the feature. The entropy of a subset can actually be computed more easily by constructing a count matrix, which tallies the class membership of the training examples by feature value. After calculating the information gain values of all features, the threshold of (0) was implemented. If the information gain values of the features are higher than the threshold, the features were selected; if not, the feature was not selected. Algorithm 1 below presents the algorithm of IGR implementation in more detail.

Correlation Selection Based Particle Swarm Optimization Search

Correlation-based feature selection (CFS) (Hall, 1999) is a subset-based filter feature selection algorithm. CFS selects features according to correlation-based function. It eliminates the irrelevant features that have low correlation with the class. In addition, it eliminates the redundant features that are highly correlated with one or more of the remaining features. The CFS feature-subset function is illustrated in Equation 19:

Particle swarm optimization (PSO) (Shi et al., 2001) is a population-based evolutionary computation technique inspired by social behavior simulation. The PSO system is initialized with a population of random solutions. This population searches for an optimal

Algorithm 1 Information Gain Ratio Algorithm

```
function IGR(F,E)
    return *Ranked Features*
    Input: Extracted Features ($f_i$) and Emotions Label ($e_i$)
    $F$ : domain of emotions
    $E$ : domain of features
    $p$ : Probability
    $H$ : Entropy
    $S = 0$;                                  // counter variable
    for each $e_i$ do:
        Calculate $p(e_i)$;                   // calculate the probability for each class
        $H_e = S + p(e_i)log_2(p(e_i))$;
        $S = H_e$;
    end for
    for each $f_j$ do:
        Calculate $p(f_j)$;                   // calculate the probability of value j for feature f
        $H_f = S + p(f_j)log_2(p(f_j))$;
        $S = H_f$;
    end for
    for each $e_i$  do:
        for each $f_j$  do:
            Calculate $p(e_i, f_j)$;
            $H_{ef} = S+p(e_i, f_j)log_2p(e_i, f_j)$; // calculate the relative entropy $e_i$ given $f_j$
            $S = H_{ef}$;
        end for
    end for
    $H(E, F) = (-1) * H_f * (-1) * H_{ef}$ ;
    $IGR = H_e - H(E, F)$
end function
```

solution by updating generations. In PSO, a potential solution is called a particle. Each particle makes use of its own memory and knowledge gained by the swarm as a whole to find the best (optimal) solution in a d-dimensional search space. The particles have a positional value and velocities which direct their movement.

CFS-PSO feature technique is used to form the fitness functions and evaluation of goodness of the reduced feature subset. For a feature subset X with m features, $x = (x_1, x_2, x_3...x_m)$, CFS first calculates a matrix of feature-class and feature-feature correlations from the training data using Equation 19 and then searches for feature subset space using PSO search. In PSO, each candidate solution of the problem is represented as a particle, which is encoded by a vector or an array. Particles move in the search space to search for the optimal solutions. During the movement each particle can recall its best experience. The entire swarm scans for the ideal (optimal) arrangement by refreshing the position of every particle based on their best understanding and its neighbouring particles. Algorithm 2 below shows the algorithm for CFS-PSO.

2.8.3 Classification

only the support vector machine classifier will discuss her.

Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm that can be employed for classification and regression purposes. It goal is to create a separating line (hyperplane) which separates the data classes. However, there are many possible hyperplanes that could be chosen. The best hyperplane is one that maximizes the margin between the separating lines.

Algorithm 2 CFS-PSO Feature Selection

```
function CFS-PSO(F)
    return *Selected Features*
    Input: Full Features set($f_i$)
    $x$ : position of each particle;
    $v$ : velocity of each particle ;
    $p$ : particle;
    $s$ : stop criteria;
    $x_p$ : personal best position;
    $x_g$ : global best position;
    $w$ : inertia weight;
    $c_1, c_2$ : acceleration constants or learning parameters;
    $r_1, r_2$ : random values between (0,1);
    $d$ : direction;
    $k$ : features in subset;
    $f_c$ : mean of features-class correlation;
    $f_f$ : average of feature-features inter-correlation;
    for each $p_i$ do:
        initialize $x_i and v_i$
    end for
    while $s$ not met do:
        $M_s = kf_c/\sqrt[2]{k + (k(k-1)f_f}$
        for each $p_i$ do:
            get $x_p$;
            get $x_g$;
        end for
        for each $p_i$ do:
            for each $d_i$ do:
                $v_i(t+1) = * v_i(t) + c_1 * r_1(x_g(t)x_i(t)) + c_2 * r_2(x_p(t)x_i(t))$; // update velocity
                $x_i(t+1) = x_i(t)+v_i(t+1)$ ; // update position
            end for
        end for
    end while
    Calculate the accuracy of the selected features subset;
    Return the selected features subset;
end function
```

A Library for Support Vector Machines (LIBSVM) is part of the open source machine learning libraries, developed at the National Taiwan University. LIBSVM implements the sequential minimal optimization (SMO) algorithm for kernelized SVM.

SMO was proposed in 1998 as a new algorithm for training SVM by Platt (Platt, 1998). It breaks the large numerical quadratic programming (QP) used by the previous SVM learning algorithms into a series of smallest possible QP problems. These are then solved analytically.

Assuming that a training set (x_1, y_1),..., (x_{n}, y_{n}), where x_{i} is the input vector y_{i} and y_{i} $\in$ -1, +1 is a binary label corresponding to it. The QP problem to train an SVM is shown in Equation 2.21 below:

$$\sum_{\alpha_i}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j \tag{2.21}$$

That subject to $0 \leq \alpha_i \leq C$, for $i = 1, 2, , n$:

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \tag{2.22}$$

Both hyperparameter and kernel functions are supplied by the user while the variables i are Lagrange multipliers.

As mentioned above, SMO breaks the QP problem into a series of smallest possible subproblems. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers α_1 and α_2, the constraints are reduced as described below:

This reduced problem can be solved analytically: one needs to find a minimum of a one-dimensional quadratic function. The letter k indicates the negative of the sum over the rest of terms in the equality constraint, which is fixed for each iteration. The algorithm proceeds as follows:

1. Find a Lagrange multiplier 1 that violates the KarushKuhnTucker (KKT) conditions for the optimization problem.
2. Pick a second multiplier α_2 and optimize the pair (α_1, α_2).
3. Repeat steps 1 and 2 until convergence occurs.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem is considered to be solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence. This is critical for large data sets since there are n(n-1)/2 possible choices for α_i and α_j.

2.9 Summary

This chapter presents the issues that faced SER features-based researchers. Some of the important existing works conducted by researches in features extraction and selection have been outlined. It was observed that the existing features extraction technique and selection methods were not sufficient to obtain the optimal set of feature; hence, a better model for features extraction and selection was required. The next chapter describes the research methodology.

CHAPTER 3

Research Methodology

3.1 Introduction

This chapter describes and discusses the research methodology for the proposed SER framework. It begins with an overview of the theoretical model. Secondly, an overview of the research framework is provided. Following this, the speech emotional datasets used in this study are discussed. Finally, the experiment environment is described along with the performance measurements.

The main research question is:

- How to extract and select a compact set of features that improve the SER performance?

The thesis project is divided into five phases with their own sub research questions. The answers to these questions will lead to a conclusion to the main research question. The research methodology is based upon constructive research and action research.

3.2 Overview of the Theoretical Model

Firstly the research questions solution will be built using constructive research. This solution is built in form of a model from the existing theories which collected by performing a comprehensive survey of a diverse source. Then the model should be tested for its practical relevance and its theoretical contribution. Figure 3.1 illustrates the constructive research process.

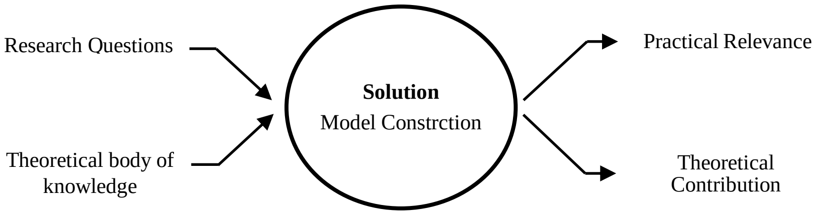


Figure 3.1: Overview of the constructive research methodology process

A theoretical body of knowledge is created using a literature survey from peer-reviewed research and published literature. The literature survey is presented in Chapter Two. The developed theoretical model seeks to be the solution to the main research question. The goal is to identify the relation between features and SER performance. Therefore, the questions of this phase are:

- *Question (1)*: What is the impact of features in SER recognition accuracy?
- *Question (2)*: What consideration should be given in designing features extraction techniques and selection methods?

3.3 Overview of the Research Framework

The action research methodology which combines theory and practice used to construct the proposed framework and practical evaluation. Figure 3.2 illustrates the action research process.

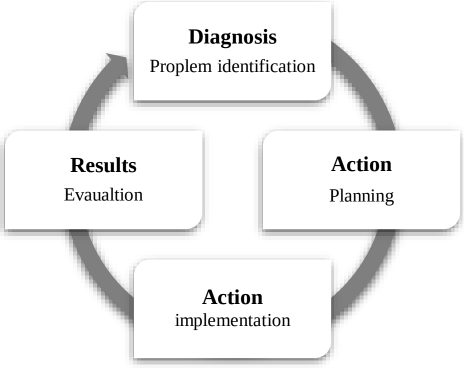


Figure 3.2: Overview of the action research methodology process

The existing features sets are diagnosed to identify problems and the current situation. Actions then are taken to construct the framework. Finally, the results are evaluated. This process is done iteratively and the phases provide feedback to each other.

3.3.1 Phase (1): Investigated SER Existing Features Extraction Techniques Capability

In this phase three acoustic features extraction techniques examine, to investigate its capability to recognize emotions. The research question is:

- *Question (3)*: What is the impact of the different types of features in emotion recognition?

Selective features extraction approach has been used for voice quality, prosodic and spectral features. This approach will be discussed in Section 1.6.2.

3.3.2 Phase (2): Identify the Best Features Extraction Technique Combination

This phase investigates different features combination to determine the best features combination. The research question is:

- *Question (4)*: What is the best feature extraction technique combination that is appropriate for emotion recognition?

3.3.3 Phase (3): Enhance the Features Extraction Technique

An enhanced features extraction technique has been proposed and examine in this phase. This enhanced technique named voice quality prosodic spectral technique (VQPS). The research question is:

- *Question (5)*: How to enhance the feature extraction technique that obtains a better representation of emotion?

The VQPS use the features combination that proves successful in the last phase and tries to enhance the recognition accuracy by using traditional and new voice quality features. the traditional features extracted using a brute force approach which will be discussed in section 1.6.2.

3.3.4 Phase (4): Reduced Features Dimensionality and Identify the Significant Features

In this phase, a hybrid selection method has been designed to reduce the features space and select the final set of features. The research question is:

- *Question (6)*: How can a better feature selection method be designed?

The hybrid selection method named hybrid filter-based feature selection method (BHFFS).

3.3.5 Phase (5): Comparison and Benchmark

In this phase, the propose SER framework performance is compared with the previous model.

3.4 Emotional Speech Dataset

The first step in developing the SER system is the selection or collection of an emotional speech dataset. In recent years, many emotional speech datasets have been built for speech emotion research; some of these datasets are standard and publicly available for researchers. While many of them used personal datasets collected by some researchers to fulfil their needs in developing SER. Three different types of datasets have been used in SER studies, namely: acted; spontaneous; and elicited datasets.

The acted (simulated) datasets were recorded with the help of professional actors. The actors were asked to act or simulate pre-defined emotions using ready-made scripts. The most popular acted datasets are the Danish emotional speech dataset (DES) (Engberg et al., 1997) and the Berlin emotional speech dataset (EMO-DB) (Burkhardt et al., 2005).

The spontaneous (natural or real) datasets convey a real emotion that is recorded in the real world usually without the knowledge of the speakers. They can be collected by recording conversations from public places such as: call centers; aircraft cockpits; TV talk shows; oral job interviews; and doctor-patient conversations, etc. The best example of the spontaneous dataset is the speech under simulated and actual stress (SUSAS) dataset (Hansen and Bou-Ghazale, 1997).

The elicited datasets were recorded form speakers who made to react in an artificial situation without their knowledge. Emotional elicitation procedures can include verbal labels, scenarios or photographs shown to a speaker. This is a tedious kind of datasets; hence, until today, there have been only a very few elicited emotional speech datasets. Examples of datasets collected in this way are the German SmartKom dataset (Schiel et al., 2002) and the German FAU Aibo Emotion Corpus (Batliner et al., 2008).

The three types of datasets mentioned above serve different purposes. The first type is suitable for theoretical researches while the second and third types can be helpful in creating real-life applications. However, the collection of spontaneous and elicited datasets is not that easy. On the other hand, the acted dataset is easy to collect and control and does not face ethical issues.

In recent years, a new type of dataset was shown which contained audio and visual emotions. The English Surrey audio-visual expressed emotion (SAVEE) (Jackson and Haq, 2014) and the SEMAINE dataset (McKeown et al., 2010) are examples of this type of dataset.

For this research, three acted datasets, namely, the Danish (DES), the German (EMO-DB) and the English (SEVAA) datasets were initially selected. Both EMO-DB and SEVAA are publicly available so they could be directly downloaded. DES is publicly available under an agreement. Many emails seeking permission to access the dataset were sent, with no response received. In the case of SEVAA, no benchmark works were found for compression. In addition, it showed a poor result; hence, EMO-DB dataset is the only dataset that was utilized in this research.

3.4.1 Berlin Dataset of Emotional Speech

This section provides details about the EMO-DB dataset which contains utterances spoken in German. This dataset is freely and publicly available and can be directly downloaded without a request. It has been used by many researches in SER, (Alonso et al., 2015).

EMO-DB recorded at the Department of Acoustic Technology of Technical University of Berlin in Germany and funded by the German Research Community. It was recorded using a Sennheiser microphone at a sampling frequency of 16 kHz, with the help of ten professional actors (five male and five female) who were asked to simulate seven emotions. These emotions were, namely: anger; anxiety; boredom; disgust; happiness; sadness; and neutrality. They did so using ten utterances as shown in Table 3.1 and Table 3.1; specifically, five short and five longer sentences that can be used in daily communication and can also be said with all the emotions. About 800 such utterances were recorded.

After recording the dataset, twenty judges were asked to listen to the utterances in a random order, in front of a computer monitor. They listened to each sample only once, before they decided which emotional state the speaker had been in. After selection, the dataset contained a total of 535 speech files.

As shown in Table 3.2, EMO-DB is an imbalanced dataset. This means that not all the emotions have the same number of recorded samples; the highest number of samples being for the emotion of anger (127), and the lowest being for the emotion of disgust (46). All

Table 3.1: EMO-DB Dataset Sentences

Code	Sentences in German	Sentences in English
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the frigde.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend knnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stck Papier befindet sich da oben neben dem Holzstck.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das fr Tten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

the available information regarding the speech dataset can be accessed via the internet.

Table 3.2: Number of Emotions in the EMO-DB Dataset

Emotion	Anger	Anxiety	Boredom	Disgust	Happiness	Sadness	Neutral
Records	127	69	81	46	71	62	79

3.4.2 Dataset Preparation

EMO-DB can be downloaded directly from the internet (http://emodb.bilderbar.info/index-1024.html). The 535 wav file was downloaded into one folder that has been separated into seven folders according to the emotion classes to facilitate labeling after the features extraction process. Every file in the dataset is named according to the same scheme. Every file name is seven (7) digits long where digits 1-2 indicate the speakers name. Digits 3-5 indicate the text code (see Table 3.1). Digit 6 indicates the emotion named by the German word for emotion. Finally, digit 7 signifies whether there is more than one version of the sentence. Figure 3.3 illustrates an example of a naming scheme for 03a01Fa.wav.

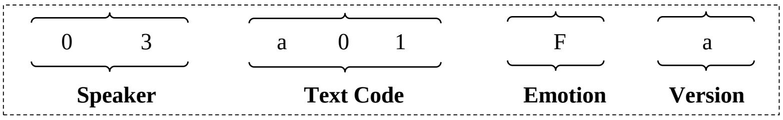


Figure 3.3: File naming scheme for (03a01Fa.wav)

After arrangement of the dataset files, the second step in dataset preparation should be noise cancellation. However, EMO-DB conceder has a standard dataset that records in a studio with high-quality equipment so it does not need that step. In fact, the researchers who wanted to test the SER system under noise conditions were required to add noise to the dataset.

3.5 Experimental Tools

For this research, a number of tools were used in developing SER as summarized in Table 3.3. The Praat tool (5.3.84)(Boersma et al., 2002) and openSMILE toolkit (2.3.0) (Eyben et al., 2013) were used to preprocess and extract the features from the speech samples; while WEKA (3.7.12) (Hall et al., 2009) was used to implement the classification and selection algorithms. In addition, it was used for dataset balancing and features normalization

Table 3.3: The Tools used in this Research

Tools	Description of usage
Praat 5.3.84	Implement the features extraction algorithms
openSMILE 2.3.0	Implement the features extraction algorithms
WEKA 3.7.12	Implement the section and classification algorithms

As mentioned above, two tools have been used for extracting acoustic features from the speech signal; the openSMILE and the Praat tool respectively. The next sections will discuss the limitations and the need for this selection. However, the literature also present works that use more than one extraction tool; for instance, (Ramakrishnan and El Emary, 2013) use Praat and openSMILE for features extraction.

3.5.1 OpenSMILE Toolkit

Open Speech and Music Interpretation by Large Space Extraction (openSMILE) is a command line tool written in C++ for signal processing and machine learning applications. It can be used with various platforms such as Linux, Windows, and MacOS; openSMILE also supports various data formats commonly used in the field of data mining and machine learning. However, regarding to this research it has a limitation on the extraction of voice quality features. According to openSMILE documentation, probability of voicing, jitter, and shimmer are the voice quality features that can be computed by openSMILE. Different works in SER make use of the openSMILE toolkit in their research; for example (Chen et al., 2012b; Tickle et al., 2013; Mariooryad et al., 2014; Cao et al., 2015; Yogesh et al., 2017; Klaylat et al., 2018).

3.5.2 Praat Toolkit

The Praat toolkit is a free graphical user interface package that is used for the recording and analysis of speech signals. Different voice quality, prosodic and spectral features can be extracted using this tool. As with openSMILE, Praat also supports various platforms including Linux, Windows and MAC. Some of the researchers who use Praat in there research studies include (Zhao et al., 2014; Tomar et al., 2014; Verma et al., 2016; Shirani and Nilchi, 2016; Liu et al., 2018). The main strength of Praat is that it has an extensive help function that is updated constantly. Praat also offers its own scripting languages which are small programs that compensate for missing functions. Different Praat scripts were developed for this research. Figure 6 following shows an example of a script used to read all audio files from a specific folder.

3.5.3 WEKA Toolkit

Waikato Environment for Knowledge Analysis (WEKA) Toolkit is an open source machine learning toolkit implemented in the Java language that contains a collection of machine learning algorithms and tools for data pre-processing. It is powerful software that contains a variety of tools for data processing and a machine learning algorithm. Various researchers use WEKA in implementing and testing their SER system; notably, the works of (Tickle et al., 2013; Oflazoglu and Yildirim, 2013; Mariooryad and Busso, 2014; Verma et al., 2016). For classification purposes, the A Library for Support Vector Machines (LIBSVM) (Chang and Lin, 2011) from WEKA was implemented.

3.6 Features Preparation

In this section, the preparation process for audio files will be present. This involves features pre-processing, features extraction approach, the features file format and features post-processing.

3.6.1 Features Pre-processing

Features pre-processing is an important step that formulates the speech signal before features extraction; it usually involves framing, windowing and removing noise from the speech signal. However, as mentioned before in Section 3.4.2, EMO-DB does not has noise.

Framing

The speech signal is segmented into several frames before extracting the features 100 frames per second ware used.

Windowing

After framing a window applied to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. Hamming window of 25 ms size was used for all features except for pitch Gaussian window of 60 ms size was used.

3.6.2 Features Extraction Approach

The features extraction approach taken in this study is a mixed approach based on selective and brute force approaches respectively. The selective approach is based on choosing the features that have been used in related work and have proved effective; while the brute force approach is based on extracting a large number of features in the assumption that some of them will be found valuable.

The Selective Approach

The selective features are chosen from INTERSPEECH features sets. Schuller et al. introduced five (5) different acoustic feature sets for SER, namely: INTERSPEECH 2009 (Schuller et al., 2009); INTERSPEECH 2010 (Schuller et al., 2010); INTERSPEECH 2011 (Schuller et al., 2013). These feature sets were used in literature studies by several researchers in SER. For the purpose of benchmarking and comparison of the result obtained from other researchers results (especially for the features selection) the INTERSPEECH 2010 feature set was used. Many researchers adopted INTERSPEECH 2010 in their work (Chen et al., 2012b; Oflazoglu and Yildirim, 2013; Mariooryad et al., 2014; Sun and Wen, 2015; Sun et al., 2015; Yogesh et al., 2017).

The feature set covers the three feature categories: voice quality (jitter, shimmer and voicing); prosodic (F0 and loudness); as well as spectral (MFCC, LSP, and log Mel freq. band) features. It uses 21 statistics functions applied to 76 LLD (38 LLD with 38 corresponding deltas). The number of pitch onsets and the total duration of the input are then appended (two (2) features). A set of 1582 acoustic features are subsequently the results. Table 1 following shows the features and the feature statistics functions.

The *emobase2010.conf* configuration file has been used to extract all features in ARFF format.

The Brute Force Approach

The brute force feature set was extracted using the Praat and openSmile toolkit. A total of 181 voice quality features have been extracted including: harmonic; voicing; jitter;

Table 3.4: The Selective Approach Features Set

Feature	Descriptors	Functions
	Voicing probability	Lin.regression error Q/A
Voice Quality	JitterLocal, jitterDDP	Percentile range (99-1)
	ShimmerLocal	Up-level time75/90
	PCM loudness	MaxPos, minPos, mean,
Prosodic	F0 envelop	Lin.regression coeff.1/2
	F0final	percentile 1/99
	MFCC[0-14]	Stddev, skewness, kurtosis,
Spectral	Log Mel freq. band[0-7]	Quartile1/2/3
	LSP frequency[0-7]	Quartile range (2-1)/(3-1)/(3-1)

shimmer; autocorrelation. Table 3.5 following shows features extracted by a brute force approach.

The New Approach

The brute force feature set was extracted using the Praat toolkit. A total of 6 voice quality features have been extracted including: Hamml; GNE; voice to total frame ratio and Do1000. Table 3.6 following shows features extracted by a brute force approach.

3.6.3 Features File Format

After features extraction, the features set were required to be saved in an appropriate file format. WEKA tool primary file format is the Attribute-Relation File Format (ARFF) which is encoded using the American Standard Code for Information Interchange (ASCII)

Table 3.5: The Brute Force Approach Features Set

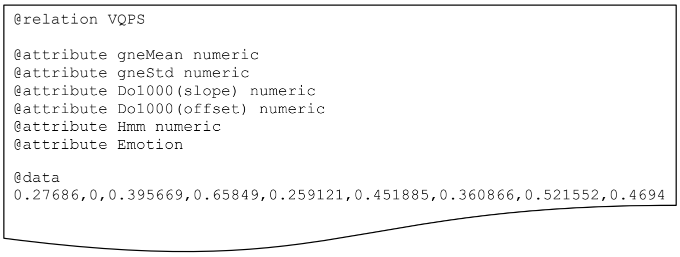
Description	Function
Harmony	min, max, range, mean, std
HNR	mean
NHR	mean
Voice break	number, degree
Period	number, mean, std
Pulses	number
Voicedunvoiced frames	ratio
Unvoicedtotal frames	ratio
Jitter	local, ddp, local absolute, rap, ppq5,, mean, std
Shimmer	Local, local dB, apq3, apq5, apq11, ddp
Autocorrelation	mean
Voicing probability	Lin.regression error Q/A

Table 3.6: The New Approach Features Set

Description	Function
GNE	mean, std
Pulses	number
Voicedtotal frames	ratio
Do1000	slope, offset
Hmmel	Hmmel

and defines a list of instances along with the individual attributes that those instances share.

The openSMILE tool has the ability to save the extracted features to ARFF file format directly. However, this is not the case for the Praat tool that saves the extracted features in a text file format which has to transfer to Excel format and subsequently to a comma-separated values (CSV) format. Finally, the file is transferred to ARFF format using WEKA tool. Figure 3.4 gives a sample of ARFF file denoting five attributes.



```
@relation VQPS

@attribute gneMean numeric
@attribute gneStd numeric
@attribute Do1000(slope) numeric
@attribute Do1000(offset) numeric
@attribute Hmm numeric
@attribute Emotion

@data
0.27686,0,0.395669,0.65849,0.259121,0.451885,0.360866,0.521552,0.4694
```

Figure 3.4: ARFF File Sample)

3.6.4 Features Post-processing

Before the extracted features can be inputted as set to the classifier, they must undergo post-processing so as to be in a better format for classification. Post-processing in the form of features scaling or normalization was used.

Missing Data Handling

After the end of the extraction process it was notes that a big number of features have many missing data. It has been handling in two way. first the features that have missing data larger than 20% it have been discarded. second for the features that have missing data less than 20% ReplaceMissingValue filter in Weka tool has been used.

Features Normalization

Some of the extracted features may vary more widely than other features. This renders the classification less effective because the decision may be made based on one feature alone. To avoid this normalization, [0-1] was applied in order to have similar distance between all features and ensure that every feature contributes as equitably as possible.

3.7 Validation

K-fold cross-validation was used in this experiment for validation purposes. This validation method has been used in many other works regarding EMO-DB (Schuller et al., 2005; **?**; **?**; Ahmad, 2016; **?**; Manolov et al., 2017). In this process, the dataset is divided into k subsets. Each time, one of the k subsets is used as the test set, and the other k-1 subsets form the training set. Error statistics are calculated across all k trials, specifically k=10 being widely used.

3.8 Evaluation Metrics

To evaluate the improvement of SER, several methods have been used to evaluate the performance in the area of emotion recognition. Two methods were used here, specifically: the confusion matrix which gave accuracy to the individual classes; and the overall accuracy. These methods are commonly used in SER evaluation.

3.8.1 Confusion Matrix

The confusion matrix is a visualization of the performance of supervised learning algorithms. It is used to show the relationships between actual and predicted classes. This is performed by presenting the number of correct and incorrect classes predicted by the model, compared with the actual classes in the test data. The confusion matrix is n-by-n, where n is the number of classes, with the rows of the matrix representing the instances in an actual class, and the column of the matrix representing the instances in a predicted class. Table 3.6 following shows an example of the confusion matrix for the classification model, which has been used to classify two classes, specifically yes and no.

Table 3.7: Example of the confusion matrix

		Predicted Classes	
		Yes	**No**
Actual Classes	**Yes**	a	b
	No	c	d

For this example, the entries in the confusion matrix have the following meaning:

- a is the number of correct predictions where an instance is yes.
- b is the number of incorrect predictions where an instance is no.
- c is the number of incorrect predictions where an instance is yes.
- d is the number of correct predictions where an instance is no.

All correct predictions are located on the diagonal of the table, so it is easy to visually inspect the table for errors. These will be represented by values outside the diagonal.

3.8.2 Classification Accuracy

The classification accuracy is the percentage of correctly classified instances over the total number of instances. It is determined using equation 3.1.

$$Accuracy = \frac{CorrectPredictedInstance}{TotalInstance} \tag{3.1}$$

Two accuracy measurements were used, including individual class accuracy, which was calculated for every emotion, and overall accuracy respectively.

3.9 Summary

This chapter outlines the theoretical model and the research framework adopted in this study and provides a guideline for detailed investigations provided in Chapters 4. Furthermore, in this chapter, the background to the datasets used in the research was described, including the tools used and how they were used. Finally, the evaluation metrics which will be used in Chapters 4 were mentioned.

CHAPTER 4

Results and Discussions

4.1 Introduction

In this chapter the results achieved by following the research methodology which described in Chapter 3 are presented. The following sections describe the output from research in each phases and seek to answer the research questions.

4.2 Creation of the SER Features model

This section presents the theoretical model to the research questions of phase 1, described in the research methodology. The model was named the SER features model and it is based on the theoretical body of knowledge gained in Chapter 2. The purpose of this model is to describe the consideration in designing features extraction techniques and selection methods that result in improving the SER recognition accuracy.

Figure 4.1 illustrates the proposed SER features model. The model blocks are built according to the theoretical knowledge to illustrate the necessary elements needed for a good foundation in SER features.

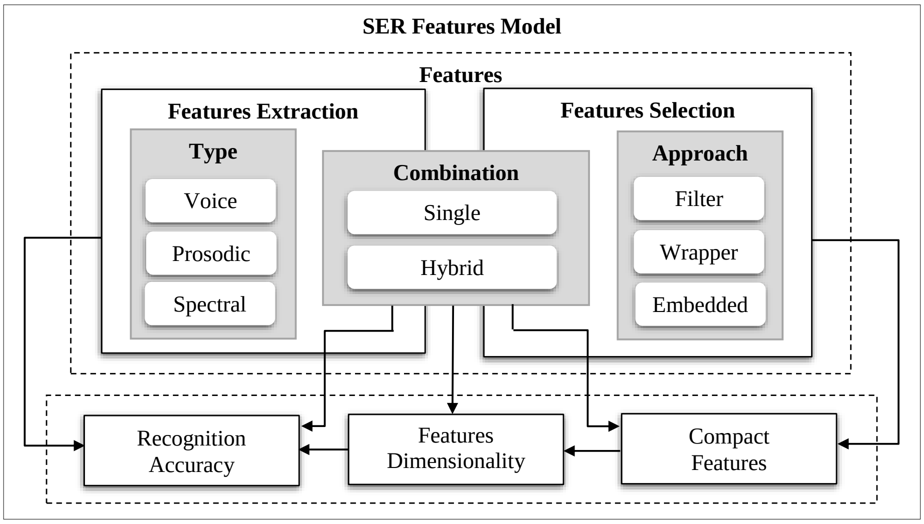


Figure 4.1: The SER Features Model

The main contribution of the model is its structuring of existing theories and concepts. The following is a description of each blocks role in the model.

1. **Features:** Features processes in recognition, mainly categorizes into two the features extraction and selection process. Therefore, this block is located at the bottom of the model to symbolize the needs to be considered in all decision made in SER developments.

   (a) **Feature Extraction:** The first process in any speech recognition system that aims to transfer the audio file to a vector of features.

   - **Feature Extraction Technique Type:** This block is for the types of features that should be used in SER. This features types affect directly the recognition accuracy. The feature extraction technique types are found in Section 2.5.1

    - **Feature Extraction Technique Combination:** This block is for the combination that made from different types of features that either to be from one single type or from hybrid type. This combination affect also the recognition accuracy directly or throw affecting features dimensional. The different feature extraction technique combination in SER are found in Section 2.7.1.

   (b) **Features Selection:** Came after features has been extracted. Feature selection aims to select significant features and reduce search space which improves the accuracy of SER.

    - **Selection Methods Combination:** Implementing single or hybrid selection method affects the final compact feature set and the dimensional of this set. The selection methods taxonomy is found in Section 2.5.2.

    - **Selection Methods Approach:** Different selection approaches have been used for SER each of them has advantage and disadvantages that could affect the final compact features set. The selection methods used in SER is found in Section 2.7.3.

2. **Features Effect:**

   - **Compact Features:** Since not all features contribute equally to emotion recognition, it is desirable to identify a compact feature set that is necessary to be analyzed. This feature set affected by the combination of the selection method single or hybrid, and it also affected by the type of the approaches that used in features selection. in the other hand, this feature set affects the features Dimensionality directly.

- **Features Dimensional:** Features Dimensional: reducing features dimensionality is always a big concern in recognition systems. the growing of the features sets size or dimensional help in improving recognition accuracies in many cases. Unfortunately, the big growth of the feature set could lead at many cases to harming the classifier performance which leads to decries the recognition accuracy. in addition to the effect of the high dimensionality in the recognition accuracy, it also wastes memory and complicates the system design.

- **Recognition Accuracy:** The accuracy is a major performance evaluation method in SER. Although some researchers use addition evaluation methods. The individual and overall emotion recognition accuracy were Remains the most important evaluation methods SER. The details about recognition accuracy are found in Section 3.5.

4.3 Creation of the SER Framework

Using the theoretical model an SER framework proposed. As illustrated in Figure 4.2. EMO-DB dataset is the input of the framework and its details is provided in Section 3.4.

The SER framework content of eleven tasks: (A) Pre-processing, (B) Features Extraction, (C) selective approach, (D) brute force approach, (E) New approach, (F) Combination, (G) Post-processing, (H) Features selection, (I) Classification, (J) Evaluation, and (K) Compar-

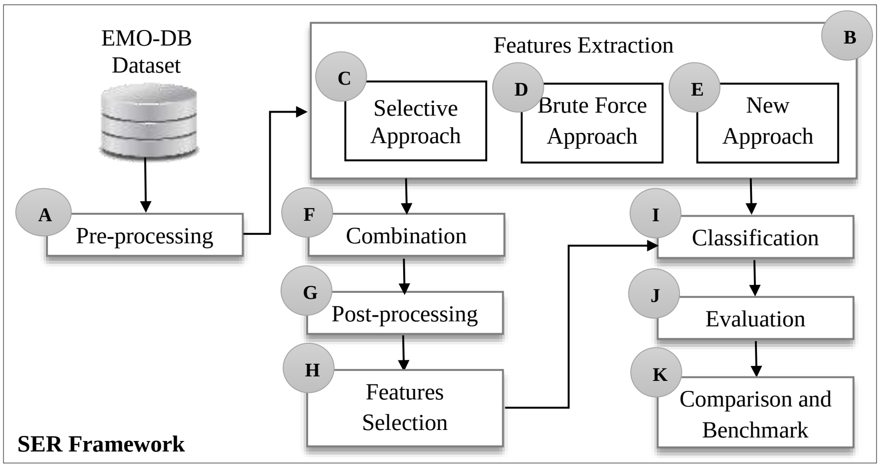


Figure 4.2: The SER Framework

ison and Benchmark. The description and explanation on each task is given below:

(A) Pre-processing: The features pre-processing involve framing and windowing as described in Section 3.6.1.

(B) Features extraction: Extracting three acoustic features namely: voice quality, prosodic and spectral features. The detail of this features types is reported in Section 2.5.1.

(C) Selective approach: A selective voice quality, prosodic and spectral features were extracted using the same method in INTERSPEECH 2010. The details about this approach and the features description are found in Section 3.6.2.1.

(D) Brute force approach: Used only for voice quality features by extraction all traditional features that prove success in previous work. The details about this approach and the features description are found in Section 3.6.2.2.

(E) New approach: Introduce six voice quality features three of them not used before in developing SER while the other not used with EMO-DB dataset. The details about this approach and the features description are found in Section 3.6.2.3.

(F) Combination: Four combination sets will examining

(G) Post-processing: The features post-processing involve normalization and missing value handling as described in Section 3.6.4.1.

(H) Features selection: Implement enhanced selection method that balances the dataset before the selection.

(I) Classification: Classify the emotion into the respective class using LIBSVM. The details about LIBSVM are found in Section 2.8.3.

(J) Evaluation: Two types of evaluation, the confusion matrix that provides the individual emotion recognition accuracy, and the classification accuracy that provides the overall recognition accuracy. The details about individual and overall recognition accuracy are found in Section 3.8.

(K) Comparison and Benchmark: A comparison with previously reported work that use different extraction technique and selection method..

Figure 4.3 show the tasks involved in first, second,third, forth and fifth phase respectively.

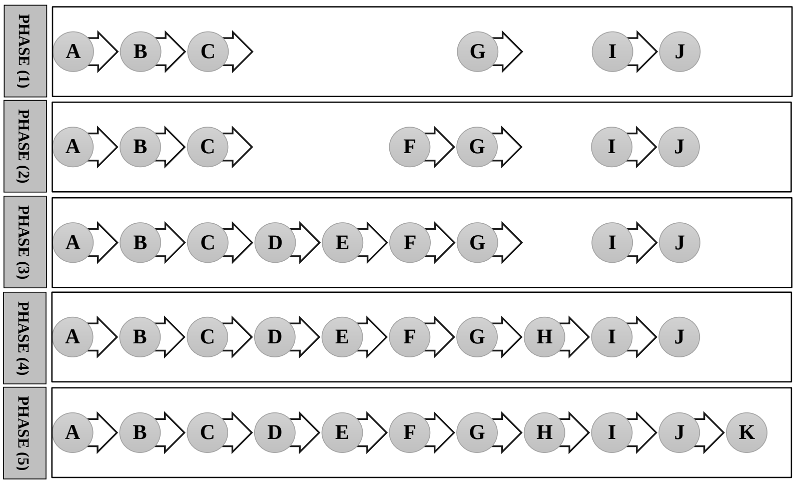


Figure 4.3: SER framework task Detail

4.3.1 Phase (1): Investigated SER Existing Features Extraction Techniques Capability

The purpose of this phase is to investigate different acoustic features extraction technique capability in recognize emotions. The phase tasks are given in Figure 4.2. Selective approach used to evaluate voice quality (VQ_S) prosodic (P_S) and spectral features (S_S) individually.

After pre-processing EMO-DB dataset audio signal as mentioned in Section 3.8, a total of 151 selective voice quality features, 114 selective prosodic features, and 1302 selective spectral features were extracted using openSMILE toolkit as shown in Table 4.1 , 4.2 and 4.3.

Table 4.1: Description of Extracted Selective Voice Quality Features

Feature Sequence	Description
31-104	Jitter (local, ddp)
105-141	Shimmer (local)
142-181	Voicing

Table 4.2: Distribution of Extracted Selective Prosodic Features

Feature Sequence	Description
182-222	loudness
223-295	Pitch(F0)

The extracted features then post-process as mentioned in Section 3.6.4 before classification. Finally, the recognition accuracy of the individual features type was evaluated by comparing the results obtained.

Table 4.3: Distribution of Extracted Selective Spectral Features

Feature Sequence	Description
296-925	MFCC
926-1261	Log Mel freq. Band
1262- 1597	LSP frequency

From the results obtained, it can be observed that the accuracy obtained using S set was found to be much higher than that of the VQ_S and P set. An overall accuracy of 87.10% could be achieved with S set; while an overall recognition of 64.11% and 69.91% could be achieved with VQ_S and P respectively. Table 4.4 below summarizes the comparison of results based on the number of features and recognition accuracy obtained from the two feature sets.

Table 4.4: Comparison of Results obtained with VQ_S , P and S Sets

Features Set	No. of Features	Accuracy (%)
VQ_S	151	64.11
P	114	69.91
S	1302	87.10

The results indicated that spectral features are more useful in improving recognition accuracy. This result is compatible with some reports in the literature that claim that spectral features technique is more useful in emotion recognition. However, given the size (dimensionality) of the extracted spectral features, it can be noted that the size of the spectral features is much larger than the other type features size. This could be the reason for the improvement progress. It also indicated that voice quality features result in the worst recognition accuracy and this also confirm the reported results in literature even with a bigger size of features than prosodic features it gives a lower result than prosodic features. However, in contrasts with some reports in literature the different between voice quality feature accuracy and prosodic features accuracy is not huge.

Amongst VQ_S emotions, sadness achieved a maximum recognition of 91.94% and the emotion of happiness achieved a minimum recognition of 26.76%. The confusion matrix of results obtained is illustrated in Table 4.5 following.

Table 4.5: Confusion Matrix obtained with VQ_S Set

Emotion	**Anxiety**	**Anger**	**Disgust**	**Boredom**	**Sadness**	**Natural**	**Happiness**
Anxiety	40	11	3	3	3	3	6
Anger	7	98	5	1	0	1	15
Disgust	1	5	21	3	1	10	5
Boredom	0	2	2	60	5	9	3
Sadness	1	0	0	3	57	1	0
Natural	6	1	3	12	6	48	3
Happiness	6	36	4	5	0	1	19

In relation to P, the emotion of sadness achieved a maximum recognition of 88.71% while the emotion of disgust achieved a minimum recognition of 43.48%. The confusion matrix of results obtained is shown in Table 4.6 following.

Table 4.6: Confusion Matrix obtained with P Set

Emotion	**Anxiety**	**Anger**	**Disgust**	**Boredom**	**Sadness**	**Natural**	**Happiness**
Anxiety	44	9	6	1	1	6	2
Anger	8	95	4	0	0	1	19
Disgust	4	6	20	3	1	10	2
Boredom	0	0	2	67	4	7	1
Sadness	1	0	1	3	55	2	0
Natural	1	0	3	6	3	62	4
Happiness	6	29	2	1	0	2	31

In relation to S_S, the emotion of sadness achieved a maximum recognition of 91.94% while the emotion of happiness achieved a minimum recognition of 73.24%. The confusion matrix of results obtained is shown in Table 4.7 following.

Table 4.7: Confusion Matrix obtained with S Set

Emotion	**Anxiety**	**Anger**	**Disgust**	**Boredom**	**Sadness**	**Natural**	**Happiness**
Anxiety	59	2	1	1	0	2	4
Anger	1	117	0	0	0	0	9
Disgust	2	1	39	1	1	1	1
Boredom	0	0	1	73	4	3	0
Sadness	0	0	0	5	57	0	0
Natural	4	0	0	4	0	69	2
Happiness	5	12	1	0	0	1	52

As can be seen from Figure 4.4 below, S set give the best recognition accuracy for all emotions. Expect for sadness emotion it gave the same results as VQ_S set. P set show also a good results In comparison with to VQ_S set. Expect for disgust where S set gave a better recognition accuracy.

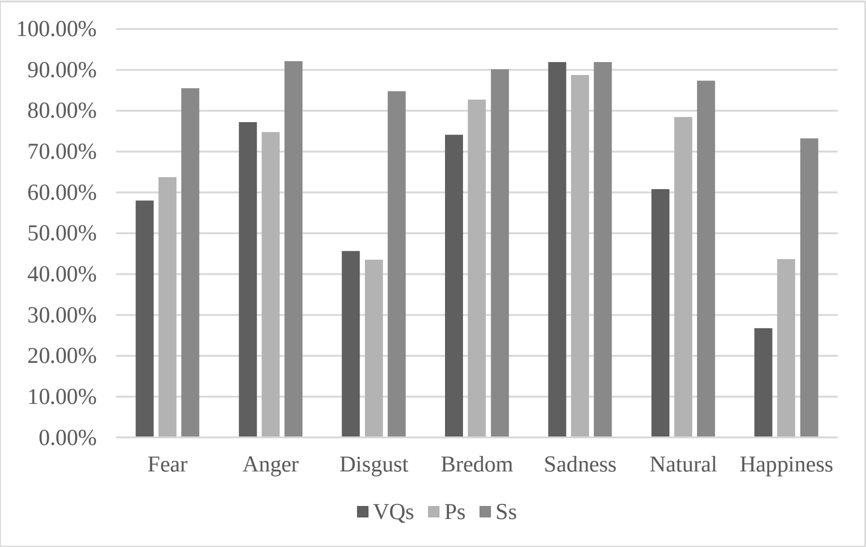


Figure 4.4: Emotion Recognition Compression For VQ_S , P and S sets

As a conclusion for this phase, spectral features are better in both overall and individual emotion recognition accuracy. Followed by prosodic features. However, the prosodic feature came after voice quality feature in disgust emotion recognition. Although voice quality features give the worth overall accuracy. It proves that it can be more useful then prosodic features on sadness and disgust emotion recognition. In addition, it shows the same capability on recognition of sadness as spectral features.

4.3.2 Phase (2): Identify the Best Features Extraction Technique Combination

The goal of this phase is to identify the best feather extraction combination. Thus the same task as phase 1 were implements in addition to task (F) Combination. Four selective features sets were examined and evaluated:

1. VQ_SP**:** combine voice quality features with prosodic features.
2. VQ_SS**:** combine voice quality features with spectral features.
3. PS*:* combine prosodic features with spectral features. -
4. VQ_SPS**:** combine voice quality features with prosodic features and spectral features.

From the results obtained, it can be observed that the accuracy obtained using VQ_SPS set was found to be much higher than that of the VQ_SP, VQ_SS and PS set. An overall accuracy of 88.04% could be achieved with VQ_SPS set; while an overall recognition of 71.40%, 87.66% and 87.48% could be achieved with VQ_SP, VQ_SS and PS respectively. Table 4.8 below summarizes the comparison of results based on the number of features and recognition accuracy obtained from the four feature sets.

Table 4.8: Comparison of Results obtained with VQ_SP, VQ_SS, PS and VQ_SPS Sets

Features Set	**No. of Features**	**Accuracy (%)**
VQ_SP	265	71.40
VQ_SS	1453	87.66
PS	1416	87.48
VQ_SPS	1567	88.04

The results indicated that the combination of the three acoustic features yields to more accuracy improvement than using a combination of two types of acoustic features. This result is compatible with literature reports that claim that it is better to use a combination of acoustic features than using one type. It is also indicated that voice quality features can improve prosodic and spectral features recognition accuracy. It even proves that voice quality features more useful when combining with spectral features than combining prosodic features with spectral features.

4.3.3 Phase (3): Enhance the Features Extraction Technique

To improve the performance of SER, this phase aims to enhance the feature extraction technique using the voice quality prosodic spectral-based feature extraction technique (VQPS). The same task as phase 2 were implements, the different will the addition of (C) Brute force and (E) New tasks. The features extraction in this phase use three features extraction approaches the selective approach which used for prosodic and spectral features. While the brute force and new approach used for voice quality features.

As a result, 181 voice quality features were extracted. Containing six new features, of which three have not been used before in SER, while the other three have not been used before with EMO-DB. In addition to 175 brute force voice quality features that have been used before in literature with EMO-DB, 151 of them are the selective voice quality features that used in phase 1 and 2. Table 4.10 following shows the description of the new and brute force voice quality set.

Table 4.9: Description of Extracted New Voice Quality Features

Feature Sequence	**Description**
1	voice to total frames ratio
2-3	GNE
4-5	Do-1000
6	Hamml

Table 4.10: Description of Extracted Brute Force Voice Quality Features

Feature Sequence	**Description**
7	HNR
8-12	Harmonic
13	NHR
14	Autocorrelation
15-17	Jitter (local absolute, rap, ppq5)
18- 22	Shimmer (local dB, apq3, apq5, apq11, ddp)
23-25	Period
26-27	voice break
28	pulses
29	voicedunvoiced frames ratio
30	unvoicedtotal frames ratio
31-104	**Jitter (local, ddp)**
105-141	**Shimmer (local)**
142-181	**Voicing**

Firstly the traditional voice quality features (VQ_T) set that extracted using the brute force approach was comberd to the proposed voice quality features (VQ_T) set that combine the extracted brute force and the new voice quality features.

From the results obtained, it can be observed that the accuracy obtained using VQ_P set was found to be slightly higher than that of the VQ_T set. An overall accuracy of 73.83% could be achieved with VQ_T set; while an overall recognition of 75.14% could be achieved with VQ_P. Table 4.11.6 below summarizes the comparison of results based on the number of features and recognition accuracy obtained from the two feature sets.

Table 4.11: Comparison of Results obtained with VQ_T and VQ_P Sets

Features Set	No. of Features	Accuracy (%)
VQ_T	175	73.83
VQ_P	181	75.14

The results indicated that the combining of the new and traditional VQ features is an effective way to improve the performance of SER. In addition, it proves that VQ features can be used alone for emotion recognition; this contrasts with some reports in literature.

Amongst VQ_T emotions, sadness achieved a maximum recognition of 90.32% and the emotion of happiness achieved a minimum recognition of 50.70%. The confusion matrix of results obtained is illustrated in Table 4.12 following.

Table 4.12: Confusion Matrix obtained with VQ_T Set

Emotion	Anxiety	Anger	Disgust	Boredom	Sadness	Natural	Happiness
Anxiety	46	5	4	1	0	7	6
Anger	3	104	3	0	0	0	17
Disgust	3	2	28	2	2	6	3
Boredom	0	0	2	66	3	9	1
Sadness	0	0	1	2	56	3	0
Natural	4	1	4	7	3	59	1
Happiness	7	20	6	2	0	0	36

In relation to VQ_P, the emotion of sadness achieved a maximum recognition of 90.32% while the emotion of happiness achieved a minimum recognition of 49.30%. The confusion matrix of results obtained is shown in Table 4.13 following.

Table 4.13: Confusion Matrix obtained with VQ_P Set

Emotion	**Anxiety**	**Anger**	**Disgust**	**Boredom**	**Sadness**	**Natural**	**Happiness**
Anxiety	48	3	5	0	3	1	9
Anger	1	109	4	0	0	0	13
Disgust	2	1	28	1	2	8	4
Boredom	0	0	1	68	1	10	1
Sadness	1	0	2	2	56	1	0
Natural	3	1	4	9	3	58	1
Happiness	9	20	4	2	0	1	35

As can be seen from Figure 4.5 below, VQ_P set enhance anxiety, anger and boredom emotion recognition compares to VQ_T set. However, for disgust and sadness, no improvement at all was show. Unfortunately the emotion recognition for natural and happiness emotion decrease.

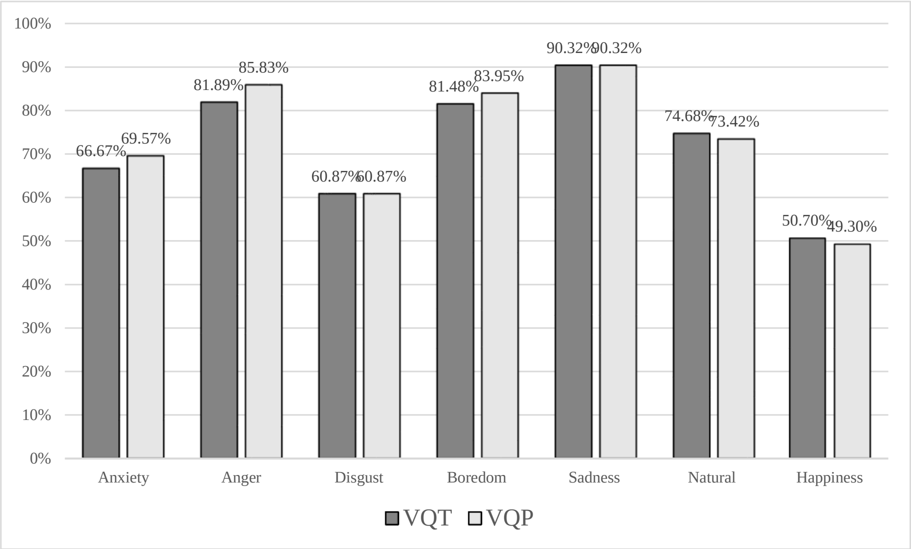


Figure 4.5: Emotion Recognitione Compressions between VQ_T and VQ_P Sets

Another observation from the result that for both features set sadness emotion was the best emotion to recognize while happens was the worst.

The proposed technique ($VQPS$) combines all the extracted features as shown in Sections 4.3.1, 4.3.2 and 4.3.3 respectively. A feature vector size of 1591 is used for classification. An overall recognition of 88.79% could be achieved with $VQPS_T$ set; while an overall recognition of 88.97% could be achieved with $VQPS_P$ set. Table 4.14 following summarizes the comparison of results based on the number of features and recognition accuracy obtained from the two feature sets.

Table 4.14: Comparison of Results obtained with $VQPS_T$ and $VQPS_P$ Sets

Features Set	No. of Features	Accuracy (%)
$VQPS_T$	1591	88.79
$VQPS_P$	1597	88.97

In relation to $VQPS_T$, the emotion of anger achieved a maximum recognition of 94.49% while the emotion of happiness achieved a minimum recognition of 77.46%. The confusion matrix for the results obtained is illustrated in Table 4.15 following.

Table 4.15: Confusion Matrix obtained with $VQPS_T$ Set

Emotion	Anxiety	Anger	Disgust	Boredom	Sadness	Natural	Happiness
Anxiety	59	2	0	0	1	4	3
Anger	1	120	0	0	0	0	6
Disgust	3	1	38	1	1	2	0
Boredom	0	0	1	76	2	2	0
Sadness	0	0	0	5	57	0	0
Natural	2	1	0	5	1	70	0
Happiness	4	11	1	0	0	0	55

In relation to $VQPS_P$, the emotion of anger achieved a maximum recognition of 94.49% while the emotion of happiness achieved a minimum recognition of 77.46%. The confusion matrix for the results obtained is shown in Table 4.16 following.

Table 4.16: Confusion Matrix obtained with $VQPS_P$ Set

Emotion	**Anxiety**	**Anger**	**Disgust**	**Boredom**	**Sadness**	**Natural**	**Happiness**
Anxiety	61	2	0	0	1	2	3
Anger	1	120	0	0	0	0	6
Disgust	3	1	37	1	2	2	0
Boredom	0	0	1	76	2	2	0
Sadness	0	0	0	5	57	0	0
Natural	2	1	0	5	1	70	0
Happiness	4	11	1	0	0	0	55

Figure 4.6 following shows a comparison of results obtained using $VQPS_P$ and $VQPS_T$. No enhancement was shown between the two sets except for the emotion of anxiety which showed an improvement from 85.51% to 88.41%. Unfortunately the emotion recognition for disgust emotion decrease

4.3.4 Phase (4): Reduced Features Dimensionality and Identify the Significant Features

As can be seen from the previous phase, SER feature extraction techniques can produced a huge amount of features. However, most of these features are irrelevant and redundant which resulted in decreasing the recognition accuracy. Therefore, this phase focuses on decreasing the features set and identify the significant features that can improve SER recognition accuracy.

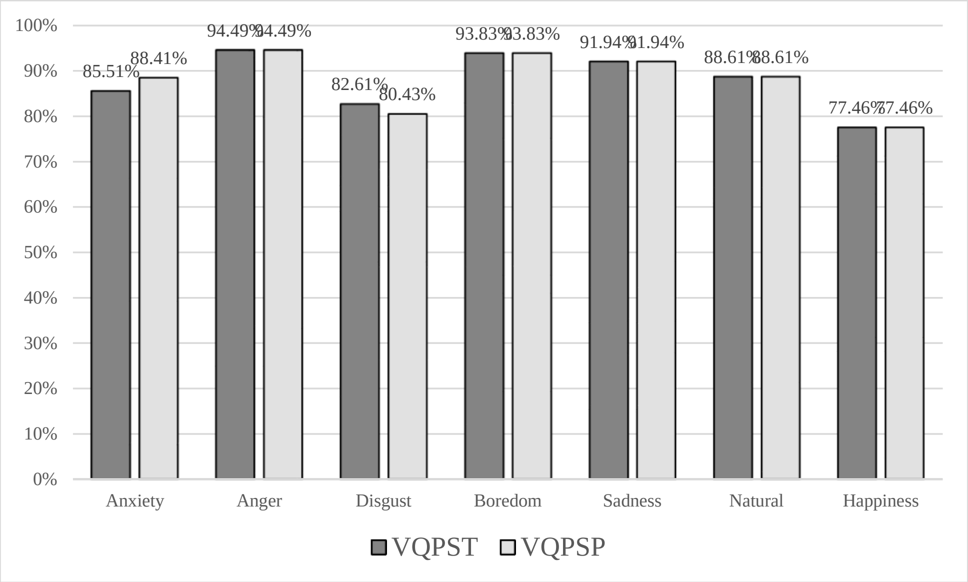


Figure 4.6: Emotion Recognitione Compressions between $VQPS_T$ and $VQPS_P$ Sets

The same task as phase 3 were implements, the different will the addition of task (H) Features selection. A two layers feature selection method is proposed to obtain the significant features. the propose method named balanced hybrid filter-based feature selection method (BHFFS).

The proposed selection method BHFFS consists of two main layers, namely: balancing layer and hybrid filter-based layer. The first layer (class balancing) implements a resembling technique which produces a random subsample of a dataset and is aimed at balancing the dataset. Meanwhile, the second layer (hybrid filter-based) aims to select significant features using hybrid filtering feature selection algorithm. A detailed description and justification of these layers are provided in the following sections.

Figure 4.7 shows the proposed approach feature selection that is used in this research. The effectiveness of the reduced feature subsets was evaluated using LIBSVM algorithms classification performance.

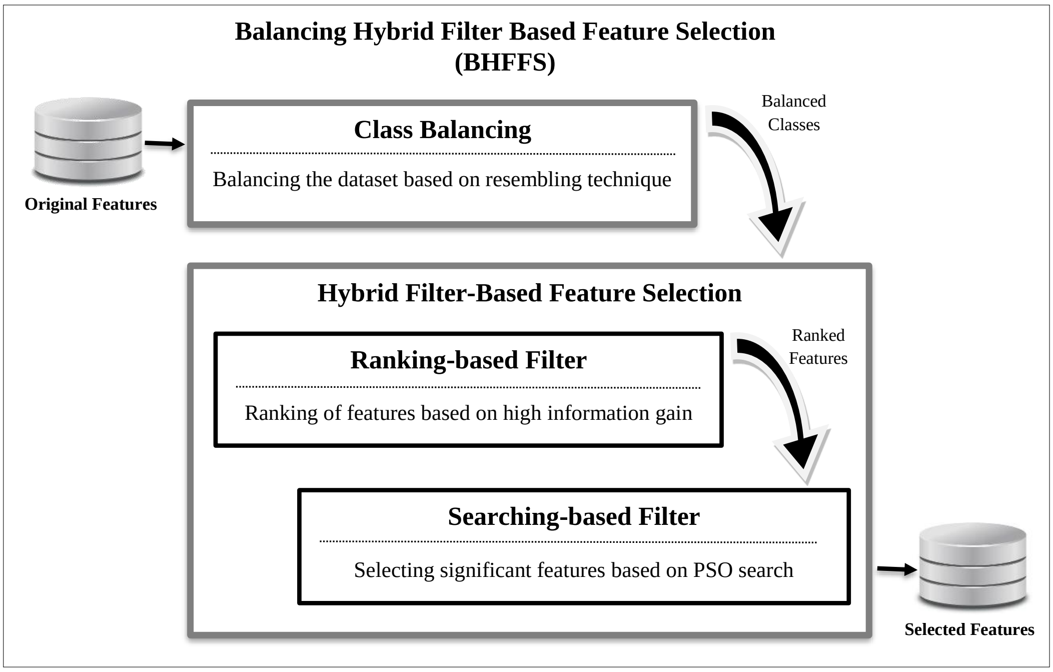

Balancing Hybrid Filter Based Feature Selection
(BHFFS)
Original Features
Class Balancing
Balancing the dataset based on resembling technique
Balanced Classes
Hybrid Filter-Based Feature Selection
Ranking-based Filter
Ranking of features based on high information gain
Ranked Features
Searching-based Filter
Selecting significant features based on PSO search
Selected Features

Figure 4.7: Balancing Hybrid Filter-based Feature Selection Architecture

Class Balancing

The dataset is classified as imbalanced if one or more of the classes have more samples than others. Due to this problem, the classification algorithms obtained good accuracy for the majority classes and poor accuracy for the minority classes. In this research, the EMO-DB is an imbalanced dataset (as mentioned in Section 5.3.1) thus, this section presents the result of the balancing process. The Resample technique was used in this research to gain a uniform distribution of the classes. Figure 4.8 following shows the dataset distribution before and after balancing.

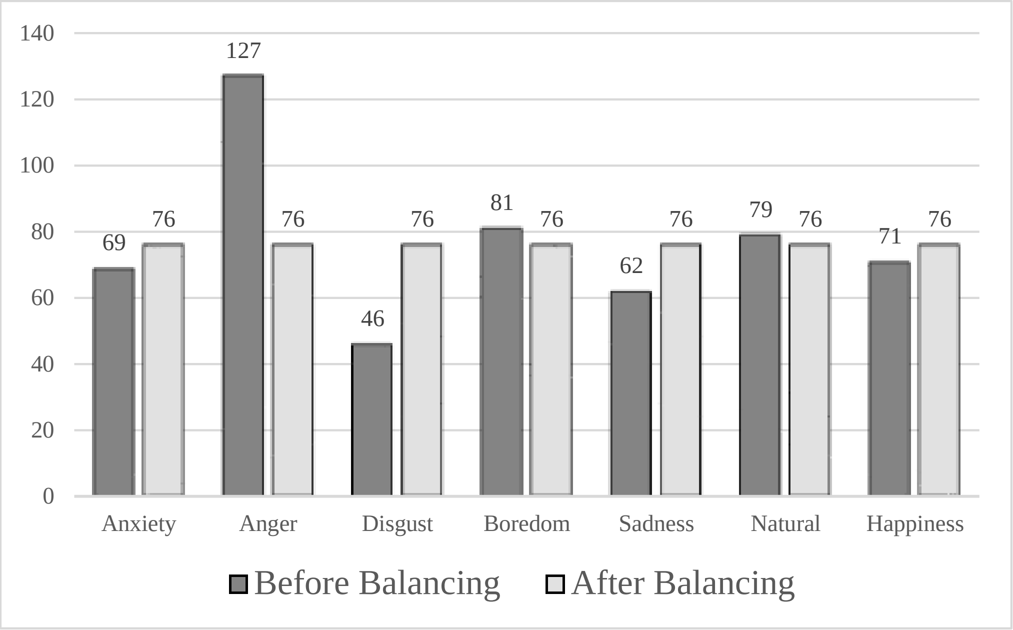


Figure 4.8: EMO-DB Distribution before and After Balancing

Hybrid Filter-Based Feature Selection

Feature selection aims to select significant features and reduce search space which improves the accuracy of SER. This section describes the two layers feature selection approach which is known as hybrid filter-based. Hybrid filter-based layer consists of two filter-based algorithms, namely, a feature filter with a ranking algorithm and features filter with search algorithm. These algorithms must be carried out in a sequential order which means that the second algorithm is dependent upon the first algorithm. The first layer (feature filter with ranking method) implements IGR algorithm which is aimed at ranking features based on high information gain entropy. The second layer (feature filter with search method) implements CFS algorithm which seeks to find features that are highly correlated with the class rather than the features that are correlated with each other using PSO as a search algorithm. Detailed descriptions of these layers are provided in the next section.

Feature Filter with Ranking

The basic premise of feature ranking is to eliminate irrelevant features, and is normally used to enable greater simplicity. A suitable ranking criterion is used to score the variables and a threshold is used to remove variables below the threshold. As mentioned earlier, feature ranking was implemented using IGR. For feature-ranking purposes, IGR was applied in all features described in Chapter 4. Table 4.17 following summarizes the IGR parameters values used.

Table 4.17: The Parameters Values used in IGR Algorithm

Description	**Value**
Number of Emotions	7
Number of Features	1597
Threshold (based on the experiment as explained in Section 4.5.1)	0

The features are ranked in decreasing order based on their relevance to class labels as shown in Appendix A. The high score rank of the feature indicates that this feature has high relevance to the class label (emotion); while a low score shows the independence of the feature from the class label. The rank score measures how much this feature relates to and contributes to class labels. Therefore, all features below the threshold (0) are removed because they have a low ranking score and could lead to poor classification accuracy.

Feature Filter with Search

Feature filter with search was implemented using CFS with PSO as a search method. IGR computed the correlation between the class label and the features individually but ignored the correlation among features. CFS looks for features that are highly correlated with the classes which have the minimum correlation between the features themselves. Table 4.18 below summarized the use of CFS-PSO parameters values.

Table 4.18: The Parameters Values used in CfS-PSO Algorithm

Description	Value
CFS Search	PSO
Population	35
Iteration	20

The result of selected features using CFS-PSO is shown in Appendix B. The result indicates that (736) number of features are the most significant, with most of these being spectral features. Table 4.19 below shows the distribution of voice quality, prosodic, and spectral features in the final selected features set.

Table 4.19: Distribution of Different Features in the Selected Features Set

Features Type	Number of Features
Voice Quality	89
Prosodic	50
Spectral	597

In addition, as noted from the results, voice quality features has a higher contribution in the selected features than prosodic features.A total of 15 Burt force traditional voice quality features were involved in the most significant features set. Furthermore, the six (6) new voice quality extracted features proposed in Section 4 were involved in the most significant features set as presented in Table 4.26 following

Table 4.20: Distribution of Different Features in the Selected Features Set

Feature Order	Features
154	Do1000(offset)
155	Hamml
162	voiced to total frames ratio
175	GNE(mean)
191	GNE(std)
194	Do1000(slope)

To evaluate the significance of these features on classification accuracy, LIBSVM algorithms have been applied with final selected features. With LIBSVM, an overall recognition of 94.74% was achieved. Table 4.26 following shows the comparison of the number of features and accuracy obtained before and after features selection.

Table 4.21: Compression of Result Before and After Feature Selection

	No. of Features	Accuracy(%)
Before features selection	1597	88.97
After features selection	736	94.74

The emotion of sadness achieved a maximum recognition rating of 100% while the emotion of happiness achieved a minimum recognition rating of 88.16%. The confusion matrix after performing the selection is shown in Table 4.22 following.

The emotion of sadness achieved a maximum recognition rating of 100% while the emotion of happiness achieved a minimum recognition rating of 88.16%. The confusion matrix after performing the selection is shown in Table 4.22 following. The emotion of sadness achieved a maximum recognition rating of 100% while the emotion of happiness achieved

Table 4.22: Confusion Matrix obtained with Selected Features

Emotion	Anxiety	Anger	Disgust	Boredom	Sadness	Natural	Happiness
Anxiety	71	1	1	0	0	2	1
Anger	1	72	0	0	0	0	3
Disgust	0	1	75	0	0	0	0
Boredom	0	0	1	71	2	2	0
Sadness	0	0	0	0	76	0	0
Natural	1	0	0	3	1	72	0
Happiness	0	8	0	0	0	1	67

a minimum recognition rating of 88.16%. The confusion matrix after performing the selection is shown in Table 4.22 following.

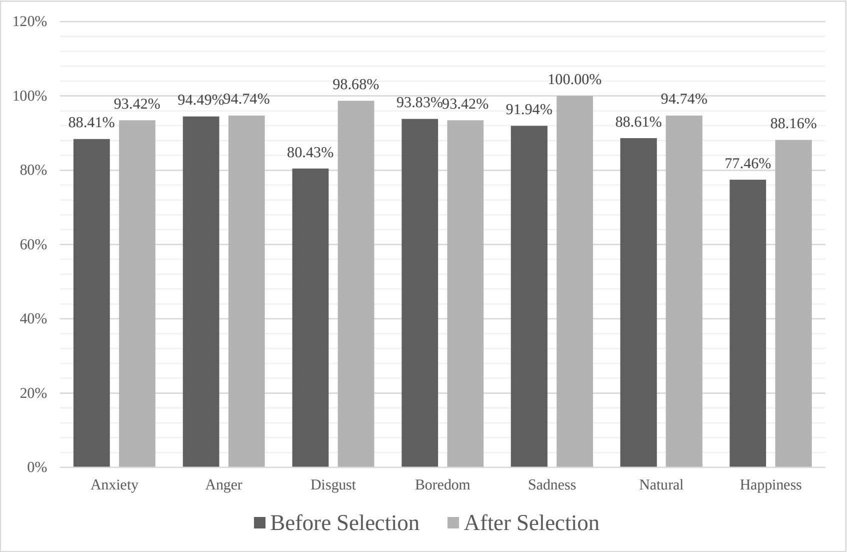


Figure 4.9: Emotion Recognition Compression Before and After Features Selection

4.3.5 Phase (5): Comparison and Benchmark

In order to improve the SER systems recognition rate, various feature extraction techniques were used in previous works. This feature extraction technique generated different combinations of voice quality, prosodic and spectral features respectively. Some of these works are provided for comparison and benchmarking. The features used by Zaho et al. (2014), Sun et al. (2015) and Shirani and Nilchi (2016) were compared. The evaluation matrix used for comparison was the LIBSVM classification accuracy based on EMO-DB dataset. Performance comparison with previous works is given in Table 4.23 following.

Table 4.23: Performance Comparison with Previous Works

Reference	**Extraction Tool**	**Features No.**	**Normalization**	**Emotion No.**	**Accuracy(%)**
(Zhao et al., 2014)	Praat	204	No	7	78.75
(Sun and Wen, 2015)	openSMILE	1582	SSMCFS	7	82.64
(Shirani and Nilchi, 2016)	Praat	68	No	6	86.53
VQPS	Praat,openSMILE	1597	[0-1]	7	88.97

Zaho et al. used Praat toolkit by which to extract 204 features including: HNR; jitter; shimmer; pitch; intensity; duration; formant; spectral energy and MFCC features from EMO-DB with seven emotions. An accuracy rate of 78.75% was obtained using LIBSVM.

Sun et al. extracted 1582 features including: HNR; jitter; shimmer; voicing probability; pitch; loudness; log Mel freq. band; as well as LSP and MFCC using an openSMILE toolkit. Using SSMCFS proposed normalization method before LIBSVM with seven emotions from EMO-DB yielded an accuracy rate of 82.64%.

Shirani and Nilchi extracted 68 features using Praat tool. These features related to: duration; pitch; intensity; formants; amplitude; energy; power; ZCR; HNR; jitter; shimmer and MFCC from EMO-DB with six emotions (anger, disgust, fear, joy, sadness, and neutral). The recognition rate of 86.53% was achieved using SVM. Table 4.24 below shows feature comparisons with previous works.

Table 4.24: Features Comparison with the Previous Works

Reference	**Voice Quality**	**Prosodic**	**Spectral**
(Zhao et al., 2014)	HNR, jitter and shimmer	pitch, intensity, duration, formant	spectral energy, MFCC
(Sun and Wen, 2015)	HNR, Jitter, shimmer, voicing probability	pitch, loudness	Log Mel freq. band, LSP, MFCC
(Shirani and Nilchi, 2016)	HNR, jitter, shimmer	pitch, duration, , intensity, formants, amplitude, energy, power, ZCR	MFCC
VQPS	HNR, jitter, shimmer, Voicing, Harmonic, Autocorrelation, Do-1000, Hammel, GNE	pitch, loudness	Log Mel freq. band, LSP, MFCC

It can be seen from all the above and from Table 4.24, that the feature extraction technique adopted in this research gave better classification accuracy compared to previous works.

In the recent past, a variety of works on SER have been reported using different selection methods. Some of these works from literature sources are given here for comparison. The selection methods results used by Sun and Wen (2015), Jassim et al. (2017) and Ding et al. (2108) respectively were compared. The evaluation metric used for comparison was the LIBSVM classification accuracy based on EMO-DB dataset with seven emotions (neutrality, anger, fear, joy, sadness, disgust, and boredom).

Two selected feature sets were used for compression, namely, the selected features form $VQPS_P$ and $VQPS_{2010}$ sets. $VQPS_{2010}$ is described in Section 4 and is used to ensure having the most similar compression parameter setting as possible. Table 4.25 following illustrates the performance comparison with reported works.

Table 4.25: Performance Comparison with Reported Works

Reference	Normalization Method	Selection	Features No.	Accuracy (%)	Selected Features No.	Accuracy (%)
(Sun and Wen, 2015)	SSMCFS	SMCFS	1582	-	86.63	88.85
(Jassim et al., 2017)	[0-1]	mRMR	1582	946	-	89.45
(Ding et al., 2018)	No	BBO	1582	-	74.29	90.13
VQPS$_{2010}$	**[0-1]**	**BHFFS**	**1567**	**588**	**88.04**	**95.11**
VQPS	**[0-1]**	**BHFFS**	**1597**	**736**	**88.97**	**94.74**

Sun and Wen. obtained an accuracy rating of 88.85% when using a proposed selection method (SMCFS) with proposed normalization (SSMCFS) method. To evaluate their proposed method, a comparison with some selection methods that have been used in SER was performed. Table 4.26 following summarizes the performance results of these methods.

Table 4.26: performance of different selection method presented by Sun and Wen (2015)

Selection Algorithm	Accuracy (%)
MCFS	88.62
PCA	86.89
LDA	71.39
mRMR	87.27
LS	87.76
DISR	85.76

Jassim et al. obtained an accuracy rating of 89.45% when using the mRMR algorithm to remove irrelevant features and reduce the features dimensionality from 1582 to 946 features. Ding et al. obtained an accuracy of 90.13% when using proposed BBO with only 40 audio files for each kind of emotion from EMO-DB. The result obtained by BBO was compared to GA results. An accuracy rating of 81.26% was obtained using GA.

From all the above, and as shown in Table 4.25, the feature selection method adopted in this research successfully presented the best classification accuracy.

4.4 Summary

in development SER system there are two important issue that have to be concedred. The suitable feature extraction technique ans selecting a reduced feature set that can significantly represent emotions by eliminating irrelevant and redundant features since they directly affects the classification performance. Thus this chapter focuses on constecting a theoretical model that answer the features related research question and help to create an SER framework.This framework implement different tasks that help in investigating the impact of type and combination of the features on recognition accuracy. it also introduce enhanced a feature extraction technique based on prosodic spectral and a new combination of voice quality features. and proposing balancing hybrid filter base feature selection method (BHFFS). The extracted and selected features are evaluated in terms of individual and overall recognition accuracy. The empirical results prove that the extracted and selected feature sets outperform the feature sets detailed in previous works. In addition, it proves that voice quality features are considerably important in developing the SER system.

CHAPTER 5

Conclusion and Future Work

5.1 Introduction

Speech emotion recognition plays an important role in HCI implementations. Unfortunately, until now, the performance of SER has not reached the maximum performance that can help the machine to understand humans completely. This thesis has proposed an SER framework which enhanced the existing features extraction technique and selection methods. Also, it has described the related literature review as well as detail methodology on designing the SER framework. After conducting the literature review we found that the performance of an SER system relies on the feature extraction technique adopted, as well as the size of the final features vectors. However, both existing extraction techniques and selection methods are having some limitations.

5.2 Summary of the Research Work

The present work created an SER features model that help in understanding the relation between features and SER recognition accuracy. This model used to create an SER framework that has five phases. The first phase is concerned with investigating the SER existing single features extraction techniques capability in recognize emotions. The results show that the spectral features are better in both overall and individual emotion recognition accuracy. While voice quality features give the worth overall accuracy. However, it shows the same capability on recognition of sadness as spectral features. In addition, it proves that it can be more useful then prosodic features on sadness and disgust emotion recognition.

The second phase is concerned with identifying the best features extraction technique combination. From the results obtained, it can be observed that the combination of the three acoustic features namely voice quality, prosodic and spectral features yields to more recognition accuracy improvement than using a combination of two types of acoustic features. It is also indicated that the voice quality features can improve both prosodic and spectral features recognition accuracy. It even proves that voice quality features more useful when combining with spectral features than combining prosodic features with spectral features.

The third phase is concerned with enhancing the features extraction technique which aims to improve the recognition accuracy. The proposed voice quality prosodic spectral-based features extraction (VQPS) technique combined new and traditional voice quality features with traditional prosodic and spectral features. The results prove that the proposed VQPS technique can improve the recognition accuracy. it also indicated that voice quality features can be used alone for emotion recognition and it is can improve the performance of SER when combined with prosodic and spectral features.

The fours phase is concerned with reducing features dimensionality and identifying the most significant features. A developed features selection method has been proposed known as balanced hybrid filter-based features selection (BHFFS) method. As a result, the features dimensionality was reduced to only (736) significant features. The spectral features constitute the majority in significant feature set followed by voice quality features. The results indicated that voice quality features have a higher contribution to the selected features than prosodic features. In addition, a total of 15 Burt force traditional voice quality features were involved in the most significant features set. Furthermore, the six (6) new voice quality extracted features were involved in the most significant features.

The final phase is concerned with comparing the results obtained using the proposed SER framework with the results obtained by previous researchers. The results indicated that the proposed framework adopted in this research gave better classification accuracy compared to previous works in extraction and selection.

5.3 Thesis Contribution

This section presents a list of contributions that are proposed in this research. These are as detailed below:

1. Creates of SER features theoretical model that presents the relation between features and SER performance (recognition accuracy).
2. Create SER framework that seeks to solve some of extraction and selection limitations.
3. The performance of three feature extraction techniques namely voice quality, prosodic and spectral in extracting the relevant features from the speech samples is evaluated in term of recognition accuracy. the evaluation is done for single and combined features set.
4. Development of an enhanced feature extraction technique called voice quality prosodic spectral based features extraction (VQPS) for improving the performance of SER. Traditional prosodic and spectral techniques were combined with new and traditional voice quality techniques to create a new extraction technique.
5. Development of an enhanced feature selection method called balanced hybrid filter-based feature selection (BHFFS). The method starts by solving the problem of imbalanced data before implementing two filter-based feature selection algorithms.

5.4 Recommendation and Future Works

1. **Different classifiers:** Different classifiers and structure can be used to evaluate the recognition accuracy in this thesis. such as using the classifier fusion method.
2. **Speaker Independent:** This work concentrates only on the speaker-dependent emotion recognition. The newly-proposed extraction technique and selection method can also be extended to speaker-independent emotion recognition.

3. **Different Dataset:** This work is designed using EMO-DB with the German language. It is envisaged that it can also be extended to different datasets with different languages, for instance, an Arabic dataset.

4. **New Voice Quality Features:** These works introduce only six voice quality features and have obtained good results. More voice quality features can also be tried in order to better improve SER performance.

5. **Linguistic Features:** This work concentrates only on acoustic features. A linguistic features technique can also be combined with the newly-proposed extraction technique to investigate their effect on SER performance.

6. **More Filter-based Feature Selection Algorithms:** This work concentrates only on two filter selection algorithms. Different filter algorithms can also be evaluated for the newly-proposed selection technique.

APPENDIX A

Results on Feature Filter with Ranking

Ranking	Feature	Ranking	Feature
0.9646	F0finEnv-sma-de-skewness	0.2504	mfcc-sma[3]-maxPos
0.8914	logMelFreqBand-sma-de[2]-iqr1-3	0.2501	F0finEnv-sma-de-maxPos
0.8733	logMelFreqBand-sma-de[4]-quartile1	0.2501	mfcc-sma-de[11]-Percentile99.0
0.8514	logMelFreqBand-sma-de[4]-linregerrA	0.2501	LSPFreq-sma-de[3]-iqr1-3
0.8512	logMelFreqBand-sma-de[3]-iqr1-3	0.2499	jitterDDP-sma-kurtosis
0.8461	mfcc-sma-de[2]-linregerrA	0.2483	jitterDDP-sma-de-iqr1-2
0.8403	logMelFreqBand-sma-de[5]-quartile1	0.2482	mfcc-sma[3]-iqr1-2
0.8254	F0final-sma-Percentile99.0	0.2475	mfcc-sma[13]-iqr1-2
0.8224	logMelFreqBand-sma-de[5]-iqr1-3	0.2459	mfcc-sma[6]-quartile1
0.8181	mfcc-sma-de[0]-iqr1-3	0.2455	shimmer(localdB)
0.814	mfcc-sma-de[0]-linregerrA	0.2454	mfcc-sma-de[7]-quartile3
0.8127	logMelFreqBand-sma-de[4]-iqr1-3	0.2454	mfcc-sma[8]-linregerrA
0.8078	logMelFreqBand-sma-de[2]-linregerrA	0.2453	logMelFreqBand-sma[0]-upleveltime75
0.8058	logMelFreqBand-sma[2]-pctlrange0-1	0.2452	logMelFreqBand-sma[7]-pctlrange0-1
0.7996	logMelFreqBand-sma-de[5]-linregerrA	0.245	voicingFinalUnclipped-sma-iqr1-3
0.7963	mfcc-sma-de[12]-linregerrQ	0.2447	mfcc-sma[8]-linregc2
0.7962	logMelFreqBand-sma-de[4]-stddev	0.2441	logMelFreqBand-sma[6]-linregerrA
0.7933	F0final-sma-de-linregerrA	0.244	mfcc-sma-de[10]-linregc2
0.7907	logMelFreqBand-sma-de[4]-iqr1-2	0.2438	LSPFreq-sma-de[5]-quartile1
0.7855	F0finEnv-sma-de-kurtosis	0.2435	logMelFreqBand-sma[2]-linregc1
0.7832	logMelFreqBand-sma-de[3]-quartile1	0.2435	mfcc-sma[12]-quartile3
0.783	logMelFreqBand-sma-de[3]-linregerrQ	0.2434	logMelFreqBand-sma[0]-amean
0.781	logMelFreqBand-sma-de[5]-quartile3	0.2429	F0final-sma-de-skewness
0.7804	F0final-sma-linregerrQ	0.2424	pcm-loudness-sma-maxPos
0.7738	mfcc-sma-de[12]-stddev	0.2422	LSPFreq-sma[1]-linregc2
0.7672	F0final-sma-linregerrA	0.2421	mfcc-sma[14]-iqr1-3
0.7672	logMelFreqBand-sma-de[3]-stddev	0.2419	logMelFreqBand-sma[2]-amean
0.7669	logMelFreqBand-sma-de[2]-iqr1-2	0.2417	mfcc-sma[12]-linregc1
0.765	mfcc-sma-de[0]-quartile3	0.2414	mfcc-sma[4]-iqr1-3
0.7645	logMelFreqBand-sma-de[4]-quartile3	0.2412	logMelFreqBand-sma[7]-quartile1
0.7645	logMelFreqBand-sma-de[2]-quartile1	0.2403	LSPFreq-sma[7]-minPos
0.7644	F0final-sma-de-linregerrQ	0.24	logMelFreqBand-sma-de[3]-quartile2
0.7632	logMelFreqBand-sma-de[2]-linregerrQ	0.2397	logMelFreqBand-sma[5]-skewness
0.7486	voicingFinalUnclipped-sma-Percentile99.0	0.2393	jitterDDP-sma-skewness
0.742	logMelFreqBand-sma-de[2]-stddev	0.2379	mfcc-sma-de[7]-linregerrA
0.7414	F0final-sma-stddev	0.2379	LSPFreq-sma-de[7]-linregerrQ
0.7377	logMelFreqBand-sma[4]-pctlrange0-1	0.2379	LSPFreq-sma-de[7]-stddev
0.7345	F0finEnv-sma-quartile3	0.2379	logMelFreqBand-sma[5]-linregc1
0.7326	mfcc-sma-de[2]-stddev	0.2377	logMelFreqBand-sma-de[5]-skewness
0.7311	logMelFreqBand-sma[3]-pctlrange0-1	0.2373	LSPFreq-sma[2]-stddev

Ranking	Feature	Ranking	Feature
0.7284	voicingFinalUnclipped-sma-pctlrange0-1	0.2372	logMelFreqBand-sma[6]-kurtosis
0.7255	logMelFreqBand-sma-de[3]-linregerrA	0.2364	H(mean)
0.7217	mfcc-sma-de[12]-linregerrA	0.2355	jitterDDP-sma-de-kurtosis
0.7201	logMelFreqBand-sma-de[4]-iqr2-3	0.2353	mfcc-sma[11]-linregerrA
0.7188	Period(mean)	0.2351	mfcc-sma[0]-linregerrA
0.717	F0finEnv-sma-de-linregerrA	0.2349	logMelFreqBand-sma[4]-quartile2
0.7166	F0finEnv-sma-de-pctlrange0-1	0.234	mfcc-sma-de[5]-maxPos
0.7128	mfcc-sma-de[14]-linregerrQ	0.2334	logMelFreqBand-sma-de[3]-upleveltime75
0.7114	logMelFreqBand-sma-de[4]-linregerrQ	0.2332	mfcc-sma[13]-linregc2
0.7047	F0final-sma-de-stddev	0.2331	mfcc-sma[0]-quartile2
0.7044	mfcc-sma-de[0]-iqr1-2	0.2329	F0final-sma-de-linregc1
0.6992	F0finEnv-sma-Percentile99.0	0.2328	F0finEnv-sma-kurtosis
0.6985	F0finEnv-sma-quartile2	0.2322	LSPFreq-sma-de[0]-linregerrA
0.6974	logMelFreqBand-sma-de[5]-iqr2-3	0.2316	mfcc-sma-de[2]-kurtosis
0.6968	mfcc-sma-de[2]-linregerrQ	0.2312	logMelFreqBand-sma-de[1]-maxPos
0.6935	LSPFreq-sma[0]-quartile1	0.2309	mfcc-sma-de[11]-pctlrange0-1
0.6933	mfcc-sma[0]-pctlrange0-1	0.2304	jitterLocal-sma-upleveltime90
0.6921	LSPFreq-sma[0]-quartile3	0.23	mfcc-sma[6]-quartile3
0.689	mfcc-sma-de[5]-iqr1-2	0.23	Autocorrelation(mean)
0.6876	logMelFreqBand-sma-de[2]-quartile3	0.2298	jitterLocal-sma-de-linregc2
0.6828	mfcc-sma-de[0]-quartile1	0.2296	mfcc-sma[9]-pctlrange0-1
0.6825	mfcc-sma[2]-quartile1	0.2291	mfcc-sma-de[6]-pctlrange0-1
0.6807	logMelFreqBand-sma-de[3]-quartile3	0.2291	mfcc-sma-de[7]-Percentile99.0
0.6807	logMelFreqBand-sma[2]-Percentile1.0	0.229	mfcc-sma[1]-iqr1-3
0.6753	logMelFreqBand-sma-de[5]-linregerrQ	0.2283	LSPFreq-sma[1]-iqr1-2
0.675	voicingFinalUnclipped-sma-quartile3	0.2282	mfcc-sma[2]-Percentile99.0
0.6748	logMelFreqBand-sma-de[3]-iqr1-2	0.2277	LSPFreq-sma-de[0]-linregc1
0.6681	mfcc-sma-de[2]-quartile3	0.2274	mfcc-sma[6]-quartile2
0.6654	logMelFreqBand-sma-de[2]-iqr2-3	0.2272	LSPFreq-sma[3]-amean
0.6631	logMelFreqBand-sma[7]-Percentile1.0	0.2272	logMelFreqBand-sma[4]-amean
0.663	mfcc-sma-de[0]-iqr2-3	0.2271	LSPFreq-sma-de[6]-linregerrA
0.6595	logMelFreqBand-sma[6]-Percentile1.0	0.2269	logMelFreqBand-sma[4]-quartile1
0.6584	F0finEnv-sma-de-Percentile1.0	0.2268	logMelFreqBand-sma[3]-quartile3
0.6574	mfcc-sma[2]-quartile2	0.2267	mfcc-sma[14]-Percentile99.0
0.6558	logMelFreqBand-sma-de[2]-linregc2	0.2266	LSPFreq-sma-de[6]-minPos
0.6531	F0finEnv-sma-de-stddev	0.2265	logMelFreqBand-sma-de[0]-quartile3
0.653	LSPFreq-sma[0]-quartile2	0.2263	mfcc-sma-de[4]-linregc1
0.6508	shimmerLocal-sma-quartile3	0.2261	mfcc-sma-de[3]-quartile1
0.6476	LSPFreq-sma[0]-amean	0.226	mfcc-sma[2]-upleveltime90
0.645	mfcc-sma-de[8]-linregerrA	0.2254	logMelFreqBand-sma[2]-upleveltime75
0.6437	F0final-sma-quartile2	0.2247	shimmer(apq5)
0.6427	mfcc-sma[2]-quartile3	0.224	mfcc-sma-de[3]-minPos
0.6408	mfcc-sma-de[5]-stddev	0.2238	voicingFinalUnclipped-sma-de-quartile2
0.6393	mfcc-sma-de[5]-linregerrQ	0.2226	LSPFreq-sma-de[6]-quartile1
0.6386	logMelFreqBand-sma-de[5]-stddev	0.2226	mfcc-sma-de[1]-iqr2-3
0.6358	F0final-sma-quartile3	0.2221	mfcc-sma[8]-quartile3
0.6357	jitterLocal-sma-de-quartile1	0.222	mfcc-sma-de[2]-minPos
0.6334	F0finEnv-sma-de-linregc2	0.2216	voicingFinalUnclipped-sma-linregerrA
0.633	logMelFreqBand-sma-de[3]-iqr2-3	0.2214	logMelFreqBand-sma[1]-upleveltime75
0.6327	F0final-sma-amean	0.2203	mfcc-sma-de[12]-linregc1
0.6313	mfcc-sma-de[5]-linregerrA	0.22	mfcc-sma[4]-Percentile1.0
0.6257	F0finEnv-sma-pctlrange0-1	0.2197	LSPFreq-sma[2]-iqr2-3
0.6245	logMelFreqBand-sma[5]-pctlrange0-1	0.2192	mfcc-sma[8]-pctlrange0-1
0.624	mfcc-sma-de[2]-iqr1-3	0.2192	mfcc-sma[1]-pctlrange0-1
0.6227	shimmerLocal-sma-de-linregerrA	0.2192	jitterLocal-sma-de-linregc1
0.6218	mfcc-sma-de[12]-iqr1-2	0.2184	logMelFreqBand-sma[6]-iqr2-3
0.6215	logMelFreqBand-sma-de[6]-iqr1-3	0.2184	LSPFreq-sma-de[5]-linregc1
0.6213	mfcc-sma-de[14]-stddev	0.2183	logMelFreqBand-sma[7]-Percentile99.0
0.6203	F0final-sma-de-Percentile99.0	0.2173	pcm-loudness-sma-de-maxPos
0.6201	mfcc-sma-de[12]-iqr1-3	0.217	pcm-loudness-sma-iqr2-3
0.6196	F0finEnv-sma-amean	0.2169	logMelFreqBand-sma-de[7]-Percentile99.0
0.6147	logMelFreqBand-sma-de[5]-iqr1-2	0.2167	GNEmean
0.6145	shimmerLocal-sma-de-quartile1	0.2165	GNEstd

Ranking	Feature	Ranking	Feature
0.6136	voicingFinalUnclipped-sma-skewness	0.2164	LSPFreq-sma[6]-linregerrA
0.6133	mfcc-sma[2]-amean	0.2159	mfcc-sma[2]-kurtosis
0.6133	mfcc-sma[1]-upleveltime75	0.2159	logMelFreqBand-sma[3]-kurtosis
0.6128	mfcc-sma-de[11]-iqr1-3	0.2159	mfcc-sma-de[10]-Percentile1.0
0.6126	mfcc-sma-de[12]-pctlrange0-1	0.2156	LSPFreq-sma-de[6]-iqr2-3
0.6118	mfcc-sma[0]-Percentile1.0	0.2152	F0final-sma-upleveltime75
0.6104	voicingFinalUnclipped-sma-iqr2-3	0.2151	shimmerLocal-sma-de-upleveltime75
0.6095	mfcc-sma[5]-quartile1	0.2148	LSPFreq-sma[1]-linregerrA
0.6081	mfcc-sma-de[9]-linregerrA	0.2141	logMelFreqBand-sma-de[6]-minPos
0.6066	jitterLocal-sma-amean	0.214	mfcc-sma[8]-iqr1-3
0.6064	mfcc-sma[1]-quartile3	0.2137	logMelFreqBand-sma[0]-linregc1
0.6055	mfcc-sma-de[11]-quartile3	0.2134	logMelFreqBand-sma[2]-iqr1-3
0.6051	mfcc-sma-de[9]-iqr1-3	0.2125	logMelFreqBand-sma[4]-skewness
0.6032	F0final-sma-iqr1-3	0.212	LSPFreq-sma[0]-maxPos
0.6024	F0finEnv-sma-de-Percentile99.0	0.2118	mfcc-sma-de[7]-iqr1-2
0.6022	voicingFinalUnclipped-sma-quartile2	0.2117	mfcc-sma[1]-maxPos
0.6006	jitter(LocalA)	0.2115	pulses
0.5988	mfcc-sma[1]-quartile2	0.2114	jitterDDP-sma-quartile1
0.5976	mfcc-sma-de[12]-Percentile99.0	0.2102	jitterDDP-sma-iqr1-2
0.597	logMelFreqBand-sma[7]-skewness	0.2101	LSPFreq-sma[3]-quartile1
0.5965	mfcc-sma-de[0]-stddev	0.2098	mfcc-sma-de[0]-skewness
0.595	voicingFinalUnclipped-sma-amean	0.2098	LSPFreq-sma-de[3]-iqr1-2
0.5938	mfcc-sma[1]-amean	0.2096	shimmerLocal-sma-de-linregc1
0.593	logMelFreqBand-sma-de[7]-linregc1	0.2087	LSPFreq-sma-de[3]-quartile1
0.5916	mfcc-sma-de[4]-quartile3	0.2084	LSPFreq-sma[2]-linregerrQ
0.5914	mfcc-sma-de[8]-stddev	0.2082	mfcc-sma-de[4]-linregc2
0.591	mfcc-sma-de[8]-linregerrQ	0.2077	LSPFreq-sma-de[6]-maxPos
0.5881	logMelFreqBand-sma-de[4]-pctlrange0-1	0.2076	LSPFreq-sma-de[5]-iqr1-3
0.5878	F0finEnv-sma-de-linregerrQ	0.2076	mfcc-sma[6]-linregc2
0.5869	logMelFreqBand-sma-de[6]-quartile3	0.2069	logMelFreqBand-sma[5]-upleveltime75
0.5867	mfcc-sma-de[12]-quartile1	0.2064	mfcc-sma[4]-kurtosis
0.5866	F0finEnv-sma-iqr1-3	0.2062	mfcc-sma-de[1]-Percentile1.0
0.5866	logMelFreqBand-sma[4]-Percentile1.0	0.2062	mfcc-sma[14]-quartile1
0.5855	mfcc-sma-de[5]-quartile3	0.2056	LSPFreq-sma[4]-minPos
0.5853	mfcc-sma-de[2]-pctlrange0-1	0.2052	mfcc-sma[12]-quartile1
0.5849	mfcc-sma[2]-iqr1-3	0.2051	mfcc-sma-de[7]-pctlrange0-1
0.5849	mfcc-sma-de[5]-iqr1-3	0.205	LSPFreq-sma-de[7]-minPos
0.5838	mfcc-sma[2]-linregerrQ	0.205	jitterLocal-sma-de-upleveltime75
0.5838	F0finEnv-sma-de-amean	0.2048	mfcc-sma[5]-maxPos
0.5822	voicingFinalUnclipped-sma-quartile1	0.2048	LSPFreq-sma[4]-kurtosis
0.5817	F0finEnv-sma-iqr2-3	0.2042	LSPFreq-sma[1]-stddev
0.5814	logMelFreqBand-sma[3]-stddev	0.204	LSPFreq-sma-de[2]-kurtosis
0.5814	mfcc-sma[2]-Percentile1.0	0.204	mfcc-sma-de[9]-quartile2
0.5808	logMelFreqBand-sma-de[4]-Percentile99.0	0.2037	mfcc-sma-de[4]-upleveltime90
0.5807	F0finEnv-sma-stddev	0.2034	LSPFreq-sma[5]-amean
0.5784	shimmerLocal-sma-amean	0.2031	mfcc-sma[9]-maxPos
0.5776	jitterLocal-sma-quartile2	0.203	mfcc-sma-de[10]-minPos
0.5742	mfcc-sma-de[8]-quartile1	0.2029	jitterDDP-sma-upleveltime90
0.5736	logMelFreqBand-sma-de[5]-linregc2	0.2028	mfcc-sma-de[6]-Percentile99.0
0.5735	shimmerLocal-sma-de-quartile3	0.2025	logMelFreqBand-sma-de[0]-iqr2-3
0.5717	mfcc-sma[5]-quartile2	0.2025	shimmer(apq11)
0.5703	mfcc-sma-de[13]-stddev	0.2025	logMelFreqBand-sma[1]-Percentile99.0
0.5702	logMelFreqBand-sma-de[6]-quartile1	0.202	logMelFreqBand-sma[7]-quartile3
0.5666	mfcc-sma-de[11]-iqr1-2	0.2016	logMelFreqBand-sma[1]-skewness
0.5643	logMelFreqBand-sma-de[7]-iqr1-2	0.2012	mfcc-sma-de[9]-minPos
0.5637	logMelFreqBand-sma-de[4]-kurtosis	0.2012	mfcc-sma[12]-iqr2-3
0.5636	mfcc-sma-de[13]-linregerrQ	0.201	logMelFreqBand-sma-de[6]-quartile2
0.5631	logMelFreqBand-sma[4]-Percentile99.0	0.2009	mfcc-sma[14]-linregc1
0.5627	mfcc-sma-de[0]-linregerrQ	0.2	logMelFreqBand-sma[7]-stddev
0.5624	F0finEnv-sma-upleveltime75	0.1997	mfcc-sma[5]-kurtosis
0.5624	logMelFreqBand-sma-de[0]-linregc1	0.1994	mfcc-sma[6]-minPos
0.5612	mfcc-sma-de[5]-quartile1	0.1983	mfcc-sma-de[9]-Percentile99.0
0.5608	logMelFreqBand-sma-de[6]-iqr2-3	0.1983	mfcc-sma-de[14]-minPos

Ranking	Feature	Ranking	Feature
0.5592	voicingFinalUnclipped-sma-upleveltime75	0.198	logMelFreqBand-sma[3]-iqr1-3
0.557	mfcc-sma-de[0]-linregc1	0.198	LSPFreq-sma[2]-amean
0.5549	shimmerLocal-sma-quartile2	0.1979	mfcc-sma[14]-iqr1-2
0.5529	jitterDDP-sma-de-linregerrA	0.1964	logMelFreqBand-sma[3]-upleveltime90
0.5529	logMelFreqBand-sma[3]-linregerrA	0.1964	LSPFreq-sma[2]-quartile1
0.5527	mfcc-sma[5]-linregerrA	0.1952	logMelFreqBand-sma[4]-upleveltime90
0.5504	logMelFreqBand-sma-de[2]-pctlrange0-1	0.1939	LSPFreq-sma[3]-linregerrQ
0.5504	mfcc-sma-de[2]-iqr2-3	0.1935	mfcc-sma[0]-linregc2
0.5502	mfcc-sma-de[14]-linregerrA	0.1933	logMelFreqBand-sma[4]-iqr1-2
0.5492	logMelFreqBand-sma-de[3]-kurtosis	0.1929	mfcc-sma[11]-maxPos
0.5482	shimmerLocal-sma-de-iqr2-3	0.1925	mfcc-sma[10]-Percentile1.0
0.5481	mfcc-sma-de[2]-quartile1	0.1923	mfcc-sma[12]-iqr1-3
0.5481	mfcc-sma[5]-linregerrQ	0.1918	pcm-loudness-sma-iqr1-2
0.5467	logMelFreqBand-sma[5]-Percentile1.0	0.1916	logMelFreqBand-sma[7]-linregerrQ
0.5466	logMelFreqBand-sma-de[7]-iqr1-3	0.1916	mfcc-sma-de[13]-minPos
0.5461	F0finEnv-sma-linregerrQ	0.1913	mfcc-sma-de[1]-pctlrange0-1
0.5425	mfcc-sma-de[2]-iqr1-2	0.19	LSPFreq-sma[4]-linregerrQ
0.5424	mfcc-sma-de[14]-iqr1-3	0.19	LSPFreq-sma-de[6]-amean
0.5419	mfcc-sma-de[9]-stddev	0.1899	LSPFreq-sma-de[5]-iqr2-3
0.5408	logMelFreqBand-sma-de[6]-iqr1-2	0.1892	mfcc-sma[7]-iqr1-3
0.5405	mfcc-sma-de[13]-linregerrA	0.1887	LSPFreq-sma[2]-linregerrA
0.5403	mfcc-sma-de[11]-linregerrA	0.1887	mfcc-sma[10]-Percentile99.0
0.5402	logMelFreqBand-sma[4]-linregerrQ	0.1884	mfcc-sma-de[1]-maxPos
0.5394	mfcc-sma-de[9]-linregerrQ	0.1881	LSPFreq-sma[3]-quartile2
0.539	shimmerLocal-sma-de-stddev	0.1878	logMelFreqBand-sma-de[3]-maxPos
0.5386	logMelFreqBand-sma-de[7]-linregc2	0.1875	pcm-loudness-sma-iqr1-3
0.5384	mfcc-sma-de[9]-linregc1	0.1873	mfcc-sma[1]-linregerrA
0.5381	mfcc-sma[2]-linregerrA	0.1871	mfcc-sma-de[5]-minPos
0.5378	Do1000(offset)	0.1862	mfcc-sma-de[2]-amean
0.5364	mfcc-sma[2]-stddev	0.1862	mfcc-sma-de[1]-linregerrQ
0.5346	shimmerLocal-sma-de-iqr1-3	0.1862	mfcc-sma-de[1]-stddev
0.5299	logMelFreqBand-sma-de[7]-iqr2-3	0.1859	mfcc-sma[6]-skewness
0.5297	logMelFreqBand-sma-de[1]-linregerrQ	0.1856	LSPFreq-sma-de[2]-amean
0.5285	mfcc-sma[3]-quartile1	0.1854	logMelFreqBand-sma[6]-quartile3
0.5272	mfcc-sma-de[8]-iqr1-3	0.1853	mfcc-sma[5]-Percentile99.0
0.5269	shimmerLocal-sma-iqr1-3	0.185	mfcc-sma-de[6]-Percentile1.0
0.5265	logMelFreqBand-sma[3]-linregerrQ	0.1848	voice break(number)
0.5259	F0finEnv-sma-quartile1	0.1846	mfcc-sma[10]-linregerrA
0.5222	mfcc-sma[5]-stddev	0.1845	LSPFreq-sma[3]-pctlrange0-1
0.5213	logMelFreqBand-sma-de[6]-linregerrA	0.1839	mfcc-sma[6]-linregerrA
0.5209	mfcc-sma-de[11]-iqr2-3	0.1834	logMelFreqBand-sma-de[0]-upleveltime75
0.5208	mfcc-sma-de[14]-quartile3	0.1827	mfcc-sma-de[8]-maxPos
0.5196	mfcc-sma[5]-Percentile1.0	0.1822	mfcc-sma[13]-linregc1
0.5186	mfcc-sma-de[9]-linregc2	0.1817	LSPFreq-sma[3]-minPos
0.5183	logMelFreqBand-sma-de[3]-linregc2	0.1816	jitterDDP-sma-de-maxPos
0.5177	logMelFreqBand-sma-de[5]-kurtosis	0.1813	LSPFreq-sma[2]-quartile2
0.5173	voice break(degree)	0.1812	LSPFreq-sma-de[3]-iqr2-3
0.5167	mfcc-sma-de[2]-linregc1	0.1811	shimmerLocal-sma-de-kurtosis
0.5167	logMelFreqBand-sma[3]-Percentile1.0	0.181	voicingFinalUnclipped-sma-de-upleveltime75
0.5166	logMelFreqBand-sma-de[6]-linregerrQ	0.1808	mfcc-sma-de[1]-quartile2
0.5149	mfcc-sma-de[12]-quartile3	0.1806	logMelFreqBand-sma-de[7]-upleveltime90
0.5143	mfcc-sma-de[0]-linregc2	0.1803	LSPFreq-sma-de[4]-minPos
0.5138	logMelFreqBand-sma[3]-Percentile99.0	0.1797	mfcc-sma[10]-quartile3
0.513	logMelFreqBand-sma-de[7]-quartile3	0.1791	logMelFreqBand-sma[3]-minPos
0.5123	logMelFreqBand-sma-de[0]-linregc2	0.1788	logMelFreqBand-sma-de[2]-quartile2
0.5107	mfcc-sma[5]-iqr1-3	0.1785	F0final-sma-minPos
0.5105	mfcc-sma-de[9]-quartile3	0.1781	jitterDDP-sma-de-linregc1
0.5082	LSPFreq-sma-de[1]-iqr1-3	0.1781	mfcc-sma-de[1]-Percentile99.0
0.508	logMelFreqBand-sma-de[5]-linregc1	0.1779	mfcc-sma[12]-Percentile1.0
0.5074	logMelFreqBand-sma-de[6]-stddev	0.1778	mfcc-sma[6]-linregerrQ
0.5072	shimmerLocal-sma-linregerrA	0.1776	logMelFreqBand-sma[0]-stddev
0.5062	logMelFreqBand-sma[4]-stddev	0.1775	mfcc-sma-de[0]-maxPos
0.5061	mfcc-sma-de[8]-quartile3	0.1774	logMelFreqBand-sma-de[2]-maxPos

Ranking	Feature	Ranking	Feature
0.5057	logMelFreqBand-sma-de[7]-linregerrA	0.1769	mfcc-sma-de[13]-maxPos
0.5055	F0finEnv-sma-linregerrA	0.1769	LSPFreq-sma-de[5]-kurtosis
0.5055	LSPFreq-sma-de[2]-iqr2-3	0.1761	mfcc-sma-de[7]-Percentile1.0
0.5043	mfcc-sma-de[6]-iqr2-3	0.1758	mfcc-sma[4]-pctlrange0-1
0.5041	mfcc-sma[13]-linregerrQ	0.1758	jitterDDP-sma-de-upleveltime75
0.5038	logMelFreqBand-sma-de[7]-quartile1	0.1756	mfcc-sma[1]-Percentile1.0
0.5038	mfcc-sma-de[12]-Percentile1.0	0.1751	mfcc-sma[6]-linregc1
0.5027	logMelFreqBand-sma-de[1]-linregerrA	0.1749	LSPFreq-sma[6]-stddev
0.5019	mfcc-sma[1]-quartile1	0.1747	logMelFreqBand-sma[5]-iqr1-2
0.5013	mfcc-sma-de[8]-iqr2-3	0.1744	LSPFreq-sma[2]-minPos
0.5001	F0finEnv-sma-iqr1-2	0.1744	LSPFreq-sma[6]-minPos
0.5	F0final-sma-iqr1-2	0.1742	logMelFreqBand-sma-de[4]-minPos
0.495	logMelFreqBand-sma-de[6]-linregc2	0.174	pcm-loudness-sma-quartile3
0.4949	mfcc-sma-de[9]-quartile1	0.1734	LSPFreq-sma-de[6]-iqr1-2
0.4928	jitter(rap)	0.1723	jitterDDP-sma-de-minPos
0.4925	Hamml	0.172	mfcc-sma-de[1]-linregerrA
0.4912	mfcc-sma-de[5]-Percentile1.0	0.1719	LSPFreq-sma-de[2]-quartile2
0.4911	logMelFreqBand-sma[1]-Percentile1.0	0.1718	voicingFinalUnclipped-sma-de-kurtosis
0.4892	F0finEnv-sma-de-linregc1	0.1715	logMelFreqBand-sma-de[7]-Percentile1.0
0.4885	logMelFreqBand-sma-de[1]-linregc2	0.1714	LSPFreq-sma[6]-quartile1
0.4866	logMelFreqBand-sma-de[1]-stddev	0.1712	logMelFreqBand-sma[3]-upleveltime75
0.4856	jitterDDP-sma-linregerrQ	0.1701	LSPFreq-sma-de[1]-linregc1
0.4843	pcm-loudness-sma-Percentile1.0	0.1695	mfcc-sma-de[9]-maxPos
0.4832	mfcc-sma-de[14]-pctlrange0-1	0.1692	jitterDDP-sma-de-quartile2
0.4831	LSPFreq-sma-de[2]-quartile3	0.1692	logMelFreqBand-sma-de[4]-quartile2
0.4831	logMelFreqBand-sma[2]-Percentile99.0	0.1688	logMelFreqBand-sma-de[2]-upleveltime75
0.4813	F0final-sma-quartile1	0.1686	mfcc-sma-de[3]-iqr1-2
0.4787	logMelFreqBand-sma-de[3]-pctlrange0-1	0.1686	mfcc-sma[1]-stddev
0.4775	jitterDDP-sma-stddev	0.1679	logMelFreqBand-sma-de[1]-Percentile1.0
0.4767	logMelFreqBand-sma-de[6]-linregc1	0.1677	mfcc-sma-de[4]-maxPos
0.4755	mfcc-sma-de[13]-iqr1-3	0.1676	LSPFreq-sma-de[7]-maxPos
0.4732	jitterDDP-sma-linregerrA	0.1676	logMelFreqBand-sma[0]-linregerrA
0.4726	logMelFreqBand-sma-de[2]-linregc1	0.1674	LSPFreq-sma-de[5]-linregc2
0.472	LSPFreq-sma-de[2]-iqr1-3	0.1673	LSPFreq-sma-de[4]-iqr1-3
0.472	mfcc-sma[4]-quartile1	0.1672	LSPFreq-sma-de[1]-quartile2
0.4697	logMelFreqBand-sma[5]-linregerrQ	0.1668	logMelFreqBand-sma[2]-skewness
0.4696	mfcc-sma-de[12]-iqr2-3	0.1665	logMelFreqBand-sma-de[4]-skewness
0.4685	mfcc-sma-de[13]-iqr2-3	0.1664	LSPFreq-sma-de[4]-linregc2
0.4683	mfcc-sma-de[11]-linregerrQ	0.1663	LSPFreq-sma-de[6]-stddev
0.4675	logMelFreqBand-sma[2]-linregerrA	0.1659	LSPFreq-sma[2]-upleveltime75
0.4667	logMelFreqBand-sma[7]-upleveltime75	0.1658	shimmerLocal-sma-de-linregc2
0.4663	logMelFreqBand-sma-de[1]-iqr1-3	0.1657	Do1000(slope)
0.4663	mfcc-sma-de[4]-linregerrQ	0.1657	LSPFreq-sma[1]-linregerrQ
0.4662	mfcc-sma-de[14]-iqr1-2	0.1652	mfcc-sma-de[4]-Percentile99.0
0.4658	mfcc-sma[2]-iqr1-2	0.1645	LSPFreq-sma-de[6]-linregerrQ
0.4654	mfcc-sma-de[11]-stddev	0.1638	LSPFreq-sma[4]-stddev
0.4639	mfcc-sma-de[11]-quartile1	0.1637	mfcc-sma[6]-stddev
0.4638	F0finEnv-sma-upleveltime90	0.1632	LSPFreq-sma-de[4]-quartile1
0.4632	mfcc-sma[10]-quartile1	0.162	mfcc-sma[0]-minPos
0.4627	mfcc-sma[4]-amean	0.1619	logMelFreqBand-sma[3]-linregc2
0.4627	mfcc-sma-de[14]-iqr2-3	0.1619	LSPFreq-sma-de[1]-maxPos
0.4627	F0final-sma-de-quartile1	0.1605	LSPFreq-sma-de[3]-kurtosis
0.4625	mfcc-sma[0]-stddev	0.1602	LSPFreq-sma-de[7]-amean
0.4624	voicingFinalUnclipped-sma-kurtosis	0.1597	logMelFreqBand-sma[6]-maxPos
0.4622	mfcc-sma-de[4]-quartile1	0.1596	mfcc-sma[14]-linregc2
0.462	logMelFreqBand-sma[0]-quartile3	0.1593	logMelFreqBand-sma-de[5]-quartile2
0.4618	logMelFreqBand-sma[0]-Percentile1.0	0.1592	logMelFreqBand-sma[3]-iqr2-3
0.4615	logMelFreqBand-sma-de[1]-quartile1	0.1591	mfcc-sma[13]-quartile3
0.4615	jitterLocal-sma-quartile3	0.1591	logMelFreqBand-sma-de[6]-skewness
0.4594	mfcc-sma[5]-amean	0.159	LSPFreq-sma[3]-Percentile1.0
0.4591	logMelFreqBand-sma-de[4]-Percentile1.0	0.1582	logMelFreqBand-sma-de[5]-amean
0.4588	mfcc-sma-de[5]-iqr2-3	0.1581	LSPFreq-sma-de[5]-minPos
0.4587	F0final-sma-linregc2	0.158	mfcc-sma[14]-maxPos

Ranking	Feature	Ranking	Feature
0.4586	jitterLocal-sma-de-quartile3	0.1573	LSPFreq-sma[3]-quartile3
0.4566	logMelFreqBand-sma-de[1]-linregc1	0.1573	mfcc-sma-de[12]-maxPos
0.4566	mfcc-sma-de[10]-linregerrA	0.1568	logMelFreqBand-sma[6]-linregc1
0.4561	logMelFreqBand-sma[0]-linregc2	0.1566	mfcc-sma[7]-quartile2
0.4555	mfcc-sma[2]-iqr2-3	0.1562	mfcc-sma[6]-iqr1-3
0.455	shimmerLocal-sma-de-Percentile99.0	0.1554	mfcc-sma[3]-Percentile99.0
0.4544	jitterDDP-sma-de-Percentile99.0	0.1553	logMelFreqBand-sma[2]-linregc2
0.4517	logMelFreqBand-sma-de[4]-linregc2	0.1551	mfcc-sma-de[9]-upleveltime90
0.4488	logMelFreqBand-sma[1]-pctlrange0-1	0.1549	LSPFreq-sma-de[7]-Percentile1.0
0.4485	mfcc-sma-de[14]-quartile1	0.1541	mfcc-sma-de[6]-upleveltime90
0.4473	logMelFreqBand-sma-de[2]-Percentile99.0	0.154	LSPFreq-sma[5]-minPos
0.4467	F0final-sma-de-iqr1-2	0.1538	jitterLocal-sma-maxPos
0.4466	mfcc-sma-de[10]-stddev	0.1538	mfcc-sma[10]-minPos
0.445	mfcc-sma-de[0]-kurtosis	0.1537	shimmerLocal-sma-de-minPos
0.4438	mfcc-sma-de[6]-linregerrQ	0.1535	mfcc-sma[7]-linregerrQ
0.4434	mfcc-sma-de[13]-pctlrange0-1	0.1528	mfcc-sma[13]-quartile2
0.4431	mfcc-sma-de[6]-stddev	0.1526	mfcc-sma-de[5]-amean
0.4429	jitterLocal-sma-de-iqr1-3	0.1525	mfcc-sma-de[11]-maxPos
0.4427	mfcc-sma-de[10]-linregerrQ	0.1525	pcm-loudness-sma-upleveltime75
0.4418	mfcc-sma-de[14]-Percentile99.0	0.1524	LSPFreq-sma[6]-linregerrQ
0.4418	logMelFreqBand-sma[2]-linregerrQ	0.1523	mfcc-sma-de[1]-quartile1
0.4412	mfcc-sma-de[10]-iqr1-2	0.1522	LSPFreq-sma-de[1]-linregc2
0.441	voicingFinalUnclipped-sma-linregc2	0.152	mfcc-sma[7]-linregerrA
0.4406	logMelFreqBand-sma[2]-stddev	0.1518	shimmerLocal-sma-de-maxPos
0.4388	mfcc-sma-de[8]-Percentile99.0	0.1517	jitterDDP-sma-maxPos
0.4385	logMelFreqBand-sma-de[2]-kurtosis	0.1515	logMelFreqBand-sma[0]-linregerrQ
0.4382	F0final-sma-de-quartile3	0.1515	logMelFreqBand-sma-de[0]-amean
0.4378	voicingFinalUnclipped-sma-de-linregc2	0.1513	mfcc-sma[2]-skewness
0.4376	mfcc-sma-de[5]-pctlrange0-1	0.1509	LSPFreq-sma[1]-kurtosis
0.4373	logMelFreqBand-sma-de[0]-Percentile99.0	0.1508	mfcc-sma[7]-amean
0.4371	logMelFreqBand-sma-de[3]-linregc1	0.1505	mfcc-sma[10]-maxPos
0.4349	mfcc-sma[5]-iqr1-2	0.1504	mfcc-sma[10]-iqr1-2
0.434	voicingFinalUnclipped-sma-de-linregc1	0.1503	LSPFreq-sma[4]-pctlrange0-1
0.4328	logMelFreqBand-sma[0]-kurtosis	0.15	LSPFreq-sma[5]-kurtosis
0.432	logMelFreqBand-sma-de[4]-linregc1	0.1499	mfcc-sma[13]-iqr2-3
0.4314	mfcc-sma-de[2]-Percentile99.0	0.1494	logMelFreqBand-sma-de[0]-maxPos
0.4311	shimmerLocal-sma-stddev	0.1489	logMelFreqBand-sma-de[2]-skewness
0.4305	mfcc-sma[13]-pctlrange0-1	0.1488	logMelFreqBand-sma-de[6]-amean
0.4302	mfcc-sma[0]-linregerrQ	0.1487	mfcc-sma[11]-amean
0.4282	mfcc-sma[13]-stddev	0.148	mfcc-sma-de[10]-maxPos
0.4273	logMelFreqBand-sma[6]-upleveltime90	0.1477	mfcc-sma[13]-Percentile99.0
0.4258	LSPFreq-sma-de[1]-quartile1	0.1467	mfcc-sma[7]-Percentile1.0
0.4254	mfcc-sma[3]-amean	0.1466	logMelFreqBand-sma[2]-upleveltime90
0.4245	jitterLocal-sma-de-iqr1-2	0.1466	LSPFreq-sma[6]-skewness
0.4234	shimmerLocal-sma-iqr2-3	0.1465	mfcc-sma[7]-maxPos
0.422	mfcc-sma[8]-amean	0.1463	F0final-sma-skewness
0.4215	logMelFreqBand-sma[4]-quartile3	0.1461	H(std)
0.42	pcm-loudness-sma-de-linregc1	0.1457	pcm-loudness-sma-de-minPos
0.4196	F0final-sma-de-iqr1-3	0.1457	mfcc-sma-de[14]-maxPos
0.4191	jitterDDP-sma-amean	0.1454	mfcc-sma[5]-minPos
0.419	mfcc-sma-de[4]-linregerrA	0.1453	mfcc-sma-de[5]-upleveltime90
0.4185	logMelFreqBand-sma-de[5]-Percentile1.0	0.1452	mfcc-sma[9]-minPos
0.4182	mfcc-sma-de[0]-Percentile99.0	0.1451	mfcc-sma-de[12]-amean
0.4178	LSPFreq-sma-de[0]-quartile1	0.145	LSPFreq-sma[5]-quartile2
0.4177	mfcc-sma-de[8]-iqr1-2	0.145	mfcc-sma-de[13]-linregc1
0.4165	logMelFreqBand-sma[4]-linregerrA	0.1448	mfcc-sma[12]-minPos
0.4162	mfcc-sma-de[10]-linregc1	0.144	mfcc-sma[9]-linregc1
0.416	mfcc-sma-de[4]-stddev	0.1438	LSPFreq-sma-de[0]-minPos
0.4159	F0final–Turn-duration	0.1432	LSPFreq-sma-de[4]-quartile3
0.4158	mfcc-sma-de[10]-quartile1	0.143	mfcc-sma[7]-stddev
0.4157	logMelFreqBand-sma[5]-linregerrA	0.1425	pcm-loudness-sma-skewness
0.4154	jitterLocal-sma-de-linregerrA	0.1424	logMelFreqBand-sma-de[3]-skewness
0.4152	mfcc-sma-de[6]-linregerrA	0.1413	LSPFreq-sma[1]-iqr1-3

Ranking	Feature	Ranking	Feature
0.4146	logMelFreqBand-sma-de[0]-stddev	0.1412	logMelFreqBand-sma-de[6]-upleveltime75
0.4134	jitterLocal-sma-de-upleveltime90	0.141	voicingFinalUnclipped-sma-linregc1
0.4129	logMelFreqBand-sma-de[7]-stddev	0.1405	jitterLocal-sma-linregc2
0.4103	jitterLocal-sma-de-iqr2-3	0.1404	F0final-sma-linregc1
0.4089	mfcc-sma[1]-linregc2	0.14	logMelFreqBand-sma-de[0]-kurtosis
0.4088	mfcc-sma[13]-linregerrA	0.1398	LSPFreq-sma-de[5]-maxPos
0.4083	F0finEnv-sma-de-upleveltime75	0.1397	logMelFreqBand-sma[6]-iqr1-3
0.4081	F0finEnv-sma-linregc2	0.1396	mfcc-sma[13]-amean
0.4046	logMelFreqBand-sma-de[1]-iqr1-2	0.139	jitterDDP-sma-de-linregc2
0.4042	mfcc-sma[0]-kurtosis	0.1389	logMelFreqBand-sma[2]-iqr1-2
0.404	LSPFreq-sma-de[1]-iqr2-3	0.1388	logMelFreqBand-sma[6]-iqr1-2
0.4038	logMelFreqBand-sma[0]-skewness	0.1385	LSPFreq-sma-de[6]-Percentile1.0
0.4022	F0finEnv-sma-linregc1	0.1381	mfcc-sma[6]-iqr2-3
0.4019	mfcc-sma[4]-quartile2	0.138	LSPFreq-sma-de[4]-iqr1-2
0.4017	LSPFreq-sma-de[0]-iqr1-3	0.137	LSPFreq-sma[6]-iqr1-3
0.4008	jitterLocal-sma-linregerrA	0.1369	mfcc-sma[1]-linregerrQ
0.4007	jitterDDP-sma-Percentile99.0	0.1368	mfcc-sma[4]-minPos
0.4002	logMelFreqBand-sma[6]-upleveltime75	0.1363	LSPFreq-sma-de[5]-linregerrA
0.3999	logMelFreqBand-sma-de[1]-iqr2-3	0.136	mfcc-sma[2]-minPos
0.3998	logMelFreqBand-sma-de[0]-linregerrA	0.1358	logMelFreqBand-sma[7]-linregerrA
0.3998	mfcc-sma-de[0]-pctlrange0-1	0.1356	LSPFreq-sma-de[4]-Percentile1.0
0.3994	shimmerLocal-sma-iqr1-2	0.1356	LSPFreq-sma[3]-iqr1-3
0.398	logMelFreqBand-sma-de[0]-pctlrange0-1	0.1356	LSPFreq-sma-de[6]-pctlrange0-1
0.3975	logMelFreqBand-sma-de[6]-kurtosis	0.1354	mfcc-sma[11]-iqr2-3
0.3965	shimmerLocal-sma-de-quartile2	0.1354	mfcc-sma-de[4]-minPos
0.3959	logMelFreqBand-sma[4]-iqr1-3	0.1353	mfcc-sma-de[8]-quartile2
0.3948	F0final-sma-de-kurtosis	0.1351	logMelFreqBand-sma[5]-maxPos
0.3943	mfcc-sma[3]-quartile2	0.1347	mfcc-sma-de[11]-linregc1
0.3941	jitterLocal-sma-iqr1-2	0.1342	LSPFreq-sma[4]-linregerrA
0.3926	jitterLocal-sma-de-linregerrQ	0.1338	LSPFreq-sma[6]-Percentile1.0
0.392	LSPFreq-sma-de[0]-iqr1-2	0.1333	LSPFreq-sma-de[4]-Percentile99.0
0.3919	mfcc-sma-de[4]-iqr1-3	0.1333	LSPFreq-sma[5]-iqr2-3
0.3919	logMelFreqBand-sma-de[1]-quartile3	0.1332	mfcc-sma-de[3]-kurtosis
0.3918	shimmerLocal-sma-Percentile99.0	0.1323	pcm-loudness-sma-de-iqr1-2
0.3911	mfcc-sma-de[9]-iqr2-3	0.1323	voicingFinalUnclipped-sma-de-linregerrA
0.3904	LSPFreq-sma[0]-linregc2	0.1322	LSPFreq-sma-de[2]-maxPos
0.3904	logMelFreqBand-sma[6]-pctlrange0-1	0.132	mfcc-sma-de[6]-maxPos
0.39	F0finEnv-sma-de-upleveltime90	0.132	jitterDDP-sma-linregc2
0.3893	mfcc-sma[14]-pctlrange0-1	0.1317	mfcc-sma[14]-quartile2
0.3888	mfcc-sma-de[8]-pctlrange0-1	0.1312	LSPFreq-sma[6]-kurtosis
0.3887	mfcc-sma-de[4]-iqr2-3	0.1308	jitterLocal-sma-de-minPos
0.3877	mfcc-sma-de[2]-Percentile1.0	0.1307	logMelFreqBand-sma[4]-upleveltime75
0.3873	LSPFreq-sma[0]-Percentile1.0	0.1301	LSPFreq-sma[3]-upleveltime75
0.3871	mfcc-sma[12]-linregerrA	0.1296	logMelFreqBand-sma-de[5]-upleveltime90
0.387	voicedunvoiced ratio	0.1294	LSPFreq-sma-de[4]-kurtosis
0.387	voicedtotal frames ratio	0.1293	logMelFreqBand-sma[2]-minPos
0.3861	mfcc-sma-de[6]-iqr1-3	0.129	LSPFreq-sma[3]-iqr2-3
0.3859	unvoicedtotal frames ratio	0.1287	LSPFreq-sma-de[5]-amean
0.3859	mfcc-sma-de[13]-quartile1	0.1286	shimmerLocal-sma-skewness
0.3832	jitterLocal-sma-iqr1-3	0.1285	logMelFreqBand-sma[5]-linregc2
0.383	F0final-sma-de-upleveltime75	0.1283	logMelFreqBand-sma-de[0]-Percentile1.0
0.3828	jitterDDP-sma-de-linregerrQ	0.1282	mfcc-sma[3]-skewness
0.3824	logMelFreqBand-sma-de[3]-Percentile99.0	0.1274	LSPFreq-sma-de[5]-quartile3
0.3821	mfcc-sma-de[7]-stddev	0.1274	LSPFreq-sma-de[4]-iqr2-3
0.3819	mfcc-sma-de[3]-linregc1	0.1268	pcm-loudness-sma-de-iqr1-3
0.3812	mfcc-sma[12]-pctlrange0-1	0.1267	mfcc-sma[4]-maxPos
0.3812	logMelFreqBand-sma-de[2]-Percentile1.0	0.1263	jitterDDP-sma-de-skewness
0.3809	logMelFreqBand-sma-de[0]-linregerrQ	0.126	LSPFreq-sma-de[1]-minPos
0.3808	mfcc-sma-de[13]-quartile3	0.1255	logMelFreqBand-sma-de[7]-minPos
0.3799	logMelFreqBand-sma-de[1]-Percentile99.0	0.1248	mfcc-sma[12]-linregc2
0.3797	mfcc-sma-de[3]-linregerrA	0.1246	logMelFreqBand-sma[7]-maxPos
0.3796	logMelFreqBand-sma-de[7]-linregerrQ	0.1243	logMelFreqBand-sma[3]-iqr1-2
0.379	logMelFreqBand-sma-de[0]-iqr1-3	0.1242	mfcc-sma[5]-skewness

Ranking	Feature	Ranking	Feature
0.3774	mfcc-sma-de[2]-linregc2	0.1241	mfcc-sma[9]-linregc2
0.3773	mfcc-sma[0]-quartile1	0.124	LSPFreq-sma[5]-quartile1
0.3772	mfcc-sma[3]-Percentile1.0	0.1239	mfcc-sma-de[6]-kurtosis
0.3747	jitterDDP-sma-de-stddev	0.1235	mfcc-sma-de[1]-minPos
0.3747	shimmerLocal-sma-quartile1	0.1235	mfcc-sma[9]-Percentile99.0
0.3738	LSPFreq-sma-de[2]-linregerrA	0.1232	logMelFreqBand-sma[2]-quartile2
0.3721	mfcc-sma[2]-linregc2	0.1231	LSPFreq-sma[7]-kurtosis
0.3716	mfcc-sma[8]-quartile2	0.1224	logMelFreqBand-sma[4]-maxPos
0.3716	F0final-sma-iqr2-3	0.122	logMelFreqBand-sma[5]-minPos
0.3708	mfcc-sma-de[8]-linregc1	0.1217	LSPFreq-sma[1]-minPos
0.3704	LSPFreq-sma[1]-iqr2-3	0.1217	mfcc-sma[0]-quartile3
0.3701	logMelFreqBand-sma[0]-pctlrange0-1	0.1214	LSPFreq-sma-de[1]-stddev
0.3695	jitterLocal-sma-quartile1	0.1214	mfcc-sma-de[0]-quartile2
0.3694	mfcc-sma[10]-quartile2	0.1211	mfcc-sma[12]-iqr1-2
0.3688	mfcc-sma-de[10]-iqr2-3	0.1211	logMelFreqBand-sma[3]-amean
0.3688	jitterDDP-sma-de-quartile3	0.1203	logMelFreqBand-sma[4]-linregc2
0.3682	shimmerLocal-sma-upleveltime75	0.1197	mfcc-sma[3]-minPos
0.3681	mfcc-sma-de[7]-iqr1-3	0.1196	mfcc-sma-de[3]-maxPos
0.3678	mfcc-sma-de[10]-iqr1-3	0.1196	LSPFreq-sma-de[2]-Percentile99.0
0.3667	mfcc-sma-de[14]-Percentile1.0	0.1194	LSPFreq-sma[6]-maxPos
0.3661	jitter(ppq5)	0.1192	mfcc-sma[4]-iqr1-2
0.365	mfcc-sma-de[8]-Percentile1.0	0.1191	LSPFreq-sma-de[0]-maxPos
0.3645	mfcc-sma[0]-upleveltime75	0.119	jitterLocal-sma-de-maxPos
0.3633	Period(std)	0.1188	voicingFinalUnclipped-sma-stddev
0.3633	mfcc-sma-de[4]-iqr1-2	0.1187	LSPFreq-sma[0]-linregerrQ
0.3622	logMelFreqBand-sma[2]-quartile3	0.1187	LSPFreq-sma-de[1]-linregerrQ
0.3622	jitterLocal-sma-de-stddev	0.1185	mfcc-sma-de[2]-maxPos
0.3617	mfcc-sma[1]-iqr1-2	0.1184	logMelFreqBand-sma[0]-quartile1
0.3611	logMelFreqBand-sma[5]-stddev	0.1177	mfcc-sma[12]-maxPos
0.3608	mfcc-sma-de[13]-Percentile99.0	0.1176	mfcc-sma[6]-Percentile99.0
0.3607	logMelFreqBand-sma[1]-linregc1	0.1175	logMelFreqBand-sma-de[4]-maxPos
0.3606	mfcc-sma-de[7]-linregerrQ	0.1173	mfcc-sma[6]-pctlrange0-1
0.36	shimmerLocal-sma-de-iqr1-2	0.1172	mfcc-sma[11]-quartile1
0.359	logMelFreqBand-sma[1]-linregc2	0.1171	mfcc-sma-de[0]-upleveltime75
0.359	mfcc-sma-de[3]-quartile3	0.117	voicingFinalUnclipped-sma-de-minPos
0.3584	LSPFreq-sma-de[2]-quartile1	0.1169	shimmerLocal-sma-de-skewness
0.3581	jitterDDP-sma-quartile3	0.1161	mfcc-sma[11]-Percentile1.0
0.3572	mfcc-sma-de[7]-linregc1	0.1159	mfcc-sma-de[14]-linregc1
0.3568	LSPFreq-sma-de[0]-quartile3	0.1159	LSPFreq-sma-de[6]-Percentile99.0
0.3565	mfcc-sma-de[6]-iqr1-2	0.1155	mfcc-sma[14]-iqr2-3
0.3549	mfcc-sma[9]-quartile2	0.1154	logMelFreqBand-sma[4]-minPos
0.3546	mfcc-sma[4]-quartile3	0.1152	mfcc-sma-de[6]-minPos
0.353	mfcc-sma-de[9]-pctlrange0-1	0.1148	LSPFreq-sma[6]-upleveltime75
0.3525	mfcc-sma-de[1]-iqr1-2	0.1147	mfcc-sma[8]-iqr1-2
0.3523	mfcc-sma[9]-amean	0.1146	logMelFreqBand-sma[1]-upleveltime90
0.3516	LSPFreq-sma[1]-skewness	0.1143	LSPFreq-sma-de[7]-linregerrA
0.3515	jitterDDP-sma-iqr1-3	0.1141	mfcc-sma-de[13]-upleveltime90
0.3515	mfcc-sma-de[8]-linregc2	0.1137	logMelFreqBand-sma[7]-iqr1-2
0.3514	mfcc-sma-de[3]-iqr1-3	0.1137	mfcc-sma[6]-iqr1-2
0.3513	LSPFreq-sma-de[1]-iqr1-2	0.1134	logMelFreqBand-sma-de[0]-skewness
0.3507	mfcc-sma[0]-maxPos	0.1133	pcm-loudness-sma-de-quartile1
0.3505	logMelFreqBand-sma[3]-quartile1	0.1132	mfcc-sma[4]-Percentile99.0
0.3499	F0final-sma-de-upleveltime90	0.113	LSPFreq-sma[0]-linregerrA
0.3497	logMelFreqBand-sma[4]-linregc1	0.1125	jitterLocal-sma-de-quartile2
0.3485	mfcc-sma[9]-linregerrA	0.1124	LSPFreq-sma[3]-maxPos
0.3473	LSPFreq-sma-de[1]-kurtosis	0.1119	LSPFreq-sma[2]-maxPos
0.3471	mfcc-sma[14]-Percentile1.0	0.1119	mfcc-sma-de[11]-linregc2
0.3467	mfcc-sma-de[6]-quartile3	0.1116	mfcc-sma-de[7]-kurtosis
0.344	mfcc-sma-de[1]-iqr1-3	0.111	LSPFreq-sma-de[7]-Percentile99.0
0.3433	F0final-sma-kurtosis	0.1109	logMelFreqBand-sma[3]-skewness
0.3428	mfcc-sma-de[9]-iqr1-2	0.1109	LSPFreq-sma-de[7]-iqr1-3
0.3418	mfcc-sma-de[10]-quartile3	0.1105	logMelFreqBand-sma-de[4]-upleveltime90
0.3406	jitterLocal-sma-Percentile99.0	0.1104	LSPFreq-sma[7]-upleveltime75

Ranking	Feature	Ranking	Feature
0.3404	mfcc-sma[12]-linregerrQ	0.1104	logMelFreqBand-sma-de[7]-upleveltime75
0.3402	jitterDDP-sma-iqr2-3	0.1102	mfcc-sma[13]-maxPos
0.3398	pcm-loudness-sma-linregc1	0.11	logMelFreqBand-sma[1]-iqr2-3
0.3394	logMelFreqBand-sma-de[5]-Percentile99.0	0.1098	mfcc-sma-de[5]-kurtosis
0.3393	voicingFinalUnclipped-sma-de-iqr1-2	0.1098	LSPFreq-sma-de[6]-linregc1
0.339	logMelFreqBand-sma-de[1]-pctlrange0-1	0.1098	LSPFreq-sma[2]-linregc2
0.339	mfcc-sma[10]-amean	0.1092	shimmerLocal-sma-maxPos
0.3387	mfcc-sma-de[10]-pctlrange0-1	0.1092	LSPFreq-sma-de[6]-linregc2
0.3375	mfcc-sma[1]-Percentile99.0	0.1089	mfcc-sma-de[7]-minPos
0.3358	logMelFreqBand-sma[0]-quartile2	0.1088	voicingFinalUnclipped-sma-linregerrQ
0.3355	mfcc-sma-de[1]-linregc2	0.1088	pcm-loudness-sma-de-linregerrA
0.3343	LSPFreq-sma[0]-iqr2-3	0.1083	mfcc-sma[11]-iqr1-2
0.3342	mfcc-sma[8]-quartile1	0.1082	logMelFreqBand-sma[2]-kurtosis
0.3336	mfcc-sma-de[3]-linregc2	0.1081	LSPFreq-sma[2]-kurtosis
0.3313	logMelFreqBand-sma[6]-skewness	0.1076	pcm-loudness-sma-de-amean
0.3309	jitterLocal-sma-de-Percentile99.0	0.1075	LSPFreq-sma-de[7]-kurtosis
0.3305	mfcc-sma[9]-linregerrQ	0.1074	logMelFreqBand-sma[7]-iqr1-3
0.3285	pcm-loudness-sma-linregc2	0.1062	mfcc-sma[7]-minPos
0.328	logMelFreqBand-sma[6]-quartile2	0.1059	logMelFreqBand-sma[0]-minPos
0.3267	mfcc-sma[0]-iqr1-3	0.1058	LSPFreq-sma[4]-maxPos
0.3266	pcm-loudness-sma-amean	0.1057	LSPFreq-sma[3]-linregc2
0.3264	jitterLocal-sma-linregerrQ	0.1057	mfcc-sma-de[7]-maxPos
0.3262	mfcc-sma[9]-quartile1	0.1057	mfcc-sma[1]-minPos
0.3259	LSPFreq-sma[0]-iqr1-2	0.1055	LSPFreq-sma[1]-maxPos
0.3257	jitterLocal-sma-stddev	0.1054	logMelFreqBand-sma[7]-amean
0.3254	voicingFinalUnclipped-sma-de-iqr1-3	0.1053	voicingFinalUnclipped-sma-de-stddev
0.3248	voicingFinalUnclipped-sma-de-quartile1	0.1053	voicingFinalUnclipped-sma-de-linregerrQ
0.3246	mfcc-sma[4]-stddev	0.1052	LSPFreq-sma-de[4]-linregerrA
0.324	logMelFreqBand-sma[5]-kurtosis	0.1052	logMelFreqBand-sma-de[2]-upleveltime90
0.3238	LSPFreq-sma-de[0]-kurtosis	0.1051	mfcc-sma-de[5]-quartile2
0.3234	mfcc-sma[12]-stddev	0.105	LSPFreq-sma[5]-quartile3
0.3233	mfcc-sma-de[13]-Percentile1.0	0.1046	LSPFreq-sma[7]-maxPos
0.3222	mfcc-sma[5]-iqr2-3	0.1044	LSPFreq-sma-de[7]-quartile1
0.3202	mfcc-sma[5]-pctlrange0-1	0.104	mfcc-sma-de[4]-kurtosis
0.3199	LSPFreq-sma[2]-quartile3	0.1037	pcm-loudness-sma-de-iqr2-3
0.3198	mfcc-sma[14]-linregerrA	0.1035	pcm-loudness-sma-de-quartile3
0.3193	F0final-sma-maxPos	0.1033	mfcc-sma[2]-maxPos
0.3174	logMelFreqBand-sma-de[4]-upleveltime75	0.103	logMelFreqBand-sma[6]-linregc2
0.3169	jitterLocal-sma-kurtosis	0.103	mfcc-sma-de[4]-upleveltime75
0.3168	mfcc-sma[14]-stddev	0.1027	LSPFreq-sma-de[1]-skewness
0.3165	mfcc-sma[5]-linregc2	0.1025	logMelFreqBand-sma-de[7]-amean
0.3165	logMelFreqBand-sma-de[6]-Percentile99.0	0.1021	mfcc-sma-de[3]-pctlrange0-1
0.3164	logMelFreqBand-sma[3]-linregc1	0.1021	jitterLocal-sma-minPos
0.3152	logMelFreqBand-sma[1]-linregerrQ	0.1016	mfcc-sma[13]-minPos
0.3148	logMelFreqBand-sma[1]-maxPos	0.1015	logMelFreqBand-sma-de[5]-minPos
0.3144	logMelFreqBand-sma[4]-iqr2-3	0.1012	mfcc-sma[8]-maxPos
0.3144	shimmerLocal-sma-de-upleveltime90	0.101	mfcc-sma[10]-pctlrange0-1
0.3138	logMelFreqBand-sma[6]-linregerrQ	0.1008	mfcc-sma[10]-upleveltime75
0.3131	logMelFreqBand-sma-de[1]-kurtosis	0.1006	LSPFreq-sma-de[7]-iqr1-2
0.3124	mfcc-sma-de[1]-quartile3	0.0998	logMelFreqBand-sma-de[5]-maxPos
0.3121	mfcc-sma[4]-linregerrQ	0.0996	LSPFreq-sma-de[6]-kurtosis
0.3113	mfcc-sma[14]-linregerrQ	0.0993	pcm-loudness-sma-upleveltime90
0.3106	mfcc-sma-de[13]-iqr1-2	0.0992	mfcc-sma-de[5]-upleveltime75
0.3103	mfcc-sma[13]-iqr1-3	0.0991	LSPFreq-sma[0]-minPos
0.3098	LSPFreq-sma-de[2]-linregerrQ	0.0988	pcm-loudness-sma-de-kurtosis
0.3098	LSPFreq-sma-de[2]-stddev	0.0986	logMelFreqBand-sma[3]-quartile2
0.3084	mfcc-sma-de[4]-Percentile1.0	0.098	mfcc-sma[11]-quartile2
0.3082	mfcc-sma[9]-stddev	0.098	logMelFreqBand-sma[1]-minPos
0.3081	logMelFreqBand-sma-de[0]-quartile1	0.0978	LSPFreq-sma-de[7]-quartile3
0.3081	F0final-sma-de-maxPos	0.0974	LSPFreq-sma-de[3]-linregerrQ
0.3078	logMelFreqBand-sma[1]-quartile3	0.0974	LSPFreq-sma-de[3]-stddev
0.3077	F0finEnv-sma-skewness	0.0973	mfcc-sma[2]-upleveltime75
0.307	mfcc-sma-de[1]-linregc1	0.0971	mfcc-sma-de[1]-upleveltime90

Ranking	Feature	Ranking	Feature
0.3068	logMelFreqBand-sma[5]-Percentile99.0	0.0965	LSPFreq-sma[4]-upleveltime75
0.3068	mfcc-sma[6]-amean	0.0962	LSPFreq-sma-de[2]-upleveltime75
0.306	jitterDDP-sma-de-iqr1-3	0.096	mfcc-sma[9]-kurtosis
0.3049	logMelFreqBand-sma[6]-quartile1	0.0959	mfcc-sma-de[7]-skewness
0.3044	mfcc-sma[0]-amean	0.0959	logMelFreqBand-sma-de[0]-upleveltime90
0.3044	mfcc-sma-de[10]-Percentile99.0	0.0954	jitterDDP-sma-minPos
0.3038	mfcc-sma-de[6]-quartile1	0.0949	LSPFreq-sma-de[0]-linregc2
0.3038	logMelFreqBand-sma[0]-Percentile99.0	0.0946	LSPFreq-sma-de[3]-minPos
0.3034	mfcc-sma[11]-iqr1-3	0.0945	mfcc-sma-de[8]-minPos
0.3033	LSPFreq-sma[2]-iqr1-3	0.0945	LSPFreq-sma-de[4]-maxPos
0.3033	mfcc-sma[10]-stddev	0.0944	mfcc-sma[12]-skewness
0.3032	logMelFreqBand-sma[5]-upleveltime90	0.0944	mfcc-sma-de[12]-minPos
0.3024	logMelFreqBand-sma[5]-iqr1-3	0.0944	LSPFreq-sma[3]-stddev
0.3021	shimmerLocal-sma-linregc2	0.0943	LSPFreq-sma-de[7]-iqr2-3
0.3013	mfcc-sma[1]-upleveltime90	0.0942	mfcc-sma[11]-Percentile99.0
0.3013	LSPFreq-sma[0]-kurtosis	0.0941	pcm-loudness-sma-kurtosis
0.3001	jitterDDP-sma-de-iqr2-3	0.094	mfcc-sma[11]-minPos
0.2996	mfcc-sma[4]-iqr2-3	0.0939	logMelFreqBand-sma-de[7]-maxPos
0.2996	mfcc-sma[9]-quartile3	0.0937	LSPFreq-sma[5]-upleveltime75
0.2994	mfcc-sma[1]-iqr2-3	0.0932	mfcc-sma[5]-upleveltime90
0.2988	LSPFreq-sma-de[1]-linregerrA	0.093	LSPFreq-sma-de[4]-upleveltime90
0.298	mfcc-sma[2]-pctlrange0-1	0.0927	shimmerLocal-sma-minPos
0.2976	mfcc-sma[12]-Percentile99.0	0.0922	LSPFreq-sma[4]-quartile2
0.2973	mfcc-sma[8]-Percentile1.0	0.0916	mfcc-sma[6]-upleveltime75
0.297	LSPFreq-sma-de[2]-iqr1-2	0.0913	mfcc-sma-de[2]-upleveltime90
0.2959	logMelFreqBand-sma[0]-maxPos	0.0912	LSPFreq-sma[0]-upleveltime75
0.2958	LSPFreq-sma-de[6]-iqr1-3	0.0909	LSPFreq-sma[6]-iqr2-3
0.2956	mfcc-sma-de[7]-iqr2-3	0.0908	LSPFreq-sma[7]-skewness
0.2954	mfcc-sma[7]-quartile1	0.0906	LSPFreq-sma[4]-iqr2-3
0.2945	logMelFreqBand-sma[5]-quartile3	0.0905	LSPFreq-sma-de[2]-pctlrange0-1
0.2945	mfcc-sma[6]-Percentile1.0	0.0904	jitterLocal-sma-de-skewness
0.2942	logMelFreqBand-sma[1]-amean	0.0903	voicingFinalUnclipped-sma-de-Percentile99.0
0.2942	mfcc-sma-de[7]-linregc2	0.0903	mfcc-sma[8]-minPos
0.2937	LSPFreq-sma[1]-quartile3	0.0902	LSPFreq-sma-de[3]-pctlrange0-1
0.2922	logMelFreqBand-sma-de[3]-Percentile1.0	0.0889	logMelFreqBand-sma-de[6]-maxPos
0.292	LSPFreq-sma[1]-quartile2	0.0889	logMelFreqBand-sma-de[1]-upleveltime75
0.292	F0final-sma-upleveltime90	0.0887	mfcc-sma-de[3]-amean
0.2916	mfcc-sma[11]-linregc1	0.0886	LSPFreq-sma-de[7]-linregc1
0.2914	LSPFreq-sma-de[3]-linregerrA	0.0883	mfcc-sma[14]-amean
0.2914	mfcc-sma[13]-quartile1	0.0882	mfcc-sma[9]-iqr1-2
0.2914	mfcc-sma-de[9]-Percentile1.0	0.0881	LSPFreq-sma[4]-amean
0.2898	voicingFinalUnclipped-sma-iqr1-2	0.0879	pcm-loudness-sma-linregerrA
0.2895	logMelFreqBand-sma[1]-linregerrA	0.0879	logMelFreqBand-sma-de[1]-minPos
0.2894	voicingFinalUnclipped-sma-upleveltime90	0.0876	mfcc-sma[10]-upleveltime90
0.2892	mfcc-sma[12]-quartile2	0.0873	logMelFreqBand-sma-de[3]-upleveltime90
0.2891	pcm-loudness-sma-de-linregc2	0.0871	logMelFreqBand-sma[0]-iqr1-2
0.2884	mfcc-sma[0]-iqr2-3	0.0869	mfcc-sma[7]-linregc2
0.2883	logMelFreqBand-sma-de[6]-pctlrange0-1	0.0864	mfcc-sma-de[7]-upleveltime90
0.2882	mfcc-sma-de[11]-Percentile1.0	0.0862	mfcc-sma-de[11]-minPos
0.288	jitterLocal-sma-upleveltime75	0.0854	LSPFreq-sma-de[5]-stddev
0.2879	LSPFreq-sma-de[1]-quartile3	0.0854	LSPFreq-sma-de[5]-linregerrQ
0.2875	logMelFreqBand-sma-de[6]-Percentile1.0	0.0853	mfcc-sma[11]-linregc2
0.2866	LSPFreq-sma-de[3]-linregc1	0.0853	LSPFreq-sma-de[3]-Percentile99.0
0.2865	logMelFreqBand-sma[7]-linregc1	0.0852	LSPFreq-sma-de[5]-skewness
0.2862	F0finEnv-sma-maxPos	0.0852	logMelFreqBand-sma[7]-linregc2
0.286	logMelFreqBand-sma[5]-amean	0.085	LSPFreq-sma-de[1]-upleveltime75
0.286	F0final-sma-de-linregc2	0.0849	logMelFreqBand-sma-de[6]-upleveltime90
0.2857	LSPFreq-sma[1]-Percentile1.0	0.0846	mfcc-sma-de[9]-amean
0.2853	jitterLocal-sma-iqr2-3	0.0842	LSPFreq-sma[6]-linregc1
0.2853	mfcc-sma-de[1]-kurtosis	0.0839	mfcc-sma-de[3]-Percentile1.0
0.2851	logMelFreqBand-sma[6]-stddev	0.0836	mfcc-sma[3]-iqr2-3
0.285	mfcc-sma[10]-linregc2	0.0834	LSPFreq-sma-de[7]-upleveltime90
0.2842	jitterDDP-sma-de-upleveltime90	0.0833	LSPFreq-sma[0]-Percentile99.0

Ranking	Feature	Ranking	Feature
0.2837	logMelFreqBand-sma[2]-maxPos	0.0829	mfcc-sma-de[7]-amean
0.2827	mfcc-sma[13]-Percentile1.0	0.0825	mfcc-sma[5]-linregc1
0.2827	logMelFreqBand-sma[2]-quartile1	0.0823	LSPFreq-sma-de[0]-amean
0.2826	logMelFreqBand-sma-de[7]-pctlrange0-1	0.0821	mfcc-sma[9]-upleveltime90
0.2822	mfcc-sma[11]-pctlrange0-1	0.0818	LSPFreq-sma-de[1]-amean
0.2822	jitterDDP-sma-quartile2	0.0816	LSPFreq-sma[4]-iqr1-3
0.282	LSPFreq-sma-de[5]-iqr1-2	0.0816	mfcc-sma[6]-kurtosis
0.2818	LSPFreq-sma[1]-amean	0.0815	logMelFreqBand-sma-de[0]-minPos
0.2814	mfcc-sma[12]-amean	0.0815	mfcc-sma-de[14]-upleveltime90
0.2811	logMelFreqBand-sma[7]-upleveltime90	0.0813	voicingFinalUnclipped-sma-de-maxPos
0.2806	logMelFreqBand-sma[5]-quartile1	0.0812	logMelFreqBand-sma-de[0]-quartile2
0.2805	mfcc-sma-de[3]-iqr2-3	0.0808	mfcc-sma-de[3]-upleveltime90
0.2805	logMelFreqBand-sma[5]-quartile2	0.0807	LSPFreq-sma[3]-linregerrA
0.2803	mfcc-sma[0]-skewness	0.0805	LSPFreq-sma[5]-linregc2
0.2803	voicingFinalUnclipped-sma-maxPos	0.0803	logMelFreqBand-sma-de[2]-minPos
0.2797	logMelFreqBand-sma-de[0]-iqr1-2	0.08	mfcc-sma-de[11]-upleveltime90
0.2796	mfcc-sma-de[3]-linregerrQ	0.0799	mfcc-sma-de[1]-upleveltime75
0.2796	LSPFreq-sma[0]-iqr1-3	0.0798	mfcc-sma[11]-skewness
0.2795	mfcc-sma-de[3]-stddev	0.0797	LSPFreq-sma[4]-Percentile1.0
0.2787	logMelFreqBand-sma[1]-stddev	0.0797	LSPFreq-sma-de[7]-upleveltime75
0.2786	mfcc-sma-de[6]-linregc1	0.0796	LSPFreq-sma-de[0]-upleveltime75
0.2785	mfcc-sma[9]-iqr2-3	0.0793	LSPFreq-sma-de[4]-linregc1
0.2782	logMelFreqBand-sma[2]-iqr2-3	0.0792	mfcc-sma-de[6]-upleveltime75
0.2779	F0final-sma-de-iqr2-3	0.0789	mfcc-sma[9]-upleveltime75
0.2771	LSPFreq-sma-de[3]-maxPos	0.0789	mfcc-sma[1]-kurtosis
0.2763	mfcc-sma-de[7]-quartile1	0.0782	mfcc-sma-de[12]-quartile2
0.2761	mfcc-sma-de[5]-linregc2	0.078	LSPFreq-sma[6]-iqr1-2
0.2754	mfcc-sma[3]-quartile3	0.0779	logMelFreqBand-sma-de[1]-skewness
0.2754	mfcc-sma[10]-linregerrQ	0.0779	logMelFreqBand-sma[6]-Percentile99.0
0.2742	LSPFreq-sma-de[3]-linregc2	0.0778	LSPFreq-sma[7]-quartile3
0.2742	mfcc-sma-de[4]-pctlrange0-1	0.0777	logMelFreqBand-sma-de[1]-quartile2
0.274	mfcc-sma[8]-linregerrQ	0.0777	LSPFreq-sma[2]-upleveltime90
0.2734	F0finEnv-sma-de-minPos	0.0775	voicingFinalUnclipped-sma-de-amean
0.2733	mfcc-sma[8]-stddev	0.0775	LSPFreq-sma[7]-iqr2-3
0.2733	logMelFreqBand-sma[5]-iqr2-3	0.0774	H(max)
0.2732	logMelFreqBand-sma-de[5]-pctlrange0-1	0.0774	logMelFreqBand-sma[6]-minPos
0.2728	jitterDDP-sma-upleveltime75	0.0774	LSPFreq-sma[4]-quartile1
0.2728	mfcc-sma-de[6]-linregc2	0.0771	logMelFreqBand-sma[1]-iqr1-2
0.2727	logMelFreqBand-sma[1]-quartile1	0.0769	LSPFreq-sma-de[7]-linregc2
0.2727	LSPFreq-sma[1]-quartile1	0.0767	mfcc-sma[14]-upleveltime75
0.2722	mfcc-sma[4]-linregerrA	0.0765	LSPFreq-sma-de[4]-stddev
0.272	mfcc-sma[0]-iqr1-2	0.0764	LSPFreq-sma-de[4]-linregerrQ
0.2719	mfcc-sma-de[5]-linregc1	0.0763	logMelFreqBand-sma-de[3]-minPos
0.2718	logMelFreqBand-sma[0]-iqr2-3	0.0762	H(min)
0.2716	logMelFreqBand-sma[1]-kurtosis	0.076	LSPFreq-sma[3]-iqr1-2
0.2711	mfcc-sma[6]-maxPos	0.076	pcm-loudness-sma-Percentile99.0
0.2708	mfcc-sma[9]-Percentile1.0	0.0759	mfcc-sma-de[11]-kurtosis
0.2707	mfcc-sma[0]-Percentile99.0	0.0759	mfcc-sma[3]-linregc1
0.2695	mfcc-sma[0]-upleveltime90	0.0759	mfcc-sma[8]-upleveltime90
0.2693	mfcc-sma[3]-linregc2	0.0753	LSPFreq-sma[5]-Percentile99.0
0.2692	mfcc-sma[8]-iqr2-3	0.075	LSPFreq-sma[6]-linregc2
0.2692	shimmerLocal-sma-upleveltime90	0.075	voicingFinalUnclipped-sma-de-upleveltime90
0.269	LSPFreq-sma-de[0]-iqr2-3	0.075	pcm-loudness-sma-de-stddev
0.269	HNR(mean)	0.075	pcm-loudness-sma-de-linregerrQ
0.2681	logMelFreqBand-sma[3]-maxPos	0.0747	LSPFreq-sma[6]-quartile2
0.2681	mfcc-sma[9]-iqr1-3	0.0747	mfcc-sma-de[0]-amean
0.2678	mfcc-sma[5]-quartile3	0.0746	mfcc-sma[14]-quartile3
0.2669	voicingFinalUnclipped-sma-de-iqr2-3	0.0745	mfcc-sma[7]-iqr2-3
0.2667	LSPFreq-sma-de[6]-quartile3	0.0738	LSPFreq-sma[4]-upleveltime90
0.2665	F0final–Turn-numOnsets	0.0736	LSPFreq-sma-de[4]-pctlrange0-1
0.2662	logMelFreqBand-sma[0]-iqr1-3	0.0735	mfcc-sma-de[5]-skewness
0.2655	jitterDDP-sma-de-quartile1	0.0735	LSPFreq-sma-de[5]-Percentile99.0
0.2642	LSPFreq-sma-de[2]-linregc1	0.0733	LSPFreq-sma[0]-stddev

Ranking	Feature	Ranking	Feature
0.264	LSPFreq-sma[0]-skewness	0.0732	mfcc-sma[8]-skewness
0.2639	logMelFreqBand-sma-de[7]-kurtosis	0.073	LSPFreq-sma[7]-linregc2
0.2627	mfcc-sma-de[0]-Percentile1.0	0.073	pcm-loudness-sma-de-quartile2
0.2625	mfcc-sma-de[12]-linregc2	0.0728	mfcc-sma-de[3]-Percentile99.0
0.2621	logMelFreqBand-sma[4]-kurtosis	0.0726	LSPFreq-sma[2]-pctlrange0-1
0.262	pcm-loudness-sma-quartile1	0.0724	LSPFreq-sma-de[3]-upleveltime90
0.2616	jitterLocal-sma-skewness	0.0723	mfcc-sma[14]-skewness
0.2616	LSPFreq-sma-de[3]-quartile3	0.0721	LSPFreq-sma[1]-upleveltime75
0.2606	shimmerLocal-sma-kurtosis	0.0717	mfcc-sma[2]-linregc1
0.2596	mfcc-sma[1]-skewness	0.0709	mfcc-sma-de[9]-upleveltime75
0.2595	mfcc-sma[10]-iqr2-3	0.0707	LSPFreq-sma[0]-pctlrange0-1
0.2595	Period(number)	0.0702	LSPFreq-sma[6]-amean
0.2585	mfcc-sma[11]-stddev	0.0702	LSPFreq-sma[3]-Percentile99.0
0.2584	mfcc-sma[10]-iqr1-3	0.07	LSPFreq-sma[1]-Percentile99.0
0.2584	mfcc-sma[0]-linregc1	0.07	mfcc-sma-de[8]-upleveltime90
0.2584	LSPFreq-sma-de[2]-minPos	0.0689	mfcc-sma[10]-kurtosis
0.2579	mfcc-sma[4]-linregc2	0.0687	mfcc-sma-de[4]-quartile2
0.2574	mfcc-sma-de[5]-Percentile99.0	0.0684	LSPFreq-sma[7]-quartile2
0.257	mfcc-sma[7]-quartile3	0.0683	LSPFreq-sma[5]-linregerrQ
0.2565	logMelFreqBand-sma[7]-kurtosis	0.0681	pcm-loudness-sma-stddev
0.2564	voicingFinalUnclipped-sma-de-quartile3	0.0678	LSPFreq-sma-de[5]-upleveltime90
0.2558	logMelFreqBand-sma[7]-quartile2	0.0674	LSPFreq-sma-de[0]-upleveltime90
0.2556	mfcc-sma[10]-linregc1	0.0671	logMelFreqBand-sma[7]-minPos
0.2549	NHR(mean)	0.067	LSPFreq-sma[5]-pctlrange0-1
0.2544	jitterLocal-sma-de-kurtosis	0.0657	mfcc-sma-de[1]-skewness
0.2538	F0final-sma-de-minPos	0.0643	mfcc-sma-de[14]-linregc2
0.2536	logMelFreqBand-sma[1]-quartile2	0.0634	mfcc-sma[3]-upleveltime75
0.2531	pcm-loudness-sma-quartile2	0.0626	LSPFreq-sma[5]-linregerrA
0.2529	logMelFreqBand-sma-de[5]-upleveltime75	0.0603	LSPFreq-sma[5]-iqr1-3
0.2527	mfcc-sma[11]-linregerrQ	0.0597	logMelFreqBand-sma-de[2]-amean
0.2519	shimmer(ddp)	0.0589	mfcc-sma-de[4]-skewness
0.2519	shimmer(apq3)	0.0573	LSPFreq-sma[6]-quartile3
0.2512	logMelFreqBand-sma[6]-amean	0.0568	logMelFreqBand-sma-de[1]-amean
0.251	LSPFreq-sma-de[7]-pctlrange0-1	0.0534	LSPFreq-sma[5]-maxPos
0.2509	LSPFreq-sma-de[2]-linregc2		

APPENDIX B

The Selected Features using CFS-PSO

Feature	Feature	Feature
F0finEnv-sma-de-skewness	mfcc-sma[0]-quartile1	mfcc-sma[12]-iqr1-3
logMelFreqBand-sma-de[2]-iqr1-3	shimmerLocal-sma-quartile1	logMelFreqBand-sma[7]-linregerrQ
logMelFreqBand-sma-de[4]-quartile1	lspFreq-sma-de[2]-linregerrA	mfcc-sma-de[13]-minPos
logMelFreqBand-sma-de[4]-linregerrA	mfcc-sma-de[8]-linregc1	mfcc-sma-de[1]-pctlrange0-1
mfcc-sma-de[2]-linregerrA	lspFreq-sma[1]-iqr2-3	lspFreq-sma-de[6]-amean
logMelFreqBand-sma-de[5]-quartile1	logMelFreqBand-sma[0]-pctlrange0-1	lspFreq-sma-de[5]-iqr2-3
F0final-sma-percentile99.0	mfcc-sma[10]-quartile2	mfcc-sma[7]-iqr1-3
logMelFreqBand-sma-de[5]-iqr1-3	shimmerLocal-sma-upleveltime75	mfcc-sma[10]-percentile99.0
mfcc-sma-de[0]-iqr1-3	mfcc-sma-de[7]-iqr1-3	mfcc-sma-de[1]-maxPos
logMelFreqBand-sma-de[4]-iqr1-3	jitter(ppq5)	logMelFreqBand-sma-de[3]-maxPos
logMelFreqBand-sma-de[2]-linregerrA	Period(std)	mfcc-sma-de[2]-amean
logMelFreqBand-sma[2]-pctlrange0-1	mfcc-sma-de[4]-iqr1-2	mfcc-sma-de[1]-linregerrQ
mfcc-sma-de[12]-linregerrQ	logMelFreqBand-sma[2]-quartile3	mfcc-sma-de[1]-stddev
logMelFreqBand-sma-de[4]-stddev	jitterLocal-sma-de-stddev	logMelFreqBand-sma[6]-quartile3
logMelFreqBand-sma-de[3]-quartile1	mfcc-sma[1]-iqr1-2	mfcc-sma[5]-percentile99.0
logMelFreqBand-sma-de[5]-quartile3	logMelFreqBand-sma[5]-stddev	mfcc-sma[10]-linregerrA
F0final-sma-linregerrQ	logMelFreqBand-sma[1]-linregc1	logMelFreqBand-sma-de[0]-upleveltime75
mfcc-sma-de[12]-stddev	shimmerLocal-sma-de-iqr1-2	mfcc-sma-de[8]-maxPos
F0final-sma-linregerrA	logMelFreqBand-sma[1]-linregc2	mfcc-sma[13]-linregc1
logMelFreqBand-sma-de[3]-stddev	mfcc-sma-de[3]-quartile3	lspFreq-sma[2]-quartile2
logMelFreqBand-sma-de[2]-iqr1-2	lspFreq-sma-de[2]-quartile1	shimmerLocal-sma-de-kurtosis
mfcc-sma-de[0]-quartile3	jitterDDP-sma-quartile3	logMelFreqBand-sma-de[7]-upleveltime90
logMelFreqBand-sma-de[4]-quartile3	mfcc-sma-de[7]-linregc1	logMelFreqBand-sma-de[2]-quartile2
logMelFreqBand-sma-de[2]-quartile1	lspFreq-sma-de[0]-quartile3	jitterDDP-sma-de-linregc1
logMelFreqBand-sma-de[2]-linregerrQ	mfcc-sma-de[6]-iqr1-2	mfcc-sma-de[1]-percentile99.0
voicingFinalUnclipped-sma-percentile99.0	mfcc-sma[9]-quartile2	logMelFreqBand-sma[0]-stddev
F0final-sma-stddev	mfcc-sma-de[1]-iqr1-2	mfcc-sma-de[0]-maxPos
logMelFreqBand-sma[4]-pctlrange0-1	mfcc-sma[9]-amean	mfcc-sma-de[13]-maxPos
F0finEnv-sma-quartile3	lspFreq-sma[1]-skewness	mfcc-sma-de[7]-percentile1.0
logMelFreqBand-sma-de[3]-linregerrA	mfcc-sma-de[8]-linregc2	lspFreq-sma[2]-minPos
Period(mean)	mfcc-sma-de[3]-iqr1-3	logMelFreqBand-sma-de[4]-minPos
F0finEnv-sma-de-linregerrA	mfcc-sma[0]-maxPos	lspFreq-sma-de[6]-iqr1-2
mfcc-sma-de[14]-linregerrQ	F0final-sma-de-upleveltime90	jitterDDP-sma-de-minPos
F0final-sma-de-stddev	logMelFreqBand-sma[4]-linregc1	mfcc-sma-de[1]-linregerrA
mfcc-sma-de[0]-iqr1-2	mfcc-sma[9]-linregerrA	voicingFinalUnclipped-sma-de-kurtosis
F0finEnv-sma-quartile2	mfcc-sma-de[6]-quartile3	lspFreq-sma[6]-quartile1
logMelFreqBand-sma-de[5]-iqr2-3	mfcc-sma-de[1]-iqr1-3	mfcc-sma-de[9]-maxPos
lspFreq-sma[0]-quartile1	F0final-sma-kurtosis	jitterDDP-sma-de-quartile2
mfcc-sma[0]-pctlrange0-1	mfcc-sma-de[9]-iqr1-2	logMelFreqBand-sma-de[2]-upleveltime75
lspFreq-sma[0]-quartile3	mfcc-sma-de[10]-quartile3	mfcc-sma[1]-stddev

Feature	Feature	Feature
mfcc-sma-de[5]-iqr1-2	mfcc-sma[12]-linregerrQ	lspFreq-sma-de[7]-maxPos
logMelFreqBand-sma-de[2]-quartile3	pcm-loudness-sma-linregc1	lspFreq-sma-de[5]-linregc2
mfcc-sma-de[0]-quartile1	logMelFreqBand-sma-de[5]-percentile99.0	logMelFreqBand-sma-de[4]-skewness
mfcc-sma[2]-quartile1	mfcc-sma-de[10]-pctlrange0-1	lspFreq-sma-de[4]-linregc2
logMelFreqBand-sma-de[3]-quartile3	mfcc-sma[1]-percentile99.0	lspFreq-sma-de[6]-stddev
logMelFreqBand-sma[2]-percentile1.0	mfcc-sma-de[1]-linregc2	shimmerLocal-sma-de-linregc2
logMelFreqBand-sma-de[3]-iqr1-2	lspFreq-sma[0]-iqr2-3	Do1000(slope)
logMelFreqBand-sma-de[2]-iqr2-3	mfcc-sma[8]-quartile1	lspFreq-sma-de[6]-linregerrQ
logMelFreqBand-sma[7]-percentile1.0	mfcc-sma-de[3]-linregc2	mfcc-sma[0]-minPos
logMelFreqBand-sma[6]-percentile1.0	mfcc-sma[9]-linregerrQ	lspFreq-sma-de[7]-amean
F0finEnv-sma-de-percentile1.0	pcm-loudness-sma-linregc2	logMelFreqBand-sma-de[5]-quartile2
lspFreq-sma[0]-quartile2	logMelFreqBand-sma[6]-quartile2	mfcc-sma[13]-quartile3
lspFreq-sma[0]-amean	mfcc-sma[0]-iqr1-3	lspFreq-sma[3]-percentile1.0
mfcc-sma-de[8]-linregerrA	pcm-loudness-sma-amean	logMelFreqBand-sma-de[5]-amean
F0final-sma-quartile2	lspFreq-sma[0]-iqr1-2	lspFreq-sma-de[5]-minPos
mfcc-sma-de[5]-stddev	voicingFinalUnclipped-sma-de-iqr1-3	mfcc-sma-de[12]-maxPos
mfcc-sma-de[5]-linregerrQ	voicingFinalUnclipped-sma-de-quartile1	logMelFreqBand-sma[2]-linregc2
logMelFreqBand-sma-de[5]-stddev	mfcc-sma[12]-stddev	lspFreq-sma-de[7]-percentile1.0
jitterLocal-sma-de-quartile1	mfcc-sma-de[13]-percentile1.0	mfcc-sma-de[6]-upleveltime90
F0final-sma-amean	mfcc-sma[14]-linregerrA	jitterLocal-sma-maxPos
mfcc-sma-de[5]-linregerrA	F0final-sma-maxPos	mfcc-sma[10]-minPos
logMelFreqBand-sma[5]-pctlrange0-1	logMelFreqBand-sma-de[4]-upleveltime75	mfcc-sma-de[5]-amean
mfcc-sma-de[2]-iqr1-3	jitterLocal-sma-kurtosis	mfcc-sma-de[1]-quartile1
shimmerLocal-sma-de-linregerrA	mfcc-sma[14]-stddev	jitterDDP-sma-maxPos
mfcc-sma-de[12]-iqr1-2	logMelFreqBand-sma[1]-linregerrQ	logMelFreqBand-sma[0]-linregerrQ
logMelFreqBand-sma-de[6]-iqr1-3	logMelFreqBand-sma[1]-maxPos	mfcc-sma[7]-amean
mfcc-sma-de[14]-stddev	logMelFreqBand-sma[4]-iqr2-3	lspFreq-sma[5]-kurtosis
mfcc-sma-de[12]-iqr1-3	shimmerLocal-sma-de-upleveltime90	logMelFreqBand-sma-de[6]-amean
logMelFreqBand-sma-de[5]-iqr1-2	logMelFreqBand-sma-de[1]-kurtosis	mfcc-sma-de[10]-maxPos
shimmerLocal-sma-de-quartile1	mfcc-sma-de[1]-quartile3	mfcc-sma[13]-percentile99.0
voicingFinalUnclipped-sma-skewness	mfcc-sma-de[13]-iqr1-2	mfcc-sma[7]-percentile1.0
mfcc-sma-de[11]-iqr1-3	mfcc-sma[13]-iqr1-3	lspFreq-sma[6]-skewness
mfcc-sma-de[12]-pctlrange0-1	lspFreq-sma-de[2]-linregerrQ	pcm-loudness-sma-de-minPos
voicingFinalUnclipped-sma-iqr2-3	lspFreq-sma-de[2]-stddev	mfcc-sma[5]-minPos
jitterLocal-sma-amean	mfcc-sma-de[4]-percentile1.0	mfcc-sma-de[5]-upleveltime90
F0final-sma-iqr1-3	mfcc-sma[9]-stddev	mfcc-sma[9]-minPos
F0finEnv-sma-de-percentile99.0	logMelFreqBand-sma-de[0]-quartile1	mfcc-sma-de[12]-amean
mfcc-sma[1]-quartile2	logMelFreqBand-sma[5]-percentile99.0	lspFreq-sma[5]-quartile2
mfcc-sma-de[12]-percentile99.0	mfcc-sma[6]-amean	mfcc-sma[9]-linregc1
mfcc-sma-de[0]-stddev	mfcc-sma[0]-amean	pcm-loudness-sma-skewness
voicingFinalUnclipped-sma-amean	mfcc-sma-de[10]-percentile99.0	logMelFreqBand-sma-de[3]-skewness
mfcc-sma[1]-amean	mfcc-sma-de[6]-quartile1	logMelFreqBand-sma-de[6]-upleveltime75
logMelFreqBand-sma-de[7]-linregc1	logMelFreqBand-sma[0]-percentile99.0	voicingFinalUnclipped-sma-linregc1
mfcc-sma-de[8]-stddev	lspFreq-sma[2]-iqr1-3	jitterLocal-sma-linregc2
mfcc-sma-de[8]-linregerrQ	logMelFreqBand-sma[5]-upleveltime90	F0final-sma-linregc1
logMelFreqBand-sma-de[4]-pctlrange0-1	logMelFreqBand-sma[5]-iqr1-3	logMelFreqBand-sma-de[0]-kurtosis
logMelFreqBand-sma-de[6]-quartile3	shimmerLocal-sma-linregc2	logMelFreqBand-sma[6]-iqr1-3
F0finEnv-sma-iqr1-3	mfcc-sma[1]-upleveltime90	mfcc-sma[13]-amean
logMelFreqBand-sma[4]-percentile1.0	lspFreq-sma[0]-kurtosis	logMelFreqBand-sma[2]-iqr1-2
mfcc-sma-de[5]-quartile3	jitterDDP-sma-de-iqr2-3	lspFreq-sma-de[6]-percentile1.0
mfcc-sma-de[2]-pctlrange0-1	mfcc-sma[4]-iqr2-3	mfcc-sma[6]-iqr2-3
mfcc-sma-de[5]-iqr1-3	mfcc-sma[9]-quartile3	mfcc-sma[1]-linregerrQ
mfcc-sma[2]-linregerrQ	mfcc-sma[1]-iqr2-3	mfcc-sma[4]-minPos
F0finEnv-sma-de-amean	lspFreq-sma-de[1]-linregerrA	lspFreq-sma-de[4]-percentile1.0
voicingFinalUnclipped-sma-quartile1	mfcc-sma[2]-pctlrange0-1	lspFreq-sma[3]-iqr1-3
F0finEnv-sma-iqr2-3	mfcc-sma[8]-percentile1.0	mfcc-sma-de[4]-minPos
logMelFreqBand-sma[3]-stddev	lspFreq-sma-de[2]-iqr1-2	mfcc-sma-de[8]-quartile2
mfcc-sma[2]-percentile1.0	logMelFreqBand-sma[5]-quartile3	logMelFreqBand-sma[5]-maxPos
shimmerLocal-sma-amean	mfcc-sma-de[7]-linregc2	mfcc-sma-de[11]-linregc1
jitterLocal-sma-quartile2	lspFreq-sma[1]-quartile3	lspFreq-sma[4]-linregerrA
mfcc-sma-de[8]-quartile1	lspFreq-sma[1]-quartile2	lspFreq-sma-de[4]-percentile99.0
logMelFreqBand-sma-de[5]-linregc2	mfcc-sma[11]-linregc1	lspFreq-sma[5]-iqr2-3
mfcc-sma[5]-quartile2	mfcc-sma-de[9]-percentile1.0	pcm-loudness-sma-de-iqr1-2

Feature	Feature	Feature
logMelFreqBand-sma-de[6]-quartile1	logMelFreqBand-sma[1]-linregerrA	voicingFinalUnclipped-sma-de-linregerrA
mfcc-sma-de[11]-iqr1-2	voicingFinalUnclipped-sma-upleveltime90	jitterDDP-sma-linregc2
logMelFreqBand-sma-de[4]-kurtosis	mfcc-sma[12]-quartile2	mfcc-sma[14]-quartile2
logMelFreqBand-sma[4]-percentile99.0	logMelFreqBand-sma-de[6]-pctlrange0-1	lspFreq-sma[6]-kurtosis
F0finEnv-sma-upleveltime75	mfcc-sma-de[11]-percentile1.0	logMelFreqBand-sma[4]-upleveltime75
logMelFreqBand-sma-de[6]-iqr2-3	lspFreq-sma-de[1]-quartile3	logMelFreqBand-sma-de[5]-upleveltime90
mfcc-sma-de[0]-linregc1	logMelFreqBand-sma-de[6]-percentile1.0	lspFreq-sma-de[5]-amean
logMelFreqBand-sma[3]-linregerrA	logMelFreqBand-sma[5]-amean	mfcc-sma[3]-skewness
logMelFreqBand-sma-de[2]-pctlrange0-1	lspFreq-sma[1]-percentile1.0	lspFreq-sma-de[5]-quartile3
mfcc-sma-de[14]-linregerrA	jitterLocal-sma-iqr2-3	lspFreq-sma-de[4]-iqr2-3
logMelFreqBand-sma-de[3]-kurtosis	mfcc-sma-de[1]-kurtosis	pcm-loudness-sma-de-iqr1-3
shimmerLocal-sma-de-iqr2-3	logMelFreqBand-sma[6]-stddev	lspFreq-sma-de[1]-minPos
mfcc-sma-de[2]-quartile1	mfcc-sma[10]-linregc2	logMelFreqBand-sma[3]-iqr1-2
mfcc-sma[5]-linregerrQ	jitterDDP-sma-de-upleveltime90	logMelFreqBand-sma[2]-quartile2
logMelFreqBand-sma-de[7]-iqr1-3	mfcc-sma[13]-percentile1.0	lspFreq-sma[7]-kurtosis
mfcc-sma-de[2]-iqr1-2	logMelFreqBand-sma[2]-quartile1	mfcc-sma[0]-quartile3
mfcc-sma-de[13]-linregerrA	jitterDDP-sma-quartile2	lspFreq-sma-de[1]-stddev
mfcc-sma-de[9]-linregerrQ	lspFreq-sma-de[5]-iqr1-2	mfcc-sma[3]-minPos
logMelFreqBand-sma-de[7]-linregc2	lspFreq-sma[1]-amean	lspFreq-sma-de[2]-percentile99.0
mfcc-sma-de[9]-linregc1	logMelFreqBand-sma[7]-upleveltime90	voicingFinalUnclipped-sma-stddev
mfcc-sma[2]-linregerrA	logMelFreqBand-sma[5]-quartile2	mfcc-sma[6]-pctlrange0-1
Do1000(offset)	mfcc-sma-de[3]-linregerrQ	voicingFinalUnclipped-sma-de-minPos
shimmerLocal-sma-de-iqr1-3	lspFreq-sma[0]-iqr1-3	lspFreq-sma-de[6]-percentile99.0
logMelFreqBand-sma-de[7]-iqr2-3	logMelFreqBand-sma[1]-stddev	logMelFreqBand-sma[4]-minPos
mfcc-sma[3]-quartile1	mfcc-sma[9]-iqr2-3	mfcc-sma[8]-iqr1-2
logMelFreqBand-sma[3]-linregerrQ	logMelFreqBand-sma[2]-iqr2-3	mfcc-sma[4]-percentile99.0
F0finEnv-sma-quartile1	F0final-sma-de-iqr2-3	lspFreq-sma[0]-linregerrA
mfcc-sma[5]-stddev	mfcc-sma-de[5]-linregc2	mfcc-sma-de[11]-linregc2
logMelFreqBand-sma-de[6]-linregerrA	mfcc-sma[3]-quartile3	logMelFreqBand-sma[3]-skewness
mfcc-sma-de[14]-quartile3	lspFreq-sma-de[3]-linregc2	logMelFreqBand-sma-de[4]-upleveltime90
mfcc-sma[5]-percentile1.0	F0finEnv-sma-de-minPos	logMelFreqBand-sma-de[7]-upleveltime75
mfcc-sma-de[9]-linregc2	logMelFreqBand-sma[5]-iqr2-3	lspFreq-sma[2]-linregc2
logMelFreqBand-sma-de[3]-linregc2	logMelFreqBand-sma-de[5]-pctlrange0-1	shimmerLocal-sma-maxPos
logMelFreqBand-sma-de[5]-kurtosis	mfcc-sma-de[6]-linregc2	lspFreq-sma-de[6]-linregc2
mfcc-sma-de[2]-linregc1	mfcc-sma[0]-iqr1-2	mfcc-sma-de[7]-minPos
logMelFreqBand-sma[3]-percentile1.0	mfcc-sma[9]-percentile1.0	voicingFinalUnclipped-sma-linregerrQ
logMelFreqBand-sma-de[6]-linregerrQ	mfcc-sma[3]-linregc2	mfcc-sma[11]-iqr1-2
mfcc-sma-de[0]-linregc2	shimmerLocal-sma-upleveltime90	logMelFreqBand-sma[2]-kurtosis
logMelFreqBand-sma-de[0]-linregc2	HNR(mean)	lspFreq-sma[2]-kurtosis
mfcc-sma-de[9]-quartile3	logMelFreqBand-sma[3]-maxPos	lspFreq-sma[3]-linregc2
logMelFreqBand-sma-de[6]-stddev	lspFreq-sma-de[2]-linregc1	mfcc-sma-de[7]-maxPos
logMelFreqBand-sma[4]-stddev	logMelFreqBand-sma-de[7]-kurtosis	logMelFreqBand-sma-de[2]-upleveltime90
mfcc-sma-de[8]-quartile3	mfcc-sma-de[0]-percentile1.0	lspFreq-sma[5]-quartile3
lspFreq-sma-de[2]-iqr2-3	logMelFreqBand-sma[4]-kurtosis	lspFreq-sma[7]-maxPos
mfcc-sma-de[6]-iqr2-3	pcm-loudness-sma-quartile1	pcm-loudness-sma-de-iqr2-3
mfcc-sma[13]-linregerrQ	lspFreq-sma-de[3]-quartile3	mfcc-sma-de[4]-upleveltime75
logMelFreqBand-sma-de[7]-quartile1	mfcc-sma[1]-skewness	lspFreq-sma-de[1]-skewness
mfcc-sma[1]-quartile1	mfcc-sma[10]-iqr2-3	mfcc-sma[13]-minPos
F0final-sma-iqr1-2	mfcc-sma[11]-stddev	logMelFreqBand-sma-de[5]-minPos
mfcc-sma-de[9]-quartile1	mfcc-sma[0]-linregc1	mfcc-sma[8]-maxPos
jitter(rap)	lspFreq-sma-de[2]-minPos	mfcc-sma[10]-pctlrange0-1
Hmm	mfcc-sma[4]-linregc2	mfcc-sma[10]-upleveltime75
mfcc-sma-de[5]-percentile1.0	voicingFinalUnclipped-sma-de-quartile3	lspFreq-sma-de[7]-iqr1-2
logMelFreqBand-sma[1]-percentile1.0	mfcc-sma[10]-linregc1	logMelFreqBand-sma-de[5]-maxPos
mfcc-sma-de[14]-pctlrange0-1	NHR(mean)	pcm-loudness-sma-upleveltime90
lspFreq-sma-de[2]-quartile3	jitterLocal-sma-de-kurtosis	mfcc-sma-de[5]-upleveltime75
F0final-sma-quartile1	F0final-sma-de-minPos	logMelFreqBand-sma[3]-quartile2
logMelFreqBand-sma-de[3]-pctlrange0-1	pcm-loudness-sma-quartile2	logMelFreqBand-sma[1]-minPos
jitterDDP-sma-stddev	logMelFreqBand-sma-de[5]-upleveltime75	mfcc-sma[2]-upleveltime75
logMelFreqBand-sma-de[6]-linregc1	shimmer(ddp)	mfcc-sma-de[1]-upleveltime90
mfcc-sma-de[13]-iqr1-3	lspFreq-sma-de[7]-pctlrange0-1	mfcc-sma[9]-kurtosis
lspFreq-sma-de[2]-iqr1-3	mfcc-sma-de[11]-percentile99.0	mfcc-sma-de[7]-skewness
logMelFreqBand-sma[5]-linregerrQ	lspFreq-sma-de[3]-iqr1-3	logMelFreqBand-sma-de[0]-upleveltime90

Feature	Feature	Feature
logMelFreqBand-sma[7]-upleveltime75	mfcc-sma[3]-iqr1-2	jitterDDP-sma-minPos
logMelFreqBand-sma-de[1]-iqr1-3	shimmer(localdB)	lspFreq-sma-de[3]-minPos
mfcc-sma-de[14]-iqr1-2	mfcc-sma-de[7]-quartile3	mfcc-sma-de[8]-minPos
mfcc-sma[4]-amean	mfcc-sma[8]-linregerrA	lspFreq-sma-de[4]-maxPos
F0final-sma-de-quartile1	logMelFreqBand-sma[0]-upleveltime75	mfcc-sma[12]-skewness
mfcc-sma-de[4]-quartile1	voicingFinalUnclipped-sma-iqr1-3	mfcc-sma-de[12]-minPos
logMelFreqBand-sma[0]-quartile3	mfcc-sma[8]-linregc2	pcm-loudness-sma-kurtosis
logMelFreqBand-sma-de[1]-quartile1	logMelFreqBand-sma[6]-linregerrA	mfcc-sma[5]-upleveltime90
jitterLocal-sma-quartile3	lspFreq-sma-de[5]-quartile1	lspFreq-sma-de[4]-upleveltime90
logMelFreqBand-sma-de[4]-percentile1.0	logMelFreqBand-sma[2]-linregc1	shimmerLocal-sma-minPos
mfcc-sma-de[5]-iqr2-3	logMelFreqBand-sma[0]-amean	mfcc-sma[6]-upleveltime75
F0final-sma-linregc2	F0final-sma-de-skewness	mfcc-sma-de[2]-upleveltime90
logMelFreqBand-sma-de[1]-linregc1	lspFreq-sma[1]-linregc2	lspFreq-sma[0]-upleveltime75
mfcc-sma-de[10]-linregerrA	logMelFreqBand-sma[2]-amean	lspFreq-sma[6]-iqr2-3
logMelFreqBand-sma[0]-linregc2	mfcc-sma[4]-iqr1-3	lspFreq-sma[7]-skewness
logMelFreqBand-sma[1]-pctlrange0-1	logMelFreqBand-sma[7]-quartile1	lspFreq-sma-de[2]-pctlrange0-1
mfcc-sma-de[14]-quartile1	lspFreq-sma[7]-minPos	mfcc-sma[8]-minPos
logMelFreqBand-sma-de[2]-percentile99.0	logMelFreqBand-sma[5]-linregc1	mfcc-sma-de[3]-amean
mfcc-sma-de[0]-kurtosis	lspFreq-sma[2]-stddev	mfcc-sma[14]-amean
mfcc-sma-de[6]-linregerrQ	logMelFreqBand-sma[6]-kurtosis	lspFreq-sma[4]-amean
mfcc-sma-de[6]-stddev	H(mean)	pcm-loudness-sma-linregerrA
mfcc-sma-de[10]-linregerrQ	logMelFreqBand-sma-de[3]-upleveltime75	mfcc-sma[10]-upleveltime90
logMelFreqBand-sma[2]-linregerrQ	mfcc-sma[0]-quartile2	logMelFreqBand-sma-de[3]-upleveltime90
mfcc-sma-de[10]-iqr1-2	F0finEnv-sma-kurtosis	mfcc-sma[7]-linregc2
voicingFinalUnclipped-sma-linregc2	mfcc-sma-de[11]-pctlrange0-1	mfcc-sma-de[11]-minPos
logMelFreqBand-sma[2]-stddev Autocorrelation(mean)	lspFreq-sma-de[5]-stddev	
logMelFreqBand-sma-de[2]-kurtosis	mfcc-sma[9]-pctlrange0-1	lspFreq-sma-de[5]-linregerrQ
voicingFinalUnclipped-sma-de-linregc2	mfcc-sma-de[7]-percentile99.0	mfcc-sma[11]-linregc2
mfcc-sma-de[5]-pctlrange0-1	mfcc-sma[1]-iqr1-3	lspFreq-sma-de[5]-skewness
mfcc-sma[5]-iqr1-2	lspFreq-sma[1]-iqr1-2	mfcc-sma-de[9]-amean
voicingFinalUnclipped-sma-de-linregc1	lspFreq-sma-de[0]-linregc1	lspFreq-sma[6]-linregc1
logMelFreqBand-sma[0]-kurtosis	lspFreq-sma-de[6]-linregerrA	mfcc-sma-de[3]-percentile1.0
mfcc-sma-de[2]-percentile99.0	logMelFreqBand-sma[4]-quartile1	mfcc-sma[3]-iqr2-3
shimmerLocal-sma-stddev	mfcc-sma[14]-percentile99.0	lspFreq-sma-de[7]-upleveltime90
logMelFreqBand-sma[6]-upleveltime90	mfcc-sma-de[4]-linregc1	mfcc-sma[5]-linregc1
mfcc-sma[3]-amean	mfcc-sma-de[3]-quartile1	lspFreq-sma-de[1]-amean
shimmerLocal-sma-iqr2-3	logMelFreqBand-sma[2]-upleveltime75	lspFreq-sma[4]-iqr1-3
mfcc-sma[8]-amean	voicingFinalUnclipped-sma-de-quartile2	mfcc-sma-de[14]-upleveltime90
jitterDDP-sma-amean	mfcc-sma-de[1]-iqr2-3	voicingFinalUnclipped-sma-de-maxPos
mfcc-sma-de[4]-linregerrA	mfcc-sma[8]-quartile3	logMelFreqBand-sma-de[0]-quartile2
logMelFreqBand-sma-de[5]-percentile1.0	mfcc-sma-de[12]-linregc1	lspFreq-sma[3]-linregerrA
mfcc-sma-de[0]-percentile99.0	mfcc-sma[4]-percentile1.0	lspFreq-sma[5]-linregc2
mfcc-sma-de[8]-iqr1-2	mfcc-sma[1]-pctlrange0-1	logMelFreqBand-sma-de[2]-minPos
logMelFreqBand-sma[4]-linregerrA	logMelFreqBand-sma[6]-iqr2-3	mfcc-sma-de[1]-upleveltime75
mfcc-sma-de[10]-linregc1	GNEmean	mfcc-sma[11]-skewness
mfcc-sma-de[4]-stddev	GNEstd	lspFreq-sma[4]-percentile1.0
F0final–Turn-duration	lspFreq-sma[6]-linregerrA	lspFreq-sma-de[7]-upleveltime75
mfcc-sma-de[10]-quartile1	logMelFreqBand-sma-de[6]-minPos	mfcc-sma-de[6]-upleveltime75
logMelFreqBand-sma[5]-linregerrA	mfcc-sma[8]-iqr1-3	logMelFreqBand-sma-de[1]-skewness
mfcc-sma-de[6]-linregerrA	logMelFreqBand-sma[0]-linregc1	logMelFreqBand-sma[6]-percentile99.0
jitterLocal-sma-de-iqr2-3	logMelFreqBand-sma[4]-skewness	voicingFinalUnclipped-sma-de-amean
mfcc-sma[0]-kurtosis	mfcc-sma-de[7]-iqr1-2	H(max)
lspFreq-sma-de[1]-iqr2-3	mfcc-sma[1]-maxPos	lspFreq-sma-de[7]-linregc2
logMelFreqBand-sma[0]-skewness	Pulses	lspFreq-sma-de[4]-linregerrQ
F0finEnv-sma-linregc1	jitterDDP-sma-quartile1	logMelFreqBand-sma-de[3]-minPos
jitterLocal-sma-linregerrA	lspFreq-sma[3]-quartile1	lspFreq-sma[3]-iqr1-2
jitterDDP-sma-percentile99.0	mfcc-sma-de[0]-skewness	mfcc-sma-de[11]-kurtosis
logMelFreqBand-sma[6]-upleveltime75	lspFreq-sma-de[3]-iqr1-2	pcm-loudness-sma-de-stddev
logMelFreqBand-sma-de[0]-linregerrA	lspFreq-sma[2]-linregerrQ	lspFreq-sma[6]-quartile2
mfcc-sma-de[0]-pctlrange0-1	lspFreq-sma-de[6]-maxPos	mfcc-sma-de[0]-amean
shimmerLocal-sma-iqr1-2	lspFreq-sma[4]-minPos	lspFreq-sma-de[4]-pctlrange0-1
shimmerLocal-sma-de-quartile2	mfcc-sma-de[7]-pctlrange0-1	mfcc-sma-de[5]-skewness

Feature	Feature	Feature
logMelFreqBand-sma[4]-iqr1-3	lspFreq-sma-de[7]-minPos	mfcc-sma-de[3]-percentile99.0
mfcc-sma[3]-quartile2	lspFreq-sma[1]-stddev	lspFreq-sma-de[3]-upleveltime90
jitterLocal-sma-de-linregerrQ	lspFreq-sma-de[2]-kurtosis	mfcc-sma[2]-linregc1
mfcc-sma-de[4]-iqr1-3	mfcc-sma-de[4]-upleveltime90	lspFreq-sma[1]-percentile99.0
logMelFreqBand-sma-de[1]-quartile3	lspFreq-sma[5]-amean	mfcc-sma[10]-kurtosis
logMelFreqBand-sma[6]-pctlrange0-1	mfcc-sma-de[10]-minPos	mfcc-sma-de[4]-quartile2
mfcc-sma-de[8]-pctlrange0-1	mfcc-sma-de[6]-percentile99.0	lspFreq-sma[5]-linregerrQ
mfcc-sma-de[4]-iqr2-3	shimmer(apq11)	pcm-loudness-sma-stddev
mfcc-sma-de[2]-percentile1.0	logMelFreqBand-sma[1]-percentile99.0	lspFreq-sma[5]-pctlrange0-1
mfcc-sma[12]-linregerrA	logMelFreqBand-sma[7]-quartile3	mfcc-sma-de[1]-skewness
Voicedunvoiced ratio	mfcc-sma[12]-iqr2-3	mfcc-sma[3]-upleveltime75
Voicedtotal frames ratio	logMelFreqBand-sma-de[6]-quartile2	lspFreq-sma[5]-linregerrA
unvoicedtotal frames ratio	logMelFreqBand-sma[7]-stddev	lspFreq-sma[5]-iqr1-3
F0final-sma-de-upleveltime75	mfcc-sma-de[14]-minPos	mfcc-sma-de[4]-skewness
logMelFreqBand-sma-de[3]-percentile99.0	logMelFreqBand-sma[3]-upleveltime90	lspFreq-sma[6]-quartile3
logMelFreqBand-sma-de[0]-linregerrQ	lspFreq-sma[2]-quartile1	logMelFreqBand-sma-de[1]-amean
logMelFreqBand-sma-de[0]-iqr1-3	lspFreq-sma[3]-linregerrQ	
mfcc-sma-de[2]-linregc2	mfcc-sma[11]-maxPos	

Bibliography

M. A. Ahmad. Artificial neural network vs. support vector machine for speech emotion recognition. *Tikrit Journal of Pure Science*, 21(6), 2016.

E. M. Albornoz, D. H. Milone, and H. L. Rufiner. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, 25(3):556–570, 2011.

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

F. Alonso-Martín, M. Malfaz, J. Sequeira, J. F. Gorostiza, and M. A. Salichs. A multimodal emotion detection system during human–robot interaction. *Sensors*, 13(11): 15549–15581, 2013.

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

C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43 (2):155–177, 2015.

V. Anoop, P. Rao, and S. Aruna. An effective speech emotion recognition using artificial neural networks. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, pages 393–401. Springer, 2018.

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

B. Basharirad and M. Moradhaseli. Speech emotion recognition methods: A literature review. In *AIP Conference Proceedings*, volume 1891, page 020105. AIP Publishing, 2017.

S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin. A review on emotion recognition using speech. In *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on*, pages 109–114. IEEE, 2017.

A. Batliner, S. Steidl, and E. Nöth. Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. In *Proc. of a Satellite Workshop of LREC*, volume 2008, page 28, 2008.

A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir. The automatic recognition of emotions in speech. In *Emotion-Oriented Systems*, pages 71–99. Springer, 2011.

D. Bitouk, R. Verma, and A. Nenkova. Class-level spectral features for emotion recognition. *Speech communication*, 52(7-8):613–625, 2010.

P. Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

M. Borchert and A. Dusterhoft. Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pages 147–151. IEEE, 2005.

F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

S.-O. Caballero-Morales. Recognition of emotions in mexican spanish speech: An approach based on acoustic modelling of emotion-specific vowels. *The Scientific World Journal*, 2013, 2013.

H. Cao, R. Verma, and A. Nenkova. Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer speech & language*, 29(1):186–202, 2015.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

B. Chen, Q. Yin, and P. Guo. A study of deep belief network based chinese speech emotion recognition. In *2014 Tenth International Conference on Computational Intelligence and Security (CIS)*, pages 180–184. IEEE, 2014.

C. Chen, M. You, M. Song, J. Bu, and J. Liu. An enhanced speech emotion recognition system based on discourse information. In *International Conference on Computational Science*, pages 449–456. Springer, 2006.

L. Chen, X. Mao, P. Wei, Y. Xue, and M. Ishizuka. Mandarin emotion recognition combining acoustic and emotional point information. *Applied Intelligence*, 37(4):602–612, 2012a.

L. Chen, X. Mao, Y. Xue, and L. L. Cheng. Speech emotion recognition: Features and classification models. *Digital signal processing*, 22(6):1154–1160, 2012b.

C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.

N. Ding, N. Ye, H. Huang, R. Wang, and R. Malekian. Speech emotion features selection based on bbo-svm. In *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, pages 210–216. IEEE, 2018.

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

I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard. Design, recording and verification of a danish emotional speech database. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Z. Esmaileyan and H. Marvi. A database for automatic persian speech emotion recognition: collection, processing and evaluation. *International Journal of Engineering*, 27:79–90, 2014.

F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

L. Fu, X. Mao, and L. Chen. Speaker independent emotion recognition using hmms fusion system with relative features. In *Intelligent Networks and Intelligent Systems, 2008. ICINIS'08. First International Conference on*, pages 608–611. IEEE, 2008.

M. Ghai, S. Lal, S. Duggal, and S. Manik. Emotion recognition on speech signals using machine learning. In *Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference on*, pages 34–39. IEEE, 2017.

D. Gharavian, M. Sheikhan, and F. Ashoftedel. Emotion recognition improvement using normalized formant supplementary features by hybrid of dtw-mlp-gmm model. *Neural Computing and Applications*, 22(6):1181–1191, 2013.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

M. A. Hall. Correlation-based feature selection for machine learning. 1999.

J. H. Hansen and S. E. Bou-Ghazale. Getting started with susas: A speech under simulated and actual stress database. In *Fifth European Conference on Speech Communication and Technology*, 1997.

A. Hassan and R. I. Damper. Classification of emotional speech using 3dec hierarchical classifier. *Speech Communication*, 54(7):903–916, 2012.

N. A. Hendy and H. Farag. Emotion recognition using neural network: A comparative study. In *Proceedings of World Academy of Science, Engineering and Technology*, number 75, page 791. World Academy of Science, Engineering and Technology (WASET), 2013.

P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and J. R. Orozco-Arroyave. Nonlinear dynamics characterization of emotional speech. *Neurocomputing*, 132:126–135, 2014.

C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha, and L. Zhao. Practical speech emotion recognition based on online learning: From acted data to elicited data. *Mathematical Problems in Engineering*, 2013, 2013.

A. B. Ingale and D. Chaudhari. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1):235–238, 2012.

P. Jackson and S. Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

A. Jalili, S. Sahami, C.-Y. Chi, and R. Amirfattahi. Speech emotion recognition using cyclostationary spectral analysis. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

W. A. Jassim, R. Paramesran, and N. Harte. Speech emotion classification using combined neurogram and interspeech 2010 paralinguistic challenge features. *IET Signal Processing*, 11(5):587–595, 2017.

A. Joshi. Speech emotion recognition using combined features of hmm & svm algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(8), 2013.

A. Joshi and R. Kaur. A study of speech emotion recognition methods. *Int. J. Comput. Sci. Mob. Comput.(IJCSMC)*, 2(4):28–31, 2013.

P. Kabal and R. P. Ramachandran. The computation of line spectral frequencies using chebyshev polynomials. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6):1419–1426, 1986.

N. Kamaruddin, A. Wahab, and C. Quek. Cultural dependency analysis for understanding speech emotion. *Expert Systems with Applications*, 39(5):5115–5133, 2012.

K. Khanchandani and M. A. Hussain. Emotion recognition using multilayer perceptron and generalized feed forward neural network. 2009.

S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout. Enhancement of an arabic speech emotion recognition system. *International Journal of Applied Engineering Research*, 13 (5):2380–2389, 2018.

M. Kockmann, L. Burget, et al. Application of speaker-and language identification state-of-the-art techniques for emotion recognition. *Speech Communication*, 53(9-10):1172–1185, 2011.

S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

S. G. Koolagudi, Y. S. Murthy, and S. P. Bhaskar. Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*, 21(1):167–183, 2018.

M. Kotti and F. Paternò. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *International journal of speech technology*, 15(2):131–150, 2012.

S. Kuchibhotla, H. Vankayalapati, R. Vaddi, and K. R. Anne. A comparative analysis of classifiers in emotion recognition through acoustic features. *International Journal of Speech Technology*, 17(4):401–408, 2014.

N. Kurpukdee, S. Kasuriya, V. Chunwijitra, C. Wutiwiwatchai, and P. Lamsrichan. A study of support vector machines for emotional speech recognition. In *Information and Communication Technology for Embedded Systems (IC-ICTES), 2017 8th International Conference of*, pages 1–6. IEEE, 2017.

S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh. Speech emotion recognition. In *Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on*, pages 1–4. IEEE, 2014.

R. B. Lanjewar and D. Chaudhari. Comparative analysis of speech emotion recognition system using different classifiers on berlin emotional speech database. *International Journal of Electrical and Electronics Engineering Research (IJEEER)*, 3(5):145–56, 2013a.

R. B. Lanjewar and D. Chaudhari. Speech emotion recognition: a review. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(4):68–71, 2013b.

R. B. Lanjewar, S. Mathurkar, and N. Patel. Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k-nearest neighbor (k-nn) techniques. *Procedia Computer Science*, 49:50–57, 2015.

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

C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005.

Y.-L. Lin and G. Wei. Speech emotion recognition based on hmm and svm. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 8, pages 4898–4901. IEEE, 2005.

Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280, 2018.

K. Lopez-de Ipiña, J. Alonso-Hernández, J. Solé-Casals, C. M. Travieso-González, A. Ezeiza, M. Faundez-Zanuy, P. M. Calvo, and B. Beitia. Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of alzheimer? s disease. *Neurocomputing*, 150:392–401, 2015.

I. Luengo, E. Navas, and I. Hernáez. Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6):490–501, 2010.

M. Lugger and B. Yang. The relevance of voice quality features in speaker independent emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–17. IEEE, 2007.

M. Lugger and B. Yang. Psychological motivated multi-stage emotion classification exploiting voice quality features. *Speech Recognition, In-Tech*, pages 395–410, 2008.

S. E. E. Lupu. Improving speech emotion recognition using frequency and time domain acoustic features. *Proceedings of SPAMEC, Cluj-Napoca, Romania EURASIP 2011*, 2011.

A. Manolov, O. Boumbarov, A. Manolova, V. Poulkov, and K. Tonchev. Feature selection in affective speech classification. In *Telecommunications and Signal Processing (TSP), 2017 40th International Conference on*, pages 354–358. IEEE, 2017.

S. Mariooryad and C. Busso. Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Communication*, 57:1–12, 2014.

S. Mariooryad, R. Lotfian, and C. Busso. Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

A. H. Marpaung and A. Gonzalez. Toward building automatic affect recognition machine using acoustics features. In *FLAIRS Conference*, 2014.

G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE, 2010.

A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. Di Natale. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems*, 63:68–81, 2014.

A. Milton, S. S. Roy, and S. T. Selvi. Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, 69(9), 2013.

C. Monzo, I. Iriondo, and J. C. Socoró. Voice quality modelling for expressive speech synthesis. *The Scientific World Journal*, 2014, 2014.

I. R. Murray and J. L. Arnott. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer Speech & Language*, 22(2):107–129, 2008.

M. B. Mustafa, M. A. Yusoof, Z. M. Don, and M. Malekzadeh. Speech emotion recognition research: an analysis of research focus. *International Journal of Speech Technology*, 21 (1):137–156, 2018.

H. Muthusamy, K. Polat, and S. Yaacob. Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. *Mathematical Problems in Engineering*, 2015, 2015.

D. Neiberg, K. Elenius, and K. Laskowski. Emotion recognition in spontaneous speech using gmms. In *Ninth International Conference on Spoken Language Processing*, 2006.

C. Oflazoglu and S. Yildirim. Recognizing emotion from turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):26, 2013.

C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014.

J. N. Padmaja and R. R. Rao. Analysis of speaker independent emotion recognition system using principle component analysis (pca) and gaussian mixture models (gmm). *Analysis*, 2017.

Y. Pan, P. Shen, and L. Shen. Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108, 2012.

T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and W.-Y. Liao. Combining acoustic features for improved emotion recognition in mandarin speech. In *International conference on affective computing and intelligent interaction*, pages 279–285. Springer, 2005a.

T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and W.-Y. Liao. Detecting emotions in mandarin speech. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 3, September 2005: Special Issue on Selected Papers from ROCLING XVI*, 10(3):347–362, 2005b.

T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and P.-J. Li. Mandarin emotional speech recognition based on svm and nn. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1096–1100. IEEE, 2006.

P. Partila, M. Voznak, and J. Tovarek. Pattern recognition methods and features selection for speech emotion recognition system. *The Scientific World Journal*, 2015, 2015.

H. Pérez-Espinosa, C. A. Reyes-García, and L. Villaseñor-Pineda. Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model. *Biomedical Signal Processing and Control*, 7(1):79–87, 2012.

J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

C. S. Ram and R. Ponnusamy. An effective automatic speech emotion recognition for tamil language using support vector machine. In *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*, pages 19–23. IEEE, 2014.

S. Ramakrishnan and I. M. El Emary. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52(3):1467–1478, 2013.

K. S. Rao and S. G. Koolagudi. Identification of hindi dialects and emotions using spectral and prosodic features of speech. *IJSCI: International Journal of Systemics, Cybernetics and Informatics*, 9(4):24–33, 2011.

A. A. Razak, R. Komiya, M. Izani, and Z. Abidin. Comparison between fuzzy and nn method for speech emotion recognition. In *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, volume 1, pages 297–302. IEEE, 2005.

A. P. Reddy and V. Vijayarajan. Extraction of emotions from speech-a survey. *International Journal of Applied Engineering Research*, 12(16):5760–5767, 2017.

S. Renjith and K. Manju. Speech based emotion recognition in tamil and telugu using lpcc and hurst parametersa comparitive study using knn and ann classifiers. In *Circuit, Power and Computing Technologies (ICCPCT), 2017 International Conference on*, pages 1–6. IEEE, 2017.

J. Rong, G. Li, and Y.-P. P. Chen. Acoustic feature selection for automatic emotion recognition from speech. *Information processing & management*, 45(3):315–328, 2009.

A. K. Samantaray, K. Mahapatra, B. Kabi, and A. Routray. A novel approach of speech emotion recognition with prosody, quality and derived features using svm classifier for a class of north-eastern languages. In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*, pages 372–377. IEEE, 2015.

R. San-Segundo, R. Cordoba, J. Ferreiros, J. Macias-Guarasa, J. M. Montero, F. Fernández, L. Dharo, R. Barra, and R. Barra. Speech technology at home: enhanced interfaces for people with disabilities. *Intelligent Automation & Soft Computing*, 15(4):647–666, 2009.

S. T. Saste and S. Jagdale. Emotion recognition from speech using mfcc and dwt for security system. In *Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of*, volume 1, pages 701–704. IEEE, 2017.

F. Schiel, S. Steininger, and U. Türk. The smartkom multimodal corpus at bas. In *LREC*. Citeseer, 2002.

B. Schuller and G. Rigoll. Timing levels in segment-based speech emotion recognition. In *Proc. INTERSPEECH 2006, Proc. Int. Conf. on Spoken Language Processing ICSLP, Pittsburgh, USA*, 2006.

B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

B. Schuller, R. Müller, M. Lang, and G. Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*, 2005.

B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. In *Proc. Speech Prosody 2006, Dresden*, 2006.

B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl. Towards more reality in the recognition of emotional speech. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–941. IEEE, 2007.

B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2794–2797, 2010.

B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, 2011.

B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

M. C. Sezgin, B. Gunsel, and G. K. Kurt. Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):16, 2012.

M. Shami and W. Verhelst. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3):201–212, 2007.

A. Shaw, R. K. Vardhan, and S. Saxena. Emotion recognition and classification in speech using artificial neural networks. *International Journal of Computer Applications*, 145 (8), 2016.

P. Shen, Z. Changjun, and X. Chen. Automatic speech emotion recognition using support vector machine. In *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*, volume 2, pages 621–625. IEEE, 2011.

Y. Shi et al. Particle swarm optimization: developments, applications and resources. In *evolutionary computation, 2001. Proceedings of the 2001 Congress on*, volume 1, pages 81–86. IEEE, 2001.

A. Shirani and A. R. N. Nilchi. Speech emotion recognition based on svm as both feature selector and classifier. *International Journal of Image, Graphics & Signal Processing*, 8 (4), 2016.

S. S. Stevens. The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, 69(1):1–25, 1956.

V. Štruc, F. Mihelic, et al. Multi-modal emotion recognition using canonical correlations and acoustic features. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4133–4136. IEEE, 2010.

R. S. Sudhkar and M. C. Anil. Emotion detection of speech signals with analysis of salient aspect pitch contour. *Emotion*, 3(10), 2016.

R. Sun, E. Moore, and J. F. Torres. Investigating glottal parameters for differentiating emotional categories with similar prosodics. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4509–4512. IEEE, 2009.

Y. Sun and G. Wen. Emotion recognition using semi-supervised feature selection with speaker normalization. *International Journal of Speech Technology*, 18(3):317–331, 2015.

Y. Sun, G. Wen, and J. Wang. Weighted spectral features based on local hu moments for speech emotion recognition. *Biomedical signal processing and control*, 18:80–90, 2015.

M. Swain, A. Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

M. Tahon, G. Degottex, and L. Devillers. Usual voice quality features and glottal features for emotional valence detection. In *Speech Prosody 2012*, 2012.

W. Tarng, Y.-Y. Chen, C.-L. Li, K.-R. Hsie, and M. Chen. Applications of support vector machines on smart phone systems for emotional speech recognition. *World Academy of Science, Engineering and Technology*, 72:106–113, 2010.

A. Tickle, S. Raghu, and M. Elshaw. Emotional recognition from the speech signal for a virtual education agent. In *Journal of Physics: Conference Series*, volume 450, page 012053. IOP Publishing, 2013.

D. Tomar, D. Ojha, and S. Agarwal. An emotion detection system based on multi least squares twin support vector machine. *Advances in Artificial Intelligence*, 2014:8, 2014.

I. Trabelsi, D. B. Ayed, and N. Ellouze. Comparison between gmm-svm sequence kernel and gmm: application to speech emotion recognition. *Journal of Engineering Science and Technology*, 11(9):1221–1233, 2016.

A. S. Utane and S. Nalbalwar. Emotion recognition through speech. *International Journal of Applied Information Systems (IJAIS)*, pages 5–8, 2013.

M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.

D. Verma, D. Mukhopadhyay, and E. Mark. Role of gender influence in vocal hindi conversations: A study on speech emotion recognition. In *Computing Communication Control and automation (ICCUBEA), 2016 International Conference on*, pages 1–6. IEEE, 2016.

D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1500–1503. IEEE, 2005a.

D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models. In *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pages 2871–2874. IEEE, 2005b.

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

D. Ververidis and C. Kotropoulos. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition. *signal processing*, 88(12):2956–2970, 2008.

L. Vidrascu and L. Devillers. Real-life emotion representation and detection in call centers data. In *International Conference on Affective Computing and Intelligent Interaction*, pages 739–746. Springer, 2005.

T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 474–477. IEEE, 2005.

T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*, 2006.

T. Vogt, E. André, and J. Wagner. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In *Affect and emotion in human-computer interaction*, pages 75–91. Springer, 2008.

H. K. Vydana, P. P. Kumar, K. S. R. Krishna, and A. K. Vuppala. Improved emotion recognition using gmm-ubms. In *Signal Processing And Communication Engineering Systems (SPACES), 2015 International Conference on*, pages 53–57. IEEE, 2015.

V. B. Waghmare, R. R. Deshmukh, P. P. Shrishrimal, and G. B. Janvale. Emotion recognition system from artificial marathi speech using mfcc and lda techniques. In *Fifth International Conference on Advances in Communication, Network, and Computing–CNC*, 2014.

K. Wang, N. An, and L. Li. Speech emotion recognition based on wavelet packet coefficient model. In *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, pages 478–482. IEEE, 2014.

K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, 2015.

K. Wang, Z. Chu, K. Wang, T. Yu, and L. Liu. Speech emotion recognition using multiple classifiers. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, pages 84–93. Springer, 2017.

C.-H. Wu and W.-B. Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2011.

S. Wu, T. H. Falk, and W.-Y. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768–785, 2011.

J. Xiaoqing, X. Kewen, L. Yongliang, and B. Jianchuan. Noisy speech emotion recognition using sample reconstruction and multiple-kernel learning. *The Journal of China Universities of Posts and Telecommunications*, 24(2):1–17, 2017.

B. Yang and M. Lugger. Emotion recognition from speech signals using new harmony features. *signal processing*, 90(5):1415–1423, 2010.

N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman, and M. Sturge-Apple. Speech-based emotion classification using multiclass svm with hybrid kernel and thresholding fusion. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 455–460. IEEE, 2012.

C. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, C. Berkai, and K. Polat. A new hybrid pso assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, 69:149–158, 2017.

M. You, C. Chen, J. Bu, J. Liu, and J. Tao. Emotional speech analysis on nonlinear manifold. In *null*, pages 91–94. IEEE, 2006.

Q. Zhang, N. An, K. Wang, F. Ren, and L. Li. Speech emotion recognition using combination of features. In *Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on*, pages 523–528. IEEE, 2013.

X. Zhao, S. Zhang, and B. Lei. Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications*, 24(7-8):1539–1553, 2014.

W. Zheng, M. Xin, X. Wang, and B. Wang. A novel speech emotion recognition method via incomplete sparse least square regression. *IEEE Signal Processing Letters*, 21(5): 569–572, 2014.

J. Zhou, G. Wang, Y. Yang, and P. Chen. Speech emotion recognition based on rough set and svm. In *Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on*, volume 1, pages 53–61. IEEE, 2006.

Y. Zhou, Y. Sun, J. Zhang, and Y. Yan. Speech emotion recognition using both spectral and prosodic features. In *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, pages 1–4. IEEE, 2009.

Y. Zhou, J. Li, Y. Sun, J. Zhang, Y. Yan, and M. Akagi. A hybrid speech emotion recognition system based on spectral and prosodic features. *IEICE TRANSACTIONS on Information and Systems*, 93(10):2813–2821, 2010.