Detecting Emotions from Speech using Machine Learning Techniques

A DISSERTATION PRESENTED

BY

TANMOY ROY

TO

THE DEPARTMENT OF ELECTRICAL & ELECTRONIC ENGINEERING SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

ELECTRONIC ENGINEERING

UNIVERSITY
OF
JOHANNESBURG

JOHANNESBURG, GAUTENG

MAY 2019

Thesis advisors: Professor Tshilidzi Marwala and Professor Snehashish Chakraverty

Detecting Emotions from Speech using Machine Learning Techniques

Abstract

Speech is an extremely effective form of communication method that makes us unique among all the species on earth. Modern-day Artificial Intelligence (AI) systems are now capable enough to strike a spoken communication with us using the Automatic Speech Recognition (ASR) system because ASR systems are presently in a very advanced stage. However, this human-machine speech communication is still not natural enough as between two humans, because the AI agents can not efficiently identify the emotional states of the speaker.

Speech propagates as a waveform and carries various information along with it apart from the intended message of the speaker. Moreover, emotion is the state of our mind which usually gets reflected in speech sound and different physical expressions. So, speech signals carry information regarding the emotional state of the speaker. The study of Speech Emotion Recognition (SER) explores various attributes of the speech signal and employs different Machine Learning (ML) techniques to identify the human emotions concealed in the speech signal efficiently.

The task of identifying emotions from the speech is difficult, and it is deceiving even for human ears. More than twenty years of research could not bring consensus among researchers regarding feature sets for SER. Even employing most powerful classification techniques such as Support Vector Machine (SVM), Neural Networks (NN), and Deep Learning (DL) techniques could not provide sat-

Thesis advisors: Professor Tshilidzi Marwala and Professor Snehashish Chakraverty

isfactory classification accuracy. So, following three significant difficulties in SER research has come out during this research which are: 1. the lack of a standard feature set; 2. considering SER as a sequence classification technique similar to ASR which should not be the case; and 3. the speech signal preprocessing still needs more effective algorithms for better classification results.

This thesis tried to address the issues listed above. First of all, a new algorithm viz. Wavelet Convolution based Speech Endpoint Detection (WCSED) is proposed for more precise detection of speech endpoints to enhance classification results. Secondly, a new feature set, named as Subjective Emotional Gap Reduction Technique (SEGRT), is developed which is designed specifically for SER. The existing feature sets are mostly borrowed from ASR feature sets. However, it is found that the ASR and SER problems are different so the feature sets should also be specific and that is why this new feature set is proposed. The SEGRT is an attempt to reduce the *subjective gap* between features extracted from the utterance of different speakers. And finally, the classification models used in SER so far has mostly considered SER as a sequence classification problem, such as Hidden Markov Model, but it is established here that SER should be a regular classification problem. So, in order to support the above, results are also provided in detail by applying most successful classification problems like Gaussian Naive Bayes, Support Vector Machines, k-Nearest Neighbors, and Feed-forward Neural Networks.

Contents

Listing of figures

List of Tables

I dedicate this dissertation to my Father, Mother, Wife, Brother, and Son. Without their sacrifice and support, I could not have achieved this.

Acknowledgments

I am grateful to my supervisor Prof. Tshilidzi Marwala for providing me this opportunity to work under him and continuous patronage for me in this endeavor. I am sincerely thankful to my co-supervisor Prof. Snehashish Chakraverty for his encouragement and valuable advice.

Without the sacrifices and support of my whole family especially of my wife Samadrita Roy, things would have been difficult. I cannot forget the teachings of my teachers and the blessings of my elders. I also cannot forget the encouragement and support received from all my friends specially Pramod, Satyakama and Dr. BS Paul. And finally, I would like to sincerely thank every person who at some point and some way helped me to achieve this.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in the University of Johannesburg, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

This dissertation contains fewer than 34,000 words including appendices, bibliography, footnotes, tables and equations.

Tanmoy Roy
May 2019

Nomenclature

Acronyms and Abbreviations

ANN	Artificial Neural Network
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CWT	Continuous Wavelet Transform
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
DWT	Discrete Wavelet Transform
FNN	Feed-forward Neural Network
ELM	Extreme Learning Machine
GMM	Gaussian Mixture Model
GNB	Gaussian Naive Bias
HMM	Hidden Markov Model
i.i.d.	independent and identically distributed
KNN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
ML	Machine Learning
MFCC	Mel-Frequency Cepstral Coefficients

RNN	Recurrent Neural Network
RBF	Radial Basis Function
SED	Speech Endpoint Detection
SEGRT	Subjective Emotional Gap Reduction Technique
SER	Speech Emotion Recognition
SVC	Support Vector Classifier
SVM	Support Vector Machine
WCSED	Wavelet Convolution based Speech Endpoint Detection

Everything is energy. Match the frequency of the reality you want and you cannot help but get that reality. It can be no other way. This is not philosophy. This is physics.

Albert Einstein

1 Introduction

Emotions are ubiquitous to human life, and studies have shown that 90% time of their everyday life, humans feel at least one emotion [1]. Once human experience any emotion, it guides their thoughts and behaviors [2], and to communicate those thoughts, they express them in different forms, and speech is one of those forms. Moreover, since emotions provoke thoughts to invoke speech, speech reflects the emotional state of the speaker. Speech Emotion Recognition (SER) field is the study of the methodologies for identifying the emotions concealed within the speech signals.

Speech is an extremely efficient form of communication method that makes us unique among all the species on earth. That is why present Artificial Intelligence (AI) systems' unique characteristic is their ability to execute spoken conversation with humans efficiently. Examples like Siri, Cortana, and Alexa are the technical marvels which not only can hear us comfortably; but they can reply to us with equal comfort as well. However, are these spoken conversations with AI agents natural enough? To be specific, are these AI agents intelligent enough to read our moods or feelings? The answer is still 'NO'. AI agents cannot read emotions expressed in our speech. The first significant work on SER was reported way back in 1978 by Williamson [3] but the challenge to detect emotions from the speech in realtime is still on.

1.1 Emotion: The hidden force which drives us

Recent research findings suggest that only ten percent of the life experience of humankind is devoid of any emotional experience, and when an emotion is experienced, it guides peoples thoughts and behavior. Darwin [4] was one of the first to make the case that the study of emotional expressions are beneficial and also discussed the effects of emotions. However, only in the 1990s, Darwin's [4] study was recognized as an essential aspect of studies related to humankind. Subsequent studies revealed that emotional states have diversified effect on our everyday life [1]. In our life from the first day, different emotions play a significant role in our survival and further progress, which researchers like Freedman et al. [5], Frank [6], and Damasio [7] explored in detail.

1.1.1 How emotion influences our every move?

The human brain controls every move they make from blinking of eyes to applying breaks while driving a car. Every external or internal stimulation (also called interrupt) has to be managed by the brain. The interrupt could be an audio-visual input, a change in the surrounding environment, an occurrence of a disease, or any other event. The human brain has three different layers of neural anatomy laid down, one layer on top of another during different phases of evolution [8]. The oldest and deepest layer is called the *reptilian brain* (fig:1.1.1 [9]) and is responsible for autonomic functions, such as heartbeat and breathing, as well as instinctual behaviors, such as the sucking reflex. The next layer is called the *limbic system*, and the *amygdala* is the vital organ in this system. The *amygdala* is responsible for registering emotional stimuli and storing emotional memories. The *limbic system* is also called the *emotional brain* since it is responsible for handling emotions. The outermost layer is *neocortex* also known as *rational* brain consisting of a layer of gray matter and responsible for conscious processing of sensory stimuli. Figure 1.1.1 shows three primary layers of the human brain and the location of the amygdala along with the mention of the major functionalities of them in brief.

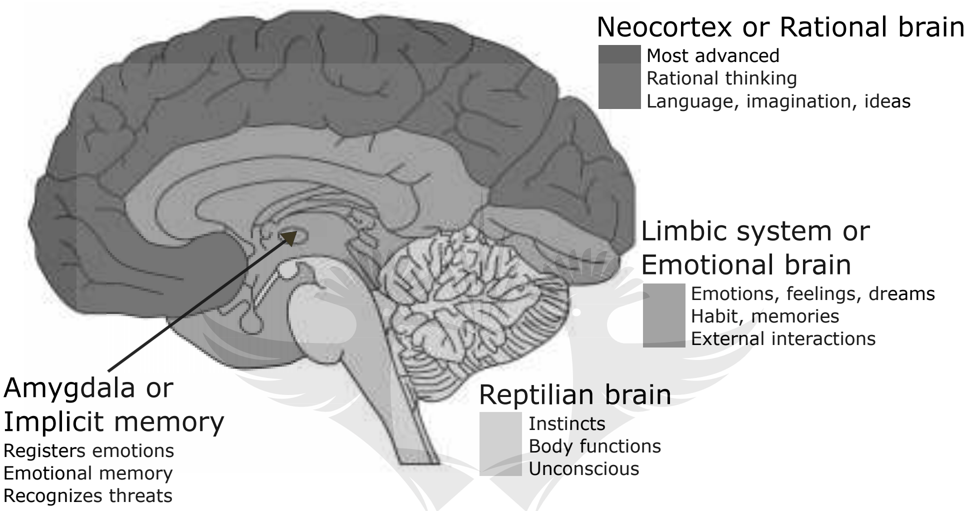

Figure 1.1.1: The figure shows three primary layers of the human brain.

Now the question is why human emotional states influence every move they make? The internal structure of the human brain has the most vital clue [10–12].

- The *limbic system* or *emotional* brain and the *neocortex* or the *rational* brain work in collaboration to respond to an incoming stimulation. Both the brains are interconnected, but the number of neural connections running from the limbic system to the cortex is far greater than the number connecting the cortex to the limbic system [10, 12] which results in a transfer of the influence of emotional states unconsciously from the emotional brain to the rational brain. Figure 1.1.2 shows how the emotional brain receives the stimuli before the rational brain, and due to a much stronger neural connection from emotional to rational brain, the collective response gets significantly

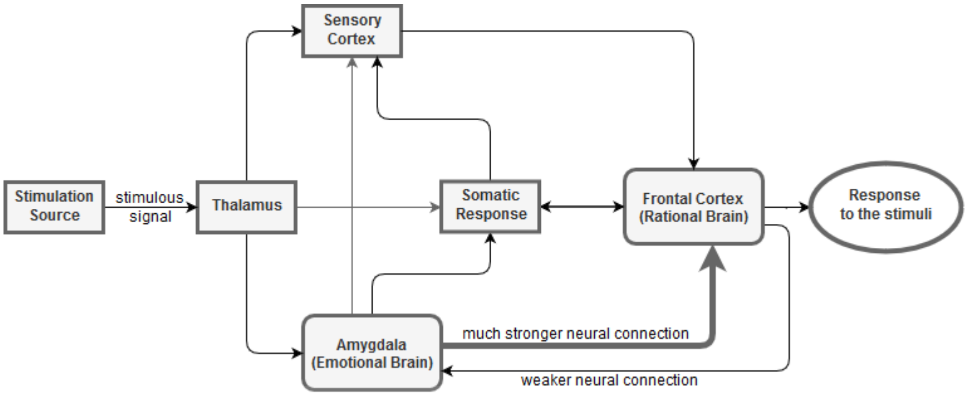

Figure 1.1.2: Figure shows the block diagram of the emotional brain influences the rational brain.

affected by the underlying emotional state.

- Moreover, the timing of neural signal's arrival in the emotional and rational brains is different; the emotional brain receives it first and gets activated relatively early [10, 11].

These two structural arrangements of the human brain explain the influence of emotional states in our every move.

1.1.2 Why is emotion detection important?

Emotion has a substantial influence on the cognitive processes in humans, including perception, attention, learning, memory, reasoning, and problem-solving [13, 14]. So, detecting the emotional state of a person involved in critical activities like flying an aircraft or driving a car can be a very significant input towards crisis management. Recent studies on Emotional Intelligence and its impact on leadership and organization is an exciting new topic of research [15, 16].

Research confirms that different emotional states have a significant impact on our health. Negative emotions like anger, anxiety, and depression not only affect the functioning of the heart but also increase the risk of heart disease [17]. On the other hand, positive emotions can help recover from cardiovascular ailments

[18]. McAllister et al. [19] have shown that patients with genetic diseases can be facilitated by managing emotional conditions. New evidence shows that positive emotions may help limit cancer growth [20] while negative emotions could contribute to cancer incidence [21, 22]. Research even demonstrated that emotions could influence the onset, course, and remission of the disease [23].

Effects of specific emotions on human behavior during financial decision making or economic judgment is an exciting topic of research. Research shows that human emotional states limit human rationality [7, 24, 25], and that leads to the proposal of artificial agents in the human decision making process [26].

So, the situations mentioned above are few of the life examples where it is crucial to track emotional states of the people for their benefit or benefit of the masses.

1.2 Speech Emotion Recognition: A Brief Overview

The Speech signal is the fastest and most natural method of communication between humans. That is why most AI systems incorporate various speech processing systems like Automatic Speech Recognition (ASR), Speaker Recognition, and Speech Synthesis. However, these systems are not natural or realistic enough because they can not detect the underlying emotional state of the speaker. The speech signal is a complex signal which contains information about the speaker and the speaker's emotional state, language, and much more apart from the intended message. Humans extensively use emotions to express their intentions through speech [27]. So, it is required to track the underlying emotional states of the speaker to build better AI systems and other relevant systems to function more naturally [28].

1.2.1 Speech Production System

Different organs like the nose, throat, vocal chord, trachea, and lungs work collectively to produce speech. While exhaling, the air within the lungs is pushed through vocal chord using the trachea. The vocal chord vibrates in that process to generate various sounds and those sounds advance through the air available in the vocal tract. That sound vibration then reaches the oral cavity through the pharyngeal cavity and depending on the position of the velum; those sounds come out from the system either through the mouth or both the nose and mouth. When the velum is in the closed position, only oral sounds are produced, and when it is in open position, both oral and nasal sounds are produced. Figure 1.2.1 [29] shows the human speech production system and what are the organs involved in the process. The lung creates the air pressure to vibrate the vocal chord, and the sound produced by the vocal chord vibration comes out as oral or nasal sounds are called human speech. Within the oral cavity, the tongue plays a vital role in generating different sounds by moving extensively within the oral cavity while speaking. So, this is, in brief, the human speech production system, and it is important to note here that the underlying emotions drive the speaker to modulate their vocal chord vibrations differently for different emotions.

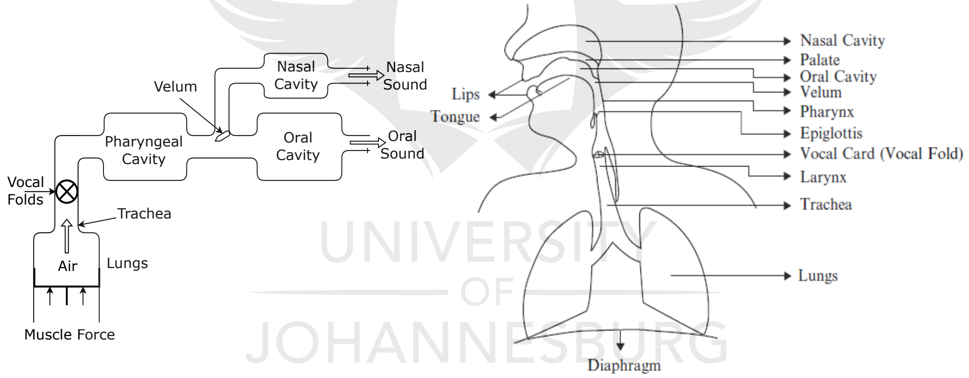

Figure 1.2.1: The figure shows the human speech production system.

1.2.2 Speech Emotion Recognition: The Technique

SER process involves multiple steps before the underlying emotional state of a spoken utterance can be predicted. Figure 1.2.2 depicts the steps involved in SER. Speech produced by an actor or recording of a natural conversation is the primary input to an SER system. The recorded speech is then appropriately labeled based on the psychological definition of different emotions. Labeled speech signals need to be processed further to make those suitable for feature extraction. Noise reduction, endpoint detection, and silence removal are the pre-processing steps involved.

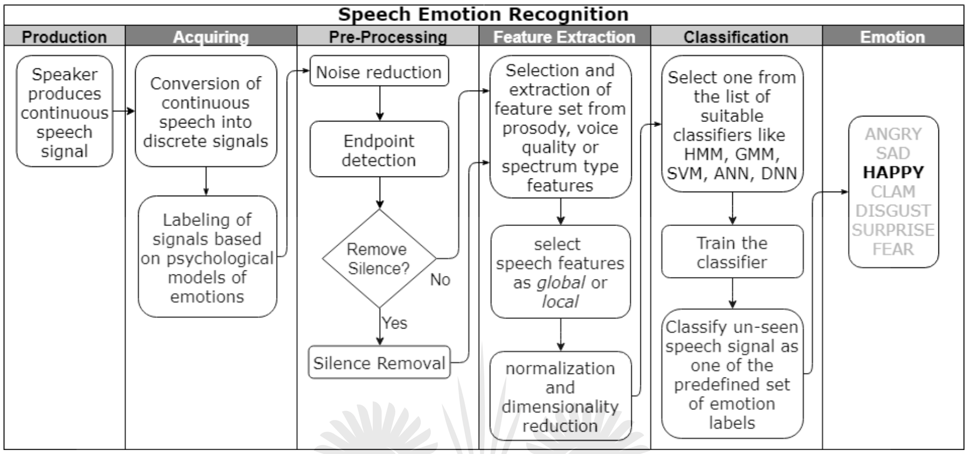
Speech Emotion Recognition
Production
Acquiring
Pre-Processing
Feature Extraction
Classification
Emotion
Speaker produces continuous speech signal
Conversion of continuous speech into discrete signals
Labeling of signals based on psychological models of emotions
Noise reduction
Endpoint detection
Remove Silence?
No
Yes
Silence Removal
Selection and extraction of feature set from prosody, voice quality or spectrum type features
select speech features as *global* or *local*
normalization and dimensionality reduction
Select one from the list of suitable classifiers like HMM, GMM, SVM, ANN, DNN
Train the classifier
Classify un-seen speech signal as one of the predefined set of emotion labels
ANGRY
SAD
HAPPY
CLAM
DISGUST
SURPRISE
FEAR

Figure 1.2.2: The figure shows the steps involved in speech emotion recognition.

Speech Features

Speech is a complex signal which contains enormous amounts of information called speech features which can describe the signal. There are three broad categories of speech features, which are prosody, voice quality, and spectral [30]. Prosody features are also called continuous features and include the pitch or fundamental frequency(F0), the signal energy, the articulation rate, the formants, and all other variants of these features. Several research results concluded that prosodic features provide a reliable indication of emotions [31–37].

The voice quality features such as harsh, breathy, or tense voice and type of voice, correlate with different emotions [35, 37, 38]. However, there is disagreement among researchers over different emotions expressed during different voice quality, for example, some researchers suggested that tense voice has an association with anger, joy, and fear; and soft voice has some association with sadness [38]. On the other hand, Murray and Arnott [35] associated that breathy voice with both anger and happiness, whereas sadness is associated with 'resonant' voice quality.

The spectral features can be extracted as the linear predictive coding (LPC) [39], cepstral coefficients (like mel-frequency cepstral coefficients (MFCC) [40, 41], and linear predictive cepstral coefficients (LPCC)). These are the short-term representation of the speech signal. MFCC and its advanced variants are very popular among SER researchers [42, 43].

Classification Techniques

Researchers have applied different classification techniques for SER. Hidden Markov Model (HMM) was the workhorse for speech related applications, and it was eventually used for SER as well [44–49]. However, the classification accuracy achieved by them was not satisfactory enough. So, other powerful classification techniques like Gaussian Mixture Model (GMM) [50–55] , Support Vector Machine (SVM) [52, 56–60], k-Nearest Neighbors (kNN) [61–64], Artificial Neural Network (ANN) [52, 65–67] and more recently Deep Learning (DL) architectures [43, 68–72] came into the picture to be used in SER. Classification performance has improved but a real life solution is still not achieved. The main reason is that classification performance on cross databases varies too much, which means that a model trained using one dataset performs poorly when tested with another dataset and the accuracy is significantly low.

Challenges in SER

Emotional feeling and their expressions in humans are very subjective and can vary significantly from one person to another. This reality is reflected in research results as well as when speaker specific systems perform better than generic systems. Recognition rates are low in cross-corpora and cross-lingual research initiatives which need to be addressed for more acceptable and industry usable SER systems. Recognition performance is high when the systems are designed on the simulated speech database compared to the systems built on the natural or semi-natural databases. Binary classification accuracy is high while classifying high-arousal and low-arousal emotions and multi-class classification with all the emotions is still challenging. The scarcity of high-quality datasets with enough data points for training is a significant concern for the SER field.

1.2.3 Applications of SER

SER has an extensive scope of applicability in human life, ranging from medical treatments, AI to entertainment.

1. Researchers are trying to improve the familiarity of HMI spoken communications through SER [52, 61].

2. Researchers are trying to read the emotions of autistic people [73, 74] for better communication with them.

3. Patient assistant systems are being developed to help patients according to their emotional needs.

4. Emotional state tracking through speech can help doctors or relatives to manage patient's condition remotely [75].

5. Smart calling agents could decide on the need to transfer a customer call to a human agent depending on the mood of the customer [66, 76].

6. Consumer feedback could be more realistic when there is prior information about the emotional state of the customer so that feedback time and practice can be reviewed.

7. SER could immensely help drivers [52] and pilots [77] to manage critical circumstances by reporting in real-time their emotional states to respective authorities.

8. Virtual assistants like Siri, Alexa, and Google home will be more sensitive to our moods and can become a more sensible human companion.

9. Entertainment like gaming, virtual reality, and augmented reality will be more lively with the introduction of the player's emotion tracking feature and concepts like interactive movie [66] and story-telling [78] will be more popular.

10. Advent of virtual learning platforms will be beneficial if learners' emotions can be adequately tracked to regulate the learning content.

1.3 Motivations and Contributions

Albert Einstein once said - "Everything in life is vibration". According to the Quantum Field Theory (QFT) of particle physics, absolutely everything is made of one or a combination of more than one fundamental fields, and elementary particles are tiny vibrations in these fields [79, 80]. The particles defining our whole existence are also continually vibrating, and that is why we exist with all our attributes [81]. Change in emotional state changes the vibration pattern, primarily the frequency of vibration [81]. One of the vibrations that we humans can generate consciously is through our vocal chord as speech. So, the speech signal is one of the closest form perceived by us, which can contain the existential vibration of humans along with the respective emotional state. That is why in this work the speech signal is considered as the most authentic source for tracking the emotional states of human beings.

Present human-computer interfacing (HCI) systems are at a very advanced state concerning interactivity and artificial intelligence. Present day humanoids and voice assistants can converse with us fluently and intelligently. However, they are not advanced enough to read our feelings or emotions. So, the initial motivation of this study was the ongoing research worldwide to detect human emotions from their speech so that HCI can become more natural.

SER has a wide range of applications spread across diversified fields such as healthcare, transportation, customer relationship management, and entertainment. The research in SER has been ongoing for more than twenty years, but the break-through is yet to arrive mostly because of the lack of consensus on speech features and difficulty in getting enough well-labeled data points. The primary motivation of this work was the challenge posed by the complexity of the SER systems. The deficiency in SER specific feature set was the motivation behind coming up with a new feature set. During this work, it is found that there is a lack of robust speech endpoint detection algorithms which motivated a new endpoint detection algorithm.

1.3.1 Contributions

One of the outstanding problems in trying to recognize emotions is that different individuals may express the same emotion differently. Which means it would be better if a separate reference point can be defined for individual speakers so that emotion classification accuracy can be improved. This work tried to implement a similar approach by using the neutral emotional state of the speakers as a reference point for individual speakers and then identifying other emotions around that neutral state. The concept is new, and it is named as *subjective emotional gap reduction technique (SEGRT)*, which tries to reduce the subjective gap of the utterances of different speakers. A novel feature extraction method has been proposed here to incorporate the SEGRT concept mentioned above. Feature extraction for SER is a challenging task, and there is no consensus among researchers on a single set of features which works best. Different speech features such as pitch, energy, formants, Mel-Scale Coefficients, and Predictive Coding are used for classification, but results are still not satisfactory enough. In the proposed feature extraction method, the digital speech signal is first transformed using discrete wavelet transform (DWT) technique, and then distance or dissimilarity is computed on the transformed coefficients between neutral and other emotional states. The new feature set is used for emotion classification using three different classification techniques to establish that the feature set is giving better or competitive results compared to the contemporary features. This novel feature set is then used with deep learning architecture to compare the emotion classification accuracy with contemporary SER system results, and the proposed model fared well.

Speech pre-processing is one of the most crucial stages of SER and other speech-based applications like ASR. Precise detection of speech endpoints is an essential pre-processing step which affects the performance of the systems where speech utterances need to be extracted from the speech signal. Existing endpoint detection (EPD) methods mostly use Short-Term Energy (STE), Zero-Crossing Rate (ZCR) based approaches, and their variants. However, STE and ZCR based EPD algorithms often fail in the presence of Non-speech Sound Artifacts (NSAs) produced by the speakers. Pattern recognition and classification techniques are also applied, but those methods require labeled data for training. In this work, a novel algorithm is proposed to extract speech endpoints, and the algorithm is termed as Wavelet Convolution based Speech Endpoint Detection (WCSED). WCSED decomposes the speech signal into high-frequency, and low-frequency components using wavelet convolution and then computes information-entropy based thresholds for the two frequency components. The low-frequency thresholds are then used to extract voiced speech segments, whereas the high-frequency thresholds are used to extract the unvoiced speech segments by filtering out the NSAs. WCSED does not require any labeled data for training and can automatically extract speech segments. Experiments are carried out on two speech databases, and the results are promising even in the presence of NSAs.

1.4 Organization of this thesis

This thesis is organized into six chapters. The first chapter is this present chapter where a brief introduction to SER is presented, touching upon the key aspects of this work.

The second chapter is a detailed survey of the SER study taken up by researchers. In this survey, different components of an SER system are described. This chapter includes the design criteria for emotional speech databases and description of some prominent databases, description of features extracted for SER, prominent classification techniques used for SER, and finally, some challenges in SER.

The third chapter proposes a novel speech endpoint detection (SED) algorithm named Wavelet Convolution based Speech Endpoint Detection (WCSED). Here the problem is defined, and then the mathematics of the proposed WCSED algorithm is discussed. After that, the databases used and the results are discussed. This proposed method is already published as an article in a reputed journal [82].

The fourth chapter introduces a novel feature set for SER, which is based on discrete wavelet transform (DWT). This feature extraction technique is described in details, and then it goes on to discuss the classification techniques, experiment environment, and results. This work is going to be published in IEEE conference proceedings [83].

The fifth chapter presents an SER system which uses a neural network as a classification technique. This chapter describes the model in detail and then goes on to discuss the experimental setup and results. This work is going to be published in Springer conference proceedings [84].

Finally, the sixth chapter concludes this thesis with remarks and future directions. This thesis has produced the following publications:

Journals:

1. T Roy, T Marwala, and S Chakraverty. Precise detection of speech endpoints dynamically: A wavelet convolution based approach. *Communications in Nonlinear Science and Numerical Simulation*, 2018. doi:https://doi.org/10.1016/j.cnsns.2018.07.008.

Conferences:

1. T. Roy, T. Marwala, and S. Chakraverty. Introducing New Feature Set based on Wavelets for Speech Emotion Classification. *1st IEEE Conference on Applied Signal Processing (ASPCON) 2018*. Accepted for IEEE Conference proceedings book.

2. T. Roy, T. Marwala, and S. Chakraverty. Speech Emotion Recognition using Neural Network and Wavelet Features. *8th Wave Mechanics Vibrations Conference 2018, NIT RKL*. Accepted for Springer Conference proceedings book.

Book Chapters:

1. T. Roy, T. Marwala, and S. Chakraverty. Novel Advancements of Automatic Emotion Recognition and its Role in the 4th Industrial Revolution. Accepted In The Disruptive Fourth Industrial Revolution: Technology, Society and Beyond, Edited by T Marwala, BS Paul. Springer.

2. T. Roy, T. Marwala, and S. Chakraverty. A Survey of Classification Techniques in Speech Emotion Recognition. Accepted In Mathematical Methods in Interdisciplinary Sciences, Edited by S Chakraverty. Wiley.

3. T. Roy, T. Marwala, and S. Chakraverty. Deep Learning in Speech Emotion Recognition: A Review. Proposed In Mathematical Methods and Vibrations, Edited by S Chakraverty. Elsevier.

The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions.

Marvin Minsky, AI Scientist

2

Machine Learning Paradigms for Speech Emotion Recognition: An Overview

The initial study on emotion started as a study of psychology and acoustics of emotions. The first detailed study on emotions was reported way back in 1872 by Charles Darwin [4]. Fairbanks and Pronovost [85] was among the first who studied pitch of voice during simulated emotion. Since the late fifties, there has been a significant increase in interest by researchers regarding a psychological and acoustic aspect of emotion [3, 5, 86, 87]. However, in the year 1995 Picard [73] introduced the term ”affective computing”, and after that, the study of emotional states has become an integral part of artificial intelligence (AI) research. In this chapter, a detail description of the critical aspects of speech emotion recognition (SER) and present state of the SER research is provided in a few sections.

The first discussion (2.1) of this chapter is on psychological models of emotions which are essential for labeling different emotions to speech utterances. These psychological models are extensively used during emotion speech database preparation. Section (2.2) reviews the prominent speech databases used in SER. Then, there is a detailed review of the speech features which have contributed significantly to the SER research in Section 2.3. Section (2.4) discusses four of the most prominent classification techniques used in SER, namely HMM, GMM, SVM, and DNN. Section (2.5) depicts the difficulties faced by the SER researchers. Finally, Section 2.6 summarizes this chapter and concludes.

2.1 Labeling Emotions: Psychological Models

Psychology of emotions can be viewed as a complex experience of consciousness (psychology), bodily sensation (physiology), and behavior (action-speech). Emotions are relatively brief episodes of synchronized responses that produce noticeable changes in the functioning of an organism. Such changes are brought about by triggering significant events [88].

There are around 300 emotions identified by researchers [89, 90], but most researchers agree on at least on some emotions, including anger, sadness, joy, fear, shame, pride, surprise, disgust, and guilt [24, 91, 92] which are very strong and easily identifiable. Psychological emotion models are used as background for labeling emotional data, and the two most prominent of such models are *Basic Emotions* model and *Valence-Arousal-Dominance* model.

Basic Emotions model was initially proposed by Ekman [91] and later enhanced by other researchers [93, 94]. According to this model, any emotion is a combination of six primary emotions *anger, disgust, fear, happiness, sadness, and surprise*. These six emotions are also referred to as *archetypal emotions*. Plutchik [95] extended the *basic emotions* model and added two more emotions as primary emotions: *anticipation* and *trust*. His model is interestingly designed like a wheel of emotions such that similar emotions are located side by side and very different (bipolar) emotions on the opposite side. This model is called Plutchik's wheel, which is like a color wheel where the intensities of the different emotions are displayed by color saturation. Figure 2.1.1 [96] shows opened representation of Plutchik's emotion model: the eight bipolar emotions are arranged according to their similarity. The color saturation accentuates the intensities of the emotions, and the combinations of the basic emotions are written in between. There is another categorical model called the *Geneva Emotion Wheel* (GEW) where two axes valence and control split the emotions into four separate groups and neutral is at the center [97, 98].

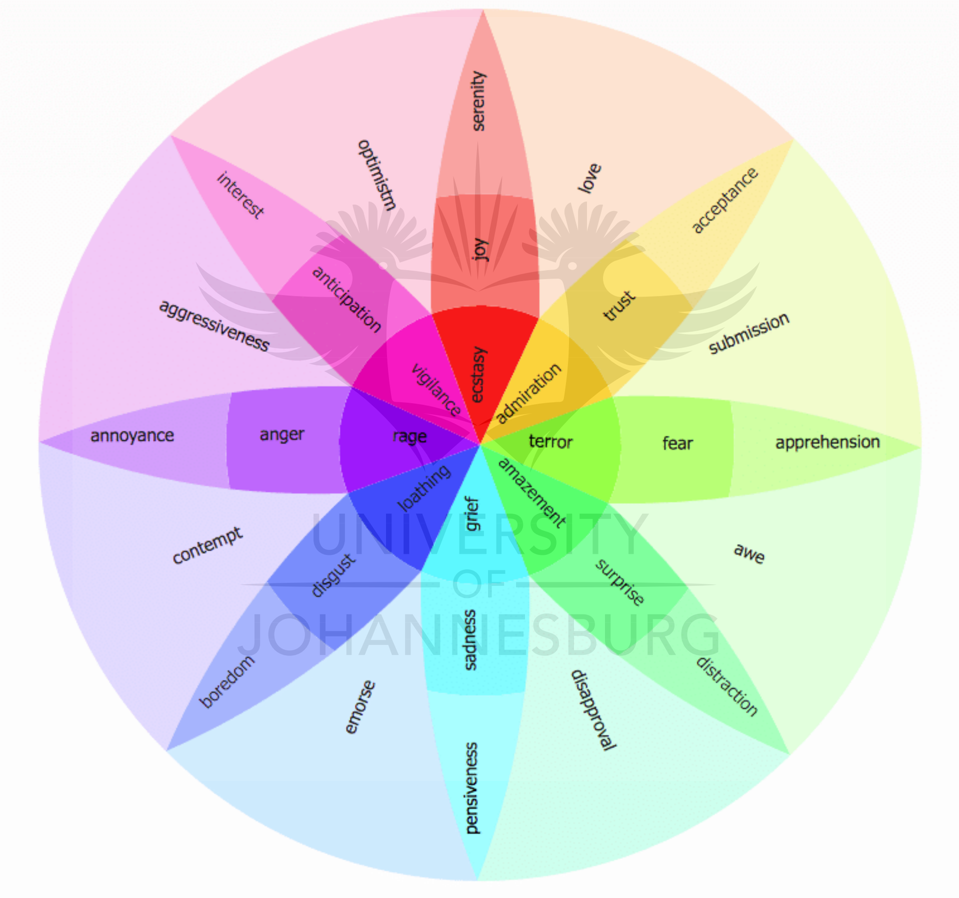

Figure 2.1.1: The figure shows opened representation of Plutchik's wheel.

The basic emotion model was built on the assumption that an independent neural sub-system serves every basic emotion. However, neuroimaging and physiological studies have failed to establish this theory [99]. More recently discrete dimensional models of emotion are gaining more importance. A two-dimensional circumplex model proposes that all affective states arise from two independent neurophysiological systems: one related to valence (a pleasure – displeasure continuum) and the other to arousal (activation-deactivation). That is varying degrees of both valence and arousal represents different emotions [100, 101]. In another approach, different underlying dimensions of affect are chosen: energetic arousal and tense arousal (fig:2.1.2 [99]).

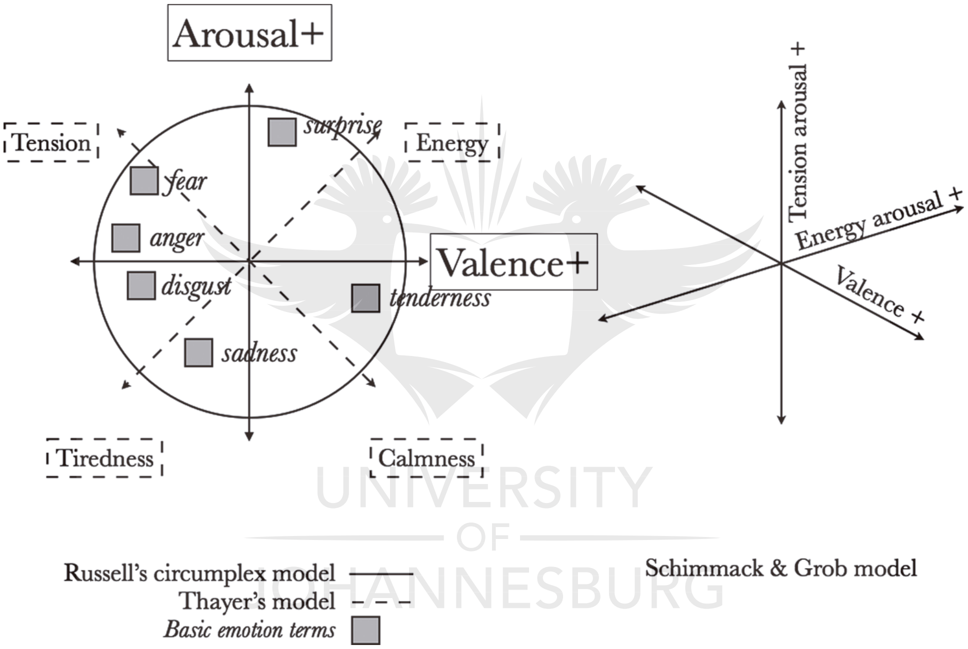

Figure 2.1.2: The figure shows the schematic diagram of the dimensional models of emotions with common basic emotion categories overlaid.

The introduction of another dimension "stance", which defines attention-rejection, into the 2-D model (fig:2.1.3 [102]) resulted in a 3-D representation of the emotions [102, 103]. One interesting point raised by Schimmack and Grob [104] is that in the 3-dimensional model, the axes are not necessarily orthogonal to each other (fig:2.1.2) in actual affect data.

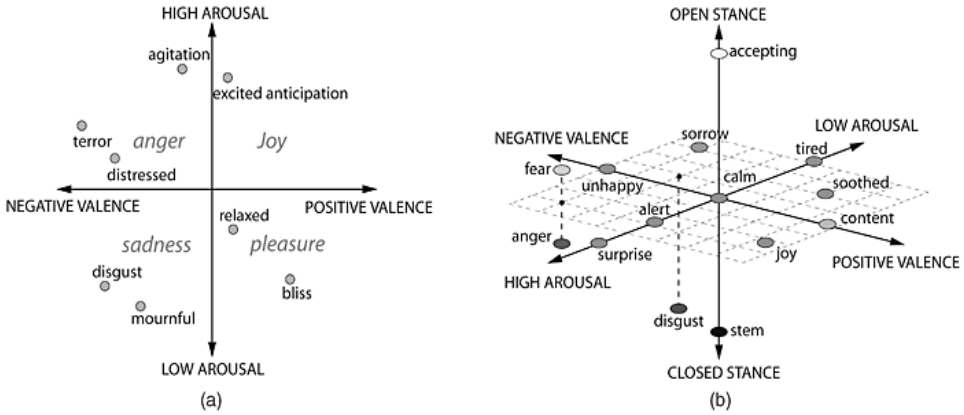

Figure 2.1.3: The section (a) shows the valence-arousal 2-dimensional model, and the 2-dimensional is extended to 3-dimensional section (b).

This psychological aspect of emotion is an essential step towards dataset creation or selection for SER research. The synthetic data creation process is required to decide on the model to follow for labeling purposes. Two-dimensional emotional models are extensively used in most of the synthetic SER datasets, and mostly the underlying emotions are re-created (enacted). The synthetic speech datasets used for this work also use the 2-dimensional model where the basic emotions are enacted.

2.2 Speech Databases

Researchers are trying to solve SER as a machine learning (ML) problem, and ML approaches are data-driven. That is why the SER research depends heavily on emotional speech databases [105, 106] because there is no mechanism till date, which can label a natural speech recording with proper emotion tag. Thus, any random speech cannot be used directly for research. Moreover, the database naturalness, quality of recordings, number and type of emotions considered, and speech collection strategy are critical inputs for the classification stage because those features of the database will decide the classification methodology [107–110].

2.2.1 Dataset design

The design of the speech database has different factors [76, 108]. First of all, the existing databases can be categorized into three categories: (1) simulated by actors, (2) semi-natural, and (3) natural. The simulated databases, created by enacting emotions by actors, are usually well annotated, adequately labeled, and are of better quality since the recordings are performed in a controlled near noise-free environments. The number of recordings is also usually high for simulated databases. However, acted emotions are not natural enough, and sometimes an expression of the same emotion varies a lot depending on the actor, which makes the feature selection process very difficult. A brief tabular description of 22 popular datasets in the SER research is provided in table.2.2.1.

Semi-natural databases are also the collection of the enactions by professional or non-professional actors, but here the actors are trying to keep it as natural as possible. Natural emotional databases are difficult to label because manually labeling a big set of speech recording is a daunting task, and there is no method available yet to label the emotions automatically. As a result, the number of emotions covered in a natural dataset is low, and the number of data points is also low. Natural recordings usually depict continuous emotions, which can create hurdles during the classification phase because of the presence of overlapping emotions.

2.2.2 Problems with emotional databases

The Following are some problems faced by SER researchers with emotional databases.

- Most available corpora do not supply enough material, and there is no standardization regarding various emotion elicitation and emotion annotation methods [129]; as a result, classification performance varies drastically across databases.

- Emotional datasets designed for SER or other research and applications are mostly private, and there are price and privacy clauses involved. This scenario is a big problem for SER research since well benchmarked suitable

Table 2.2.1: Brief details of some emotional databases used in SER research sorted as per publication dates.

Corpus Name	Type	Emotions	Lang	Description
DES [111]	sim	AN,JO,SA,SU,NE	dan	4 actors × 5 emotions (2 words+ 9 sentences + 2 passages)
Noam [112]	sim	AN,DI,FE,JO,NE,SA	heb	60 Hebrew and 1 Russian actors
Pereira [113]	sim	AN,JO,NE,SA	eng	2 actors × 5 emotions ×8 utterances
INTERFACE [114]	sim	AN,DI,FE,JO,NE,SU,SA	eng,slv spa,fre	English (186 utterances), Slovenian (190 utterances), Spanish (184 utterances), French (175 utterances)
KISMET [107]	sim	AP,AT,PRO,SO,NE	eng	1002 utterances, 3 female speakers, 5 emotions
FERMUS III [115]	sem	AN,DI,JO,NE,SA,SU	ger,eng	2829 utterances, 7 emotions,13 actors
LDC2002S28 [116]	sim	NE,PA,ANX,AN,DE,SA, EL,JO,IN,BO,SH,PR,CO	eng	7 actors × 15 emotions × 10 utterances
AIBO [117]	nat	AN,BO,JO,NE,SA	ger	14 Speakers (7 Male + 7 Female) × 40 Commands
ESMBS [118]	sim	AN,JO,SA,DI,FE,SU	chi	720 utterances, 12 speakers, 6 emotions
BabyEars [110]	sim	AP,AT,PRO	eng	509 utterances, 12 actors (6 males + 6 females)
Emo-DB [119]	sim	AN,JO,SA,FE,DI,BO,NE	ger	800 utterances (10 actors, 10 utterances)
MPEG-4 [120]	mov	JO,AN,DI,FE,SA,SU,NE	eng	2440 utterances, 35 speakers
Call centers[76]	nat	AN,FR,JO,NE	eng	7200 utterances
CLDC [121]	sim	JO,AN,SU,FE,NE,SA	chi	1200 utterances, 4 actors
Natural [122]	nat	AN,NE	chi	388 utterances, 11 speakers, 2 emotions
KES [123]	sim	NE,JO,SA,AN	kor	5400 utterances, 10 actors
IEMOCAP [124]	sem	HA,AN,SA,FR,NE	eng	Audio visual and motion capture data of
VAM [125]	nat	HA,AN,SA,DI,FE,SU,NE	ger	12 hours of audio-visual recordings of the German TV
IITKGP-SESC [126]	sim	AN,COM,DI,FE, JO,NE,SAR,SU	tel	12000 utterances (15 sentences,10 artists, 10 sessions)
				the scripted dialog of 10 actors
NIMITEK [105]	sem	JO,SA,AN,FE,DI,NE	ger	15 hours of audio visual data of 10 speakers
Belfast [127]	sem	FR,FE,DI,SU,AM,AN	eng spa	Contains 3 sets of data, each containing 570, 650 and 180 video clips of 114, 82, and 60 actors respectively
				talk show “Vera am Mittag”
RAVDESS [128]	sim	CA,HA,SA,AN,FE,DI,SU,NE	eng	24speakers(12male,12female)×2sentences×8emotions ×2repeatations

Abbreviations {emotions# AM:amused,AN:anger,ANX:anxiety,AP:approval,AT:attention,BO:boredom,CA:calm,COM:compassion,CO:contempt, DE:despair,DI:disgust,EL:elation,FE:fear,FR:frustrated,IN:interest,JO:joy,NE:neutral,PA:panic,PR:pride,PRO:prohibition,SA:sadness, SAR:sarcastic,SH:shame,SO:soothing,SU:surprise}, {*types#* sim:simulated,sem:semi natural,nat:natural}, {*languages#* chi:chinease,dan:danish, eng:english,fre:french,ger:german,heb:hebrew,kor:korian,slv:slovenian,spa:spanish,tel:telugu} dataset is scarce. That is why public datasets like Emo-DB become very popular among SER researchers.

- In most simulated databases, enacted emotions are not natural enough, even human recognition rates in some databases [118] are around 65%, which is very low.
- Some datasets do not have quality recordings and do not provide critical details like phonetic transcriptions, which create further hindrance towards the usability of those databases.

2.3 Speech Features for SER

Speech signals carry an enormous amount of information apart from the intended message. Researchers agree that speech signals also carry vital information regarding the emotional state of the speaker [27]. However, researchers are still undecided over the right set of features of the speech signals, which can represent the underlying emotional state. This section contains the details of feature sets which are heavily used so far in SER research and performed well in the classification stage. There are three prominent categories in speech features used in SER : (1) the prosodic features, (2) the spectral or vocal tract features, and (3) the excitation source features. The following sub-sections will discuss these features in detail.

2.3.1 Prosody Speech Features

The human speech production system is a very sophisticated apparatus. Humans, while speaking can utilize different tools available in this system for varying the duration, pitch, and intensity of the spoken utterances, called prosody alteration, to express their various feelings in words. Prosody features are the characteristics of the speech sound generated by the human speech production system, for example, pitch or fundamental frequency (F_0) and energy. Researchers used different derivatives of pitch and energy as various prosody features [130–132]. These are also called continuous features and can be grouped into the following categories

[35, 76, 93, 94]: (1) pitch-related features; (2) formants features; (3) energy-related features; (4) timing features; and (5) articulation features. Several studies tried to establish the relationship between prosody speech features and the underlying patterns of different emotions [33–37, 94, 133, 134].

Most of the early studies of SER considered the fundamental frequency (F_0) as the most prominent attribute which represents different emotions [86, 87, 135–137]. After that, other important features for SER like *energy, speech duration, formants* are also introduced by researchers along with F_0 and their derivatives [61, 122, 135, 138, 138–140]. Several studies tried to establish the relationship between prosody speech features and the underlying patterns of different emotions [33–37, 133].

2.3.2 Excitation Source Features

The features used to represent glottal activity, mainly the vibration of glottal folds, are known as the source or excitation source features. These are also called voice quality features because glottal folds determine the characteristics of the voice. Some researchers believe that the emotional content of an utterance is strongly related to voice quality [38, 94, 135]. Speakers have their unique voice quality signature, and different voice qualities can convey relevant information like intentions, attitudes, and emotions.

Human vocal folds vibrate to generate quasi-periodic impulse-like excitation in the vocal tract system during speech production. Glottal vibrations or excitation source signal can be extracted by using inverse filtering (IF) technique on the speech signal to remove the vocal tract contribution [141]. The signal received after inverse filtering speech signal is also called *linear prediction* (LP) residuals, which contain only higher order relations. The relations present among the distant speech samples are treated as higher-order relations, whereas the adjacent relations are treated as lower-order relations.

Cowie et al. [94] grouped acoustic correlates, related to voice quality, are grouped into the following categories. 1. voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level; 2. voice pitch; 3. phrase, phoneme, word and feature boundaries; 4. temporal structures.

Voice quality measures for a speech signal includes harshness, breathiness, and tenseness. The relation of voice quality features with different emotions is not a well-explored area, and researchers have produced contradictory conclusions. For example, Scherer [38] associated anger with tense voice whereas Murray and Arnott [35] associated anger with a breathy voice. Many SER researchers [56, 142, 143] extracted features from the glottal waveform for emotion classifications. However, deriving the accurate transfer function by canceling out the effect of the vocal tract system, and obtaining the closed phase duration of the glottal cycle [144, 145] is a challenge.

2.3.3 Spectral Features

Spectral features are the characteristics of various sound components generated from different cavities of the vocal tract system. They are also called segmental or system features. Spectral features extracted in the form of 1. ordinary linear predictor coefficients (LPC) [39], 2. one-sided autocorrelation linear predictor coefficients (OSALPC) [146], 3. shorttime coherence method (SMC) [147], and 4. least-squares modified Yule–Walker equations (LSMYWE) [42].

However, the extracted spectrum is often needed to pass through a bank bandpass filters [93]. The filters' bandwidths are usually evenly distributed with respect to a suitable nonlinear frequency scale such as the Bark scale [40], the Mel-frequency scale [40, 148], the modified Mel-frequency scale, and the ExpoLog scale [42] because a human being does not perceive pitch in a linear scale.

Researchers claim that the sequence of shapes of the vocal tract system also carries emotion-specific information, along with the information related to the sound unit [29]. The spectrum characterized by formant frequencies and their respective bandwidths is extensively analyzed for emotional speech [36, 87, 149]. It is inferred that the first formant(F_1) for angry speech has a higher mean than the neutral speech [87]. Researchers [87, 150, 151] also observed association among changes in the spectral component and glottal source excitation; for example, higher F_0 in angry speech tend to have smaller F_1 amplitudes. Some studies [58, 152, 153] have shown that properties of formants like magnitude and shift vary across vowels for different emotional states.

There is a particular type of spectral features called the *cepstral* features which are extensively used by SER researchers. Cepstral features can be derived from the corresponding linear features like linear predictor cepstral coefficients (LPCC) is derived from LP. Mel-frequency cepstral coefficients are one such cepstral feature which along with its various derivatives is widely used in SER research [43, 72, 154–156].

2.3.4 Deep Feature Learning Methods

The advent of deep learning has proven to be a paradigm shift towards looking at feature extraction stage in the machine learning process. The ability of DL methods to learn underlying representations from data has already proven to be very robust to variability in data such as speech signals [157, 158]. One such feature extraction technique Generalized Discriminant Analysis (GerDA) is proposed by Stuhlsatz et al. [159] to learn discriminative features of low dimension. Han et al. [160] used DNN to extract high-level features from raw data. Researchers [71, 161] also employ a 1-layer CNN trained with a Sparse Auto-encoder (SAE) to extract affective features for speech emotion recognition.

End-to-end deep learning systems are becoming very popular among SER researchers, where raw speech is fed into a deep neural network model. These end-to-end SER models usually combine CNN with RNN where the CNN layer is responsible for feature learning. For example, some researchers [69, 162] proposed end-to-end models where they stacked CNN layers before Long Short-Term Memory (LSTM) layers. However, researchers suggested that shallow 1-layer or 2-layer CNN structures may not be able to learn effectively the affective features which are discriminative enough to distinguish the subjective emotions [70]. So, a deep structure is recommended.

Drawbacks of Deep Feature Learning

Feature set learned using DL methods usually needs a very high number of attributes to be provided as input. That means global speech features such as energy, *F*0, needs to be further broken down into different derivatives. Moreover, feature sets learned through DL methods usually becomes very high in dimension, and it runs into thousands sometimes [160, 163].

2.4 Classification Techniques

Speech Emotion Recognition (SER) deals with speech signals. The analog (continuous) speech signal is sampled at a specified time interval to get the discrete time speech signal. A discrete time signal can be represented as follows:

$$\begin{gathered} C = \{c_l\}_{l \in \mathbb{L}}, \quad \textit{where } \{c_l\} = \{c_1, c_2, \dots, c_l\}, \\ \{c_l\} \in \mathbb{R} \end{gathered} \tag{2.1}$$

where L is the total number of sample points in the speech signal. First, only the speech utterance section is extracted from the speech sound by using a speech endpoint detection algorithm. In this case, an algorithm proposed by Roy et al. [82] is used.

This speech signal contains various information that can be retrieved for further processing. Emotional states guide human thoughts, and those thoughts are expressed in different forms [2] such as speech. The primary objective of an SER is to find the patterns in speech signals which can describe the underlying emotions. The pattern recognition task is carried out by different machine learning (ML) algorithms. Features are extracted from the speech signal C in two forms (1) local features by splitting the signals into smaller frames and computing statistics of each frame; and (2) global features by calculating statistics on the whole utterance.

Let, there be N number of sample points after the feature extraction process. If local features are computed from C by assuming 10 splits, then there will be 10 data points from C. Now, suppose there is a total of 100 recorded utterances, and each utterance is split into 10 frames, then will be total $(100 \times 10) = 1000$ data points available. When global features are computed, then each utterance will produce one data point. The selection of local or global feature depends on the feature extraction strategy. Now, suppose, N is the number of data points such that $n = 1, 2, 3, .., N$, where n is the index. If D number of features are extracted from C then each data point is a D dimensional feature vector. Each utterance in the speech database is labeled properly, so that it can be used for supervised classification. So, the data set is denoted as

$$X = \{x_n, y_n\}_{n=1}^{N} \tag{2.2}$$

where y_n is the label corresponding to a data point x_n and $X \in \mathbb{R}^{N \times D}$. Once the data is available the next step is to find a *predictive function* called *predictor*. More specifically the task of finding a function f is called learning so that $f : X \rightarrow Y$. Different classification models take different approaches to learning.

Researchers used different types of classifiers for SER, but in most of the situations, a proper justification is not provided for choosing a particular classification model. Two apparent explanations are that classifiers which are successful in ASR are assumed to be working well in SER (like HMM), and secondly, those classifiers which perform well in most classification problems are chosen (like SVM, and GMM) [29]. There are two broad categories of classifiers (figure: 2.4.1) used in SER: the *linear classifiers* and the *non-linear classifiers*. Table.2.4.1 shows a list of classifiers commonly used in SER along with literature references.

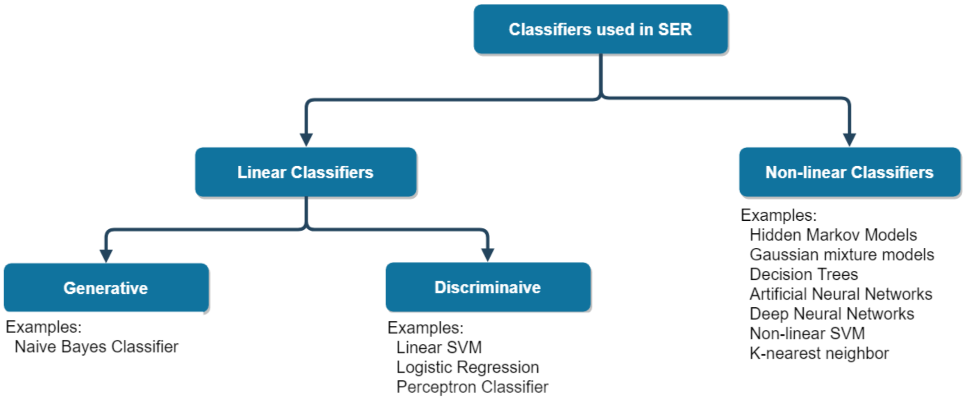

Figure 2.4.1: Categories of classifiers used in SER along with some examples.

Although in the table.2.4.1, there are eight classifiers listed, but not all of them become prominent for SER tasks. In the following subsections, four most prominent classifiers (HMM, GMM, SVM, and DNN) for SER are discussed to depict the SER specific implementation technique.

Table 2.4.1: List of literatures on SER grouped by prominent classifiers

No.	Classifiers	References
1.	Hidden Markov Model	[44–49, 58, 67, 118, 164]
2.	Gaussian Mixture Model	[50–55, 63, 107, 110, 156, 165, 166]
3.	k-Nearest Neighbor	[61–64, 167, 168]
4.	Support Vector Machine	[47, 52, 56–60, 169–174]
5.	Artificial Neural Network	[28, 52, 63, 65–67, 117, 175]
6.	Bayes classifier	[61, 67, 165, 176]
7.	Linear discriminant analysis	[64, 76, 177–179]
8.	Deep Neural Network	[43, 68–72, 154, 161–163, 180–184]

2.4.1 Hidden Markov Model

HMMs are suitable for the sequence classification problems that consists of a process that unfolds in time. That is why HMM is very successful in ASR systems where the sequence of the spoken utterances is a time-dependent process. HMM parameters are tuned in the model training phase to best explain the training data for the known category. The model classifies an unseen pattern based on the highest posterior probability.

HMM comprises two processes. The first process consists of a first-order Markov chain whose states capture the temporal structure of the data, but these states are not observable that is hidden. The *transition model*, which is a stochastic model, drives the state transition process. Each hidden state has an observation associated with it. The *observation model*, which is again a stochastic model, decides that in a given hidden state the probability of occurrence of different observations [185–187].

Figure 2.4.2 shows a generic HMM where S_i and O_i are the states and observations respectively, $i = 0, 1, \ldots T - 1$. A is the *transition model* and B is the *observation model*. Assuming the length of the observation sequence to be T so that $O = O_0, O_1, \ldots, O_{T-1}$ is an observation sequence and N is the number of hidden states. The state *transition* probability matrix is denoted by A, whereas the *observation* probability matrix is denoted by B. Also, let π_0 be the initial state probability for the hidden Markov chain.

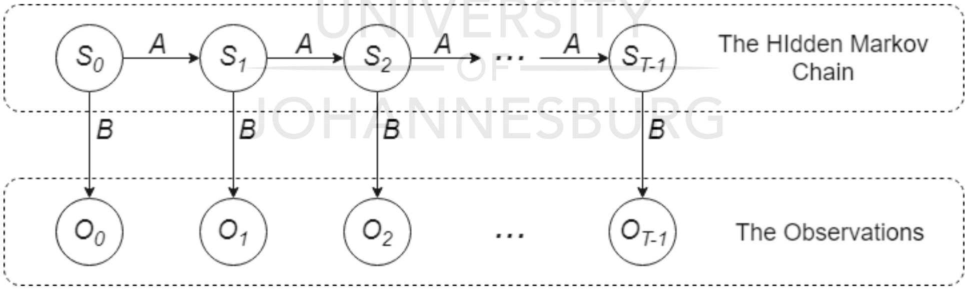

Figure 2.4.2: Schematic diagram of an HMM

In the training phase, the model parameters are determined. Here, the model is denoted by λ, which contains three parameters A, B, and π thus $\lambda = (A, B, \pi)$. The parameters are usually determined using the expectation maximization (EM) algorithm [188] so that the probability of the observation sequence Q is maximum.

Now, since the model λ is determined, the probability of an unseen sequence O_u that is $p(O_u|\lambda)$ can be found to get the sequence classification results.

SER researchers used HMM for a long time and used it with various types of feature sets. For example, some researchers [44, 45, 67, 118] used prosody features, and some others [45, 46, 118] used the spectral features. Researchers using the HMM achieves the average SER classification accuracy is between 75.5%-78.5% [47–49, 58, 118, 164], which is comparable with other classification techniques, but further improvement possibilities are low. Moreover, that is why HMM has been replaced by other classification techniques in later studies like SVM, GMM, or DNN.

Difficulties in using HMM for SER

- HMM may follow two types of topology: fully connected or left-to-right. Most ASR systems use the left-to-right topology [164], but this topology will not work for SER because a particular token can occur at any stage of the utterance. So, in the case of SER fully connected topology is more suitable [118]. However, the problem domain of SER is different from ASR, and the sequence in the utterance is the most crucial attribute towards successful ASR, but for SER sequence is not that essential. The whole utterance represents the emotional state and not the sequence of words or silence.

- the optimal number of states required for SER is hard to decide because there is no fixed rule of splitting the speech signal into smaller frames.

- The observation type could be discrete or continuous [185], but for SER, it is hard to decide on whether to consider it discrete or continuous.

- In ASR every spoken word is broken into smaller phonemes which are very neatly handled by the HMM. However, for SER, at least a word should start making some sense, and even a set of words should be reasonable, which is a very different scenario than ASR. So, applying HMM for SER poses another challenge.

2.4.2 Gaussian Mixture Model

An unknown distribution $p(x)$ can be described by a convex combination of K base distributions like Gaussians, Bernoulli's or Gammas, using mixture models. Gaussian Mixture Model (GMM) is the special case of mixture models where the base distribution is assumed to be Gaussian. GMM is a probabilistic density estimation process where a finite number of K Gaussian distributions of the form $\mathcal{N}(x|\mu_k, \Sigma_k)$ is combined , where x is a D-dimensional vector, i.e. $x \in \mathbb{R}^D$, μ_k is the corresponding *mean* vector and Σ_k is the covariance matrix, such that [189, 190]

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2.3}$$

where π_k are the mixture of weights, such that $0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$. And θ denotes the collection of parameters of the model $\theta := \{\mu_k, \Sigma_k, \pi_k : k = 1, 2, \dots, K\}$.

Now, consider the dataset $X = x_1, \dots, x_N$ and it is assumed that x_n, $n = 1, \dots, N$, are independent and identically distributed (i.i.d.) and drawn from an unknown distribution $p(x)$. The objective here is to find a good approximation of $p(x)$ by means of a Gaussian mixture model with K mixture components and for that the maximum likelihood estimate (MLE) of the parameters θ need to be obtained [190, 191]. The i.i.d. assumption allow the $p(X \mid \theta)$ to be written [190] as follows:

$$p(X \mid \theta) = \prod_{n=1}^{N} p(x_n \mid \theta) \tag{2.4}$$

where, the individual likelihood term $p(x_n \mid \theta)$ is a Gaussian mixture density as in eq.2.3. Then, it is required to get the log-likelihood [189, 190]

$$\log p(X \mid \theta) = \sum_{n=1}^{N} \log p(x_n \mid \theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \tag{2.5}$$

So, the MLE of the model parameters θ that maximize the log-likelihood defined in eq.2.5 need to be obtained. The maximum likelihood of the parameters μ_k, Σ_k, and π_k is estimated using the Expectation Maximization (EM) algorithm, which is a general iterative scheme for learning parameters in mixture models.

GMM is one of the most popular classification technique among SER researchers, and many research works are based on GMM [50–55, 63, 107, 110, 156, 165]. Although average accuracy achieved is not up to the mark, around an average 74.83%-81.94%, but least training time of GMM among the prominent classifiers made it an attractive choice as SER classifier.

GMMs are efficient in modeling multi-modal distributions [191] with much less number of data points compared to HMMs. So, when global features are extracted from speech for SER, less number of data points are available but, GMM works better in those scenarios [93]. Moreover, the average training time is minimal for GMM [93].

Difficulties in using GMM for SER

GMMs cannot model temporal structure since x_n are drawn i.i.d. GMM was integrated with the vector auto-regressive process to capture the temporal structure in SER [192]. Deciding the optimal number of Gaussian component is a difficult problem [193].

2.4.3 Support Vector Machine

SVM is fundamentally a two-class or binary classifier. The SVM provides state-of-the-art results in many applications [194]. Possible values for the label or output are usually assumed to be $\{+1, -1\}$ so that the predictor becomes $f : \mathbb{R}^D \rightarrow \{+1, -1\}$ where f is the predictor and D is the dimension of the feature vector. So, given the training data set of N datapoints $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_n \in \mathbb{R}^D$ and $y_n \in \{+1, -1\}$ and $n = 1, 2, \ldots, N$, the problem is to find the f with least classification error. Consider a linear model of the form

$$f(x, w) = w^T x + b \tag{2.6}$$

to solve this binary classification problem, where $w \in \mathbb{R}^D$ is the weight vector, and $b \in \mathbb{R}$ is the bias. Also, assume that the dataset is linearly separable in the feature space and the objective here is to find the separating hyperplane that maximizes the margin between the positive and negative examples, which means $w^T x_n + b \geq 0$ when $y_n = +1$ and $w^T x_n + b < 0$ when $y_n = -1$. Now, the requirement that the positive and the negative examples nearest to the hyperplane to be at least 1 unit away from the hyperplane yields the condition $y_n(w^T x_n + b) \geq 1$ [190]. This condition is known as the canonical representation of the decision hyperplane. Here, the optimization problem is to maximize the distance to the margin, defined in terms of w as $\|w\|^{-1}$, which is equivalent to minimizing $\|w\|^2$, that is [190]

$$\begin{aligned} \underset{w,b}{\arg\min} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_n(w^T x_n + b) \geq 1, \forall \ n = 1, \ldots, N \end{aligned} \tag{2.7}$$

Eq.2.7 is known as the *hard* margin which is an example of a quadratic programming. The margin is called *hard* because the formulation does not allow any violation of margin condition.

The assumption of linearly separable data set needs to be relaxed for better generalization of the data because, in practice, the class conditional distributions may overlap. This is achieved by the introduction of a *slack variable* $\xi_n \geq 0$ where $n = 1, \ldots, N$, with each training data point [195, 196]. Which updates the optimization problem as follows [190]

$$\begin{aligned} \underset{w,b,\xi}{\arg\min} \quad & \frac{1}{2}\|w\|^2 + C\sum_{n==1}^{N}\xi_n \\ \text{subject to} \quad & y_n(w^T x_n + b) \geq 1 - \xi_n \\ & \xi_n \geq 0 \\ & \forall\ n = 1, \ldots, N \end{aligned} \tag{2.8}$$

where $C > 0$ trades off the size of the margin and the total amount of slack that we have. This model allows some data points to be on the wrong side of the hyperplane to reduce the impact of overfitting.

Various methods have been proposed to combine multiple two-class SVMs to build a multiclass classifier. One of the commonly used approaches is *one-versus-the-rest* approach [197] where K separate SVMs are constructed where K is the number of classes and $K > 2$. The k^{th} model $y_k(X)$ is trained using the data from class C_k as the positive examples and the data from the remaining $K - 1$ classes as negative examples. There is another approach called *one-versus-one* where all possible pairs of classes are trained in $K(K-1)/2$ different 2-class SVM classifiers. Platt [198] proposed Directed Acyclic Graph SVM (DAGSVM).

Support Vector Machine (SVM) is extensively used in SER [47, 52, 56–60, 169–171]. Performance of SVM for SER task in most of the researches carried out yielded nearly close results, and accuracy is varying around 80% mark. However, Hassan and Damper [60] achieved 92.3% and 94.6% classification accuracy using linear and hierarchical kernels, respectively. They have used a linear kernel instead of non-linear RBF kernel because of very high dimensional features space [172]. Hu et al. [173] explored GMM supervector based SVM with different kernels like *linear, RBF, polynomial* and *GMM KL divergence* and found that GMM KL performed the best in classifying emotions.

Difficulties with SVM

There is no systematic way to choose the kernel functions, and hence separability of transformed features is not guaranteed. Moreover, in SER perfect separation in training data is not recommended to avoid over-fitting.

2.4.4 Deep Learning

Deep feedforward networks, also called feedforward neural networks, or multilayer perceptrons (MLPs), are the pure form of deep learning models. The objective of an MLP is to approximate some function f^* such that a classifier, $y = f^*(x)$ maps an input x to a category y. An MLP defines a mapping $y = f(x; \theta)$ and learns the value of the parameters θ that results in the best function approximation. Deep networks are represented as a composition of different functions, and a directed acyclic graph describes how those functions are composed together. For example, there might be three functions $f(1)$, $f(2)$, and $f(3)$ connected in a chain, to form $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ where $f^{(1)}$ is the first layer of the network, $f^{(2)}$ is the second layer, and so on. The length of the chain gives the *depth* of the model, and this *depth* is behind the name of *deep learning*. The last layer of the network is *output layer*.

The training phase of neural network $f(x)$ is altered to approximate $f^*(x)$. Each training example x has a corresponding label $y \approx f^*(x)$ and training decides the values for θ such that the output layer can produce values close to y say $\hat{y}$. However, the behavior of the hidden layers are not directly specified by training data, and that is why those layers are called *hidden*. The hidden layers bring the nonlinearity into the system by transforming the input $\varphi(x)$, where φ is a non-linear transform. The whole transformation process is done in the hidden layers, which provide a new representation of x. So, now it is required to learn φ, and the model now becomes $y = F(x; \theta, w) = \varphi(x; \theta)^T w$, where θ is used to learn φ and parameters w maps $\varphi(x)$ to the desired output. The φ is the so-called *activation function* of the hidden layers of the feedforward network [199].

Most modern neural networks are trained using the maximum likelihood estimate, which means the cost function $J(\theta)$ is the negative log-likelihood or the cross-entropy function between the training data and the model distribution. So, the cost function becomes [199]

$$J(\theta) = -\mathbb{E}_{x,y\sim\hat{p}_{data}} \log p_{model}(y \mid x) \tag{2.9}$$

where p_{model} is the model distribution, which varies depending on the selected model, and $\hat{p}_{data}$ is the target distribution from data. The output distribution determines the choice of the output unit. For example, Gaussian output distribution requires a linear output unit, Bernoulli output distribution requires a *Sigmoid* function, *Softmax* Units for Multinoulli Output Distributions, and so on. However, the choice of hidden unit is still an active research area but rectified linear units (ReLU) are the most versatile ones which work well in most of the scenarios. Logistic Sigmoid and Hyperbolic Tangent are other two options out of many other functions researchers are using.

So, in the forward propagation $\hat{y}$ is produced, and the cost function $J(\theta)$ is computed. Now, the information generated in the form of $J(\theta)$ is appropriately processed so that w parameters can be appropriately chosen. This task is accomplished in two phases, first computing the gradients using the famous back-propagation algorithm, and in the second phase, the w values are updated based on the gradients computed by the backprop algorithm. The w values are updated through methods like stochastic gradient descent (SGD). The backprop algorithm applies the chain rule recursively to compute the derivatives of the cost function $J(\theta)$.

Different variants of deep learning exist now, but Convolutional Neural Networks (CNNs) [200, 201] and Recurrent Neural Networks (RNNs) [202] are the most successful ones. Convolutional networks are neural networks that use convolution in place of general matrix multiplication in at least one of their layers. Whereas, when feedforward neural networks are extended to include feedback connections, they are called recurrent neural networks. RNNs are specialized in processing sequential data.

SER researchers have used CNNs ([43, 68–72, 161]), RNNs [69, 154, 163], or combination of the two extensively for SER. Shallow 1-layer or 2-layer CNN structures may not be able to learn effectively the affective features which are discriminative enough to distinguish the subjective emotions [70]. So, researchers are recommending a deep structure. Researchers [72, 154, 180, 181] have studied the effectiveness of attention mechanism.

Researchers are applying end-to-end deep learning systems in SER [69, 162, 182–184], and most of them use arousal-valence model of emotions. Although using end-to-end deep learning the average classification accuracy for arousal is 78.16%, which is decent, for valence it is pretty low 43.06%. Among other DNN techniques, very recently maximum accuracy of 87.32% is achieved by using a fine-tuned Alex-Net on Emo-DB [70]. Han et al. [160] used Extreme Learning Machine (ELM) for classification where a DNN takes as input the popular acoustic features within a speech segment and produces segment-level emotion state probability distributions, from which utterance-level features are constructed.

Drawbacks of Deep Learning

1. Implementation of Tensor operations is a complicated task, which results in limited resources [199]. There are limited set of libraries like tensorflow [203], theano [204], pytorch [205], mx-net [206], and cntk [207] which provide the service.

2. Back-propagation often involves summation of many tensors together, which makes the memory management task difficult and often requires huge computational resources.

3. Introduction of DL methods also increased the feature set dimension for SER manifolds, for example, Wöllmer et al. [163] extracted total 4843 features.

4. One crucial question is raised by Lorenzo-Trueba et al. [208] is how emotional information should be represented as labels for supervised DNN training, e.g., should emotional class and emotional strength be factorized into

separate inputs or not?

2.5 Difficulties in SER Studies

SER mystery is not yet solved, and it has proved to be difficult. Here are the prominent difficulties faced by the researchers.

- The topic called *emotion* is inherently uncertain. Because the very experience of emotion is very subjective, its expression varies largely from person to person. Moreover, there is little consensus over the definition of emotion. These are the fundamental hurdle to proceed with the research [209]. For example, several studies [49, 50, 138, 165, 210–212] reported that there is confusion between anger and happiness in emotional expression.

- SER is challenging because of the affective gap between subjective emotions and low-level features [70]. Also, the feature analysis in SER is less studied [160], and researchers are still actively looking for the best feature set.

- Speaker and language dependency of classification results are a concern for building more generic SER systems [213]. The same model gives very different classification results with different datasets. Studies reported the speaker dependency phenomenon and tried to address that issue [138, 165, 178, 214] .

- Standard speech databases are not available for SER research so that new models can be effectively benchmarked. Moreover, the absence of good quality natural speech emotional databases is hindering the real-life implementation of SER systems.

- Cross-corpora recognition results are low [56, 155, 215]. This indicates that existing models are not generalizing enough for real-life implementation.

- Classification between high-arousal and low-arousal emotions are achieved more accurately, but for other cases, it is low, which needs to be addressed.

Moreover, the accuracy of n-way classification with all the emotions in the database is still very low.

2.6 Summary and Conclusion

This chapter reviewed different phases of SER. The primary focus is on four prominent classification techniques used in SER to date. HMM was the first technique which has seen some success and then GMM and SVM propelled that progress forward. Also, presently DL techniques, mainly CNN-LSTM combination, is providing state of the art classification performance. However, things have not changed much in case of selecting a feature set for SER because the low-level descriptors (LLDs) are still one of the prominent choices, although some researchers in recent times are trying DL techniques for feature learning. The nature of SER databases is changing, and features like facial expressions and body movements are being included along with the spoken utterances. However, the availability of quality databases is still a challenge.

This chapter is a survey of the advancements that happened so far in the SER field. It is observed that the SER field is still facing many challenges, which are barring research outputs from being implemented as an industry-grade product. It is also noticed during this study that research work related to feature set enhancement is much less compared to the works done on enhancing classification techniques. However, the available classification techniques in Machine Learning (ML) field are in a very advanced state, and with the right feature set, they can yield very high classification accuracy rates. Deep learning as a sub-field of ML even achieved state-of-the-art classification accuracy in many fields like computer vision, text mining, automatic speech recognition, to name a few. So, the classification technique should not be a hindrance for SER anymore; only the appropriate feature set needs to be fed into the classification system. Efforts have been made in this work, towards finding the right feature set for SER and then applying the appropriate classification method to classify the emotions.

If you want to find the secrets of the universe, think in terms of energy, frequency and vibration.

3

Precise Detection of Speech Endpoints Dynamically: A Wavelet Convolution based approach

3.1 Introduction

Speech endpoints are the beginning and end points of the actual speech utterance within the speech signal. Speech Recognition and its related field of research has come a long way and has matured enough. However, the precise detection of speech endpoints is still an important factor affecting the recognition performance of Automatic Speech Recognition (ASR) systems. Lamel and Rabinar [216] explained the importance of accurate endpoint detection in speech recognition and has shown that speech recognition performance dramatically reduces due to an error in endpoint detection. Researchers like Li et al. [217] and Junqua et al. [218] also reflect the similar view. Background noise and other sound artifacts which are not the part of the actual speech utterance exists in the speech recordings. When a recording with noise is used for analysis, the presence of those noise distorts the results. Also, the silent sections before and after the actual utterance are not required in the analysis for most of the cases, thus the requirement for precise extraction of the speech utterance by separating it from those noises and silence sections.

Digitally recorded speech can be acquired from different sources such as telephone recordings, studio recordings, and conversations recorded in the natural environment. All these recordings contain various noise depending on the recording environment. Even the recordings in nearly noise-free environment contain sound artifacts produced by the speaker during the recording. Examples of such sound artifacts are mouth clicks and pops, heavy breathing, and lip smacking. In this chapter, these sound artifacts are referred to as Non-speech Sound Artifacts (NSAs). These NSAs need to be filtered out in most of the speech based applications for estimating good results because their effect is similar to noise in systems like ASR.

Though the quest to find a solution for End-Point Detection (EPD) problem started a long time ago in the 1970s, the search is still on because the correct solution is still not found which can cater for all the challenging scenarios. Figure 3.1.1 shows examples of the NSAs present in speech recordings such as breathing noise, mouth clicks, and pops.

Existing EPD methods frequently use Short-Term Energy (STE) and Zero-Crossing Rate (ZCR) based methods and their variants. Rabiner and Sambur [219] proposed a simple and fast algorithm to determine endpoints based on energy and ZCR. Savoji [220] also used STE and ZCR as features, and their proposed algorithm uses the knowledge-based heuristics for speech classification. Lamere et al. [221] utilized the STE based approach with three energy thresholds, two for beginning and one for ending. Energy and ZCR based algorithms work well when there is no background noise, and no NSA type noise exists in the sound recordings. Constant background noises present in speech utterances can be filtered out using a suitable noise reduction algorithm for sound. However, segregating the NSAs, present in the speech recordings, is a challenging task because STE and ZCR based attributes are not enough to segregate speech from NSAs. It is observed that the presence of NSAs nullifies the distinction in values for STE and ZCR for speech and non-speech sections. Figure 3.1.2 shows how STE and ZCR plots look like in the presence of heavy breathing noise. From the plot, it is clear that there is not much visible distinction between the values of STE and ZCR in the speech segment and noise segment. Also, Lamel and Rabinar [216] have shown that energy based explicit approaches for EPD failed in the presence of NSAs. While using a heuristic approach, they have classified the EPD problem into implicit, explicit, and hybrid with respect to the speech recognition system. In an explicit approach, EPD task is an independent module in the speech recognizer, whereas, in an implicit approach, there is no separate stage in the recognizer for EPD. The Hybrid approach has an EPD module at the initial phase, but after recognition, the initial EPD results of EPD are updated. So, when the NSA type noises are present in speech utterances, STE and ZCR based approaches are not suitable for solving the EPD problem.

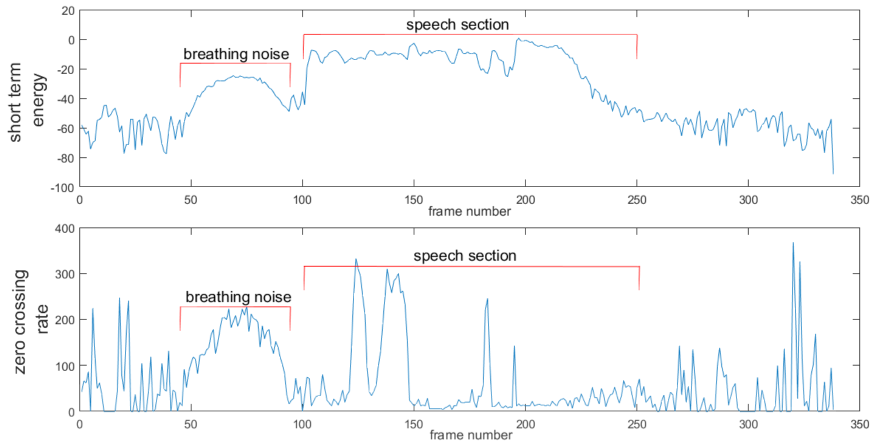

Figure 3.1.2: This figure shows how STE and ZCR plots look like in the presence of heavy breathing noise.

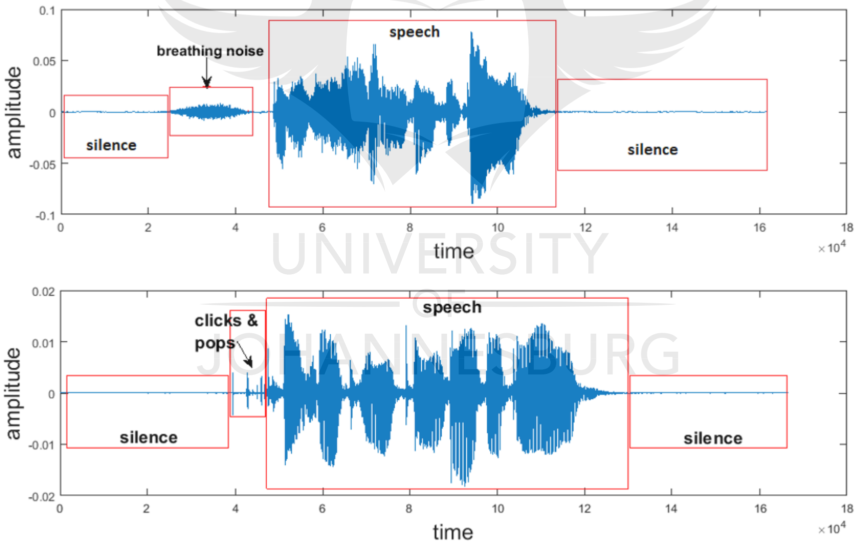

Figure 3.1.1: A speech signals containing breathing noise and mouth clicks and pops along with leading and trailing silence section.

Researchers have applied pattern recognition (PR) and machine learning (ML) techniques to solve the EPD problem. Classification techniques such as Support Vector Machine (SVM), Hidden Markov Model (HMM), Neural Network, and other suitable techniques for sequence classification are extensively used in different algorithms. Atal and Rabinar [222] considered pattern recognition approach using Energy of the signal, ZCR, Auto Correlation coefficient, First predictor coefficient, and Energy of the prediction error as a feature set. They also mentioned the limitations of using PR techniques. First of all, the algorithm needs to be trained for particular recording conditions. Second, manually locating voiced, unvoiced, and silence for preparing training data is a tedious and time-consuming process. Hidden Markov Model (HMM) classification technique is applied by Wilpon and Rabiner [223] and has shown that the HMM-based EPD approach performs significantly better in the noisy environment compared to the energy-based approach. Qi and Hunt [224] used the multilayer feed-forward network with hybrid features to classify voiced, unvoiced, and silence from the speech and achieved 96% classification rate. Kun and Wang [225] applied SVM for speech segregation in computational auditory scene analysis (CASA) problem domain and considered pitch and amplitude modulation spectrum (AMS) based features. Some researchers [218, 226] tried to solve a problem similar to the EPD, which is the problem of word boundary detection. Junqua et al. [218] used time frequency (TF) parameters along with adaptive threshold while Wu and Lin [226] used adaptive TF (ATF) features with Self-Constructing Neural Fuzzy Inference Network (SONFIN) as a classification technique. However, the presumption to work for classification techniques require properly labeled data for training, and the task of labeling data is a manual or off-line process. Since the manual intervention is required in the classification approach for endpoint detection, it will be challenging to automate the whole EPD process. Lamel and Rabinar [216] also pointed out that pattern classification approaches should not be readily applied in the EPD owing to strong overlapping between the NSAs and the speech sounds. So, these are the reasons to look for techniques other than classification.

Some researchers also proposed methods other than those based on Energy and ML for EPD and problems similar to that. Zhu and Chen [227] utilized the distance between autocorrelated functions and threshold as the feature set to find the endpoints. They have assumed that there exist some leading and trailing frames in the speech recording, which can be considered as silence section. However, this assumption might not hold for all speech databases or in real-world scenarios, and the proposed algorithm relaxed these assumptions to a great extent. Li et al. [217] proposed an optimal filter along with a three-state transition diagram for endpoint detection. Ghanbari and Karami-Mollaei [228] used adaptive wavelet packet threshold on noisy speech to enhance the speech signal for better voice activity detection. Atanas [229] has shown that mean-delta feature for trajectory-based endpoint detection of telephone speeches performs better than the energy based features. Bhowmick and Chandra [230] used wavelet decomposition for speech enhancement and designed an improved voice activity detector.

In this chapter, a new algorithm is proposed as an independent module and is named as the WCSED (Wavelet Convolution based Speech Endpoint Detection) which is developed based on the wavelet transform. It is worth mentioning here that wavelets provide a powerful and remarkably flexible set of tools for handling fundamental problems in science and engineering [231]. Researchers from various fields have applied wavelet technique effectively [232–235]. The WCSED algorithm is a deviation from the energy and ZCR based approaches. It is formulated by utilizing the simple fact that NSAs are high-frequency sound, and use the concepts of wavelet convolution and entropy as a building block. First, the input speech signal is decomposed into high-frequency (HF) and low-frequency (LF) components using the wavelet convolution method. It is observed (Fig 3.3.2) that the NSAs are much more prominent in the HF components than in the LF components.

Also, the voiced sections of a speech utterance are low-frequency sounds, whereas unvoiced sections are high-frequency sounds. Thus it can be stated that the HF components represent both the unvoiced speech and the NSAs, and the LF components represent the voiced speech. Two sets of thresholds are computed based on the entropy values for both the HF and LF components. The speech signal is broken down into manageable frames to calculate the entropy of the decomposed components. The LF thresholds extract the voiced speech segment, whereas the HF thresholds are used to segregate the unvoiced speech segments from the NSAs.

Results show that WCSED precisely extracts speech segments in the presence of NSAs. Moreover, the proposed algorithm works with unlabeled data as there is no training involved, which contributes to the easy automation of the EPD process by the proposed algorithm. Also, in WCSED, threshold computation does not assume that there exists a fixed number of leading and trailing frames, which further improves the flexibility of the algorithm as far as the use of dataset is concerned.

This chapter is organized into the following sections. Section 3.2 describes the problem in hand. Section 3.3 describes the proposed solution in detail and relevant concepts. Section 3.4 briefly describes the datasets used. In Section 3.5 results of the algorithm and observations are elaborated. Finally, Section 3.6 concludes this chapter and suggests possible directions which can be explored to extend or utilize this work.

3.2 The Problem

In this section, the problem of speech endpoint detection is elaborated.

3.2.1 Difficulties in endpoint detection

Continuous speech signals are recorded, digitized, and stored as discrete-time signals, which are mostly used for speech-based applications such as the ASR, Speech Emotion Recognition (SER). Source of continuous speech signal can be recorded in a natural environment or a specialized studio. A continuous-time signal $xc(t)$ is specified by an uncountable infinite number of signal values in every interval, whereas a discrete-time signal $S(n)$ consists of only one signal value in each sampling interval. Since computer systems can not handle continuous-time signals, $xc(t)$ needs to be discretized to get the discrete signal $S(n)$. Depending on the requirement of the applications, the continuous speech signals are sampled at specific intervals to get the digitized signals. For this work published data-sets are used which are recorded in controlled environments for research purpose. These data-sets are stored in a digitized format that is in digital signal form.

Apart from the speech segment, speech recordings contain two more segments, the silence section at the beginning and the end of the recordings and the noise section (see Fig.3.1.1). Speech databases from different projects are recorded with a different degree of background noise. Here, speech databases which are recorded in a quiet environment with negligible or no continuous background noise are considered. Although there is negligible background noise, there are some unwanted sound artifacts generated during the recording by the speakers such as lip smacking, heavy breathing, mouth clicks, and pops. Fig.3.1.1 shows the presence of NSAs in speech recording.

The problem here is to separate speech utterances from silence and noise segments. Silence can usually be separated by applying algorithms based on STE and ZCR when there is negligible continuous background noise, and no NSAs exists in the recordings. However, STE based approaches fail to segregate the energy level of speech and noise when noise exists in recordings. Moreover, noise and speech segments of a recording do not contain any standard characteristics which can distinguish them. Also, human speech contains two types of sound, Voiced sounds such as vowels (a,e,i,o,u) and unvoiced sounds such as k and p. The characteristics of unvoiced sounds are very similar to noise, and that needs to be taken care of while filtering out the noise. So the problem here is three folds:

- segregate speech from trailing and leading silence
- consider the presence of noise
- need to be careful not to consider unvoiced speech sounds as noise.

3.2.2 Problem Statement

We are considering discrete-time speech signals as input to our system. A discrete time signal X can be mathematically represented as a sequence of numbers as follows [236]:

$$\begin{aligned} X = \{x[n]\}, \quad where\ x[n] = \{x_1, x_2, \ldots, x_n\}, \\ -\infty < n < \infty, \\ (x_1, x_2, \ldots, x_n) \in \mathbb{R} \end{aligned} \tag{3.1}$$

here n is an integer and $x[n]$ is the sequence usually generated by taking a periodic sample from an analog signal.

$$x[n] = \{idle[k], speech[m], noise[l]\}, where\ \ n = k + m + l \tag{3.2}$$

This sequence $x[n]$ comprises of three sections (eq 3.2), the idle section $idle[k]$, the noise section $noise[l]$ and the speech section $speech[m]$ where $n = k + l + m$. These sections are not distinguishable by mere evaluation of the values in these sequences because no predefined ranges or thresholds of values exists.

The task here is to extract only the $speech[m]$ section from $x[n]$. It is assumed here that $speech[m]$ contains a continuous sequence extracted from $x[n]$. But the $noise[l]$ and $idle[k]$ sections can contain combination of multiple sequence fragments from $x[n]$. So, the sequence of $x[n]$ contained in $speech[m]$ cannot be found in either $noise[l]$ or in $idle[k]$.

So, the objective here is to look for pattern in $x[n]$, that can distinguish $speech[m]$ from $noise[l]$ and $idle[k]$ and finally extract the $speech[m]$ from $x[n]$.

3.3 Proposed Solution

A solution based on wavelet convolution to the problem stated in Section 3.2.2 is proposed here. The pattern has been found in the speech signals that demarcate speech utterances from a non-speech section of the recording. The concept of entropy is applied to get an approximation of information content in wavelet convolution coefficients. In the following subsections, these concepts are discussed before formulating the actual solution.

3.3.1 Convolution

Convolution is an important operation in signal and image processing domain. It is a concept extensively used in linear algebra. Convolution is one of the cornerstones of the wavelet transform concept, and continuous wavelet transform is applied to solve the endpoint detection problem. In this section, the concept of continuous convolution is briefly discussed.

Convolution operates with two functions, one is *input*, and another is *kernel*, and produces a third function. First, the *kernel* is flipped (rotation by 180) about its origin and is slid past the *input* to compute the sum of products at each displacement. Let there be an input function *f* and kernel function *g*. Then the convolution between *f* and *g*, denoted by *h*, is defined as follows [199]:

$$h(i) = (f * g)(i) = \int_{-\infty}^{\infty} f(i-j)g(j)dj \tag{3.3}$$

where the minus sign accounts for the flipping of the kernel function *g*, i is the required displacement, and j is a dummy variable that is integrated out [237].

3.3.2 Wavelets

The Concept of Wavelet decomposition is the key to solve the speech endpoint detection problem in this algorithm. This section describes important and relevant areas of the Wavelet concept in as much detail required for this work.

Why Wavelets?

Signals carry overwhelming amounts of data which needs to be extracted as information. Often, the difficulties involved in the task of extracting relevant information from those data become a hurdle for the field of study to which those signals belong. Sparse representation of signals is an efficient way to look for relevant information and patterns in signals. Sparse representation is achieved through decomposing signals over oscillatory waveforms using Fourier or wavelet bases. Speech signals too carry different types of data that need to be extracted as information for better results in various research areas and applications that use speech signals.

Non-stationary signals are the signals whose frequencies, and other statistical properties vary over time. Fourier Transform (FT) is not suitable for analyzing non-stationary signals. Short Time Fourier Transform (STFT) was introduced to overcome this shortcoming of FT. However, during the STFT process, while transforming the time domain signal into the frequency domain, vital time information is lost. This phenomenon of losing time information can be explained by Heisenberg's Uncertainty Principle [see [238]].

Wavelet analysis is best suited in this scenario where non-stationary signals are analyzed to look for a change in frequency components over time. Speech is a non-stationary signal. For this reason, wavelet decomposition is applied here to find relevant frequency components in speech signals. Wavelets define a sparse representation of well-localized piecewise regular signal through the coefficient amplitudes, and few coefficients are required to represent that transient structure. That sparse representation may include transients and singularities — this is why wavelet analysis is vital in speech processing.

Wavelet Analysis

This section describes the method of wavelet analysis. Consider a finite energy signal $x(t)$ where the energy of x is defined by its squared norm and is expressed as follows [238]:

$$\|x(t)\|^2 = \int_{-\infty}^{+\infty} |(x(t))|^2 dt < +\infty \tag{3.4}$$

So, the space on which the $\|x(t)\|^2$ norm is defined has to be square integrable because the integral $\int_{+\infty}^{-\infty} |(x(t))|^2 dt$ must exists. That space is denoted as $\mathbb{L}^2(\mathbb{R})$ is a Hilbert space and is the vector space of the finite energy functions and thus $x(t) \in \mathbb{L}^2(\mathbb{R})$.

The objective here is to decompose the signal x into a linear combination of a set of functions which belongs to $\mathbb{L}^2(\mathbb{R})$. Let us consider a function $\psi(x) \in \mathbb{L}^2(\mathbb{R})$

whose dilation and translation forms a set of functions in $\mathbb{L}^2(\mathbb{R})$ space [238]

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right), \textit{ where } \tau \in \mathbb{R}, s \in \mathbb{R}^{+} \textit{ and } s \neq 0 \tag{3.5}$$

τ and s are the translation and scaling (dilation) parameters respectively and s cannot be negative since negative scaling is undefined [238]. Normalization by $\frac{1}{\sqrt{s}}$ ensures that $\|\psi_{\tau,s}(t)\|$ is independent of s. The family of functions $\psi_{\tau,s}$ is called *wavelets* and ψ is called the *mother wavelet*.

So, now the signal x can be represented as wavelet inner-product coefficients [231]

$$\langle x, \psi_{\tau,s}\rangle = \int_{-\infty}^{\infty} x(t)\psi_{\tau,s}(t)dt \tag{3.6}$$

here both x and ψ are considered as real-valued signals. When ψ is a complex wavelet, the right hand side of equation 3.6 will have complex conjugate of ψ as $\psi^{*}_{\tau,s}(t)$. The *mother wavelet*, also referred to as the *wavelet function* or the *kernel function*, has zero average, meaning $\int_{-\infty}^{\infty}\psi(t)dt = 0$. Apart from satisfying zero average condition *wavelet functions* has to satisfy two more mathematical criteria. First one is that the wavelet function must have finite energy: $E = \int_{-\infty}^{\infty}|\psi(t)|^2 dt < \infty$, which ensures that ψ is square integrable and the inner product in eq 3.6 exist. So the second one is called the admissibility condition which eventually boils down to the condition of zero average, stated earlier, which ensure that x can be reconstructed again after decomposition. The *wavelet function* need to be selected carefully based on the type of analysis to be performed on the input signal because that will help to identify regularities and singularities. The choice of the *mother wavelet* to be used in continuous wavelet transform is restricted only to the conditions of finite energy and admissibility [239]. *Wavelet function* can be either orthogonal or nonorthogonal and only the orthogonal functions form *wavelet basis*. That is why the orthogonal wavelets give a compact representation of the signal and are useful for signal processing. On the other hand, nonorthogonal wavelets produce wavelet spectrum, which is highly redundant at large scales and are more useful for time series analysis [240].

Here continuous wavelet transform (CWT) is used for the analysis, and the focus in on CWT. However, before going into details of CWT, here are two reasons behind selecting CWT over Discrete Wavelet Transform (DWT) for this solution. A discrete sequence of τ is complex to describe, and amplitudes of wavelet coefficients are difficult to interpret since the regularity of a discrete sequence is not well-defined [238]. Moreover, the purpose of the CWT is to extract information from the signal, whereas DWT is good at reconstructing the signal. Here information needs to be extracted from speech signals, and thus CWT is chosen. The scaling parameter s in CWT can vary continuously over $\mathbb{R}$ and can take any value, whereas values s are restricted in DWT. So, signal analysis at an arbitrary scale (or frequency) is possible in CWT and not in DWT, which is an essential criterion for the current problem.

Now, CWT of $x(t)$ with respect to wavelet function $\psi(t)$ at scale s and position τ is the projection of x on ψ and is defined as inner product coefficients in eq 3.6 [238]:

$$C(\tau, s; x(t), \psi(t)) = \langle x, \psi_{\tau,s} \rangle = \int_{-\infty}^{\infty} x(t)\psi_{\tau,s}(t)dt \tag{3.7}$$

which measures the variation of x in the neighborhood of τ proportional to s [238]. Calderon [241] has shown that CWT can be defined as a convolution operation.

$$C(\tau, s; x(t), \psi(t)) = \int_{-\infty}^{\infty} x(t)\psi_{\tau,s}(t)dt = x * \bar{\psi}(\tau) \tag{3.8}$$

where

$$\bar{\psi}(\tau) = \frac{1}{\sqrt{s}}\psi\left(\frac{-t}{s}\right) \tag{3.9}$$

So, CWT extracts information by convolution and not exactly decomposes the signal into sub-signals. For CWT, the reconstruction frame is less important and problematic as well because the *inverse wavelet transform* for CWT is still not well defined. This wavelet convolution operation is the foundation of the proposed solution. CWT must be discretized to be implemented on a computer. That is what is done here by selecting a discrete set of relevant scales for analysis rather than a continuous scale. The shifting (translation) has to be done continuously over all the points of the signal to be analyzed through convolution operation as defined in Eq 3.3.

3.3.3 Information Entropy

The concept of Entropy was introduced in physics as a thermodynamic state variable. It provides an appropriate measure of randomness or disorganization in a system and increases along with the randomness of the system. Here the Information Entropy (IE) of each frame is computed for the speech utterances. IE of a frame is the expected amount of information contained in that frame. Statistically, it is defined as [242]:

$$E(X) = \sum_{i=1}^{N} p(x_i) log_{10} p(x_i), \tag{3.10}$$

where $X = \{x_1, x_2, ..., x_N\}$ is a set of random phenomena, and $p(x_i)$ is the probability of a random phenomenon x_i. IE computation for a speech signal involves breaking the signal into small frames and calculating IE for that frame. Also, for the complete signal, a set of entropy values for each frame is retrieved which is the entropy vector for that signal.

During this work, it is observed that IE of amplitude values of a signal continues to be significantly high and stable when there is a decent disturbance in the system. This observation is useful to keep track of voice activity in a signal recording and a separate voice from silence. This phenomenon is also aligned with the basic property of information entropy, which says that when the probability of the points in the system is equal, then entropy will be high. In the silence section of the speech, a particular set of values (which is 0 or approximately 0) appears very often which makes those values highly probable and other points less probable and thus silent section has low IE. Consider a silent frame $f = \{0.0001, 0.0001, 0.0001, 0.0001, 0.002, 0.004\}$

so, $p(x = 0.0001) = 0.667$ whereas $p(x = 0.002) = 0.167$ and $p(x = 0.004) = 0.167$ and we get $E(X) = 0.377$ as per eq 3.10. On the other hand consider a frame with speech utterance $f = \{0.4, 0.6, 0.9, 0.9, 0.6, 0.4\}$ where $p(x = 0.4) = p(x = 0.6) = p(x = 0.9) = 0.333$ and $E(X) = 0.477$. Here even with only six values and assumed values, IE is higher by 0.1. In the real scenario with much more values and possibilities, this difference becomes significantly higher.

So, from the current problem perspective described in Section 3.2.2 we can write

$$E(speech[m]) \gg E(idle[m]) \tag{3.11}$$

In the proposed algorithm the concept of entropy is a key component in separating speech section from silence.

3.3.4 Concept of Frame

Human speech generation apparatus that is tongue, lip and the other parts of our vocal system involved in producing sound needs approximately 25-30 milliseconds gap between two uttered words because it needs that time to prepare the system to produce the next sound. So, if it is required to break the signal into smaller frames and the size should be chosen within that range. Frames are needed for this algorithm, and it is fixed at **20ms** and is termed as *frame length*. Also, the concept of *frame shift* is used to define the actual shift of data points in the signal, which is fixed at **10ms**. The combination of *frame length* and *frame shift* is used to avoid the effect of the abrupt split of waves during frame splits, to some extent.

3.3.5 Formulation of the Solution

The first step to applying the wavelet decomposition method for analyzing a signal is to select a suitable *mother wavelet*. Here Daubechies wavelet has been selected for this algorithm, specifically DB_8 [243]. Daubechies wavelets are one of the popular wavelets among researchers for speech processing [244, 245]. The shape of a DB_8 signal is shown in Fig.3.3.1. Since continuous wavelet transform is considered here, the scaling and translation parameters s and τ can vary continuously over $\mathbb{R}$ [239]. So, from continuous scales, an arbitrary set of scales is selected to cover the possible frequency range of the human speech recording signals. Here an orthogonal wavelet function DB_8 is convolved over the discrete input signal to get the coefficient values at different scales (frequencies). Orthogonality of DB_8 helps to remove the redundancy of the wavelet coefficient.

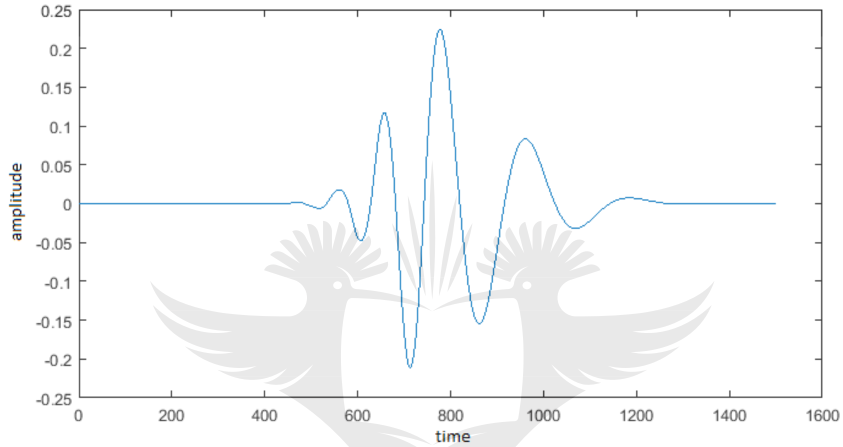

Figure 3.3.1: Figure shows DB_8 wavelet shape at scale 100

The objective here, as described in Section 3.2.2, is to find pattern in discrete sequence $x[n]$ (eq 3.1) to segregate speech segment from the rest of the sequence. The wavelet convolution operation is applied to analyze the sequence $x[n]$ and search for consistent patterns. It is observed during the experiments that the presence of NSAs are prominent in the coefficient amplitude plot when the wavelet scale is small (high-frequency) (Fig.3.3.2). It is equivalent to the fact that NSAs has similarities with high-frequency wavelets since low scale value implies high-frequency. However, it is observed that at higher scales (low-frequencies) the NSAs are almost non-existent in the plot. Figure 3.3.2 shows the coefficient amplitudes at different scales for a speech utterance with breathing noise. Scale 10 highlights 3200Hz frequency components where breathing noise is very prominent. Scale 23 highlights 1391Hz frequency components where noise is most prominent compared to a speech utterance. Scale 50 highlights 640Hz frequency components where the weak presence of noise can be seen. And finally, Scale 100 highlights 320Hz frequency components where the noise section is very weak compared to the speech section. The phenomenon is well supported by the fact that NSAs are usually high-frequency sounds and thus produce high coefficient values in convolution with a low scale (high-frequency) wavelets. This observed phenomenon is the backbone of this approach to solving the problem of speech endpoint detection.

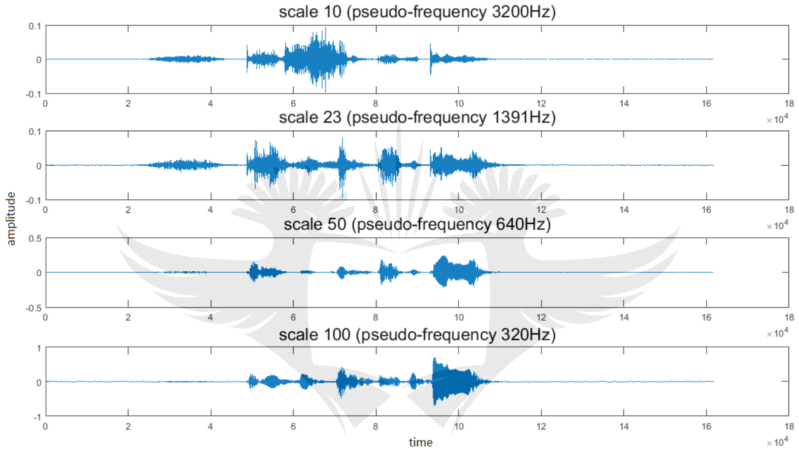

Figure 3.3.2: Coefficient Amplitudes at different scales for a speech utterance with breathing noise.

A set of scales has been selected to cover the intended range of frequency. The frequency range of the human speech signal spread approximately within the range between 250Hz and 6000Hz [246]. However, it is observed that the NSAs are prominent around 3000Hz and around 300Hz the presence of noise is very weak, so here the selected upper limit as 3000Hz and lower limit as 300Hz. Two sets of scales are selected to accommodate the selected frequency range using DB_8 mother wavelet:

- $scale_{hf}$ includes set of high frequency range (low scale values)
- $scale_{lf}$ includes set of low frequency range (high scale values)

At a low scale, wavelet coefficient values are much smaller compared to coefficient values at a high scale, and this is the reason why more scales are selected for $scale_{hf}$ than $scale_{lf}$.

It is assumed here that there exists a gap of few milliseconds between the NSAs and the speech utterances. It is highly improbable that the speaker can produce some NSAs precisely before and after the actual utterance without any time gap. For example, the noise of breathing out cannot come out while speaking because the voice is already coming out with exhalation, and if at all breathing noise comes out while speaking it would distort the speech utterance. Similarly, mouth pop and click sounds cannot be produced by the speaker while uttering a speech because that will interrupt the utterance.

Now the wavelet transform of the discrete sequence $x[n]$ (3.1) is performed, which is defined as convolution of x with a scaled and translated version of ψ the mother wavelet (DB_8) [240] to generate set of coefficients as described in eq 3.8.

$$coefs = x * \psi \tag{3.12}$$

Coefficient sets are needed to be combined to get two vectors that can be used for further processing. To achieve that sum or average strategy has been applied depending on the loudness of the actual signal X. When loudness is higher than a specific threshold value, the coefficient values are averaged; otherwise, they are summed.

After the coefficients are combined into two vectors, namely $coef_{hf}$ and $coef_{lf}$, the entropy is computed for both the vectors. The coefficient vectors are broken down into frames, and then entropy is computed using the formula defined in eq 3.10. These entropy vectors are special in a sense that they represent high-frequency entropy (say ce_h) and low-frequency entropy (say ce_l) of the wavelet coefficients.

The entropy vectors ce_h and ce_l are further used to calculate two sets of thresholds one for high-frequency and the other for low-frequency. Low-frequency thresholds are used to identify locations with the presence of speech utterance because low-frequency components are distinctly separate from $idle[k]$ and $noise[l]$ sections.

Then high-frequency thresholds are used to stretch those identified speech utterance zones with proper voiced and unvoiced trails at the beginning and end of a speech utterance.

3.3.6 The Algorithm

The proposed algorithm WCSED is designed to work independently. Systems which require to extract speech segment from speech signals can incorporate WCSED as a separate module. The steps of the proposed algorithm are listed in Algo 1 section and the flow as a block diagram in 3.3.3. Here the pseudo code is provided in the listing and the functions, in brief, are mentioned to maintain the readability of the algorithm.

WCSED algorithm consists of one main module and three submodules. The main module is called WCSED, which accepts discrete time speech signal as input and returns the extracted speech segment. The "WaveConv" module is responsible for computing the CWT on the input signal and returns the coefficients. The "GetEntropyVector" module computes entropy by breaking down the input sequence into segments and returns a vector. Finally, the points towards the edges of the end-points are selected by considering the threshold values provided.

Assumptions for this WCSED algorithm are kept at the minimum to maintain generality. The thresholding concept is applied, but the assumptions on leading and trailing silence similar to Zhu and Chen [227] are relaxed because that would restrict the scope of this algorithm to specific datasets. Thresholds are dynamically calculated.

3.4 Datasets

In this study two speech databases are used one is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [128] dataset and the other one is

Algorithm 1 WCSED algorithm

Input: Discrete-time signal S(n), where n is the length of the signal and Sampling Rate **Output:** Extracted Speech Segment S_{extr}(k), where $k <= n$

```
function WCSED($S(n), FS$)    ▷ S=discrete time signal and FS=sampling rate
    $FL \leftarrow FrameLength$
    $FSH \leftarrow FrameShift$
    $MW \leftarrow Daubechies$    ▷ mother wavelet
    $SC(m) \leftarrow [HighFrequencyScales, LowFrequencyScales]$    ▷ m number of scales
    $COEF_{m\times n} \leftarrow WaveConv(S(n), SC(m), MW)$    ▷ coefficients
    $CE \leftarrow GetEntropyVector(COEF_{m\times n}, FL, FSH)$
    $th_u, th_l \leftarrow$ compute upper and lower thresholds
    $sec_{al} \leftarrow CE \geq th_l$
    $pos_s \leftarrow IncludeEdges(CE, sec_{al}[start], back, th_u)$
    $pos_e \leftarrow IncludeEdges(CE, sec_{al}[end], front, th_l)$
    $S_{extr}(k) = S[pos_s, pos_e]$
    return $S_{extr}(k)$    ▷ The extracted speech

function WaveConv($S, SC, MW$)    ▷ signal,scales,mother wavelet
    $CD_{m\times n} \leftarrow$ output matrix
    for ($m = 1; m <= lenght(SC); m++$) do    ▷ iterate through all the scales
        $f \leftarrow$ get the reference wavelet
        $CF_{1\times n} \leftarrow S * f$    ▷ convolution gives the coefficients
        $CD_{m:} \leftarrow diff(CF)$    ▷ take approximate derivative
    return $cd$    ▷ derivative of coefficients

function GetEntropyVector($i, fl, fs$)    ▷ input sequence, frame len, frame shift
    $len \leftarrow length(i)$
    $sp \leftarrow 1$
    $ep \leftarrow sp + fl - 1$
    while $ep \leq len$ do
        $entropy_v \leftarrow Entropy(i[sp, ep])$    ▷ calculate entropy
        $sp \leftarrow sp + fs$
        $ep \leftarrow sp + fl - 1$
    return $entropy_v$    ▷ The entropy vector
```

WCSED algorithm cont...

```
function IncludeEdges($i, p, d, th$)
    $edges \leftarrow []$
    $len \leftarrow length(i)$
    if $d = back$ then
        $init \leftarrow p - 1; incr \leftarrow -1; limit \leftarrow 1$
    else
        $init = p + 1; incr \leftarrow 1; limit \leftarrow len$
    for $(cntr = init; cntr < limit; cntr = cntr + incr)$ do
        if $i[cntr] \geq th$ then
            $ec \leftarrow ec + 1$
            $edges[ec] \leftarrow cntr$
        else
            $break$
    $pos \leftarrow edges[length(edges)]$
    return $pos$                                        ▷ last point near edge
```

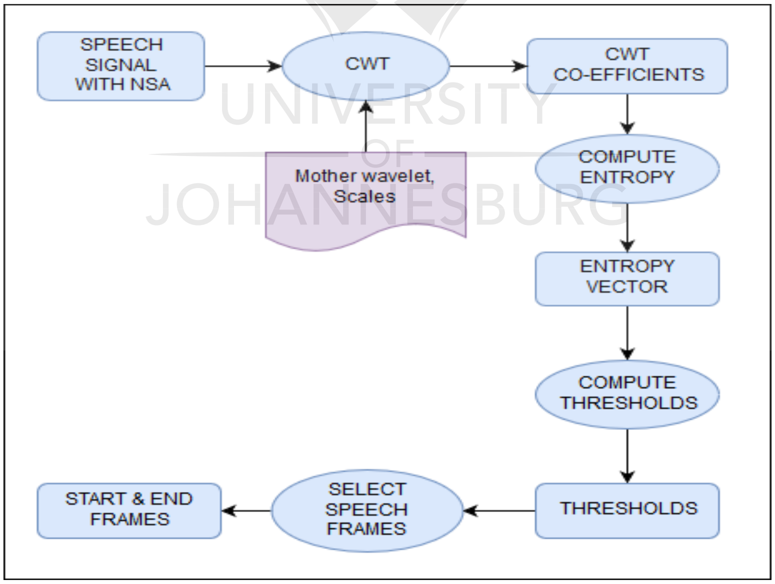

Figure 3.3.3: Block diagram of the WCSED algorithm

EMO-DB ([119]) which is a German language corpus. RAVDESS dataset was primarily created in view of research areas related to Emotion Recognition in Speech and Song. Only the speech recordings are used for this current work. While working on Speech Emotion Recognition, it is observed that the recordings contain different sound artifacts generated by the speakers such as heavy breathing, mouth clicks, pops, and lip-smacking. These sound artifacts are making endpoint detection task, and the need for a robust endpoint detection algorithm was felt. The proposed algorithm is also tested on another speech database called EMO-DB, which is a German language corpus.

3.5 Results and Observations

The primary objective of WCSED algorithm is to automate the process of extracting the speech segments precisely in the presence of the NSAs, and it has shown promising results. It has successfully extracted the speech segments from almost all the recordings. In very few cases, a significant amount of speech could not be extracted, but the algorithm performed poorly in those rare cases. The speech recordings containing NSAs are efficiently processed by separating those unwanted artifacts from actual speech.

Some speakers pause for some few milliseconds between the words. Those pauses should be included as a part of speech segment since pauses can add quality to the speech recording while extracting say emotional quotient, and the algorithm did it well in those cases too.

Fig.3.5.1 and Fig.3.5.2 show the end result of the algorithm depicting the extracted segment along with corresponding entropy values.

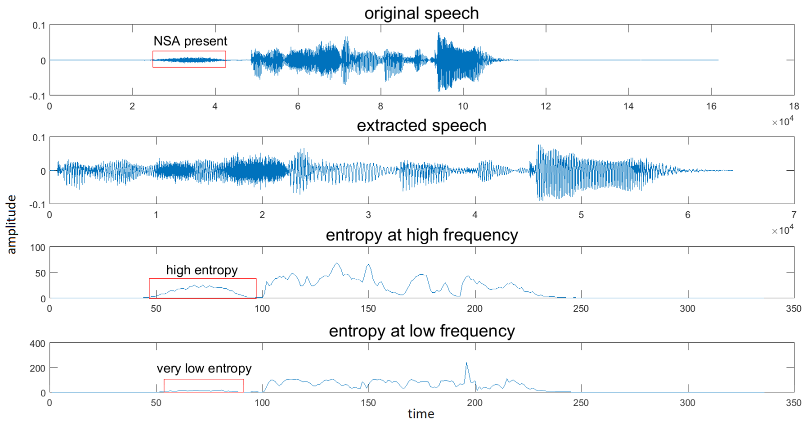

Figure 3.5.1: The Figure shows extracted speech along with corresponding entropy. The breathing noise NSA is precisely discarded.

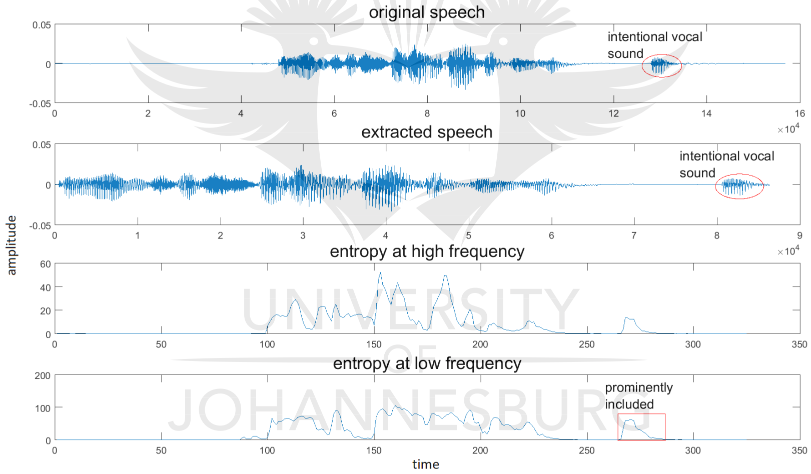

Figure 3.5.2: The Figure shows extracted speech along with corresponding entropy. Speaker's intentional voice sound is meaningfully included in the extracted speech.

The algorithm is tested on two speech databases RAVDESS and EMO-DB, and results are compared in detail. The experiment results are summarized in Table 3.5.1, where the percentage deviation is depicted. More than 20% of the total number of speech recordings are selected as sample for cross verifying with the results received by applying the WCSED algorithm. Those samples are manually checked for possible start-frames and end-frames of the speech segments in the recordings. Since WCSED algorithm extracts the speech segment based on frames, the selected samples are also processed based on start and end frames. After manually extracting the frames of the samples, it is checked whether the start and end frames are deviating from the frames as reported by the WCSED algorithm of the corresponding speech recordings.

Table 3.5.1: Deviation Percentage based accuracy measure of WCSED

	Average % of Deviation			
Speaker	RAVDESS		EMO-DB	
Gender	Start	End	Start	End
FEMALE	1.027	2.259	1.147	0.734
MALE	0.576	2.847	1.416	1.248
Average % of deviation	**0.777**	**2.584**	**1.249**	**0.93**

Simulations are executed ten times on the selected sample to check whether there is any discrepancy for different simulations. However, it is observed that in every simulation, the algorithm has produced the same results. The results are cross verified in a few stages. First, the beginning and end frames are calculated for the selected samples manually, and they are termed as manual-frames. The beginning and end frames reported by the WCSED algorithm are termed as algorithm-frames. Then the absolute deviation between the manual-frames and algorithm-frames are computed. Manual-frames are considered as a baseline for the calculation of the frame length of the extracted speech. Then the percentage of deviation in frames is computed compared to the frame length of the extracted speech.

By analyzing the deviations for RAVDESS, it is observed that the overall start-frame deviation is 0.777% (means approximately 99.3% accurate), while the end-frame deviation is 2.585% (means approximately 97.5% accurate). Thus, the algorithm extracts the start frames more accurately than the end frames. This accuracy gap occurs because different speakers end their utterance with different styles

Table 3.5.2: Frame Difference based accuracy measure of WCSED

Frame Difference %	RAVDESS		EMO-DB	
Frame Difference Ranges	Start	End	Start	End
0 (Frames)	55.33	14.43	23.01	27.43
1-5 (Frames)	40.55	59.11	55.75	61.95
6-10 (Frames)	1.37	13.06	17.7	6.19
11-20 (Frames)	1.37	13.4	3.54	4.42
above 20 (Frames)	1.37	0	0	0

and varying pause or silence between spoken words. So, the overall accuracy of the WCSED algorithm to detect start-frame is 99.3% (approx), and end-frame is 97.5%(approx). Similar way for EMO-DB the accuracy for start-frame is 98.8% (approx) and for end-frame is 99.1% (approx). These differences of accuracy are due to the recording environment and speaking style of the speakers.

Table 3.5.2 shows another measure of accuracy as coverage of the ranges of frame differences between manually detected and WCSED results. It shows that the differences are within five frames covering 95.88% and 78.76% in the start-frames whereas 73.54% and 89.38% in the end-frames for RAVDESS and EMO-DB respectively. In Table 3.5.2 the frame differences are not more than twenty frames except on one case.

Table 3.5.4 [229] compares the results obtained by Atanas [229] with the WCSED algorithm, and WCSED is doing better because 0-20 frame difference range constitutes a better share. The WCSED results are also compared with optimal filter based approach [217] in the Table 3.5.3 [217]. When up to 3 frame differences for start-frame are considered, on the RAVDESS database it is much better (94.16%) than the filter based approach (74.78%) whereas on the EMO-DB database (74.33%) it is almost the same. Overall the WCSED performs significantly better. The Figures 3.5.3 and 3.5.4 show the histograms of the frame differences for the RAVDESS and the EMO-DB respectively. It is clear from the two figures that the frame-differences vary around zero, which is a sign of high accuracy for both the databases.

It is observed during testing that the deviations are different for female and male speakers. A factor that contributed to this phenomenon is possibly the loudness variation in female and male speakers. Male voices in this recordings are usually louder and more prominent than female voices.

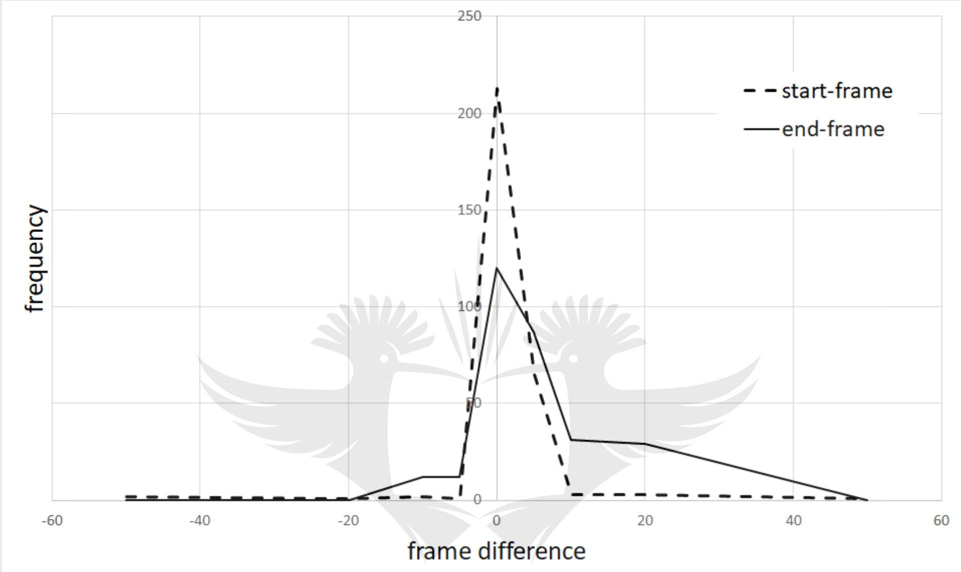

Figure 3.5.3: Histogram showing the frame differences in start and end frames computed by WCSED algorithm on RAVDESS database.

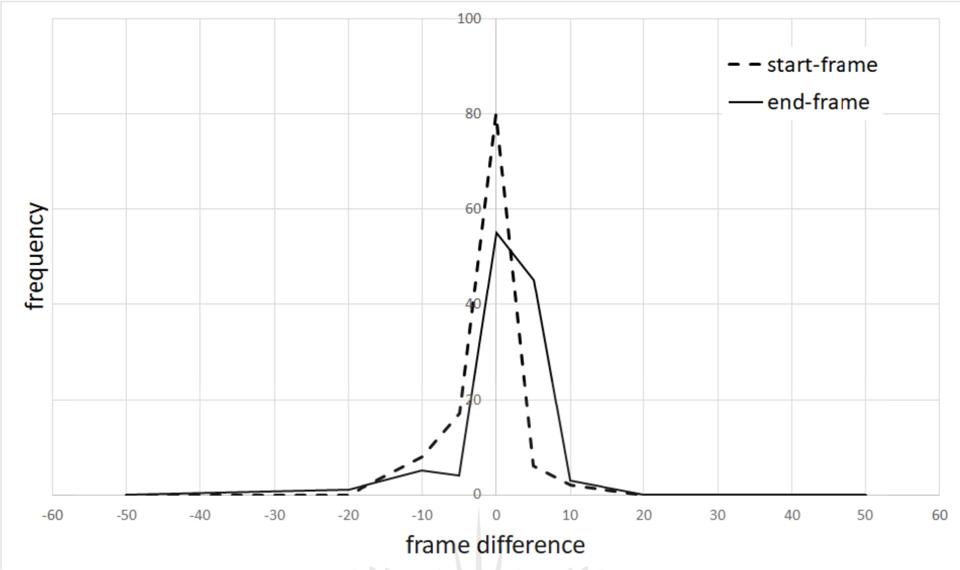

Figure 3.5.4: Histogram showing the frame differences in start and end frames computed by WCSED algorithm on EMO-DB database.

Table 3.5.3: Comparing WCSED start frame results with Optimal Filter and Energy Model (OFEM) combination approach proposed by Li et al. where they also tested with HMM for comparison.

Differences in number of frames	0	≤ 1	≤ 2	≤ 3
WCSED on RAVDESS	**55.33%**	**89.01%**	**93.13%**	**94.16%**
WCSED on EMO-DB	23.01%	60.18%	69.91%	74.33%
OFEM	25.97%	57.84%	69.21%	74.58%
HMM	22.95%	54.82%	69.9%	76.52%

Table 3.5.4: Comparison of the WCSED based results with trajectory-based EPD approach proposed by Ouzounov. Ouzounov used three features called Modified Teager Energy (MTE), Energy-Entropy (EE) and Mean-Delta (MD).

	Start point		End point	
Frame Difference Ranges	0-10	0-20	0-10	0-20
WCSED on RAVDESS DB	97.25%	98.63%	86.6%	100%
WCSED on EMO-DB	**96.46**%	**100**%	**95.58**%	**100**%
Trajectory-based with MD	61.45%	95.8%	17.55%	82.44%
Trajectory-based with MTE	54.19%	95.8%	28.62%	88.54%
Trajectory-based with EE	55.34%	96.56%	18.32%	82.06%

Finally, the time complexity of the WCSED algorithm is directly proportional to the length of the input signal. When the input signal length increases, the algorithm will take more time to extract the speech utterance from the input signal.

3.6 Conclusion and Future directions

The proposed WCSED algorithm attempted to address four issues of speech endpoint detection problem. First, automating the process of the EPD. Second, discarding the NSAs and extracting start and end points appropriately. Third, relaxing the assumptions like the availability of labeled data which could hinder this algorithm from working correctly across different speech databases and in real-world applications. And finally, extract the end-points accurately. The results discussed in Section 3.5 show great promise, and the WCSED successfully addressed the issues mentioned above. Moreover eliminating the dependency towards labeled data for EPD should significantly impact the ASR and related fields. Comparison of the WCSED accuracy results with similar works has also shown that the WCSED is performing significantly better.

This algorithm can further be applied to different speech signal based systems where utterances need to be extracted from the speech signals in the presence of different NSAs. For example, this algorithm can be applied in the preprocessing stage of an ASR or an SER system.

Wavelet convolution (CWT) based approach to find consistent patterns in a discrete time signal can be applied to solve similar problems in speech recognition domain and other domains where patterns need to be identified from signals. The CWT can be used to enhance the feature set of various classification problems.

Finally, some scope of improvement for the WCSED algorithm has been identified, which should be a subject of a future direction. It is mentioned in the result section that the level of loudness of the speaker's utterances could be an essential factor to improve the end-point selection results. Further investigation and action in that direction could yield more accuracy of this WCSED algorithm. The WCSED could further be adapted to work in the online or real-time scenario to enhance the scope and usability of the algorithm. These improvements will be taken up as future advancements of the algorithm.

Everything in Life is Vibration.

Albert Einstein

4

Introducing Novel Feature Set based on Wavelets for Speech Emotion Classification

Speech Recognition research is in a very advanced stage at present, and examples like Siri, Alexa, and Google Home show that the speech recognition task is already advanced and this field of study is surging ahead to address other unsolved issues. However, one crucial aspect of human speech, which is still not appropriately addressed in human-machine interaction is ”emotion”. Emotion recognition from speech signal is a challenging task, and researchers are still looking for a well-accepted solution.

Speech Emotion Recognition (SER) study applies different classification techniques from Machine Learning (ML) field to classify different emotions of the speech utterances. Feature extraction from speech signals is a vital task of SER systems since the selection of features eventually affects classification performance. Researchers have used different speech features and combinations of those so far but yet not identified any combination of them as best. So, the quest for the best feature set for SER is still unsolved.

Ayadi et al. [93] grouped the speech features into four categories as follows : (a) Continuous - includes features like pitch, energy, and formants (b) Qualitative - such as voice quality, harshness, breathy (c) Spectral - such as Linear Predictive Coding (LPC), and Mel-Frequency Cepstral Coefficients (MFCC) (d) TEO-based - such as TEO-FM-Var, and TEO-CB-Auto-Env. Researchers have tried most of the features in different combinations in SER systems for emotion classification. Some SER systems use continuous features [31, 32, 94], and some use spectral features [53, 160, 246, 247]. However, the adequate level of classification accuracy is not yet achieved with different classification techniques using the existing set of features.

The SER research efforts are mostly focused on finding suitable classification techniques, and various techniques are used such as the Gaussian Mixture Model (GMM) [53], Hidden Markov Model (HMM) [118, 248], Support Vector Machine (SVM) [246, 249] and very recently different Deep Learning (DL) architectures [160, 247] to name a few. However, comparatively, less focus is given to developing SER specific features and using the same feature sets which are being used for Automatic Speech Recognition (ASR) or their variants. These two problems, the ASR and SER, are different so their features should also be different.

Features like pitch, energy, MFCC, and LPC describe the properties of the speech sound well. So, these should be the right criteria when it is required to track variation in speech properties. However, the SER requires attributes of the speech utterances, which can describe the emotional states of the speakers, which is not described adequately by the existing features. That is why existing features are not performing significantly well, and there is no universal consensus on a specific feature set among researchers [160, 247]. Some researchers even suggested that MFCC features are not that effective in SER as it is in ASR [118, 250, 251].

Presently the classification techniques are in a very advanced state which can robustly classify with the right set of data, specifically the advent of Deep Neural Networks (DNN). However, applying DNNs in SER is still not showing enough promise like in other fields.

So, there is a possibility that researchers are missing some phenomenon which should be considered to represent the emotional content of speech utterances. Emotion is a subjective experience which is expressed differently by different speakers, and that makes SER feature extraction a challenging task. That is where this new feature concept comes into the picture which tries to extract the difference between the emotional state and neutral state of the speaker. Here we are introducing a set of features which is a dissimilarity measure based on discrete wavelet transform. The feature set is the speaker specific dissimilarity measure of the information content of different emotional speech utterances from the same speaker's neutral speech utterance. This feature set is named as *Subjective Emotional Gap Reduction Technique (SEGRT)* because it tries to reduce the *subjective gap* between the features extracted from the speech of two different speakers.

This proposed feature set does not preserve the sequential nature of the speech signal and is more like global features. So, using the SEGRT feature set, the SER problem is no longer mostly a sequence classification problem such as the one based on the HMM. So, conventional classification techniques such as Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Naive Bayes (NB) can be applied instead of sequence classification techniques. Here SVC, KNN, and NB are used with the SEGRT feature set to establish the relevance of the feature set and the results are supportive enough to claim that advancement of these features can bring fruitful results in improving SER accuracy rate.

4.1 Feature extraction

The proposed feature set is extracted from speech signals represented as discrete time signals, let's say S, which is defined as follows [236]:

$$\begin{gathered} S = \{s_n\}_{n \in \mathbb{N}}, \quad where \ \{s_n\} = \{s_1, s_2, \ldots, s_n\}, \\ \{s_n\} \in \mathbb{R} \end{gathered} \tag{4.1}$$

First, only the speech utterance section is extracted from the speech sound by using the speech endpoint detection algorithm proposed by Roy et al. [82] and another sequence is received

$$X = \{x_n\}_{n \in \mathbb{N}} \tag{4.2}$$

where $\{x_n\}$ is a subsequence of $\{s_n\}$ and assume N be the length of X. The proposed Feature extraction technique uses the concept of Discrete Wavelet Transform (DWT), Information Entropy, and Dissimilarity Measure. These concepts are briefly discussed in the following subsections.

4.1.1 Discrete Wavelet Transform

Discrete Wavelet Transform (DWT) decomposes a signal into trend and fluctuation sub-signals. So the number of sub-signals to be generated by the DWT depends on the transformation level [231]. There are different types of DWTs, but five Daubechies wavelet transforms *db6*, *db8*, *db10*, *db12*, and *db14*, are chosen for this work. The trend signals highlights the trends at different level of decomposition of signals which enables trends to help in finding patterns in a signal. While the fluctuation signals help in locating any abrupt changes in the signal. The combination of trend and fluctuation signals provides a robust pattern recognition system in a signal.

Here DWT of the signal X in eq.4.2 is taken till level 4. The transformation for level 1 can be described as a mapping $X \longmapsto (t1|f1)$ where $t1$ and $f1$ are the 1-level trend and fluctuation sub-signals respectively having length half of the length of

X i.e. $N/2$. At the level-2 of DWT the trend sub-signal of level-1 i.e. $t1$ is further broken down into trend and fluctuation signals so that $t1 \longmapsto (t2|f2)$. And at level-2 the transformation is defined as $X \longmapsto (t2|f2|f1)$ where $t2$ and $f2$ are the level-2 trend and fluctuation signals respectively with length $N/4$. So, in a similar way at level-4 the DWT is defined as

$$X \longmapsto (t4|f4|f3|f2|f1) \tag{4.3}$$

where $t4$ and $f4$ are level-4 trend and fluctuation sub-signals respectively with length $N/16$ and $f3$ is the level-3 fluctuation sub-signal with length $N/8$.

Thus, from each 4-level DWT of X, five sub-signals are retrieved, and from all the five DWT we have considered i.e. *db6*, *db8*, *db10*, *db12*, and *db14* we get 25 sub-signals. These sub-signals will be further processed for actual feature extraction.

4.1.2 Information Entropy

Entropy provides an approximate measure of randomness or disorganization in a system and increases along with the randomness within the system. Entropy is low when there is ordered activity (like sine waves), and entropy is high when there is random activity [252]. Information Entropy (IE) is an expected measure of the information content in a signal. IE of a sequence Q of length n is defined as (see. [242])

$$E(Q) = \sum_{i=1}^{n} p(q_i) log_{10} p(q_i), \tag{4.4}$$

where $Q = \{q_1, q_2, ..., q_n\}$ is a set of random phenomena, and $p(q_i)$ is the probability of a random phenomenon q_i.

Here IE of the sub-signals, produced by DWT (eq.4.3), is computed to generate IE sequence. IE sequence generation process involves breaking the sub-signal into small frames and calculating IE for a frame using the formula shown in eq.4.4. Figure 4.1.1 shows how IE of 4-level *db8 trend* sub-signal for Happy and Angry emotions are varying around the Neutral state for the same in the EMODB dataset.

Figure 4.1.2 shows how IE of 1-level *db8 fluctuation* sub-signal for Happy and Angry emotions are varying around the Neutral state for the same in the EMODB dataset. The entropy sequences are further used for dissimilarity measurement.

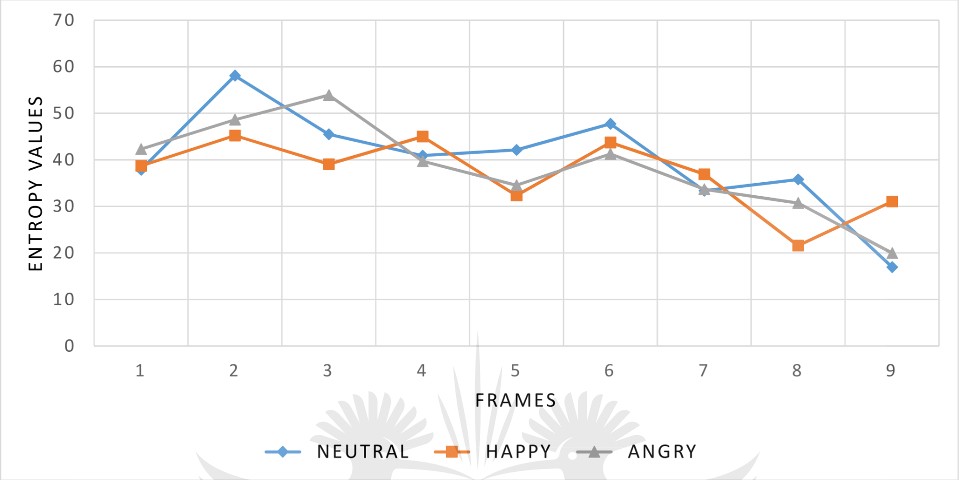

Figure 4.1.1: The figure shows how *trend* sub-signal for Happy and Angry emotions are varying around the Neutral sub-signal.

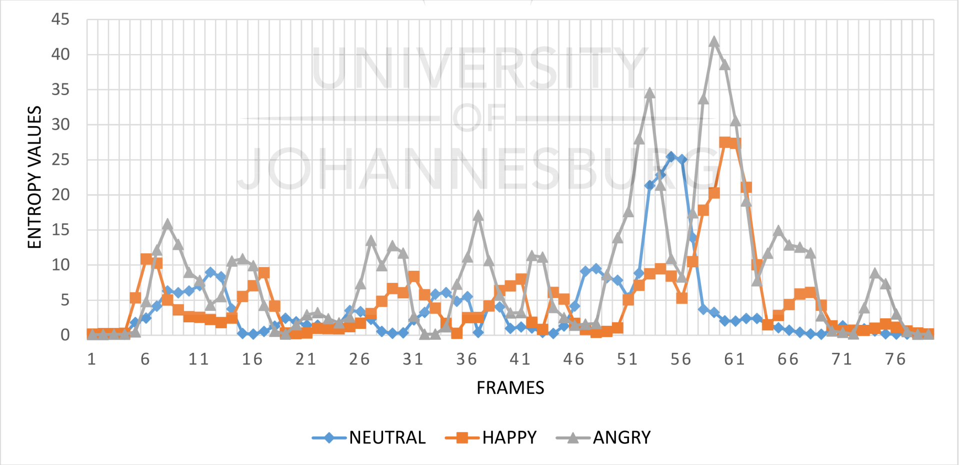

Figure 4.1.2: The figure shows how *fluctuation* sub-signal for Happy and Angry emotions are varying around the Neutral sub-signal.

And for the complete sub-signal, a set of entropy values for each frame is retrieved which is the entropy sequence for that sub-signal.

$$t4 \mapsto \{te4_k\}, fj \mapsto \{fej_k\} \tag{4.5}$$

where $j = \{1, 2, 3\}$, m is the number of frames into which the sub-signals are broken down, $k = \{1, 2, ..., m\}$, and $te4_k$, as well as fej_k are the entropy values at the *kth* frame of the trend and fluctuation sub-signals respectively.

4.1.3 Dissimilarity Measure

The feature set for SER proposed in this chapter takes a unique approach by measuring the dissimilarity between each emotional state of a speaker from the corresponding neutral state of the same speaker. Here dissimilarity is computed as the Euclidean distance. Suppose, features are being extracted from an angry emotional state utterance of speaker S1. For that, the neutral and angry state utterances of speaker S1 are first broken down till IE sequences (similar to eq.4.5). So, for both angry and neutral utterances, there will be 25 IE sequences each. Then, Euclidean distance is computed between every 25 sequences of neutral and angry utterances. Eq.4.6 shows an example of the Euclidean distance computation between two trend IE sequences [189].

$$d(\{te4_k\}_{NU}, \{te4_k\}_{AN}) = \sqrt{\sum_{1}^{m} (\{te4_k\}_{NU} - \{te4_k\}_{AN})^2} \tag{4.6}$$

where *NU* denotes *neutral* state and *AN* denotes *angry* state, So, we have now 25 distance measures. Two more values are computed from the IE sequences; one is the covariance, and another is the correlation coefficient, which gives 50 more features. Thus there are 75 values computed along with the gender of speaker and length ratio between neutral and angry utterance and making the total number of features to 77. So, for each emotional utterance of the speakers, there will be 77 attributes to be extracted as features and used for emotion classification.

4.1.4 Feature Enhancement and Dimensionality Reduction

The speech signal is initially pre-processed with a noise reduction technique to enhance the signal quality. Here for this work spectral subtraction method is used. The speech signal is further enhanced using advanced speech endpoint detection technique. Here, the endpoint detection technique used is the WCSED [82].

The SEGRT technique is applied to the pre-processed speech signal. However, no dimensionality reduction technique is used in SEGRT process. Two most widely used dimensionality reduction techniques *principle component analysis*(PCA) and *independent component analysis* (ICA) are tried with SEGRT, but both of them failed to enhance the dataset. When PCA is applied with SEGRT, the minuscule variations at the same wavelet transform levels are overlooked, and the data points of the same levels of different emotions are merged into single components. On the other hand, applying ICA in SEGRT results in very few components which are markedly different from one another. So, ICA is also not suitable for this analysis.

4.2 Classification Techniques

Here the problem of classifying emotions is broken down into multiple binary classification problems. Three widely used classification techniques Support Vector Classifier (SVC) with Radial Basis Function (RBF) kernel, Gaussian Naive Bayes (GNB) and K-Nearest Neighbor (KNN) are deployed for this task. These three classifiers take different approaches to solve the same two-class classification problem.

Linear SVC separates two groups (classes) by determining the hyperplane, which maximizes the margin between two classes provided the two classes are linearly separable in the feature space. However, that kind of linearly separable data are rare, and mostly we deal with data which can be efficiently separated by nonlinear region. For these situations, the SVC is applied with nonlinear kernel functions to transform the data into higher dimensional space where the classes can be linearly separable and perform the classification. Gaussian RBF is a commonly used kernel for SVC [172]. SVC with RBF kernel tends to overfit but the introduction of the ”soft margin” concept, proposed by Cortes and Vapnik [196], added the flexibility to regulate the extent of over-fitting on the model where harder margins are more likely to overfit. For this work, SVC with RBF kernel is selected for classification with carefully chosen hyper-parameters to manage model overfitting.

The Naive Bayes (NB) classifier is based on the Bayes theorem, and it is easy to build a classification model using NB. NB classifier usually works with categorical data and Bayes theorem equation is used to find the posterior probability of the classes by computing the likelihood from the frequency. However, for numerical data where attribute values are not categorical, numerical attributes can be assumed to follow Gaussian distribution, and then the posterior probability can be computed by finding the likelihood using the assumed Gaussian distribution. This approach of assuming a Gaussian distribution is called the Gaussian Naive Bayes. Here the features are numerical, and GNB is applied.

The K-Nearest Neighbor is a simple but effective classification technique based on distance or similarity measures. It is a lazy-learning method because there is no pre-models existing, and almost all computations take place while classifying new data points. A majority vote of its neighbors classifies a new data point, and the number of neighbors to consider is selected before applying the KNN as k parameter. Selecting the right value for k is an essential criterion of using the KNN because that affects classification performance. The new data point is assigned to the class, which gets majority votes from its neighbors. There are distance measures like *Euclidean*, *Manhattan*, and *Minkowski* for continuous variables while *hamming* is for categorical variables to choose from depending on the feature set.

4.3 Experiment

Two emotion data sets are used for this experiment one is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [128] which is an English language dataset and the other one is the EMODB [119] which is a German language corpus. Eight emotional states are considered for this work and the following short-codes for the emotions are used throughout this chapter: clam(CA), happy(HA), sad(SA), angry(AN), fear(FE), disgust(DI), surprise(SU) and boredom(BO). The EMODB dataset does not contain utterances with *calm* and *surprised* emotions whereas the RAVDESS dataset does not have utterance with boredom state.

Experiments are conducted by taking two emotion labels at a time and applying binary classification techniques to classify the two emotions. For example, happy and sad emotions are taken for classification and checked the performance. Stratification method is used to split the data into training and testing sets with ten splits. This train/test split is required to reduce the extent of over-fitting menace on the classification model. Stratification helps to rearrange the data into folds so that each fold can represent the complete data well and is a better approach compared to regular cross-validation to manage the bias and variance. Then the stratified folds are used for cross-validation to measure the performance of the classifiers. Among the three classifiers used for this experiment, SVC and KNN, have few hyper-parameters to be appropriately selected for better classification results. GNB classifier, on the other hand, does not have hyper-parameters. Two important hyper-parameters for SVC with RBF kernel are C and *gamma*, and they are selected as $C = 10$ and $gamma = 0.001$. The values of C and *gamma* are carefully chosen to fit the model well because on the one hand very high C and *gamma* values can lead to over-fitting on the other hand very low values can lead to under-fitting. For KNN the distance metric is chosen as *Manhattan* and $neighbors = 5$. These hyper-parameters for both SVC and KNN are selected based on the Grid Search cross-validation results.

4.4 Results and Discussions

The accuracy of the three classification techniques SVC, GNB, and KNN using two databases, RAVDESS and EMODB is shown in fig.4.4.1 and fig.4.4.2. Average accuracy rates of the SVM, GNB, and KNN are 73.67%, 77.71%, and 69.41% respectively using the RAVDESS while 73.74%, 80.88%, and 72.75% respectively using the EMODB. Some accuracy percentages are over 90% such as sad-angry classification using the EMODB reaches 93.98% whereas calm-angry classification using the RAVDESS reaches 93.21%. One interesting observation has been depicted in fig.4.4.3 which shows how SER performance varies with the two databases used for this work. The figure 4.4.3 shows the deviation in the SER performance of KNN, SVM, and GNB classifiers. The deviations are the difference between the accuracy using the RAVDESS database and the accuracy using the EMODB database. The emotional states considered here are present in both the databases.

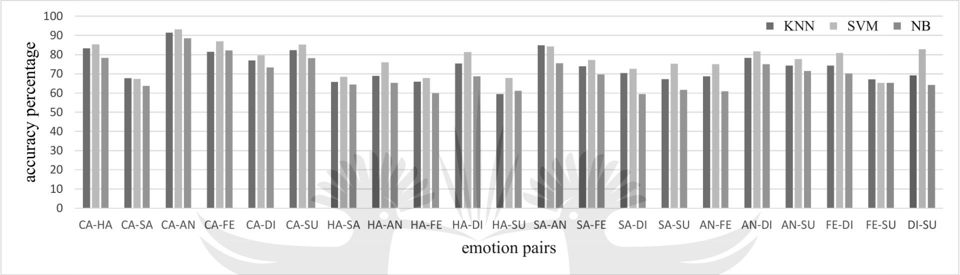

Figure 4.4.1: The figure shows the performance of KNN, SVM and GNB classifier on RAVDESS dataset.

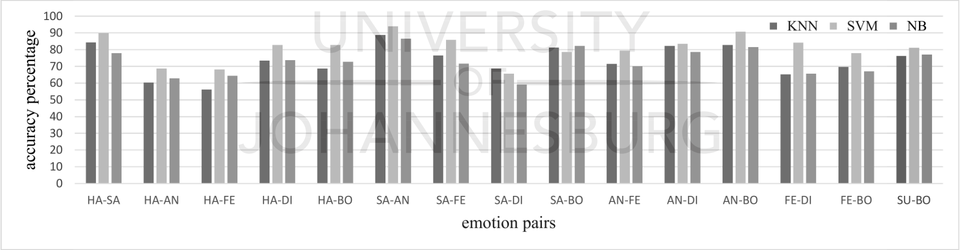

Figure 4.4.2: The figure shows the performance of KNN, SVM and GNB classifier on EMODB dataset.

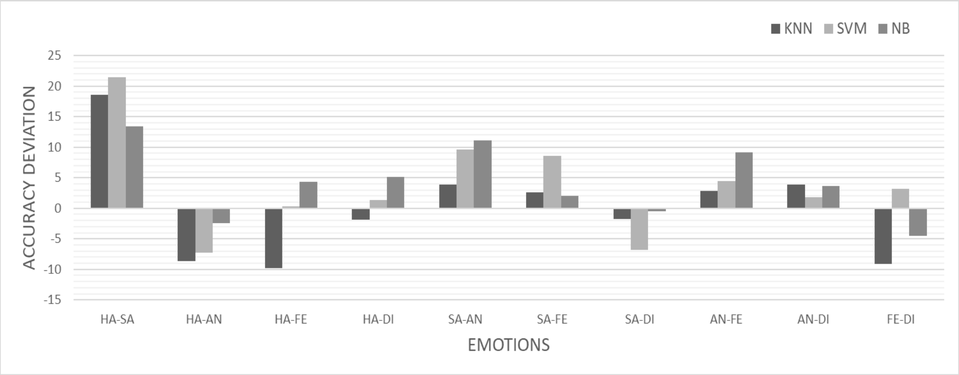

Figure 4.4.3: The figure shows the deviation in the SER performance of KNN, SVM, and GNB classifiers.

Table 4.4.1 shows the computation time of the three models on RAVDESS dataset. Here datasets are split into ten folds for cross-validation, and on average there are 383 rows. Since the KNN does the bulk of computations at the prediction time, it has maximum score time. The GNB model is the quickest among the three classifiers as far as the score time is concerned. The resource used for the experiment is a standard one and has an AMD CPU (4 cores) of 1.9Gz clock speed and 6GB of memory.

It is observed that some of the emotion pairs are hard to separate using the SEGRT, and as a result, the classification accuracy is low. For example, happy and fear in the EMODB is hard to separate with 56.12%, 68.14%, and 64.29% accuracies and nearly similar result 65.92%, 67.79%, and 59.95% in RAVDESS using KNN, SVM, and GNB respectively. This observation supports the fact that some emotions are hard to separate from speech utterance only and even the human ear can be deceived in those scenarios, and research results [109, 110] also support the fact.

Table 4.4.1: Computation Time

	Average	Average	Average Classification
	Training	Classification	Time for individual
	Time (ms)	Time (ms)	Data Points (ms)
KNN	1.698	3.957	0.103
SVM	16.113	1.859	0.049
GNB	1.403	0.553	0.014

Compared to existing research findings [160, 246, 247, 253], the classification performance of the proposed model is very promising. Here the average classification accuracy achieved is more than 70%, and in a few cases, it is more than 90% whereas similar researches [160, 247] have achieved 63.89% and 57.91% maximum accuracy respectively. However, the proposed model is not precisely comparable to [247] and [160] since they considered multi-class classification. The SEGRT can address the high dimensionality issue of the SER because it consists of 77 features, which is much smaller than the recent research works reported [160, 247, 251]. Computation time and processing power required is also significantly low compared to the DL architectures presently in use for the SER since the training process needs negligible time and resource in comparison.

4.5 Conclusion

In this chapter, a novel feature set (SEGRT) for SER is proposed and used it with prominent classification techniques to establish that the feature set can produce better results than existing feature sets. The specialty of SEGRT is that it is designed specifically for the SER, which is not proposed earlier. Three classification techniques with different working principles are selected because it is required to verify that the feature set is responding well when different classification techniques are applied so that the novelty of the feature set can be established. From the discussion in Section-4.4, it can be concluded that the proposed method could be able to address four critical issues of SER: accuracy, high-dimensionality, high computation time, and high-end computation resources.

For future work, the feature set will be further fine-tuned so that the feature extraction time could be reduced and the classification performance can improve further. New attributes could be added to the feature set to make it robust. Presently researchers are applying Deep Learning (DL) architectures for SER, so, this feature set can also be tested with DL classification techniques in the future.

Understanding of life begins with the understanding of patterns.

Fritjof Capra, Theoretical Physicist

5 Speech Emotion Recognition using Neural Network and Wavelet Features

5.1 Introduction

Recent advancements of Automatic Speech Recognition (ASR) has made a significant impact on the Human Machine Interaction (HCI) systems by making it possible for us humans to strike nearly natural spoken conversations with the machines (e.g. the robots like Sophia, and Erica). However, are those conversations natural? The answer is 'no' because an essential factor called 'emotion' is missing in those conversations. Human speech is the glottal wave generated due to the vibration of the vocal folds and speech signals get affected by emotion states of the speaker. Speech emotion recognition (SER) is a field of study where methods are being developed to extract human emotions concealed in the speech signals so that machines can understand our emotions from the speech itself. However, after more than 20 years of research, a satisfactory level of accuracy is not yet achieved.

SER study classifies different emotions concealed in the speech utterances by applying classification techniques from the Machine Learning (ML) field. Speech signal, like other signals, contains various types of information which needs to be extracted as features for further processing. In SER, those extracted features play a significant role in the performance of the SER systems since the selection of features eventually affects classification performance. Different speech features and combinations of those have been used by researchers so far, but there is no consensus on a specific feature set that can be considered best. As a result, a wide range of features are being used for SER, and researchers are proposing new feature sets.

Existing speech features are categorized into five groups based on the previous work of Ayadi et al. [93] (see figure.5.1.1). Researchers have tried most of the features in different combinations in SER systems for emotion classification. Some researchers [31, 32, 94] used continuous features and recommended it, whereas some other researchers [53, 160, 246, 247] recommend spectral features. However, the classification accuracy achieved so far with various feature-sets is not up to the mark and not ready for industrial use. So, Siri or Alexa is not ready to read the mood from our voices.

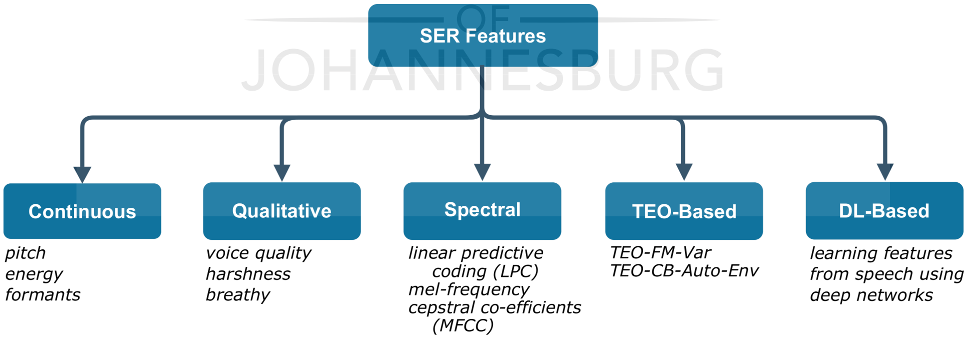

Figure 5.1.1: Category of features used for the SER.

Proper selection of the classification model is an important phase in SER systems because it will be selected based on the feature set. While it is required to maintain the sequence of the speech signals then it is required to deploy classification techniques like Hidden Markov Model (HMM) [118, 248] or Long Short-Term Memory (LSTM) based Deep Learning (DL) [160, 247] methods. Researchers also deployed the Gaussian Mixture Model (GMM) [53], and Support Vector Machine (SVM) [246, 249]. Following the research works, it is observed that finding suitable classification technique was the main focus of SER researchers, and less priority was given to develop SER specific feature sets. Interestingly most of the standard feature sets used so far are successfully used in automatic speech recognition (ASR), either directly or derived. However, ASR and SER should not be considered as a similar problem because a human can express their emotions in many ways, and they could have a unique sequence of events. Thus, in SER, there are no grammars like in case of ASR. So, those speech characteristics should be considered as features which can represent the emotions concealed in human speech rather than considering it as a sequence. ASR features are very good at tracking the variations in speech properties, but SER requires features which can represent the emotional states of the speakers well. Researchers also mentioned the need to divert from regular ASR features for SER [118, 250, 251, 254].

In this work, a new feature set used to overcome the difficulties faced in SER feature selection. Efforts have been made to find those characteristics of speech utterances, which could be able to represent the emotional content more prominently. Experiences in different emotional states of human are very speaker specific, which can be expressed uniquely by different speakers. The new feature set tries to overcome this specification by extracting the differences of different emotional states from the corresponding speaker's neutral state. With the application of discrete wavelet transform (DWT) and dissimilarity measure, this new feature set is developed.

Artificial Neural Network (ANN) is used for the classification task in this work. Since the feature set is not sequential, a simple ANN architecture can be deployed to demonstrate that the feature set is competent enough and no DL architecture is required to achieve comparable classification accuracy. Also, ANN results are compared with other prevalent classification techniques like Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Naive Bayes (NB) to establish the relevance of the feature set.

5.2 Description of the proposed SER model

SER involves many stages which are shown in figure 5.2.1. In the initial stage, speech data-set needs to be acquired, which is recorded to capture different human emotions. The speech recordings need to be processed to suit the actual processing. Then possible features are selected and if required, some features are engineered according to the requirement. So based on the feature selection scheme, features are extracted. In the next step, the classification technique is selected based on the feature set, and finally, classification is performed to identify the emotional state of the speaker. The feature set and the classification technique are discussed in detail in the following sections.

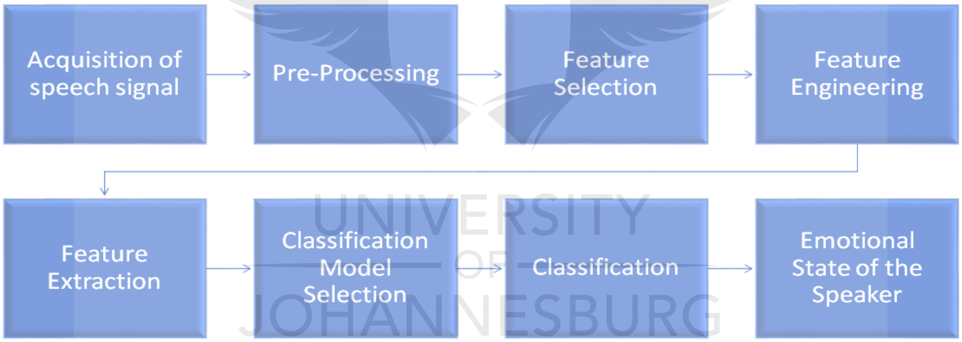
Acquisition of speech signal
Pre-Processing
Feature Selection
Feature Engineering
Feature Extraction
Classification Model Selection
Classification
Emotional State of the Speaker

Figure 5.2.1: Steps involved in a SER system.

5.2.1 Feature set description

This work is based on a novel feature set which is specifically developed for the SER. That feature set is described in detail in this section. A discrete speech signal S, which is defined as follows [236]:

$$\begin{gathered} S = \{s_n\}_{n \in \mathbb{N}}, \quad where\ \{s_n\} = \{s_1, s_2, ..., s_n\}, \\ \{s_n\} \in \mathbb{R} \end{gathered} \tag{5.1}$$

and assume N to be the length of S.

The signal in eq.5.1 is decomposed into trend and fluctuation sub-signals using discrete wavelet transform (DWT) [231]. Five Daubechies wavelet transforms *db*6, *db*8, *db*10, *db*12, and *db*14 till level 4 are used for this work. The transformation for level 1 can be described as a mapping $S \longmapsto (t1|f1)$ where $t1$ and $f1$ are the 1-level trend and fluctuation sub-signals respectively whose length is half of the length of S i.e. $N/2$. At level-2 of DWT the trend sub-signal of level-1 i.e. $t1$ is further broken down into trend and fluctuation signals so that $t1 \longmapsto (t2|f2)$. Level-2 transformation is defined as $S \longmapsto (t2|f2|f1)$ where $t2$ and $f2$ are level-2 trend and fluctuation signals respectively with length $N/4$. So, in similar way at level-4 the DWT is defined as ([231])

$$S \longmapsto (t4|f4|f3|f2|f1) \tag{5.2}$$

where $t4$ and $f4$ are level-4 trend and fluctuation sub-signals respectively with length $N/16$ and $f3$ is the level-3 fluctuation sub-signal with length $N/8$. Thus, from each 4-level DWT of S, five sub-signals are retrieved and from all the five DWTs we have considered i.e. *db*6, *db*8, *db*10, *db*12, and *db*14 we will get 25 sub-signals.

In the next step, Information Entropy (IE) of the trend and fluctuation signals are computed. Information Entropy (IE) is an expected measure of the information content in a signal. IE is low when there is ordered activity (like sine waves), and entropy is high when there is random activity [252]. IE of a sequence Q of length n is defined as (see. Kullback [242]) $E(Q) = \sum_{i=1}^{n} p(q_i) log_{10} p(q_i)$ where $Q = \{q_1, q_2, ..., q_n\}$ is a set of random phenomena, and $p(q_i)$ is the probability of a random phenomenon q_i. A set of IE values for each frame of the sub-signal are retrieved as the entropy sequence for that sub-signal [231].

$$t4 \mapsto \{te4_k\}, fj \mapsto \{fej_k\} \tag{5.3}$$

where $j = \{1, 2, 3\}$, m is the number of frames into which the sub-signals are broken down, $k = \{1, 2, ..., m\}$, and $te4_k$, as well as fej_k are the entropy values at the kth frame of the trend and fluctuation sub-signals respectively. Figure 5.2.2

shows how IE of happy and angry utterances vary around neutral state for trend sub-signal.

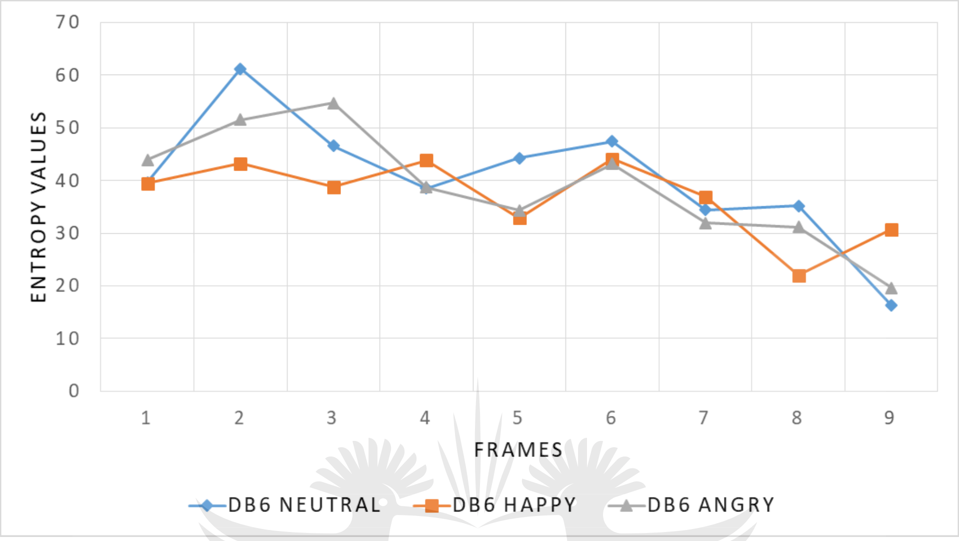

Figure 5.2.2: The figure shows how IE values of 4-level *db6 trend* sub-signal for Happy and Angry emotions are varying around the Neutral state.

The feature set derivation takes a unique approach by computing the similarity between IE values of each emotional states of a speaker from the corresponding neutral state of the same speaker. Suppose, features are being extracted from an angry emotional state utterance of the speaker S1. For that, the neutral and angry state utterances of speaker S1 are first broken down till IE sequences (similar to eq.5.3). So, for both angry and neutral utterances, there will be 25 IE sequences each. Then, the Euclidean distance is computed between every 25 sequences of neutral and angry utterances. So, there are now 25 distance measures.

There are two more attributes computed, one is covariance, and another is the correlation coefficient of the IE sequences, each giving 25 values making it 50 more values. Apart from 75 values, the gender of a speaker and the length ratio between neutral and angry utterance is also considered, which makes the total number of features 77.

5.2.2 Classification Technique

The classification technique used in this system is a Feed-forward Neural Network (FNN). FNNs are quintessential deep neural networks. Figure.5.2.3 depicts the functioning of an FNN. The hidden layers transform the input data in the search for a mapping function, and finally, the predicted results are produced. Gradients of the loss function provide essential input to the optimizer to update the weights further to find the best possible weights.

The architecture of the neural network (NN) [255] is described in this section. The input layer dimension is 77 because there are 77 features, so, each data point will have 77 dimensions. The input vector X is defined as $X \in \mathbb{R}^{77\times 1}$. Then there is a single hidden layer with 100 nodes. Thus, the weight matrix W^h of the hidden layer is defined as $W^h \in \mathbb{R}^{77\times 100}$. So, the input to the hidden layer I^h can be defined as

$$I^h = (W^h)^T X, \ \ where\ I^h \in \mathbb{R}^{100\times 1} \tag{5.4}$$

The input to hidden layer I^h now needs to be transformed using a non-linear function called the activation function. This activation function introduces non-linearity to the model so that important patterns can be identified to classify the data-points.

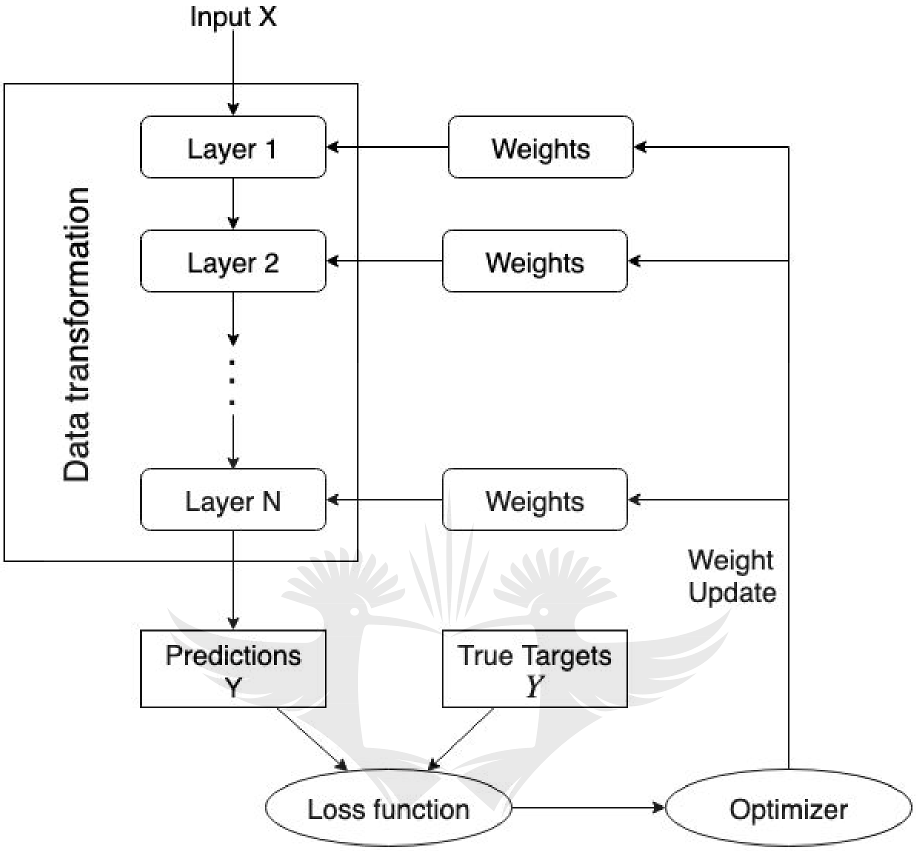

Figure 5.2.3: Schematic diagram of a feed-forward neural network.

The hidden layer activation function for this model is chosen to be Rectifier Linear Unit (ReLU) (see fig.5.2.4) which is mathematically defined as

$$ReLU(x) = max(0, x)$$

ReLU is a simple yet effective activation function which is widely used in various NN and Deep Neural Network architectures.

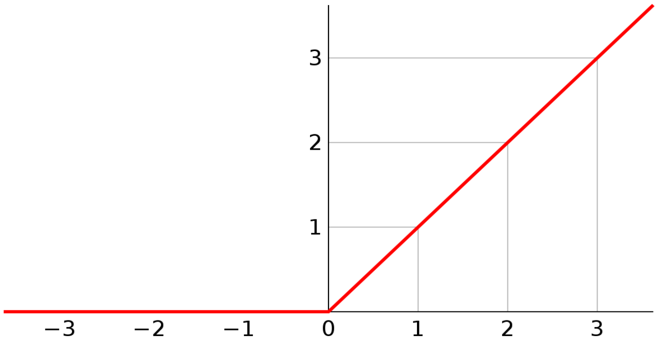

Figure 5.2.4: Shape of a Rectifier Linear Unit (ReLU) function

So, the output of the hidden layer (L^h) is then the transformation of I^h in eq.5.4 by ReLU function:

$$L^h = relu(I^h), \ where \ L^h \in \mathbb{R}^{100\times1} \tag{5.5}$$

The weight matrix for the output layer (W^o) will be defined as $W^o \in \mathbb{R}^{100\times1}$, and the input to the output layer (I^o) can be defined as

$$I^o = (W^o)^T L^h, \ where \ I^o \in \mathbb{R} \tag{5.6}$$

The input to the output layer I^o again needs to be transformed with a non-linear function to get the predicted output of the model, and the probability value is required for further verification with actual labels.

In the model, the output layer activation function is Sigmoid (see fig.5.2.5) which is mathematically defined as [199]

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

This is a widely used for output layers in binary classification problems because it gives results which can be considered as a probability that is within the possible range of probability values.

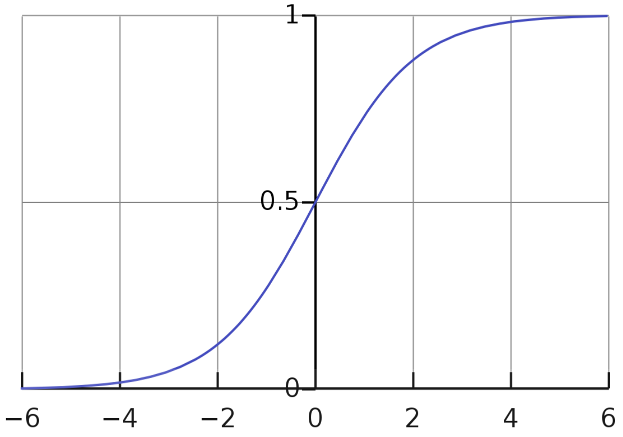

Figure 5.2.5: Shape of a sigmoid function

So, the final output of the output layer (L^o) is defined as follows ([189]):

$$L^o = sigmoid(I^o), \ \ where\ L^o \in \mathbb{R} \tag{5.7}$$

Now, the desired output for the output layer (Y) is the corresponding label of the data-point. Here the labels are discretized to be either 0 or 1. So, the actual deviation from expected result needs to be computed and here comes the concept of cost functions, which measures the deviation. Here, the binary cross entropy or negative log-loss function is used as a cost function (C) defined as follows ([199]):

$$C(L^o, Y) = -\sum_i L_i^o \ log(Y_i) \tag{5.8}$$

Next, it is required to measure how sensitive is the cost function (C) with respect to the weights. This is required because weights are the only components of the system which could be tweaked to get the best classification prediction. So, the partial derivative of C with respect to the weights W^o i.e. $\partial C/\partial W^o$ is computed. This partial derivative can be further broken down based on the chain rule as follows ([190]):

$$\frac{\partial C}{\partial W^o} = \frac{\partial C}{\partial L^o} \times \frac{\partial L^o}{\partial I^o} \times \frac{\partial I^o}{\partial W^o} \tag{5.9}$$

where $\partial I^o / \partial W^o = L^h$, $\partial L^o / \partial I^o = sigmoid'(I^o)$ and $\partial C / \partial L^o = L^o - Y$

So, eq.5.9 shows that some small change in W^o will affect the I^o which in turn will affect the L^o and eventually the cost C gets affected. So, $\partial C / \partial W^o$ is dependent on all the weights of the hidden layer as well because I^o is a function of L^h which in turn depends on W^h. So, in this way, the impacts of the previous layer weights on the cost function can be extracted by applying the chain rule and this is termed as the *error back-propagation*. Based on the propagated errors, weights are updated to check in next iteration how much the C is changing and whether C is reducing since the objective is to *minimize* C.

All the steps described so far in this section was for a single data point to track the error in prediction and update the weights accordingly towards achieving cost minimization. When the whole dataset is considered for optimization, there is a strong need for a standard algorithm rather than doing the whole process on an ad-hoc basis. That is why a popular optimization technique specialized for NN called *adam* is used to update the weights based on the errors propagated backward. The *adam* is a gradient-descent based optimization method which uses *stochastic gradient descent*.

Three widely used classification techniques the Support Vector Classifier (SVC) with Radial Basis Function (RBF) kernel, Gaussian Naive Bayes (GNB) and K-Nearest Neighbor (KNN) are also tested on the feature set for comparing NN classification results. These three classifiers take different approaches. The SVC is based on the principle of finding the separating hyperplane. The GNB is based on Bayes theorem using the concepts of posterior probability and likelihood. The KNN computes distance or similarity and majority voting to classify different classes.

5.3 Experiment

For the experiment, an English language dataset called Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [128] is used. Seven emotional states are considered for this work and short-codes for the emotions are used throughout this chapter and these are clam(CA), happy(HA), sad(SA), angry(AN), fear(FE), disgust(DI), and surprise(SU).

The problem is approached as a binary classification problem by considering two emotion labels at a time. Data has been split into training, testing, and validation data sets using stratification with 10 splits. Stratification split is used to reduce the extent of over-fitting in the classification model, and moreover, the stratification split is done in such a way that each fold can represent the whole data set. The folds are used for cross-validation to measure the performance of the classifier.

Figure 5.3.1 shows the ANN model used for this work. The model has 77 input nodes, 100 hidden nodes, and one output node.

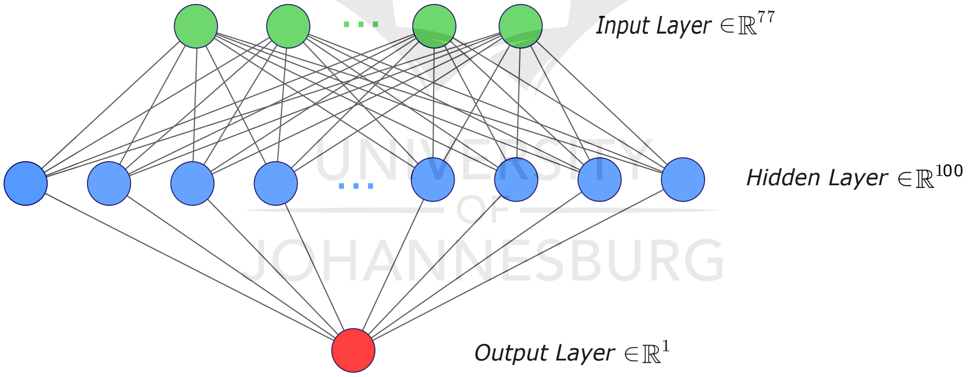
...
Input Layer $\in \mathbb{R}^{77}$
...
Hidden Layer $\in \mathbb{R}^{100}$
Output Layer $\in \mathbb{R}^{1}$

Figure 5.3.1: Design of the neural network.

The design is explained in detail in Section 5.2.2. Here, some of the hyper-parameters of the model are explained. First of all, the *epoch* is set to 100, that is the model sees the whole dataset 100 times. Next, the *batch size* is 40 so that after every 40 records the weights get updated. How the model loss and accuracy has converged with each epoch both for training, and testing can be observed in fig.5.3.2. The *tensorflow* [203] deep learning framework is used as back-end while the front-end library was *keras* [256].

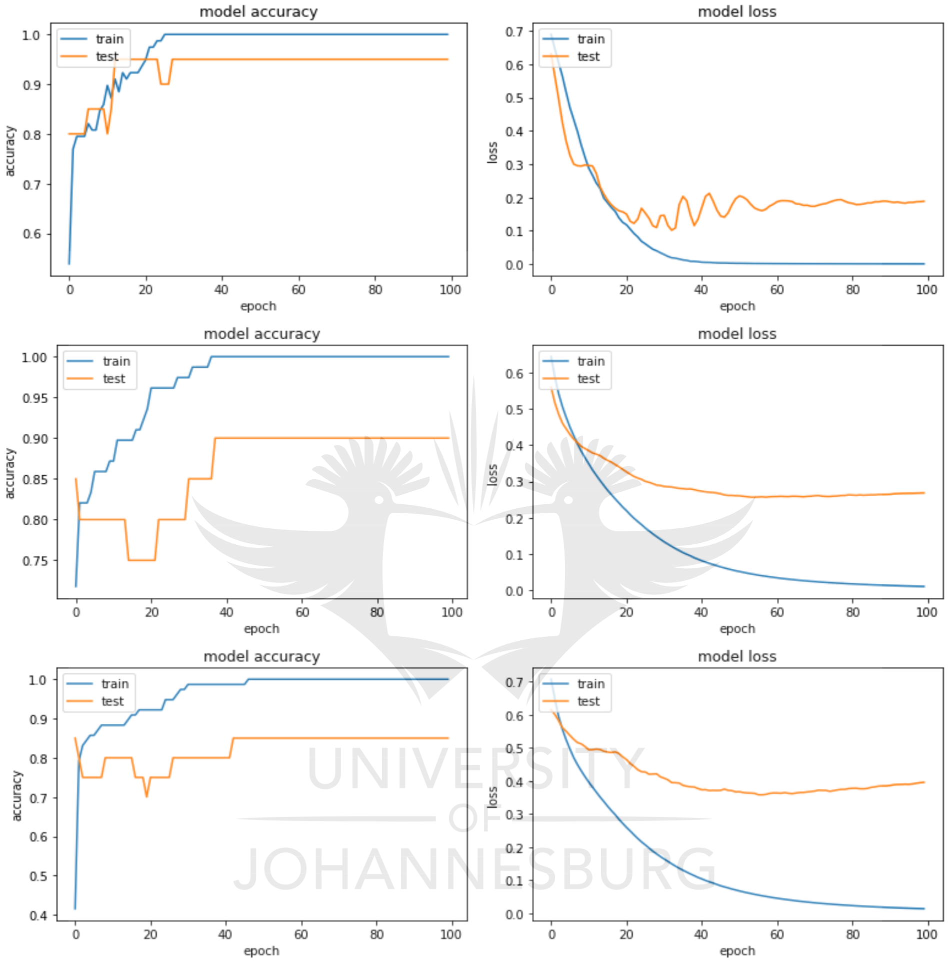

Figure 5.3.2: The figure shows how model accuracy and loss changes with each epoch for testing and training.

The other three classifiers also have some hyper-parameters which need to be mentioned here. For KNN the distance metric is chosen as $neighbors = 5$ and *Manhattan*. For SVC with RBF, kernel hyper-parameters are selected as $C = 10$ and $gamma = 0.001$. GNB is a straightforward model and does not have hyper-parameters. Proper selection of hyper-parameters is vital for getting optimal results from the classification models.

5.4 Results and Discussions

The classification accuracy of the model is shown in fig.5.4.1. The average accuracy achieved is more than 80% while some accuracies cross 90%. The results are compared with three well-known classification methods: SVC, GNB, and KNN, and the results are shown in fig.5.4.2. These results show that the NN model performs better than other techniques.

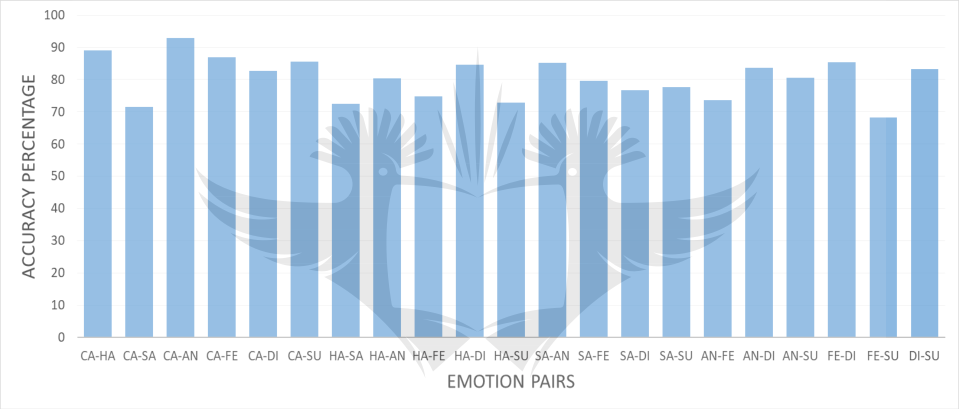

Figure 5.4.1: ANN classification accuracy.

It is observed that some of the emotion pairs are hard to separate using this feature set, and as a result, the classification accuracy is low. For example, happy and fear is hard to separate with 74.83%, 65.92%, 67.79%, and 59.95% using ANN, KNN, SVM, and GNB, respectively. Research results also support the fact that human ear can get deceived sometimes in separating emotions from speech [109, 110].

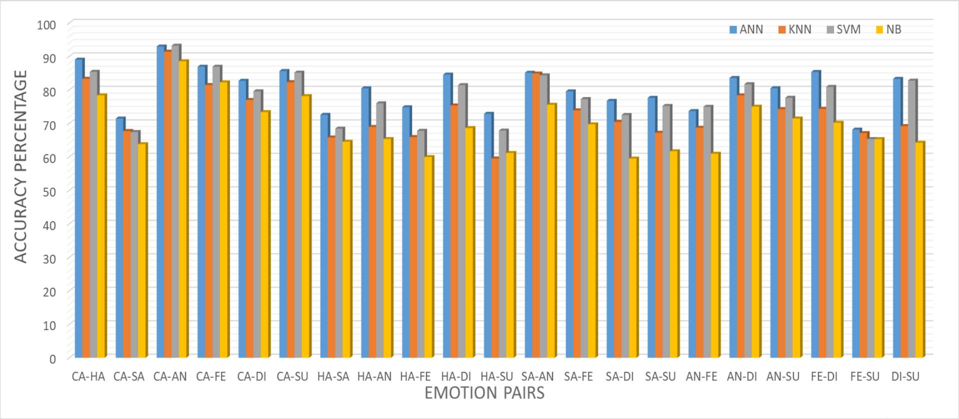

Figure 5.4.2: Accuracy comparison between ANN, SVM, KNN, and GNB.

The proposed model demonstrates promising results compared to contemporary research results [160, 246, 247, 253]. The average accuracy achieved is more than 80%, and in some cases, it exceeded 90% whereas [160, 247, 251] achieved 63.89% and 57.91% maximum accuracy. Very high feature space dimension is a real challenge while applying the DL architectures for SER and the number of features sometimes run into thousands [160, 163]. The proposed model can address this high dimensionality issue by reducing the number of features to 77. Finally, computation and processing capability was much less when compared to deep learning models. The resources used for this work are AMD 1.9Gz CPU (4 cores) processor and 6GB of memory.

5.5 Conclusion

A novel classification model for SER is proposed in this work. Application of neural network model using the new wavelet-based feature set produced promising results compared to contemporary research results. The NN model developed here is a simple one to show the novelty of the feature set that the feature set can perform reasonably well even with simple models and no complicated DL architecture is required. Applying DL architecture should improve classification accuracy even more. The proposed model is capable of addressing high-dimensionality, high computation time, and resource issues of SER along with high accuracy.

There is a plan to apply deep learning architecture to improve the classification accuracy even further so that the industrial implementation of this model is possible in the near future. There is plenty of scopes left to enhance the feature set further by adding new features and possibly by introducing generative methods to create a good volume of artificially created data to carry this trend of SER research further for better results.

Modern physics had shown that the rhythm of creation and destruction is not only manifest in the turn of the seasons and the birth and death of living creatures, but is also the very essence of inorganic matter. For modern physicists, Shiva's dance is the dance of subatomic matter.

Fritjof Capra, Theoretical Physicist

6 Conclusion and Future Directions

6.1 Conclusion

Researchers have identified and proposed various applications of the technique of recognizing emotional states from the speech signals [52, 61, 73, 74]. Introduction of sensitivity to AI systems through emotion detection is going to revolutionize many sectors like medical, and entertainment. Thus, SER is important, and it is going to be an integral part of AI systems. That is the reason why there is a significant thrust towards solving the problem of SER in the last few years. Although the task of SER has proved to be challenging, the enthusiasm of the researchers is increasing in this endeavor and an increasing number of research papers support this. The problems in this field are old, and numerous proposals are available. However, both the problems are still unsolved. This thesis is an endeavor towards solving a few problems in this field. Chapter 1 of this thesis first explains the reason behind the influence of emotions in human life and why emotion detection is essential. Then there is a very brief overview of an SER and including the various application of SER. After that, the motivations behind this research has been explained, and it is pointed out that the lack of well-defined feature set for the SER has been the biggest motivation behind this work. The contributions are also mentioned in this chapter in brief.

Chapter 2 of this thesis reported the present state of SER research and concluded that cross-corpus classification accuracy is significantly low, which means that present models are not generalizing well for the unseen data. This makes the present classification models unfit for industry use and this is the reason why SER systems have not yet been incorporated in various consumer applications. Low accuracy in cross-corpus classification scenario could be because of three reasons: 1. the database parameters and attributes vary significantly. 2. the features extracted from the speech databases are not adequately reflecting the characteristics of emotions. 3. a combination of both the previous reasons. During this work also, it is observed that the classifiers are not responsible for low generalization of the SER models, because powerful classification techniques like the deep learning models performed very well with complex problems like image classification or automatic speech recognition, but the DL models are not that successful in solving the SER problem. This research is an endeavor towards solving the two main issues with SER which have been identified during the initial study: 1. speech endpoint detection in pre-processing stage is not robust enough to remove the endpoints precisely; 2. existing features are not describing the emotional states of the speech signals well.

Chapter 3 proposes a novel speech end point detection algorithm named the WCSED. Here a wavelet convolution based approach is adopted to remove the endpoints from speech utterances. The WCSED produced very good results compared to contemporary research results. Precise endpoint detection contributes significantly to SER or other speech processing systems. So, this algorithm is a vital contribution towards successful speech emotion recognition.

Deficiency of the SER feature sets in describing the underlying emotional state of the speaker has been identified as the most critical issue. A novel feature set for SER is proposed in Chapter 4. Researchers have tried different feature sets as described in Chapter 2, but most of them are borrowed from the ASR features. Which means the features are not specifically designed for SER. So, an SER specific feature set is required for the SER models to perform well. The feature set is based on a novel concept named the *subjective emotional gap reduction technique* (SEGRT). The SEGRT is based on the discrete wavelet transform method. The SEGRT is unique in the sense that for the first time a dissimilarity measure is calculated between a neutral state and an emotional state of the speaker. It is mentioned in Chapter 1 that the subjective nature of emotional expression makes it difficult to segregate different emotional states. The dissimilarity measure tried to mitigate this *subjective gap* so that the extracted features for different speakers can be put into nearly similar scale. The proposed feature set is used with different classifiers like GNB, SVM, and KNN, and the results are very encouraging. The concept of the SEGRT is a novel concept, first introduced through this work. This concept is new and need further investigation and improvement to establish it as a new standard for SER feature set. SEGRT is capable of addressing four critical issues of SER: classification accuracy, high-dimensionality, high computation time, and high-end computation resources.

The fifth chapter proposes a *deep learning* classification model for SER using the SEGRT feature set. A single hidden layer *feed-forward neural network* (FNN) model is used and the results are promising compared to contemporary research results. This neural network model uses the cross-entropy loss function as a cost function. Which means entropy value between the distribution of target emotions and predicted emotions are used as a measure to evaluate the performance of the neural network. Here low cross-entropy value means better fit but very low cross-entropy would mean an overfit. The subjective gap of different emotions from the corresponding neutral state at different wavelet decomposed layers is weighted with the respective weight of the neural network model. So, for different emotions, different wavelet transform level is given importance depending on the frequency of the speaker's voice. Thus, the performance of this model will depend on how aptly the dataset depicts the subjective gap. Moreover, the stratification split of the training data helps to prevent the cross-entropy values to fall very low by randomly selecting wavelet decomposed signal from different levels. The FNN results are better compared to other classifiers like GNB, SVM, and KNN. However, the training time using the FNN is high because the model is trained in 100 *epochs* with 40 batch size. The SEGRT feature set contains 77 attributes which is much lower than regular feature sets and so, it is capable of addressing the high-dimensionality issue of the SER.

6.2 Future Directions

The proposed WCSED algorithm demonstrated encouraging results and can be further tested with other speech processing systems as well. Some scope of improvement for the WCSED algorithm have been identified as: 1. level of loudness of the speaker's utterances could be an essential factor to improve the end-point selection results; 2. WCSED could be further adapted to work in the online or real-time scenario to enhance the scope and usability of the algorithm.

The proposed SEGRT feature set should be made more robust by further reducing redundant attributes and adding more relevant attributes. There is a scope to apply deep learning architecture to improve the classification accuracy further so that the industrial implementation of this model can be developed in the near future. A generative model can be developed to generate a good volume of artificially created data based on the SEGRT so that there is more relevant data available for research.

Research and study of a subject can never be complete unless the proposed contributions get implemented to solve real-life problems for humankind. So, my constant endeavor will also be to see those proposals to be implemented in real-life.

Appendix A

Publications

Journals

1. T Roy, T Marwala, and S Chakraverty. Precise detection of speech endpoints dynamically: A wavelet convolution based approach. *Communications in Nonlinear Science and Numerical Simulation*, 2018. doi:https://doi.org/10.1016/j.cnsns.2018.07.008.

Conferences

1. T. Roy, T. Marwala, and S. Chakraverty. Introducing New Feature Set based on Wavelets for Speech Emotion Classification. *1st IEEE Conference on Applied Signal Processing (ASPCON) 2018*. Accepted for IEEE Conference proceedings book.

2. T. Roy, T. Marwala, and S. Chakraverty. Speech Emotion Recognition using Neural Network and Wavelet Features. *8th Wave Mechanics Vibrations Conference 2018, NIT RKL*. Accepted for Springer Conference proceedings book.

Book Chapters

1. T. Roy, T. Marwala, and S. Chakraverty. Novel Advancements of Automatic Emotion Recognition and its Role in the 4th Industrial Revolution. Accepted In The Disruptive Fourth Industrial Revolution: Technology, Society and Beyond, Edited by T Marwala, BS Paul. Springer.

2. T. Roy, T. Marwala, and S. Chakraverty. A Survey of Classification Techniques in Speech Emotion Recognition. Accepted In Mathematical Methods in Interdisciplinary Sciences, Edited by S Chakraverty. Wiley.

3. T. Roy, T. Marwala, and S. Chakraverty. Deep Learning in Speech Emotion Recognition: A Review. Proposed In Mathematical Methods and Vibrations, Edited by S Chakraverty. Elsevier.

References

[1] Debra Trampe, Jordi Quoidbach, and Maxime Taquet. Emotions in everyday life. *PLOS ONE*, 10(12):1–15, 12 2015. doi: 10.1371/journal.pone.0145450. URL https://doi.org/10.1371/journal.pone.0145450.

[2] G. L. Clore and J. R. Huntsinger. How emotions inform judgment and regulate thought. *Trends in cognitive sciences*, 11(9):393–9, 2007.

[3] J Williamson. Speech analyzer for analyzing pitch or frequency perturbations in individual speech pattern to determine the emotional state of the person, 1978.

[4] C Darwin. The expression of emotion in man and animals. *Watts*, 1948.

[5] D.G. Freedman, C.B. Loring, and R.M. Martin. *Emotional behavior and personality development. In: Infancy and Early Childhood*. New York: Free Press, 1967.

[6] Robert H. Frank. *Passions Within Reason: The Strategic Role of Emotions*. Norton, 1988.

[7] A Damasio. *Descartes' Error*. New York: Grosset/Putnam, 1994.

[8] Paul D MacLean and V. A Kral. A triune concept of the brain and behaviour. Toronto ; Buffalo] : published for the Ontario Mental Health Foundation by University of Toronto Press, 1973. ISBN 0802032990. Includes bibliographies.

[9] Douglas S. Massey. A brief history of human society: The origin and role of emotion in social life: 2001 presidential address. *American Sociological Review*, 67(1):1–29, 2002. ISSN 00031224. URL http://www.jstor.org/stable/3088931.

[10] Rita Carter. *Mapping the Mind*. University of California Press, Berkeley, CA, 1998.

[11] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press, New York, 1998.

[12] Joseph LeDoux. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon and Schuster, New York, 1996.

[13] C. M. Tyng, H. U. Amin, M. Saad, and A. S. Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 2017. doi: 10.3389/fpsyg.2017.01454.

[14] James Marcum. The role of emotions in clinical reasoning and decision making. *The Journal of medicine and philosophy*, 38, 08 2013. doi: 10.1093/jmp/jht040.

[15] Bassem Maamari and Joelle Majdalani. Emotional intelligence, leadership style & organizational climate. *International Journal of Organizational Analysis*, 25, 05 2017. doi: 10.1108/IJOA-04-2016-1010.

[16] Hyun Jung Lee. How emotional intelligence relates to job satisfaction and burnout in public service jobs. *International Review of Administrative Sciences*, 84(4):729–745, 2018. doi: 10.1177/0020852316670489.

[17] Peter J Gianaros, Anna L Marsland, Dora C.-H Kuan, Brittney L Schirda, J. Richard Jennings, Lei K Sheu, Ahmad R Hariri, James J Gross, and Stephen B Manuck. An inflammatory pathway links atherosclerotic cardiovascular disease risk to neural activity evoked by the cognitive regulation of emotion. *Biological Psychiatry*, 75:738–745, 2014. doi: doi: 10.1016/j.biopsych.2013.10.012.

[18] B. L. Fredrickson, R. A. Mancuso, C. Branigan, and M. M. Tugade. The undoing effect of positive emotions. *Motivation and emotion*, 24(4):237–258, 2000.

[19] Marion McAllister, Linda Davies, Katherine Payne, Stuart Nicholls, Dian Donnai, and Rhona MacLeod. The emotional effects of genetic diseases: Implications for clinical genetics. *American Journal of Medical Genetics Part A*, 143A(22):2651–2661, 2007. doi: 10.1002/ajmg.a.32013. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.32013.

[20] Tamar L. Ben-Shaanan, Maya Schiller, Hilla Azulay-Debby, Ben Korin, Nadia Boshnak, Tamar Koren, Maria Krot, Jivan Shakya, Michal A. Rahat, Fahed Hakim, and Asya Rolls. Modulation of anti-tumor immunity by the brain's reward system. *Nature Communications*, 9(1), 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05283-5.

[21] Michael H. Antoni, Susan K. Lutgendorf, Steven W. Cole, Firdaus S. Dhabhar, Sandra E. Sephton, Paige Green McDonald, Michael Stefanek, and Anil K. Sood. The influence of bio-behavioural factors on tumour biology: pathways and mechanisms. *Nature Reviews Cancer*, 6:240–248, Mar 2006. URL https://doi.org/10.1038/nrc1820.

[22] Yoichi Chida, Mark Hamer, Jane Wardle, and Andrew Steptoe. Do stress-related psychosocial factors contribute to cancer incidence and survival? *Nature Clinical Practice Oncology*, 5:466–475, May 2008. URL https://doi.org/10.1038/ncponc1134. Review Article.

[23] Jon Parsons and Nicholas Marcer. *Osteopathy: Models for Diagnosis, Treatment and Practice*. Elsevier Health Sciences, 2005.

[24] Nico H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[25] K.R Scherer. Approaches to Emotion, ed. K.R. Scherer and P. Ekman, chapter On the nature and function of emotion: A component process approach., pages 293–318. Hillsdale, NJ: Erlbaum, 1984.

[26] Tshilidzi Marwala and Evan Hurwitz. Artificial intelligence and asymmetric information theory. *CoRR*, abs/1510.02867, 2015. URL http://arxiv.org/abs/1510.02867.

[27] Dimitrios Ververidis and Constantine Kotropoulos. A state of the art review on emotional speech databases. In *In Proceedings of the 1st Richmedia Conference*, pages 109–119, 2003.

[28] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. In *ICONIP'99. ANZIIS'99 ANNES'99 ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378)*, volume 2, pages 495–501 vol.2, Nov 1999. doi: 10.1109/ICONIP.1999.845644.

[29] K. Sreenivasa Rao and Shashidhar G. Koolagudi. *Emotion Recognition using Speech Features*. Springer-Verlag New York, 2013.

[30] Gangamohan P., Kadiri S.R., and Yegnanarayana B. Analysis of emotional speech — a review. In *Toward Robotic Socially Believable Behaving Systems - Volume I*, Intelligent Systems Reference Library, pages 205–238. Springer, Cham, 2016. doi: 10.1007/978-3-319-31056-5_11.

[31] Louis Bosch. Emotions, speech and the asr framework. *Speech Communication*, 40(1):213–225, 04 2003. doi: 10.1016/S0167-6393(02)00083-3.

[32] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17:582 – 596, 06 2009. doi: 10.1109/TASL.2008.2009578.

[33] Martin Borchert and A Dusterhoft. Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. pages 147 – 151, 01 2005. ISBN 0-7803-9361-9. doi: 10.1109/NLPKE.2005.1598724.

[34] Suzanne Beeke, Ray Wilkinson, and Jane Maxim. Prosody as a compensatory strategy in the conversations of people with agrammatism. *Clinical linguistics & phonetics*, 23:133–55, 03 2009. doi: 10.1080/02699200802602985.

[35] Iain Murray and John Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93:1097–108, 03 1993. doi: 10.1121/1.405558.

[36] Rainer Banse and Klaus Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70:614–36, 04 1996. doi: 10.1037/0022-3514.70.3.614.

[37] Roddy Cowie and E Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. pages 1989 – 1992 vol.3, 11 1996. doi: 10.1109/ICSLP.1996.608027.

[38] Klaus Scherer. Vocal affect expression. a review and a model for future research. *Psychological bulletin*, 99:143–65, 04 1986. doi: 10.1037//0033-2909.99.2.143.

[39] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall signal processing series. Prentice-Hall, 1978. ISBN 9780132136037. URL https://books.google.co.za/books?id=YVtTAAAAMAAJ.

[40] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-015157-2.

[41] Jr. John R. Deller, John H. L. Hansen, and John G. Proakis. *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 1999. ISBN 978-0-7803-5386-2.

[42] S. E. Bou-Ghazale and J. H. L. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442, July 2000. ISSN 1063-6676. doi: 10.1109/89.848224.

[43] M. Chen, X. He, J. Yang, and H. Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, Oct 2018. ISSN 1070-9908. doi: 10.1109/LSP.2018.2860246.

[44] D. Bitouk, R. Verma, and A. Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52:613–625, 2010.

[45] G. Zhou, J. H. L. Hansen, and J. F. Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Trans. Speech and Audio Processing*, 9: 201–216, 2001.

[46] Norhaslinda Kamaruddin and Abdul Wahab. Features extraction for speech emotion. *J. Comp. Methods in Sci. and Eng.*, 9(1,2S1):1–12, apr 2009. ISSN 1472-7978. URL http://dl.acm.org/citation.cfm?id=1608790.1608791.

[47] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. 01 2003.

[48] David Philippou-Hübner, Bogdan Vlasenko, Tobias Grosser, and Andreas Wendemuth. Determining optimal features for emotion recognition from speech by applying an evolutionary algorithm. In *INTERSPEECH*, pages 2358–2361, 01 2010.

[49] J. Lin, W. Wei, C. Wu, and H. Wang. Emotion recognition of conversational affective speech using temporal course modeling-based error weighted cross-correlation model. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–7, Dec 2014. doi: 10.1109/APSIPA.2014.7041621.

[50] J. H. Jeon, R. Xia, and Y. Liu. Sentence level emotion recognition based on decisions from subsentence segments. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4940–4943, May 2011. doi: 10.1109/ICASSP.2011.5947464.

[51] H. Atassi and A. Esposito. A speaker independent approach to the classification of emotional vocal expressions. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 147–152, Nov 2008. doi: 10.1109/ICTAI.2008.158.

[52] 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 *Proc. IEEE ICASSP*, volume 1, pages I–577–80 vol.1, 2004. doi: 10.1109/ICASSP.2004.1326051.

[53] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. Emotion recognition in spontaneous speech using gmms. In *Proc. INTERSPEECH*, 2006.

[54] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan. Speech emotion recognition using both spectral and prosodic features. In *2009 International Conference on Information Engineering and Computer Science*, pages 1–4, Dec 2009. doi: 10.1109/ICIECS.2009.5362730.

[55] Iker Luengo, Eva Navas, Inma Hernáez, and Jon Sánchez. Automatic emotion recognition using prosodic parameters. In *INTERSPEECH*, 2005.

[56] Rui Sun and Elliot Moore. A preliminary study on cross-databases emotion recognition using the glottal features in speech. In *INTERSPEECH*, 2012.

[57] Siqing Wu, Tiago H. Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768 – 785, 2011. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2010.08.013. URL http://www.sciencedirect.com/science/article/pii/S0167639310001470. Perceptual and Statistical Audition.

[58] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition based on phoneme classes. In *Proc. ICSLP*, pages 889–892, 2004.

[59] Oudeyer Pierre-Yves. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1):157 – 183, 2003. ISSN 1071-5819. doi: https://doi.org/10.1016/S1071-5819(02)00141-6. URL http://www.sciencedirect.com/science/article/pii/S1071581902001416. Applications of Affective Computing in Human-Computer Interaction.

[60] Ali Hassan and Robert I. Damper. Multi-class and hierarchical svms for emotion recognition. In *INTERSPEECH*, 2010.

[61] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proc. Fourth International Conference on Spoken Language*, volume 3, pages 1970–1973, Oct 1996. doi: 10.1109/ICSLP.1996.608022.

[62] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, and Wen-Yuan Liao. Combining acoustic features for improved emotion recognition in mandarin speech. In Jianhua Tao, Tieniu Tan, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 279–285, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32273-3.

[63] Yongjin Wang and Ling Guan. An investigation of speech-based human emotion recognition. In *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, pages 15–18, Sep. 2004.

[64] C. M. Lee, S. Narayanan, and R. Pieraccini. Recognition of negative emotions from the speech signal. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, pages 240–243, Dec 2001. doi: 10.1109/ASRU.2001.1034632.

[65] Aiqin Zhu and Qi Luo. Study on speech emotion recognition system in e-learning. In Julie A. Jacko, editor, *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, pages 544–552, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-73110-8.

[66] R. Nakatsu, A. Solomides, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 804–808 vol.2, June 1999. doi: 10.1109/MMCS.1999.778589.

[67] R. Fernandez and R. W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40:145–159, 2003.

[68] Srinivas Parthasarathy and Ivan Tashev. Convolutional neural network techniques for speech emotion recognition. September 2018.

[69] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, March 2016. doi: 10.1109/ICASSP.2016.7472669.

[70] S. Zhang, S. Zhang, T. Huang, and W. Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, June 2018. ISSN 1520-9210. doi: 10.1109/TMM.2017.2766843.

[71] Q. Mao, M. Dong, Z. Huang, and Y. Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, Dec 2014. ISSN 1520-9210. doi: 10.1109/TMM.2014.2360798.

[72] Michael Neumann and Ngoc Thang Vu. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *CoRR*, abs/1706.00612, 2017. URL http://arxiv.org/abs/1706.00612.

[73] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997. ISBN 0-262-16170-2.

[74] Stefanie Schelinski and Katharina von Kriegstein. The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development. *Journal of Autism and Developmental Disorders*, 49(1):68–82, Jan 2019. ISSN 1573-3432.

doi: 10.1007/s10803-018-3681-z. URL https://doi.org/10.1007/s10803-018-3681-z.

[75] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, July 2000. ISSN 0018-9294. doi: 10.1109/10.846676.

[76] Chul Min Lee and Shrikanth Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13:293–303, 2005.

[77] John H.L. Hansen and Douglas A. Cairns. Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments. *Speech Communication*, 16(4):391 – 422, 1995. ISSN 0167-6393. doi: https://doi.org/10.1016/0167-6393(95)00007-B. URL http://www.sciencedirect.com/science/article/pii/016763939500007B.

[78] Fred Charles, David Pizzi, Marc Cavazza, Thurid Vogt, and Elisabeth Andre. Emoemma: Emotional speech input for interactive storytelling. In *8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, volume 2, pages 1381–1382, 2009.

[79] S. Carroll. *The Particle at the End of the Universe: How the Hunt for the Higgs Boson Leads Us to the Edge of a New World*. Plume, 2013. ISBN 9780142180303. URL https://books.google.co.za/books?id=R-PYCwAAQBAJ.

[80] Meera Raghu. A study to explore the effects of sound vibrations on consciousness. *International Journal of Social Work and Human Services Practice*, 6(3):75–88, 2018. doi: 10.13189/ijrh.2018.060302.

[81] C.B. Pert. *Molecules of Emotion: The Science Behind Mind-Body Medicine*. Scribner, 1999. ISBN 9780684846347. URL https://books.google.co.za/books?id=T6VII6nulvEC.

[82] T. Roy, T. Marwala, and S. Chakraverty. Precise detection of speech endpoints dynamically: A wavelet convolution based approach. *Communications in Nonlinear Science and Numerical Simulation*, 2018. doi: https://doi.org/10.1016/j.cnsns.2018.07.008.

[83] T. Roy, T. Marwala, and S. Chakraverty. Introducing novel feature set based on wavelets for speech emotion classification. In *Proc. IEEE ASPCON*, 2018.

[84] T. Roy, T. Marwala, and S. Chakraverty. Speech emotion recognition using neural network and wavelet features. In *Proc. 8th Wave Mechanics and Vibrations Conference*, 2018.

[85] Grant Fairbanks and Wilbert Pronovost. Vocal pitch during simulated emotion. *Science*, 88(2286):382–383, 1938. ISSN 0036-8075. doi: 10.1126/science.88.2286.382. URL http://science.sciencemag.org/content/88/2286/382.

[86] Carl E. Williams and Kenneth N. Stevens. On Determining the Emotional State of Pilots During Flight: An Exploratory Study. *Aerospace Med*, 40: 1369–1372, dec 1969.

[87] Carl E. Williams and Kenneth N. Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52 (4B):1238–1250, 1972. doi: 10.1121/1.1913238.

[88] Klaus R. Scherer. Series in affective science., chapter Psychological models of emotion., pages 137–162. Oxford University Press, New York, NY, US, 2000. ISBN 0-19-511464-7.

[89] J Connor and G Arnold. *Intonation of Colloquial English*. Longman UK, 1973.

[90] M Schubiger. *English intonation: its form and function*. Niemeyer Germany, 1958.

[91] P Ekman. An argument for basic emotions. *Cogn. Emot.*, 6:169–200, 1992.

[92] Carroll E Izard. *The Psychology of Emotions*. Springer US, 1991.

[93] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[94] R Cowie, E Douglas-Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, and J G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001. doi: 10.1109/79.911197.

[95] R. Plutchik and H. Kellerman.

[96] Radoslaw Nielek, Miroslaw Ciastek, and Wiesław Kopeć. Emotions make cities live: towards mapping emotions of older adults on urban space. pages 1076–1079, 08 2017. doi: 10.1145/3106426.3109041.

[97] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005. doi: 10.1177/0539018405058216.

[98] Klaus R. Scherer, Vera Shuman, Johnny J. R. Fontaine, and Cristina Soriano. The grid meets the wheel: Assessing emotional feeling via self-report. In *Components of emotional meaning: A sourcebook.*, Series in affective science, pages 281–298. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199592746.003.0019.

[99] Tuomas Eerola and Jonna K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011. doi: 10.1177/0305735610362821.

[100] J.A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[101] J. Posner, J.A. Russell, and B.S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005.

[102] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 30:2067–2083, 01 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.26.

[103] H Scholsberg. Three dimensions of emotion. *Psychological Rev.*, 61:81–88, 1954.

[104] U. Schimmack and A. Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000.

[105] M. Gnjatovic and D. Rosner. Inducing genuine emotions in simulated speech-based human-machine interaction: The nimitek corpus. *IEEE Transactions on Affective Computing*, 1(2):132–144, July 2010. ISSN 1949-3045. doi: 10.1109/T-AFFC.2010.14.

[106] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162 – 1181, 2006. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2006.04.003. URL http://www.sciencedirect.com/science/article/pii/S0167639306000422.

[107] Cynthia Breazeal and Lijin Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12(1):83–104, Jan 2002. ISSN 1573-7527. doi: 10.1023/A:1013215010749. URL https://doi.org/10.1023/A:1013215010749.

[108] Nick Campbell. Databases of emotional speech, 01 2000.

[109] Inger S. Engberg, Anya V. Hansen, Ove Andersen, and Paul Dalsgaard. Design, recording and verification of a danish emotional speech database. In *Proc. 5th European Conference on Speech Communication and Technology*, 1997.

[110] Malcolm Slaney and Gerald McRoberts. Babyears: A recognition system for affective vocalizations. *Speech Communication*, 39:367–384, 2003.

[111] I. Engberg and A. Hansen. Documentation of the danish emotional speech database (des), 1996. URL http://kom.aau.dk/~tb/speech/Emotions/des.pdf.

[112] Noam Amir, Samuel Ron, and Nathaniel Laor. Analysis of an emotional speech corpus in hebrew based on objective criteria. In *Proc. ISCA Workshop (ITRW) Speech and Emotion*, 01 2000.

[113] H. Hu, M. Xu, and W. Wu. Dimensions of emotional meaning in speech. In *Proceedings of the ISCA ITRW on Speech and Emotion*, 2000.

[114] Vladimir Hozjan, Zdravko Kacic, Asunción Moreno, Antonio Bonafonte, and Albino Nogueiras. Interface databases: Design and collection of a multilingual emotional speech database. In *LREC*, 2002.

[115] B. Schuller. Towards intuitive speech interaction by the integration of emotional aspects. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 6, pages 6 pp. vol.6–, Oct 2002. doi: 10.1109/ICSMC.2002.1175635.

[116] Murray Grossman Nii Martey John Bell Mark Liberman, Kelly Davis. Emotional prosody speech and transcripts ldc2002s28, 2002.

[117] Raquel Tato, Rocío Santos, Ralf Kompe, and José Manuel Pardo. Emotional space improves emotion recognition. In *INTERSPEECH*, 2002.

[118] Tin Lay Nwe, Say Wei Foo, and Liyanage C. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, 2003.

[119] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Proc. 9th European Conference on Speech Communication and Technology*, volume 5, pages 1517–1520, 2005.

[120] Björn W. Schuller, Stephan Reiter, Ronald Müller, Marc Al-Hames, Manfred K. Lang, and Gerhard Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, pages 864–867, 2005. doi: 10.1109/ICME.2005.1521560. URL https://doi.org/10.1109/ICME.2005.1521560.

[121] J. Zhou, G. Wang, Y. Yang, and P. Chen. Speech emotion recognition based on rough set and svm. In *2006 5th IEEE International Conference on Cognitive Informatics*, volume 1, pages 53–61, July 2006. doi: 10.1109/COGINF.2006.365676.

[122] Donn Morrison, Ruili Wang, and Liyanage C. De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98 – 112, 2007. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2006.11.004. URL http://www.sciencedirect.com/science/article/pii/S0167639306001713.

[123] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak. Speech emotion recognition using eigen-fft in clean and noisy environments. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive*

Communication, pages 689–694, Aug 2007. doi: 10.1109/ROMAN.2007.4415174.

[124] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Lang Resources & Evaluation*, 2008. doi: https://doi.org/10.1007/s10579-008-9076-6.

[125] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE International Conference on Multimedia and Expo*, pages 865–868, June 2008. doi: 10.1109/ICME.2008.4607572.

[126] Shashidhar G. Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K. Sreenivasa Rao. Iitkgp-sesc: Speech database for emotion analysis. In Sanjay Ranka, Srinivas Aluru, Rajkumar Buyya, Yeh-Ching Chung, Sumeet Dua, Ananth Grama, Sandeep K. S. Gupta, Rajeev Kumar, and Vir V. Phoha, editors, *Contemporary Computing*, pages 485–492, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-03547-0.

[127] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, Jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.26.

[128] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess). *Public Library of Science*, 13(5):1–35, 05 2018. doi: 10.1371/journal.pone.0196391.

[129] Ingo Siegert, Ronald Böck, and Andreas Wendemuth. Using a pca-based dataset similarity measure to improve cross-corpus emotion recognition. *Computer Speech & Language*, 51:1 – 23, 2018. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2018.02.002. URL http://www.sciencedirect.com/science/article/pii/S0885230816302650.

[130] Roddy Cowie and Randolph R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1):5 – 32, 2003. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(02)00071-7. URL http://www.sciencedirect.com/science/article/pii/S0167639302000717.

[131] Tanja Bänziger and Klaus R. Scherer. The role of intonation in emotional expressions. *Speech Communication*, 46(3):252 – 267, 2005. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2005.02.016. URL http://www.sciencedirect.com/science/article/pii/S0167639305000890. Quantitative Prosody Modelling for Natural Speech Description and Generation.

[132] K. S. Rao and B. Yegnanarayana. Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):972–980, May 2006. ISSN 1558-7916. doi: 10.1109/TSA.2005.858051.

[133] A. Oster and A. Risberg. The identification of the mood of a speaker by hearing impaired listeners, 1986.

[134] Jianhua Tao, Yongguo Kang, and Aijun Li. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1145–1154, July 2006. ISSN 1558-7916. doi: 10.1109/TASL.2006.876113.

[135] J.R. Davitz and M. Beldoch. *The Communication of Emotional Meaning*. McGraw-Hill series in psychology. Greenwood Press, 1964. ISBN 9780837185279. URL https://books.google.co.za/books?id=1ggRAQAAIAAJ.

[136] Grant Fairbanks and LeMar W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs*, 8(1):85–90, 1941. doi: 10.1080/03637754109374888.

[137] Grant Fairbanks and Wilbert Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, 6(1):87–104, 1939. doi: 10.1080/03637753909374863.

[138] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10):787 – 800, 2007. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2007.01.010. URL http://www.sciencedirect.com/science/article/pii/S0167639307000040. Intrinsic Speech Variations.

[139] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *INTERSPEECH*, 2002.

[140] Richard Huber, Anton Batliner, Jan Buckow, Elmar Nöth, Volker Warnke, and Heinrich Niemann. Recognition of emotion in a realistic dialogue scenario. In *INTERSPEECH*, 2000.

[141] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63 (4):561–580, April 1975. ISSN 0018-9219. doi: 10.1109/PROC.1975.9792.

[142] Matti Airas and Paavo Alku. Emotions in short vowel segments: Effects of the glottal flow as reflected by the normalized amplitude quotient. In Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp, editors, *Affective Dialogue Systems*, pages 13–24, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[143] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer. Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing*, 2(3):162–174, July 2011. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.14.

[144] G Fant, Qiguang Lin, and C Gobl. Notes on glottal flow interaction. *KTH, Speech Transmission Laboratory, Quarterly Report 2-3*, pages 21–45, 01 1985.

[145] C Gobl. Voice source dynamics in connected speech. *KTH, Speech Transmission Laboratory, Quarterly Report 2-3*, pages 123–159, 1988.

[146] J. Hernando and C. Nadeu. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5(1):80–84, Jan 1997. ISSN 1063-6676. doi: 10.1109/89.554273.

[147] R. Le Bouquin. Enhancement of noisy speech signals: Application to mobile radio communications. *Speech Communication*, 18(1):3 – 19, 1996. ISSN 0167-6393. doi: https://doi.org/10.1016/0167-6393(95)00021-6. URL http://www.sciencedirect.com/science/article/pii/0167639395000216.

[148] John R. Deller, Jr., John G. Proakis, and John H. Hansen. *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 1993. ISBN 0023283017.

[149] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1):227 – 256, 2003. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(02)00084-5. URL http://www.sciencedirect.com/science/article/pii/S0167639302000845.

[150] Helen Hanson. Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, 101:466–81, 02 1997. doi: 10.1121/1.409206.

[151] V. K. Mittal and Bayya Yegnanarayana. Effect of glottal dynamics in the production of shouted speech. *The Journal of the Acoustical Society of America*, 133 5:3050–61, 2013.

[152] J. H. L. Hansen and B. D. Womack. Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 4(4):307–313, July 1996. ISSN 1063-6676. doi: 10.1109/89.506935.

[153] Bogdan Vlasenko, Dmytro Prylipko, David Philippou-Hübner, and Andreas Wendemuth. Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In *INTERSPEECH*, 2011.

[154] S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231, March 2017. doi: 10.1109/ICASSP.2017.7952552.

[155] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, July 2010. ISSN 1949-3045. doi: 10.1109/T-AFFC.2010.8.

[156] O. M. Mubarak, E. Ambikairajah, and J. Epps. Analysis of an mfcc-based audio indexing system for efficient coding of multimedia sources. In *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005.*, volume 2, pages 619–622, Aug 2005. doi: 10.1109/ISSPA.2005.1581014.

[157] Y Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 08 2013. doi: 10.1109/TPAMI.2013.50.

[158] Dong Yu, Michael Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide. Feature learning in deep neural networks - studies on speech recognition tasks. *arXiv preprint arXiv:1301.3605*, 2013.

[159] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5688–5691, May 2011. doi: 10.1109/ICASSP.2011.5947651.

[160] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Proc. INTERSPEECH*, 2014.

[161] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 801–804, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2654984. URL http://doi.acm.org/10.1145/2647868.2654984.

[162] P. Tzirakis, J. Zhang, and B. W. Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093, April 2018. doi: 10.1109/ICASSP.2018.8462677.

[163] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn W. Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, 2008.

[164] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 1, pages 401–404, July 2003. doi: 10.1109/ICME.2003.1220939.

[165] M. Lugger and B. Yang. The relevance of voice quality features in speaker independent emotion recognition. In *2007 IEEE International Conference*

on Acoustics, Speech and Signal Processing - ICASSP '07, volume 4, pages IV–17–IV–20, April 2007. doi: 10.1109/ICASSP.2007.367152.

[166] Pramod Aeluri and V Vijayarajan. Extraction of emotions from speech-a survey. *International Journal of Applied Engineering Research*, 12:5760–5767, 01 2017.

[167] Feng Yu, Eric Chang, Ying-Qing Xu, and Harry Shum. Emotion detection from speech to enrich multimedia content. In *IEEE Pacific Rim Conference on Multimedia*, 2001.

[168] Valery Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. In *ICSLP 2000*, pages 222–225, 01 2000.

[169] H. Pérez Espinosa, C. A. Reyes García, and L. Villaseñor Pineda. Features selection for primitives estimation on emotional speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5138–5141, March 2010. doi: 10.1109/ICASSP.2010.5495031.

[170] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad. Emotion recognition using acoustic and lexical features. In *INTERSPEECH*, 2012.

[171] Lan-Ying Yeh and Tai-Shih Chi. Spectro-temporal modulations for robust speech emotion recognition. In *INTERSPEECH*, 2010.

[172] Chih Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification, 2003.

[173] H. Hu, M. Xu, and W. Wu. Gmm supervector based svm with spectral features for speech emotion recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–413–IV–416, April 2007. doi: 10.1109/ICASSP.2007.366937.

[174] M. Ghai, S. Lal, S. Duggal, and S. Manik. Emotion recognition on speech signals using machine learning. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pages 34–39, March 2017. doi: 10.1109/ICBDACI.2017.8070805.

[175] Valery Petrushin. Emotion in speech: Recognition and application to call centers. *Proceedings of Artificial Neural Networks in Engineering*, 01 2000.

[176] Y. Wang, S. Du, and Y. Zhan. Adaptive and optimal classification of speech emotion recognition. In *2008 Fourth International Conference on Natural Computation*, volume 5, pages 407–411, Oct 2008. doi: 10.1109/ICNC.2008.713.

[177] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso. An acoustic study of emotions expressed in speech. In *INTERSPEECH*, 2004.

[178] Sinead McGilloway, Roddy Cowie, Cowie ED, Stan Gielen, M Westerdijk, and Sybert Stroeve. Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of the ISCA ITRW on Speech and Emotion*, 01 2000.

[179] D. Ververidis, C. Kotropoulos, and I. Pitas. Automatic emotional speech classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–593, May 2004. doi: 10.1109/ICASSP.2004.1326055.

[180] C. Huang and S. S. Narayanan. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 583–588, July 2017. doi: 10.1109/ICME.2017.8019296.

[181] Che-Wei Huang and Shrikanth Narayanan. Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *INTERSPEECH*, 2016.

[182] J. Han, Z. Zhang, F. Ringeval, and B. Schuller. Reconstruction-error-based learning for continuous emotion recognition in speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2367–2371, March 2017. doi: 10.1109/ICASSP.2017.7952580.

[183] J. Han, Z. Zhang, F. Ringeval, and B. Schuller. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5005–5009, March 2017. doi: 10.1109/ICASSP.2017.7953109.

[184] Maximilian Schmitt, Fabien Ringeval, and Björn W. Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *INTERSPEECH*, 2016.

[185] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, Jan 1986. ISSN 0740-7467. doi: 10.1109/MASSP.1986.1165342.

[186] Yariv Ephraim and Neri Merhav. Hidden markov processes. *Information Theory, IEEE Transactions on*, 48:1518 – 1569, 07 2002. doi: 10.1109/TIT.2002.1003838.

[187] Mark Stamp. A revealing introduction to hidden markov models. *Science*, pages 1–20, 01 2004.

[188] Arthur Dempster, Natalie Laird, and D.B. Rubin. Maximum likelihood from incomplete data via em algorithm. *J. Royal Statistical Soc., Series B*, 39: 1 – 38, 09 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

[189] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2007. ISBN 978-0-387-31073-2.

[190] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics forMachine Learning*. Cambridge University Press, 2019.

[191] M. Bishop. *Neural Networks For Pattern Recognition*, volume 227. 01 2005.

[192] Moataz Ayadi, Mohamed S. Kamel, and Fakhri Karray. Speech emotion recognition using gaussian mixture vector autoregressive models. volume 4, pages IV–957, 05 2007. doi: 10.1109/ICASSP.2007.367230.

[193] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3:72–83, 1995.

[194] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[195] Kristin Bennett and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1, 01 2002. doi: 10.1080/10556789208805504.

[196] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. doi: 10.1007/BF00994018.

[197] Vladimir N. Vapnik. *Statistical learning theory*. Wiley, 1998.

[198] John Platt. Probabilities for sv machines. In A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Shuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–73. MIT Press, 2000.

[199] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[200] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

[201] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, Dec 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541.

[202] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[203] https://github.com/tensorflow/tensorflow.

[204] https://github.com/Theano/Theano.

[205] https://github.com/pytorch/pytorch.

[206] https://github.com/apache/incubator-mxnet.

[207] https://github.com/Microsoft/CNTK.

[208] Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai. Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis. *Speech Communication*, 99:135 – 143, 2018. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2018.03.002. URL http://www.sciencedirect.com/science/article/pii/S0167639317303096.

[209] Marc Schröder and Roddy Cowie. Issues in emotion-oriented computing – towards a shared understanding. In *Workshop on Emotion and Computing, HUMAINE*, 2006.

[210] Noam Amir, Ori Kerret, and Dimitry Karlinski. Classifying emotions in speech: a comparison of methods. In *INTERSPEECH*, 2001.

[211] Jangwon Kim, Sungbok Lee, and Shrikanth Narayanan. An exploratory study of manifolds of emotional speech. pages 5142–5145, 2010.

[212] D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models. In *2005 IEEE International Symposium on Circuits and Systems*, pages 2871–2874 Vol. 3, May 2005. doi: 10.1109/ISCAS.2005.1465226.

[213] S. G. Koolagudi and K. S. Rao. Real life emotion classification using vop and pitch based spectral features. In *2010 Annual IEEE India Conference (INDICON)*, pages 1–4, Dec 2010. doi: 10.1109/INDCON.2010.5712728.

[214] J. Kim, J. Park, and Y. Oh. On-line speaker adaptation based emotion recognition using incremental emotional information. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4948–4951, May 2011. doi: 10.1109/ICASSP.2011.5947466.

[215] Mohammad Shami and Werner 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. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2007.01.006. URL http://www.sciencedirect.com/science/article/pii/S016763930700009X.

[216] Lori Lamel and L Rabinar. An improved endpoint detector for isolated word recognition. *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, 29(4):777–785, 1981.

[217] Qi Li, Jinsong Zheng, Augustine Tsai, and Qiru Zhou. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Treans on Speech and Audio Processing*, 10(3):146–157, 2002.

[218] J. C. Junqua, B. Mak, and B. Reaves. A robust algorithm for word boundary detection in the presence of noise. *IEEE Trans. Speech Audio Processing*, 2:406–412, 1994.

[219] L. R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell Syst. Tech. J.*, 54(2):297–315, 1975.

[220] M.H. Savoji. A robust algorithm for accurate endpointing of speech signals. *Speech Communication*, (8):45–61, 1989.

[221] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf. The cmu sphinx-4 speech recognition system. In *Proc. of the ICASSP*, 2003.

[222] BS Atal and L Rabinar. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, 24(3):201–212, 1976.

[223] J. G. Wilpon and L. R. Rabiner. Application of hidden markov models to automatic speech endpoint detection. *Computer Speech and Language*, pages 321–341, 1987.

[224] Yingyong Qi and Bobby R. Hunt. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 1(2):250–255, 1993.

[225] H. Kun and D. Wang. A classification based approach to speech segregation. *Journal of Acoustical Society of America*, 2012.

[226] G. Wu and Ch. Lin. Word boundary detection with male-scale frequency bank in noisy environment. *IEEE Trans. on Speech and Audio Processing*, 8 (5):541–554, 2000.

[227] J. Zhu and F. Chen. The analysis and application of a new endpoint detection method based on distance of autocorrelated similarity. In *Proc. of the EUROSPEECH*, pages 105–108, 1999.

[228] Yasser Ghanbari and M.R. Karami-Mollaei. A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets. *Speech Communication*, 48(1):927–940, 2006.

[229] Ouzounov Atanas. Telephone speech endpoint detection using mean-delta feature. In *Cybernetics and Information Technologies*, volume 14, 2014.

[230] Anirban Bhowmick and Mahesh Chandra. Speech enhancement using voiced speech probability based wavelet decomposition. *Computers Electrical Engineering*, 62:706–718, 2017. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2017.01.013. URL http://www.sciencedirect.com/science/article/pii/S0045790617300873.

[231] JS Walker. *A Primer on WAVELETS and Their Scientific Applications*. Taylor and Francis Group, LLC, 2008.

[232] Marie Farge. Wavelet transforms and their applications to turbulence. *Annual Review Fluid Mechanics*, 24:395–457, 1992.

[233] Zoran Nenadic and J W Burdick. Spike detection using the continuous wavelet transform. *IEEE Trans on Biomedical Engineering*, 52(1):74–87, 2005.

[234] S Nayak and S Chakraverty. Interval wavelet method for solving imprecisely defined diffusion equations. In Sunil Jacob John, editor, *Handbook of Research on Generalized and Hybrid Set Structures and Applications for Soft Computing*, chapter 21, pages 457–472. IGI Global, 2016.

[235] I Turkoglu, A Arslan, and E Ilkay. An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural networks. *Computers in Biology and Medicine*, 33:319–331, 2003.

[236] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing*. Pearson, 2007.

[237] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Pearson, 2008.

[238] Stephane Mallat. *A wavelet tour of signal processing: The sparse way*. Addison-Wesley, 1986.

[239] I. Daubechies. *Ten Lectures on Wavelets*. SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 1992.

[240] C. Torrence and G.P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.

[241] A. P. Calderon. Intermediate spaces and interpolation:the complex method. *Stud. Math*, pages 113–190, 1964.

[242] S. Kullback. *Digital Signal Processing*. Wiley, 1959.

[243] Erdal Dinç and Dumitru Baleanu. Multidetermination of thiamine hcl and pyridoxine hcl in their mixture using continuous daubechies and biorthogonal wavelet analysis. *Talanta*, 59(4):707 – 717, 2003. ISSN 0039-9140. doi: https://doi.org/10.1016/S0039-9140(02)00611-2.

URL http://www.sciencedirect.com/science/article/pii/S0039914002006112.

[244] B. T. Tan, Minyue Fu, A. Spray, and P. Dermody. The use of wavelet transforms in phoneme recognition. In *Proc. Fourth International Conference on Spoken Language*, volume 4, pages 2431–2434 vol.4, Oct 1996. doi: 10.1109/ICSLP.1996.607300.

[245] D. Campo, O.L. Quintero, and M. Bastidas. Multiresolution analysis (discrete wavelet transform) through daubechies family for emotion recognition in speech. In *Proc. Journal of Physics: Conference Series*, volume 705, 2016.

[246] P. Shen, Z. Changjun, and X. Chen. Automatic speech emotion recognition using support vector machine. In *Proc. International Conference on Electronic Mechanical Engineering and Information Technology*, volume 2, pages 621–625, 2011. doi: 10.1109/EMEIT.2011.6023178.

[247] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Proc. INTERSPEECH*, 2015.

[248] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. Speech emotion recognition using hidden markov models. In *Proc. INTERSPEECH*, pages 2679–2682, 01 2001.

[249] E. Mower, M. J. Mataric, and S. Narayanan. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070, July 2011. ISSN 1558-7916. doi: 10.1109/TASL.2010.2076804.

[250] Marko Lugger and Bin Yang. Psychological motivated multi-stage emotion classification exploiting voice quality features. In F.Mihelic and J. Zibert, editors, *Speech Recognition, Technologies and Applications*, chapter 22. I-Tech, 2008.

[251] B Yang and M Lugger. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90:1415–1423, 2010.

[252] R Q Quiroga, O A Rosso, E Basar, and M Schurman. Wavelet entropy in event-related potentials: a new method shows ordering of eeg oscillations. *Biological Cybernetics*, 84:291–299, 2001.

[253] Haytham M Fayek, Margaret Lech, and Lawrence Cavedonb. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

[254] N. Fragopanagos and J.G. Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(5):389–405, 2005. ISSN 0893-6080. doi: doi:10.1016/j.neunet.2005.03.006.

[255] S. Chakraverty and Susmita Mall. *Artificial Neural Networks for Engineers and Scientists: Solving Ordinary Differential Equations*. CRC Press, Taylor & Francis Group, 2017.

[256] https://github.com/keras-team/keras.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.