Addis Ababa University

College of Natural Sciences

Emotion Identification from Spontaneous Communication

Mikiyas Kebede Dorry

A Thesis Submitted to the Department of Computer Science in

Partial Fulfilment for the Degree of Master of Science in

Computer Science

Addis Ababa, Ethiopia

March, 2016

Addis Ababa University

College of Natural Sciences

Mikiyas Kebede Dorry

Advisor: Fekade Getahun (Ph.D.)

This is to certify that the thesis prepared by Mikiyas Kebede, titled: *Emotion Identification from Spontaneous Communication* and submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Name	Signature	Date

Advisor: _____

Examiner: _____

Examiner: _____

Abstract

This thesis work aimed to design a model for automatic identification of emotion from spontaneous communication using the acoustic characteristics of human speech. For this purpose, an experimental setup to collect and annotate call center Amharic telephone dialogs containing natural emotions is presented. These dialogs, involving 35 subjects (18 male and 17 female), are first manually decomposed into speaker turns and then segmented into intermediate chunks to be used as the analysis unit for feature calculation. Open class annotation is carried out by 3 professional human experts and the various emotional states are mapped onto 4 cover classes before a Majority Voting (MV) technique is applied to decide the perceived emotion in each chunk. Then, a total of 170 acoustic features consisting of prosodic, spectral and voice quality features are extracted from each chunk. An optimal feature set is selected through the use of generic algorithm and used to train Multilayer Perceptron Neural Network (MLPNN) classifier.

The classification performance is based on extracted features. The experimental results showed that a combined feature vector containing 33 features conveys more emotional information in a natural and spontaneous speech communication. Our speech emotion recognition model exhibits an accuracy of 72.4% in identifying Anger, Fear, Positive and Sadness emotions. Hence, it can be used for real world emotion recognition applications or can be used in combination with other speech processing technologies such as speech recognition and speaker identification to improve their performance. To demonstrate this, the proposed speech emotion recognition model is implemented using a prototype application that performs emotion identification close to real-time.

Keywords: Speech emotion recognition; Spontaneous speech emotion; Acoustic features; Feature extraction; Feature selection; Classifier; Multilayer Perceptron Neural Network

Acknowledgments

Completing a thesis is impossible to achieve without support. I would like to thank Dr. Fekade Getahun, my thesis advisor, for his guidance and support through the process. His influence in instruction, support, and supervision is invaluable. I really appreciate his friendly approach, encouragement, and constructive advices.

I would like to give a special thanks to Mekuanent Birara for his provision of resources and sharing ideas. I acknowledge my gratitude to psychologists (Moges, Nafkot and Jemal) for their professional support and participation in the preparation of the speech emotion corpus.

I owe special gratitude to my family and friends for their continuous encouragement during the course of this thesis. Finally, I would like to thank my lovely wife, Nafkot, and my kids, Naomi and Lukas, who have supported me and put up with my absence throughout my time as a graduate student.

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

ANNs	Artificial Neural Networks
DNN	Deep Neural Network
GMM	Gaussian Mixture Model
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HNR	Harmonic Noise Ration
LFCC	Linear Frequency Cepstrum Coefficients
LPC	Linear Predictive Coding
MFCC	Mel-frequency Cepstrum Coefficient
MLPNN	Multilayer Perceptron Neural Network
MV	Majority Voting
PCA	Principal Component Analysis
SER	Speech Emotion Recognition
SFS	Sequential Feature Selection
SVM	Support Vector Machines

Chapter 1: Introduction

Speech is the vocalized form of a Language that humans use to communicate or share thoughts, idea, and emotions. Several researches are conducted on speech production and speech perception of sounds used in vocal languages [1]. Speech production refers to how speech organs involved in making a sound whereas speech perception refers to the processes by which humans are able to interpret and understand the sounds used in language. In speech, each word is created out of the phonetic combination of a limited set of vowel and consonants speech sound units. These speech sound units can be represented as speech signals in digital form.

According to Nwe *et al.* [2], information in speech can be broadly categorized into the semantic and paralinguistic information. The semantic part carries linguistic information whereas the paralinguistic information refers to the implicit message such as the emotional state of the speaker. Individual speech can vary based on different timing and amplitude of the movement of speech articulators. Besides, the physical mechanism of speech undergoes changes, which can affect the nasal cavity resonance and the mode of vibration of the vocal cords. Speech can be characterized by its message content or the signal carrying the message information of i.e., the acoustic waveform. Some specific information hidden in speech signal can be detected using advanced signal processing method only. As presented by Sigmund [3], information such as emotional and mental state of the speaker can be identified from the facial expression, speech, brainwaves and other biological features of the speaker. For instance, stress has physiological consequences on heart beat rate, respiration and muscular tension. The muscular tension of vocal cords and vocal tract may have an adverse effect on the quality of speech. This indicates that there are several mental and emotional state hints carried within the speech signal. These speaker-specific information can be further analyzed from the speech signal to understand the intention of the speaker and will have a great relevance especially in the absence of a face to face communication.

1.1 Background

There has been an increase interest in speech emotion recognition to improve the capability of current speech technologies. Significant efforts are devoted in emerging methods for automatic human emotion recognition, which is an attractive research issue in Human Computer Interactions (HCIs) [4]. Researches such as, [5, 6, 7], show that different

emotional state of a speaker can be predicted by analyzing the speech signal. Identifications of basic speech emotions such as happiness, surprise, fear, sadness and neutral were the main focus of most of the studies on this area. Moreover, taking a broader view of emotion as a mental state, signal processing researchers [6, 8] have explored the possibilities of automatically detecting other mental states which share some characteristics with emotion, such as stress, depression and cognitive load.

Speech emotion recognition is the process of identifying the emotional state of the speaker out of the speech sample. Automatic emotion recognition could provide improved human-to-human communication as well as HCI by being adaptive to the observed emotions. In HCI, if the emotional state of the speaker identified and presented the machine could understand the emotional state of the user and act accordingly to create a natural communication. In human-to-human communication such as call center service, the support staff can handle the conversation in a more adjusting manner if the emotion as well as the intention of the caller is identified earlier (i.e., as soon as the conversation started).

The speech emotion classification has three phases: the first one is the construction of speech emotion database. The second is the speech processing phase containing feature extraction and selection that involve extracting relevant information associated with the given speech unit. The next phase is the classification processes in which, based on the training and testing dataset, the observed emotion gets analyzed.

1.2 Motivation

Free call center telephone conversations are exposed to abuses and pranks. Due to the difficulty of identifying the intension of the caller, the prank may continue for extended period of time. Hence, the importance of recognition model that automatically predict the intents and emotions of the caller from speech signal assists to control such calls. To achieve this, some mental and emotional states, such as intention and pretension, should be taken into consideration. However, recognizing the intention of the caller has more challenges than just detecting basic emotions as tried to be addressed in related literatures [3]. Additionally, expressions and intonation for emotions and mental state slightly differ from language to language.

In Ethiopia, there are a number of free call centers established to serve the society by providing information and counseling on different sectors of social and health related

issues. Governmental and non-governmental organizations have a significant role on creating awareness and providing counseling on critical, fatal and hot social issues, such as HIV/AIDS, women and children abuses, and others. One of the delivery channels for such cases is the use of free telephone line which is indirectly paid by the service owner organization. A method that progressively alerts on the emotional state of the speaker can help to determine and recommend the trust level of the caller, so as to enhance the counseling process.

This thesis proposes an automatic emotional state identification method dedicated to Amharic speakers which takes into account the users' intentions and their emotions in realistic conditions. In this study naturalistic, non-simulated Amharic speeches with interferences, such as background noises that are common in real world scenarios will be considered.

1.3 Statement of the Problem

Identification of a speaker's emotion or mental state is a challenging problem, in view of the significant variability in its expression posed by linguistic, contextual, and speaker specific characteristics associated to the speech [8]. It is not clear which speech features are most powerful in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch and energy. As stated in Cowie *et al.* [9], emotion expression depends on culture, gender and group membership so that, the emotion recognizer needs to be able to go beyond the obvious basic emotional states.

Though there are significant efforts devoted in developing methods for automatic human emotion recognition from speech signal, most of the existing works [2, 10, 11] focus only on some basic emotions such as happiness, anger, fear, sadness and neutral. This is due to their universal properties and the availability of the relevant training and test materials [4]. However, there are other emotional states that need to be addressed in order to get the most out of speech signal. According to Cowie and Cornelius [12], speech in daily life tends to express moderate emotional states rather than full-blown basic or primary emotions.

Besides, most research works use simulated speech corpus to train and test the model. However, such approach has major drawbacks. One of the downsides is that simulated speeches tend to contain exaggerated emotions. As stated in Cowie *et al.* [9], although the

use acted database appears to be an easy approach, actors tend to over-act the emotion they are supposed to portray, thus restricting the available material to full-blown emotions. Therefore, the result obtained with acted emotion would not apply with accuracy to natural emotional expressions. In addition, simulating a real emotion is not possible, even professional actors cannot simulate real emotion perfectly with their speech [3]. However, the task of speech emotion recognition from natural spontaneous speech is very challenging for the following reasons.

- More than one emotion can be observed in a single utterance and it is very difficult to determine the boundaries between the portions containing a single emotion.
- One may undergo a certain emotional state such as sadness for days, weeks, or even months. In such a case, other emotions will be temporary and not last for more than a few minutes. As a consequence, it is not clear which emotion the automatic emotion recognizer will detect, the long-term emotion or the transient one.
- How a certain emotion is expressed generally depends on the speaker, culture and environment. Most research works have focused on monolingual emotion classification, making an assumption there is no cultural difference among speakers. However natural emotions involve inconsistencies in the expression of emotion due to social norm, deceptive purposes as well as natural ambiguity of emotional expression [9].
- Emotions occurring in spontaneous speech seem to be more difficult to recognize compared to acted speech. A difficulty with spontaneous emotions is in their labeling as the actual emotion of the speaker is almost impossible to know with certainty [13].

Even though the aforementioned challenges indicate how difficult to use natural emotions in conducting a research in the area of emotion recognition, it is still worth working on natural emotions in comparison with the disadvantages of emotion portrayal. This is because, the better way to construct an applicable system that is capable of recognizing natural emotions in the real world is to train the classifier with natural emotions.

This study hypothesizes that, one could use speech signal extracted from natural and spontaneous conversations to construct a model capable of detecting the emotional state of a speaker. The study will address the following research questions:

- Which speech signal features characteristically deal with emotions?

- How to model emotion using best speech features?
- How can we estimate the quality of emotion conveyed in spontaneous speech signal?

1.4 Objectives

General Objective

The general objective of this thesis is to design a model that recognizes the mental and emotional state of a speaker from spontaneous speech signal.

Specific Objectives

The specific objectives of the study are:

- Conduct detail literature review on issues related to emotion detection and recognition from speech signal.
- Prepare a speech emotional corpus that represents different mental and emotional states.
- Study the acoustic characteristics of speech signal and extract the appropriate prosodic, spectral and voice quality features.
- Determine significant features that are relevant for emotion detection.
- Develop an appropriate speech emotion recognition model.
- Develop a prototype which demonstrates and evaluates the recognition of emotions from spontaneous speech signal.

1.5 Methods

Literature Review

Extensive literature review will be conducted in this thesis work in order to identify the problem, research variables, methodologies and approaches; to find out what others have done in the area, their limitations and constraints. Documents including books, previous research thesis works, articles, journals and other publications will be assessed.

Data Collection

Various data collection strategies will be used such as interview, questionnaires and other documents, forms to acquire the required data. Moreover, recorded real-world telephone dialogs will be collected to build speech corpus that will have emotion annotations and documents to allow reuse of data. The collected speech corpus will be divided randomly

into training and test dataset. The training dataset is used to train the model and will take 80% of the total dataset, and the remaining 20% will be used as a validation and test dataset.

Tools and Techniques

For the purpose of achieving the research objectives, a number of tools are required. A free audio editing tool called Audacity will be used for preprocessing the speech signal as well as to prepare the frequency domain representation of the speech signal for analysis. Praat signal processing tool that is capable of supporting scripting will be used for feature extraction from speech signal and then feature selection and classification will be done using Weka machine learning tool. Finally, the prototype of the proposed model will be developed in Java programming language using Eclipse Java JDK.

Experimentation and Testing

The developed model will be experimented in various conditions to test its performance and accuracy. This will help to see the strengths and weakness of the model.

1.6 Scope and Limitations

This research work aims on recognition of four different mental and emotional states from acoustic signal of speeches in spontaneous telephone conversation.

We focus on the acoustic characteristics of Amharic language telephone dialogs to train and test the model. The work does not consider other factors such as semantic features and facial expressions. Although emotion identification from natural spontaneous speech requires big speech corpus, the limited size of speech corpus forced us to focus only on emotions that appear frequently in the material at hand.

1.7 Application of Results

Speech emotion identification is particularly useful for applications which require Human-Computer Interaction (HCI) such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion. Other technologies like in-car board systems can take the mental state information of the driver so as to initiate safety strategies [14]. It may also be useful to enhance automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. Speech emotion recognition has also been used in agent-client call center applications and mobile communication [6, 8, 15]. Moreover, the

technology for monitoring speech to discover patterns is essential for intelligence and law enforcement organizations, surveillance tasks as well as behavioral health informatics [3]. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice.

The application of emotion recognition is not limited to HCI. Human-human communications can also benefit from its application. In call center operations for instance, identification of the emotional state of a caller in real-time can suggest the communicating person either to adapt himself/herself with the mood of the caller so as to provide a better service or end the call before the call duration takes long. Hence, the application assists in providing an improved customer service as well as save call cost and time. It can also be used to monitor recorded messages and provide offline emotional assessment for more effective call-back prioritization, customer satisfaction assessment and so on [16].

1.8 Organization of the Rest of the Thesis

The rest of this thesis is organized as follows. In Chapter 2, we will present different literatures reviewed in this thesis work in order to gain a concrete understanding on speech processing in general and speech emotion processing in particular. In Chapter 3, recent research works that are related to speech emotion recognition will be investigated for the purpose of understanding how far the speech emotion recognition is explored so far and what are the limitations and gaps that need to be addressed. Then, a systematic approach to build a speech emotion identification model will be presented in detail in Chapter 4. In Chapter 5, the experimental setup and results of the proposed model will be presented in depth. Finally, Chapter 6 will conclude the thesis with recommendations and directions for future works.

Chapter 2: Literature Review

Extensive literature review is carried out for this thesis work to get a deeper understanding of speech processing in general and speech emotion recognition in particular. These include the characteristics of speech and the various speech signal features that carry relevant emotional cues and their extraction and selection methods as well as the different classification techniques used in speech emotion recognition.

2.1 Speech

Speech is the most powerful and natural form of communication to share thoughts, ideas, and emotions. Human voice is the key tool that humans use to communicate. In addition to the intended message, a significant part of information contained in speech signal refers to the speaker [3].

Vowels and consonants are the basic segments of speech and together they form syllables, words and eventually utterances. Utterance is a continuous piece of speech beginning and ending with a clear pause. In the case of oral languages, it is generally but not always bounded by silence. Utterance is a chunk that is syntactically and semantically meaningful, intermediate unit between words and turns obtained by splitting turns at manually labeled higher syntactic boundaries and at lower syntactic boundaries in combination with pauses. In a dialog, a turn consists of all utterances of one speaker from the moment this speaker starts speaking to the moment he/she hands over to the dialog partner [17].

2.2 Amharic Speech

Amharic is a Semitic language that is being used as the working language of Ethiopia with the greatest number of speakers after Arabic. It is estimated to be the highly spoken language in Ethiopia next to Oromiffa, with more than 17 million mother tongue and at least an additional 5 million of second language speakers [18]. Only few attempts have been made in the area of Amharic natural language technologies due to the limitations of resources and the morphological complexity of the language. For instance, a model for Amharic Speech Recognition was introduced by Solomon Teferra Abate and W. Menzel in [19] using HMMs. The authors used an existing medium size read speech corpus to build Triphone- and Syllable-based models. The authors also stated that besides resource scarcity, phonological and morphological complexity as well as dialect diversity of the

language are the major challenges in developing Amharic Speech Recognition Systems language model.

Husain Seid and B. Gambäck [18] proposed a hybrid HMM/ANN approach for speaker-independent continuous speech recognition system for Amharic. The authors used frame vectors consisting 20 features that were generated automatically by CSLU toolkit and trained the recognizer using phonemes as base units from an existing speech corpus to construct the proposed model.

As presented in Solomon Teferra *et al.* [20], most of the attempts done on Amharic speech recognition focused on linguistic, lexical and acoustic models in order to develop Amharic speech recognizer. The authors added that the speech corpus they used cannot be used to develop recognizers for spontaneous speech or telephone-based applications.

According to Sebsibe Hailemariam and Kishore in [21], acoustic information can be exploited with minimal language model to build improved quality speech systems. This can benefit languages with limited linguistic resources such as Amharic. The authors proposed an approach for extraction of linguistic information with the aid of acoustic data in order to build improved speech systems. The authors conducted grapheme based speech synthesis and recognition experiments for Amharic language and perceptual test was conducted on the grapheme based Amharic voice to evaluate the performance of the system. The proposed Transcription correction algorithm used 39 different feature variables consisting 13 MFCC, 13 delta coefficients and 13 coefficients to train HMM. The acoustic features used as an evidence for selection of the best pronunciation unit at that acoustic segment. This implies that acoustic features can be used to overcome the scarcity of vital linguistic resource as well as improve the lacks of intelligibility and/or naturalness of speech systems.

To the best of our knowledge, identifying and classifying speech signal patterns for the purpose of emotional state detection for Amharic speakers is an untouched study area and there are no attempt to identify features for emotion representation. We believed that these limitations can be tackled by focusing on the acoustic characteristics of a speech corpus collected from Amharic speakers in order to train and test a statistical classifier for the purpose of emotion identification. The proposed approach in turn can be used to enhance other speech systems.

2.3 Speech Processing

Speech processing focuses on studying speech signals and their processing methods. Since signals are usually processed in a digital representation, speech processing can be viewed as a special case of Digital Signal Processing (DSP), applied to speech signals [1].

There are three phases common to most speech processing applications. The first step is speech preprocessing, which provides signal operations such as digitalization, pre-emphasis, frame blocking, and windowing. The second step is feature extraction, which represents the process of transforming sequences of preprocessed speech samples to feature vectors representing characteristics of the time-varying speech signal. Finally, the extracted features from speech signal are put together into feature vector corresponding to the final aim of the speech processing called classification [22].

Speech processing research focused on mining of specific information from speech signal to develop analyzers that are task, speaker, and vocabulary independent so as to easily adapt to a variety of applications for different languages. The main applications of speech processing includes the following domains [1]:

- Speech recognition, which deals with analysis of the linguistic content of a speech signal and its conversion into a computer-readable format, which is like understanding the speech for the computer.
- Speaker recognition, where the aim is to recognize the identity of the speaker.
- Speech synthesis: the artificial synthesis of speech, which usually means computer-generated speech. Advances in this area improve the computer's usability.

As research in speech processing has matured, attention has gradually shifted from linguistic-related applications such as speech recognition towards paralinguistic speech processing problems, in particular the recognition of speaker identity, emotion, gender, and age [8].

2.4 Emotion

Emotion does not have a commonly agreed theoretical definition, even though a wide variety of definitions have been proposed. However, people know emotions when they feel them. A research work by Kleinginna and Kleinginna [22] tried to resolve the resulting terminological confusion by compiling 92 definitions and 9 controversial statements from a variety of sources in the literature of emotion and proposed the following definition:

"Emotion is a complex set of interaction among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generating cognitive process such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive."

Emotional states are often correlated with particular physiological states [5], which in turn present predictable effects on speech features, especially on pitch, timing, and voice quality. For this reason, researchers were able to study and define different aspects of emotions [15, 23]. It is widely thought that emotion can be characterized in two dimensions: activation and valence [12].

Activation refers to the amount of energy required to express a certain emotion. According to a physiological study of the emotion production mechanism made by Williams and Stevens [24], it has been found that when someone is in state of anger, fear, or joy, the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry, and there is an occasional muscle tremors. As a consequence, speech is then loud, fast and articulated with strong high-frequency energy. On the other hand, when someone is bored or sad, the parasympathetic nervous system is aroused, the heart rate and blood pressure decrease, and salivation increases, which results in slow, low-pitched speech with a weak high-frequency energy [25]. Thus, acoustic features such as the pitch, timing, voice quality, and articulation of the speech signal highly correlate with the underlying emotion [23].

However, emotions cannot be distinguished using only activation. For example, both the anger and the happiness emotions correspond to high activation but they express different affect. This difference is characterized by the valence dimension. Unfortunately, there is no agreement within researchers on how, or even if, acoustic features correlate with this dimension [26]. Therefore, while classification between high-activation (also called high-arousal) emotions and low-activation emotions can be achieved at high accuracies, classification between different emotions is still challenging [27].

On the other hand, as stated by Nwe *et al.* [2], emotional states have a definite temporal structure based on their broad or narrow sense effect. The broad sense reflects the underlying long-term emotion and the narrow sense refers to the short-term excitation of

the mind that prompts people to action. In automatic recognition of emotion, a machine would not distinguish if the emotional state were due to long-term or short-term effect so long as it is reflected in the speech or facial expression.

An important issue in speech emotion recognition is the need to determine a set of important emotions to be classified by an automatic emotion recognizer. Psychologists and Linguists have defined inventories of the emotional states, most encountered in our lives. Atypical set is given by some researchers which contains 300 emotional states [28, 29]. However, classifying such a large number of emotions is very difficult. The best known theoretical idea in emotion research is that certain emotion categories are primary and others are secondary [12]. Other researches described emotions as tree-structured lists containing primary, secondary and tertiary emotions [30]. According to Ayadi *et al.* [31], many researchers agree with the 'palette theory', which states that any emotion can be decomposed into six primary emotions, also known as, archetypal emotions. The primary emotions include Anger, Disgust, Fear, Joy, Sadness, and Surprise [9].

According to Cowie and Cornelius [12], at least 60 emotional state classes are required to define emotions that occur with a reasonable frequency in everyday life. A mapping onto cover class approach is investigated in emotion classification research by Batliner *et al.* [32]. The significance of the cover class mapping approach is not limited to emotional state class reduction, it also used to combine less frequently appearing emotional states together onto a broader cover class to balance the distribution of the emotional states in the natural speech emotion corpus as it used in Steidl [17].

2.5 Speech Emotion Recognition

Speech emotion recognition refers to the process of extracting the high-level affective status of a speech utterance from the low-level features [33]. The human voice carries a wealth of information about emotion, mood and mental states. Types of information that can be found in speech are broadly classified as semantic and paralinguistic (prosodic) information. The semantic part of the speech carries linguistic information insofar that the utterances are made according to the rules of pronunciation of the language. On a semantic-syntactic level the spoken words and phrases can transmit clear reference to the emotional state of the speaker. On the other hand, paralinguistic information refers to the implicit messages such as the emotional state of the speaker. For speech emotion recognition, the identification of the paralinguistic features that represent the emotional state of the speaker

is an important first step [2]. Prosodic features such as pitch, timing and energy of the speech signal contribute to the recognition of emotions [7]. In general, prosodic is useful to capture non-verbal expressions. However, recognition of emotions in speech is a complex task that is furthermore complicated by the fact that there is no unambiguous answer to what the "correct" emotion is for a given speech sample [13].

2.6 Components of Speech Emotion Recognition

Speech Emotion Recognition (SER) system is an application of speech processing in which the patterns of extracted speech features are mapped by the classifier during the training and testing session using pattern recognition algorithms to detect the emotions from each of their corresponding patterns. Since the emotion is to be detected from the input speech signal, the whole signal processing revolves around the speech signal for the extraction and selection of speech features corresponding to emotions [34]. The next is generating a database for training and testing of extracted speech features followed by the last stage of emotion detection by the classifier section using pattern recognition algorithms, as shown in Figure 2.1 [34].

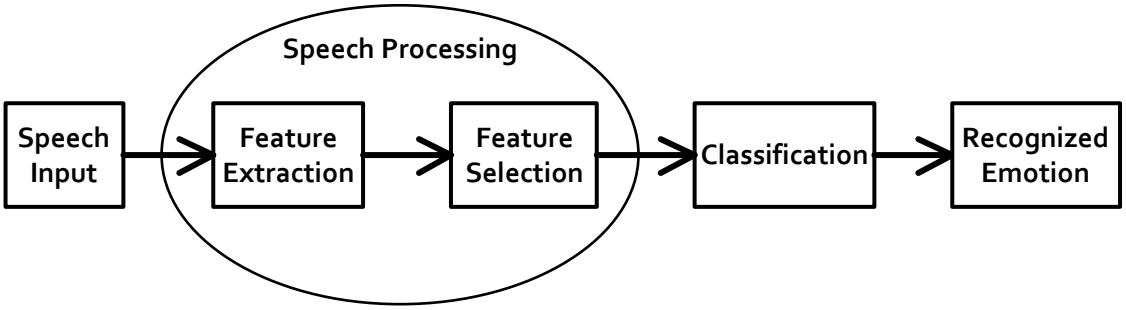

Figure 2.1: Basic Speech Emotion Recognition System

The emotion recognition technique is synonymous to speaker recognition system but its different approach to detect emotions makes it intelligent and adds performance to achieve better service in various applications [1].

Speech Emotion Recognition includes preprocessing (such as audio segmentation and normalization), extraction and selection of relevant features, and training and testing of potential classifiers for the purpose of emotions classification. Each of these principal parts are reviewed and discussed in the next consecutive sections.

2.6.1 Preprocessing

Preprocessing refers to all operations required to be performed on the time samples of speech signal before extracting features [31]. It can also be done for removing noise presented in the collected samples. Preprocessing may include digitalization, frame blocking and windowing [3].

- **Digitalization:** the whole process of speech signal processing that starts by converting an analog speech signal to a digital form is called digitalization.
- **Frame blocking:** short-time analysis is the most common approaches in speech signal processing. The pre-emphasized signal is blocked into frames of N samples with a typical range between 10-30 msec. Overlapping frames are commonly used to control how quickly parameters can change from frame to frame, i.e., to smooth the extracted contours.
- **Windowing:** a signal observed for a finite interval of time may have distorted spectral information. To minimize this distortion, a signal is multiplied by a window-weighting function before feature extraction is performed.

Having extracted the suitable speech features from the preprocessed time samples, some post-processing may be necessary before the feature vectors are used to train or test the classifier. For example, the extracted features may be of different units and hence their numerical values have different orders of magnitude. In addition, some of them may be biased. This can cause some numerical problems in training some classifiers. Therefore, feature normalization may be necessary in such cases.

2.6.2 Speech Segmentation

Several researches have been conducted to compare the various speech units related to their relevance in speech emotion [17]. As stated in the paper by Vogt and Andr´e [35], the time segments have to be chosen very carefully as they have to fulfil two conflicting conditions:

- Emotion changes can occur very quickly, but the segment length sets the temporal resolution of recognizable changes,
- Reliable statistical features can often only be computed over longer segments.

Classification experiments on the word, chunk, and turn level are carried out by Steidl in [17], and the best results are obtained on the chunk level. This is because, chunk level

segmentation compromises between the length of the unit and the homogeneity of the emotional state within this unit. This means that even though the possibility of obtaining an isolated emotion in a word is high, the word level segmentation is quite complex as the model need to have a word recognition capability. Whereas, the utterance level segmentation is quite easy however, multiple emotions can be reflected within a single utterance.

2.6.3 Feature Extraction

Speech features are the acoustics information usually derived from the analysis of speech in both time as well as frequency domains. The goal of feature extraction is to the extract relevant features from speech signals with respect to emotions [36]. An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions. Moreover, the classification performance largely relies on the kind of features we can extract [37]. Some of the design decisions that need to be considered are discussed below:

Region of Analysis (local vs global features)

Region of speech used for feature extraction is one of the design issue that needs to be considered in speech emotion recognition.

- **Local features:** speech signals are not stationary. Due to this reason, it is common in speech processing to divide a speech signal into small segments called frames. Within each frame the signal is considered to be approximately stationary [38]. Prosodic speech features, such as pitch and energy, are then extracted from each frame.
- **Global features:** are calculated as statistics of all speech features extracted from an utterance.

There has been a disagreement on which of these features are more suitable for speech emotion recognition. The global features have been broadly used in speech emotion recognition and the majority of researchers have agreed that they are better in terms of classification accuracy and classification time [39, 40]. Another advantage of global features is that their number is much less than local features. Therefore, the application of cross validation and feature selection algorithms to global features are executed much faster than if applied to local features [31].

However, researchers have claimed that global features are efficient only in distinguishing between high-arousal emotions, e.g., anger, fear, and joy, versus low-arousal ones, e.g., sadness [2]. They claim that global features fail to classify emotions which have similar arousal, e.g., anger versus joy. Another disadvantage of global features is that temporal information present in speech signals is completely lost [39]. Moreover, it may be unreliable to use complex classifiers such as the HMM with global speech features since the number of training vectors may not be sufficient for reliably estimating model parameters. On the other hand, complex classifiers can be trained reliably using the large number of local feature vectors and hence their parameters will be accurately estimated. This may lead to higher classification accuracy than that achieved if global features are used.

Feature Types

Another important issue in speech emotion recognition is the extraction of speech features that efficiently characterize the emotional content of speech and to represent them in n-dimensional feature vector. Unfortunately, it is quite impossible to compare features across published works, since conditions vary a lot and even minor changes in the general setup can make results incomparable [41]. Though there is no general agreement on which features are the most appropriate ones so far, good features seem to be highly data dependent [35, 42].

Speech features can be grouped into three categories: Prosody continuous features, qualitative features, and spectral features.

- **Prosodic speech features:** Prosody continuous features such as pitch and energy are the most commonly applied features to distinguish and classify emotion state [15, 22]. Speech prosody is useful for utterance-based emotion detection because, the arousal state of the speaker (high activation versus low activation) affects the overall energy, energy distribution across the frequency spectrum and the frequency and duration of pauses of speech signal [12].
- **Voice Quality Features:** Emotional content of an utterance is strongly related to its voice quality [9]. Experimental studies with listening human subjects demonstrated a strong relation between voice quality and the perceived emotion [43]. A wide range of phonetic variables contribute to the subjective impression of voice quality.

- **Spectral-based speech features:** Spectral features are often selected as a short-time representation for speech signal. According to Nwe *et al.* [2], it is recognized that the emotional content of a speech has an impact on the distribution of the spectral energy across the speech range of frequency. For example, it is reported that utterances with happiness emotion have high energy at high frequency range while utterances with the sadness emotion have small energy at the same range.

Some of the acoustic features that are useful for emotion recognition are discussed in the following section.

- **Pitch:** according to Ververidis and Kotropoulos [39], a pitch signal has information about emotion because it depends on the tension of the vocal folds and the sub-glottal air pressure. The pitch signal is produced from the vibration of the vocal folds. Two features related to pitch are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. The time elapsed between two successive vocal fold openings is called pitch period T, while the vibration rate of the vocal folds is the fundamental frequency of the phonation F0 or pitch frequency. Pitch frequency value can be calculated in each speech frame and the statistics of pitch can be obtained in the whole speech sample. These statistical values reflect the global properties of characteristic parameters [44].
- **Formants:** formants represent the amplifications of certain frequencies in the spectrum resulting from resonance of the vocal tract independent of the perceived pitch [17]. Tracking formants over time is used to model the change in the vocal tract shape.
- **Mel Frequency Cepstrum Coefficient (MFCC):** MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system [44]. The signal is cut into short overlapping frames, and for each frame, a feature vector is computed, which consists of Mel Frequency Cepstrum Coefficients. The cepstrum is the inverse Fourier transform of the log-spectrum [45].
- **Log Frequency Power Coefficients (LFPCs):** are simply derived by filtering each short-time spectrum with 12 bandpass filters having bandwidths and center frequencies corresponding to the critical bands of the human ear [39].
- **Linear Predictive Coding (LPC):** represents the characteristics of particular channel of speech. A speaker with different emotional speech will have different

channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech [44]. The computational method of LPCC is usually a recurrence of computing the Linear Prediction Coefficients (LPC), which is according to the all-pole model used by Joshi and Zalte [44].

- **Jitter and Shimer:** the term jitter denotes cycle-to-cycle variations of the fundamental frequency. On the other hand, shimmer denotes variations of the energy from one cycle to another [17].
- **Harmonics-to-Noise Ratio:** is a measure for the degree of periodicity of a voiced signal, which can be found from the relative height of the maximum of the autocorrelation function [17].

A feature set comparison done by Vogt and Andr´e [35] demonstrated that pitch-related features play a dominant role in acted speech, whereas for spontaneous emotions, the focus lies more on MFCC. A research work by Ververidis and Kotropoulos [39] also concluded that MFCCs provide a better representation of a speech signal since they additionally exploit the human auditory frequency response. However, LFPCs are better features than MFCCs for emotion classification in practice. It was shown that features based on cepstral analysis such as LPCC and MFCC clearly outperform the performance of the linear-based features of LPC in detecting stress in speech signal. However, a study by Nwe *et al.* [2] compared a linear-based feature, namely LFPC, and two traditional cepstral-based features, namely LPCC and MFCC. Their result shows that LFPC is a better choice as feature parameters for emotion classification than the traditional feature parameters for the classification of the six archetypal emotions: anger, disgust, fear, joy, sadness, and surprise.

2.6.4 Database for Training and Testing

According to a study by Vogt *et al.* [41], there are no standard databases that could be used for benchmarking. Basically, a good database is as important as the desired results. Speech emotion recognition research deals with databases of acted, induced or completely natural spontaneous emotions (real-life emotions), though the complexity of the task increases with the naturalness, as shown in Figure 2.2 [41].

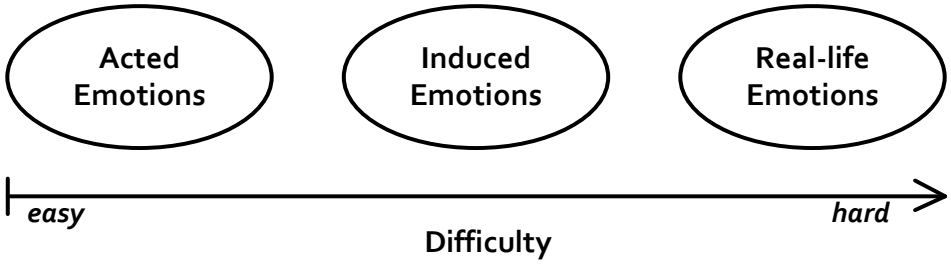

Figure 2.2: Types of Databases used for Emotion Recognition and their Difficulty

There are different databases created by the speech processing community with the help of professional actors which is widely used in research work. The results are uncompromising though the emotions are acted rather than spontaneous or natural. Some of the famous databases listed by Lanjewar and Chaudhari [34] are: The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database available.

Induced data, such as the SmartKom corpus and the German Aibo emotion corpus, where people were recorded in a lab setting fulfilling a certain task that was intended to elicit emotions without the subjects knowing that their emotional state was of interest [41]. Spontaneous speech from actual telephone services could be counted as a source for spontaneous emotions [13]. For instance, the call center communication dealt with by Devillers *et al.* [42] is fully realistic as it is obtained from live recordings of real-life spoken dialogs from call center services.

As per the review made by Ayadi *et al.* [31], almost all the existing emotional speech databases have some limitations for assessing the performance of proposed emotion recognizers. Their review stated that most speech emotional databases do not well enough simulate emotions in a natural and clear way. In some databases such as KISMET, the quality of the recorded utterances is not so good. Moreover, in a research work by Zhang *et al.* [36], the researchers mentioned the simulated database they used, which was the popular Berlin database, as a limitation of their study. They added that, it will be an interesting and challenging task to develop an online real-time speech emotion recognition system in their future work.

2.6.5 Feature Selection

Feature selection is the process of searching a minimum set of the most relevant features for the purpose of enhancing the classification precision and learning runtime. Features are selected in terms of their significance in characterization of the emotions, and also to control the dimensionality of combined features by representing them into n-dimensional feature vector which will further be classified to determine the emotion in speech [34]. For this purpose, various feature selection techniques can be used in finding an appropriate linear or nonlinear mapping from the original feature space d to another sub-space of dimensionality m (i.e., $m <= d$) while preserving as much relevant classification information as possible [46].

It is also common to use dimensionality reduction techniques in speech emotion recognition applications in order to reduce the storage and computation requirements of the classifier and to have an insight about the discriminating features, which in turn increase the classification accuracy [46]. There are two approaches for dimensionality reduction: feature selection and feature reduction.

Feature selection determines which features are the most beneficial as most classifiers are negatively influenced by redundant, correlated or irrelevant features. Thus, in order to reduce the dimensionality of the input data, a feature selection algorithm is implemented to choose the most significant features of the training data for the given task. Alternatively, a feature reduction algorithm such as Principal Component Analysis (PCA) can be used to encode the main information of the feature space more compactly [41].

2.6.6 Classification Techniques

There are various types of classifiers available intended for their specific usage based on types of features to be classified. As stated in Vogt *et al.* [41], any statistical classifier that can deal with high-dimensional data can be used, but static classifiers like Support Vector Machines (SVM), Artificial Neural Networks (ANN), and decision trees for global statistics features, and Hidden Markov Model (HMM) for short-term features as a dynamic modeling technique are most commonly found in the literature on emotional speech recognition [2, 5, 47].

Static classifiers have proved to be successful for acted data, but for natural data, recognition accuracy is only useful in a problem with very few emotion classes. Whereas dynamic classification with HMMs is used less often than static classification, but it is

thought to be advantageous for better capturing the temporal activity integrated in speech. Though, HMMs might even better be suited for natural emotions, they have almost exclusively been applied to acted data [41]. According to Wagner *et al.* [48], dynamic classification is very promising, but currently, more feature types can be exploited for static classification. However, when the feature set is restricted to the same feature type, for instance only MFCCs and energy, HMM often outperforms static modeling techniques.

ANN is another common classifier used for many pattern recognition applications. Though, the classification accuracy of ANN is fairly low compared to other classifiers, ANN is known to be more effective in modeling nonlinear mappings. Besides, their classification performance is usually better than HMM and Gaussian Mixture Model (GMMs) when the number of training examples is relatively low [44], which is the case for most spontaneous speech emotion databases. A research work by Unluturk *et al.* [49] stated that neural networks are quick to respond. Artificial neural network classifier is reviewed in the next section.

Artificial Neural Networks (ANNs)

Neural Network is a network structure consisting of a number of nodes connected through directional links. Each node in the network represents an information processing unit called neuron and the links between nodes specify the causal relationship between connected neurons. Neural network has the ability to learn from experiential knowledge expressed through inter-unit connection strengths, and can make such knowledge available for use [50]. The following section discusses a kind of ANNs called multilayer perceptron.

Multilayer Perceptron Neural Networks (MLPNNs)

Multilayer perceptron is a kind of neural networks that can be monitored and modified during training time. MLPNNs are layered feed forward (FF) network typically trained with static back propagation in order to classify static pattern [51]. The algorithm consists of two major steps. In the forward pass, the predicted outputs are calculated corresponding to the given input. In the backward pass, partial derivatives of the cost function with respect to the different parameters are propagated back through the network. A typical multilayer perceptron consists of a set of input nodes forming the input layer, one or more hidden layers of computation nodes, and a set of output classes also called output layer of nodes [52].

In general, the architecture of a neural network is defined by the characteristics of a node and the characteristics of the node's connectivity in the network. Network architecture is specified by the number of inputs to the network, the number of outputs, the total number of elementary nodes that are usually equal processing elements for the entire network, and their organization and interconnections [50]. MLP neural networks (MLPNNs) and recurrent neural networks (RNN) are the most common types of ANNs used in speech emotion recognition. MLPNNs are relatively popular in speech emotion recognition due to the ease of implementation and their well-defined training algorithm once the structure of ANN is completely specified [31]. MLPNN is suitable in data-rich environments and has an advantage over other types of machine learning algorithms for scaling.

2.7 Summary

This Chapter with various aspects of speech modelling and speech processing components for the purpose of understanding the acoustic characteristics of the speech signals along with their processing methods and applications. As the application of speech processing such as speech recognition and speaker identification has matured, the attention of speech processing researches has gradually shifted to paralinguistic research problems such as emotion recognition. Emotions are often correlated with particular physiological states which in turn present predictable effects on speech features. This makes human speech the most powerful form of communication to share thoughts, ideas and emotions. Speech processing techniques such as preprocessing, feature extraction and selection along with various classification methods enable the mapping of low-level information (features) onto high-level states (emotions) of a speech signal. On the other hand, the nature of the emotion corpus has a great impact on the performance as well as on the applicability of the speech emotion recognition model. Acted emotion corpuses containing emotion portrayals are the widely used type of corpus because of its simplicity and availability. Induced corpus contains emotional data where human subjects were recorded in a lab setting fulfilling a certain task that was intended to elicit emotion without the knowledge of the subjects. Whereas, natural corpuses contains real-life spontaneous communications such as live recordings of spoken dialogs from call center. Unfortunately natural corpuses are not widely available for researches. The difficulty of emotion recognition task increases with the naturalness of the database. Finally, we also reviewed some classification techniques relevant to emotion recognition such as SVM, HMM, ANN as well as a special kind of neural network called MLPNN.

Chapter 3: Related Work

Speech emotion detection has become a fundamental task in human-computer interaction systems. Given a set of features, various types of classification models have been used for this task, such as HMM [5, 53], GMMs [7, 54, 55] and neural networks [10, 25, 33]. The challenge of speech emotion recognition is strictly correlated with the type of speech emotion corpus used to train a particular classifier.

This Chapter identified the research variables, methodologies and approaches followed by recent published research works in order to identify the gaps and limitations in the speech emotion detection domain. The following three sections discuss recent research works that used simulated, induced, and natural speech emotion corpus respectively. Section 3.4 will assess some of the comparisons made in relation to speech emotion classification. Finally, Section 3.5 will conclude the Chapter by summarizing the type of speech corpus, feature extraction and selection methods used as well as the classifier techniques applied along with the reported performance in recognizing various types of emotional state classes.

3.1 Speech Emotion Recognition using Simulated Emotions

Schuller *et al.* [7] used a continuous one-state HMM with GMMs as a classifier to recognizes the emotional user state classified as joy, anger, irritation, fear, disgust, sadness and neutral inner state by analyzing spoken utterance on both the semantic and signal level. Acted emotional speech corpus was used in this work and the researchers admitted it as the weakness of the study since users tend to exaggerate when acting. The total recognition rate for understanding emotional phrases using the proposed approach was 88.1%.

Another work by Schuller *et al.* [5] tried to address the performance of global statistics versus instantaneous features. Utterances collected from 5 speakers in 7 emotional states (acted and spontaneous) were modelled using HMMs with up to 64 states and up to 4 mixtures per state. A left-right topology was used with an additional jump limit of two states at most. Performance was compared to the outcome of a GMM trained on global statistics of the utterances. While the latter achieved an average recognition accuracy of 86.8%, classification with instantaneous features reached only 77.8%.

Nwe *et al.* [2] run experiments on discrete HMMs. Short acted utterances in 6 archetypal emotions obtained from 12 non-professional speakers were modelled by HMMs with up to 8 states. The choice of an ergodic (fully connected) topology instead of a left-right

structure was deduced from the assumption that emotional cues contained in an utterance may not occur strictly sequentially. Best results were achieved for 4 states.

A research work by Tao *et al.* [4] introduced the Fused Hidden Markov Model which is trained in neutral expressed audio-visual corpus to reduce the utterance influences in visual parameters for the audio-visual-based emotion recognition by combining audio and visual features with a Multi-stream HMM (MHMM). The research focuses on basic emotions, known as happiness, surprise, fear, anger, sadness and neutral and achieved a mean accuracy value of 85.2%.

A speaker-dependent algorithm that allows a robot to express its emotion by modulating the intonation of its voice using neural network and SVM was introduced in [11]. The research used imitated basic emotional speeches to pronounce short sentences or phrases such as, “Hello”, “How are you?”, and “Great” and achieved an average of 95% accuracy.

A research work by Khanchandani and Hussain [51] explored NN methods to recognize basic human emotions using Berlin Emotional Speech database. The research trained extracted prosodic features of speech using Multilayer Perceptron NNs (MLPNNs) and Generalized Feed Forward NNs (GFFNNs) to categorize the emotions. Finally, the performance of MLPNN and GFFNN for recognition of emotions was examined. GFFNN (98.08%) recognized emotional test patterns with more accuracy than MLPNN (89.62%).

Han *et al.* [33] proposed to utilize Deep Neural Networks (DNNs) to produce segment-level emotion state probability distributions from high level features consisting of MFCC, pitch-based and HNR. The research used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database containing audio data from 10 actors to train the model. Then, the authors split the database into training data, consisting of utterances from 8 speakers, and the remaining 2 speakers for testing. The result demonstrated that the accuracy of 54.3% is obtained to identify five different emotion states: Excitement, Frustration, Happiness, Neutral and Sadness.

On the other hand, a neural network designed to classify the power spectrum of the voice signals was developed by Unluturk *et al.* [49] and achieves an average recognition performance of 100%. This performance is impressive and statistically reliable because 97932 data segments are used in training the neural network and a new set of 24483 testing sets is utilized to test the performance. The samples tested for voice recognition are acquired from the movies “Anger Management” and “Pick of Destiny”. The research also

stated that the optimal values for neural networks weights are estimated using the back propagation algorithm.

A study by Dai *et al.* [37] analyzed 2442 utterances from the Emotional Prosody Speech and Transcripts corpus and extracted 62 features from each utterance. A neural network classifier was built to recognize different emotional states of these utterances. They obtained over 90% accuracy in distinguishing hot anger and neutral states, over 80% accuracy in distinguishing happiness and sadness as well as in distinguishing hot anger and cold anger. They also achieved 62% and 49% accuracy for classifying 4 and 6 emotions respectively. They had 20% accuracy in classifying all 15 emotions in the corpus.

3.2 Speech Emotion Recognition using Induced Emotions

An approach based on biologically inspired methods is proposed in Caponetti *et al.* [25] using a Long Short-Term Memory recurrent neural network (LSTM) classifier and applied to two different representation models of emotion speech signal namely Mel-Frequency Cepstral Coefficient (MFCC) and the Lyon Cochlear to unveil differences between the two models in terms of emotion recognition rate. Speech Under Simulated and Actual Stress (SUSAS) corpus was used for this experiment and the researchers found out that combining the LSTM classifier with the Lyon Cochlear representation gives better recognition results, 75.5%, than combining the same classifier with the traditional MFCC representation, 71.5%.

A smoothed nonlinear energy operator (SNEO)-based Amplitude Modulation Cepsral Coefficients (AMCC) for recognizing basic emotions from speech signal using GMM was proposed by Alam *et al.* [54]. An average of 44.5% of accuracy was achieved to recognize happiness, emphatic, anger, bored, and neutral emotions from existing spontaneous speech corpus.

Callejas *et al.* [6] propose a method for predicting the user mental state for the development of spoken dialogue systems. A corpus of spontaneous Spanish speech dialogue was acquired to recognize the emotional state of the user. Depending on the emotional state and the classifier used, a range of values between 76 – 97% were exhibited in the evaluation of the model. The researchers evaluated four alternatives: a multinomial Naïve Bayes classifier, n-gram based classifier, a classifier based on grammatical inference techniques and neural networks. They obtained a classification accuracy of 88.5%, 51.2%, 75.7% and 97.5%. As the best result were obtained using MLPNN, they used MLPs as

classifiers for their experiments, where the input layer received the current situation of the dialogue and the values of the output layer used as the posteriori probability of selecting the different user intention given the current situation of the dialogue.

Unlike the previous research works discussed above, the author in [3] created and used its own speech corpus, known as ExamStress. The speech was collected from University students during their final state examination to represent stressed speech, and a few days later another speech on some issues collected from the same group of students again to represent the neutral speech to correlate with stress in influencing the voice. For this study, vowel duration, formants and fundamental frequency features are used and a Mahalanobis distance measure was used as a classifier. The model achieved 88% and 72% of accuracy for speaker dependent and speaker independent recognitions respectively.

Steidl [17] also presented FAU Aibo emotion corpus containing emotionally colored speech of children at the age of 10 to 13 years interacting with the Sony robot Aibo. Firstly, this induced corpus is labeled with 11 emotion-related states on the word, turn and intermediate chunk level. The author then applied a map onto cover class approach to fold the 11 emotion classes to a 4-class problem containing anger, emphatic, neutral, and motherese. Both acoustic and linguistic features are used in the experiment and the result indicates that the linguistic features performed slightly worse that the acoustic features. A performance comparison among the 3 statistical classifiers, GMM, LDA and ANN is also presented. As a result, LDA achieved the highest class-wise averaged recognition rate of 68.9% using energy and duration based prosodic features and spectral MFCC features.

3.3 Speech Emotion Recognition using Natural Emotions

A German speech database containing authentic emotional expressions from a TV talk-show was used to train Support Vector Regression (SVR) in the research work by Grimm and Kroschel [56]. A generalized three-dimensional emotion space method motivated by emotion psychology is proposed to define emotions as three basic attributes called “emotion primitives”. In order to estimate the three emotion primitives (valence: positive-negative axis, activation: calm-excited axis, and dominance: week-strong axis), utterance-based feature vector was built containing the statistics of fundamental frequency, energy and the MFCC. Then, two different techniques, PCA and SFS, were used to reduce the feature vector size to 20. The experiments were done using 10-fold cross-validation and

correlation coefficients of 0.46, 0.86, and 0.79 were obtained for valence, activation, and dominance respectively.

Vidrascu and Devillers [57] used paralinguistic cues extracted from 10 hours dialog corpus recorded in a French Medical emergency call center to train SVM and Logistic Model Tree (LMT) classifiers. The authors tried to identify two high-level emotion groups, Negative and Positive, using a total of 800 turn-based segments; 400 for each emotion group. Paralinguistic cues such as prosodic (F0 and energy), spectral (formants), and disfluency features were used and a correct detection rate of about 82% was obtained. Even though the result seems very high for natural emotional speech, the model is limited to a two-class classification problem capable of identifying only Negative and Positive emotions.

3.4 Comparisons of Speech Emotion Classification

A research work by Shuller *et al.* [58] tried to compare various classifiers by applying each of them on two acted and one induced databases. The authors stated that the two classifiers with the highest performance for emotion recognition are artificial neural networks and support vector machines. However, comparison of the various research works on automatic emotion recognition with different classifiers, feature vectors, target classes, or emotion databases, is difficult.

According to a study by Vogt *et al*. [41], accuracy of 50% may be excellent for 4-class problem for one database, while for another database, recognition rates of 70% to 80% can be reached. This does not mean that the database in the former case was not well designed, but rather that it is a harder task and that can be considered to many factors that potentially affect the performance of the model. These factors might include the nature of the database used (acted, induced, or natural), the number and types of emotional state classes, and so on. The authors also added that, the rule of thumb for natural emotions is that recognition rate is not much more than twice chance level. This means, for a 4 class problem (chance level of 25%), accuracy of 50% is a good result.

3.5 Summary

In this Chapter, recent research works related to our thesis are assessed in detail. In order to highlight the impacts of various types of speech emotion databases on the performance of recognition, we grouped the works into 3 major categories based on the kind of corpus

used in each research. Table 3-1 summarizes some of the related works reviewed regarding speech emotion recognition.

Table 3-1: Summary of Related Works

Paper	**Corpus**	**Features**	**Classifier**	**Emotion Classes**	**Evaluation (RR)**
Schuller *et al.* [7]	Acted	Semantic and Acoustic	HMM with GMMs	7 basic emotions Joy, Anger, Irritation, Fear, Disgust, Sadness and Neutral	88.1%
New *et al.* [2]	Acted	Short time LFPC	HMM	6 basic emotions Anger, Disgust, Fear, Joy, Sadness and Surprise	78%
Grimm and Kroschel [56]	Natural corpus from TV talk-show	F0 Energy MFCCs	SVR	3 emotion primitives Valence, Activation and Dominance	Valence 46% Activation 82% Dominance 79%
Alam *et al.* [54]	Acted	AMCC	MLPNN GFFNN	5 basic emotions Anger, Emphatic, Neutral, Positive and Rest	44.5%
Han *et al.* [33]	Acted	MFCC Pitch HNR	DNN	5 emotion states Excitement, Frustration, Happiness, Neutral and Sadness	0.54%

Paper	Corpus	Features	Classifier	Emotion Classes	Evaluation (RR)
Vidrascu and Devillers [57]	Natural call center dialogs	Prosodic Spectral Disfluency	SVM LMT	2 emotion groups Negative Positive	82%
Steidl [17]	Induced FAU Aibo corpus	Prosodic Spectral	GMM LDA ANN	4 cover classes Anger, Emphatic Neutral and Motherese	68.9%

According to the related works reviewed in this Chapter, most of the researches share common drawbacks such as using simulated speech corpus in a controlled environment, using existing preprocessed noise free acted speech corpus, or focusing on the simplest and most common emotional states. Even though different researchers managed to achieve high recognition accuracy from acted database, the gap would apparently be larger for spontaneous emotions, and such accuracy level is more difficult to attain from natural speech corpus.

After studying the related works, it can be identified that the feature set which is mostly employed comprised Pitch, MFCCs and Energy. Additionally, the HMM technique is widely used by the researchers due to its effectiveness for acted, preprocessed and large size training and test dataset. However, voice signals are random signals and are not readily measurable and lack uniquely recognizable features. Therefore, the neural network becomes interesting for classifying these signals because of its practical advantages such as real-time processing, adaptability and training capability. Besides, results from the reviewed literatures are encouraging and suggest that neural networks are potentially useful for spontaneous emotion recognition. Neural networks also exhibit high performance which is one of the requirements as emotion should be recognized almost instantly for real-time spontaneous emotion recognition.

On top of that, the main constraint of spontaneous emotion recognition researches is the availability of spontaneous speech emotion databases. From the literatures we assessed for the purpose of this thesis work, we found that neural network provides a better result with limited training data, which again suits our objective. Taking all these facts into consideration, we selected neural networks as our classification approach for this thesis.

Chapter 4: Spontaneous Speech Emotion Identification

The ultimate goal of speech emotion identification is to automatically identify the emotional state of a speaker from the acoustic feature set of his/her speech. In this Chapter, we deal with the modeling aspects of speech emotion identification from spontaneous communication. Identification of speech emotion is the process of analyzing speech features in order to get association between the various speech features and the underlying emotional state of the speaker. It determines mainly which speech features can be used to accurately identify emotions that people experience in a day to day conversations.

As discussed in Chapter 3, most of the previous research works focused on modeling emotion recognition from acted speech which are exaggerated and can be easily identified by any listener. However, in real situation people do not tend to express their emotion in a way that can be identified easily. This fact indicates that studies involving emotion portrayal is not an ideal approach on building applicable speech emotion classification model even though it has a lot of significances in identifying the relationship between speech signals and emotional state of the speaker.

In a nutshell, this thesis work focuses on three major components in order to achieve emotion identification from spontaneous communication. These components include corpus preparation from call center spontaneous telephone dialogs, construction of speech emotion recognition model and developing a prototype application to demonstrate the speech emotion classification.

A natural speech emotion corpus preparation focuses on the process of collecting spontaneous telephone dialogs from a call center and performing multi-levels of segmentation and then, annotation of the segmented speech data using human experts. The second component, speech emotion model, is composed of signal processing, feature extraction and selection as well as classifier training. Signal processing involves preprocessing (such as audio format and channel conversion), windowing and normalizing. Feature extraction deals with computing the acoustic features from each segmented speech data. The feature selection task identifies the most relevant feature set in relation to emotions out of the extracted features. Finally, classification maps feature vectors onto emotional state classes through classifier training. The last component uses the speech emotion model as an input to recognize the emotional state of a separate test speech data (which are not part of the training dataset) in near real-time.

The high-level architecture for the overall spontaneous speech emotion classification process is shown in Figure 4.1.

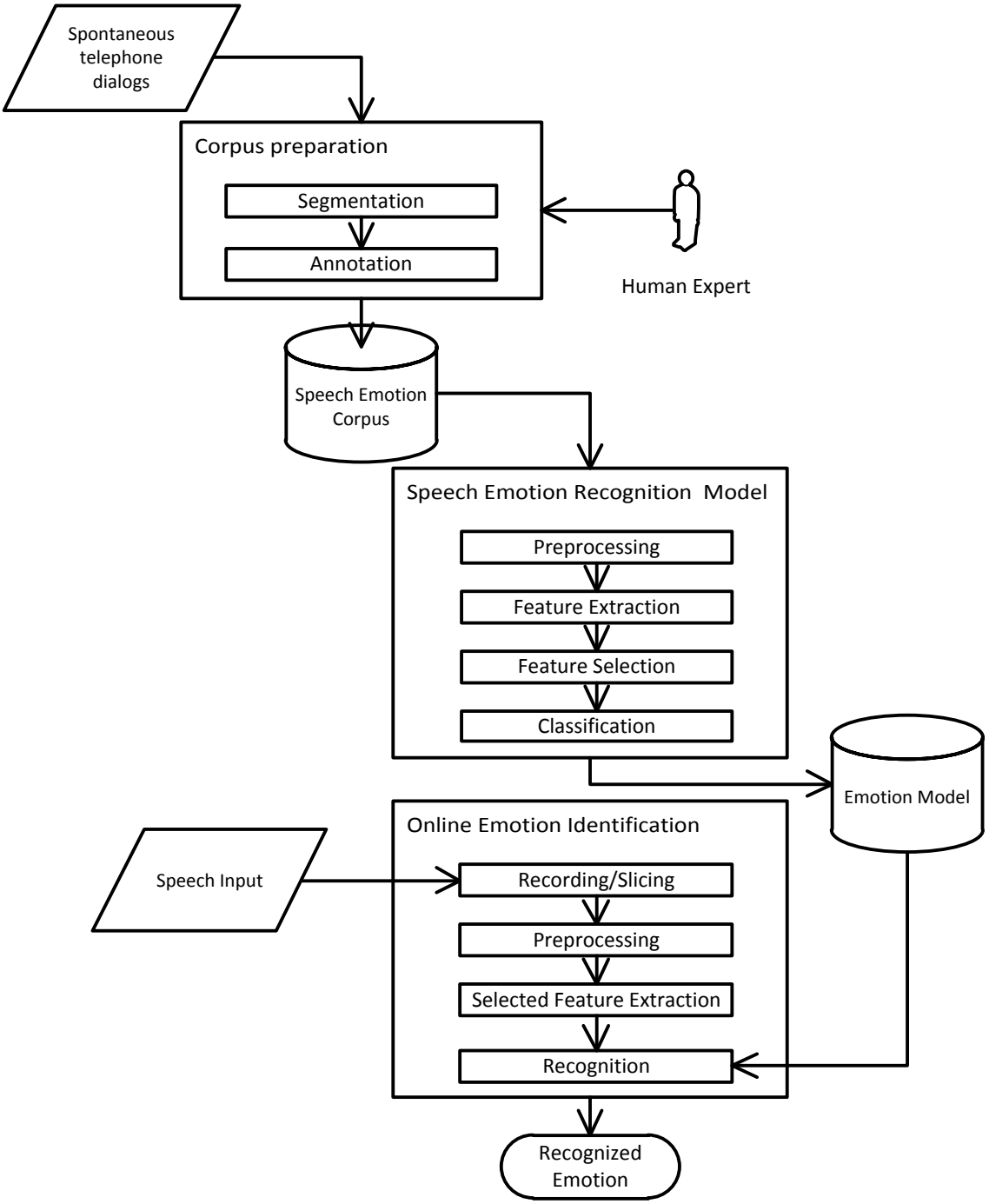

Figure 4.1: High-level Architecture for Spontaneous Speech Emotion Identification

4.1 Corpus Preparation

In order to tackle the limitations of acted emotions, we focused on the preparation of a natural spontaneous speech corpus having naturalistic emotions. The motivation behind this is the limited number of emotion corpuses available for research purpose and most of

them are composed of acted emotions called emotion portrayals. Even though no linguistic speech features are considered in this thesis, preparing a natural speech corpus containing Amharic speakers can enable us to address all aspects of emotions that are not universal across several cultures.

For the purpose of this thesis work, a natural spontaneous speech emotion corpus is collected from recorded archives of counseling sessions in a call center. Most of the recordings used as input for our corpus had background noises and poor audio quality. Among the recording archives, we have carefully chosen 35 dialog sessions with relatively good recording quality. The corpus preparation task includes multiple levels of segmentation and annotation of the segmented speech with emotion classes that potentially represent the emotional state of the speech. The use of these speech data respected ethical considerations securing the anonymity of the speakers. The overall process to build the spontaneous speech emotion corpus is shown in Figure 4.2.

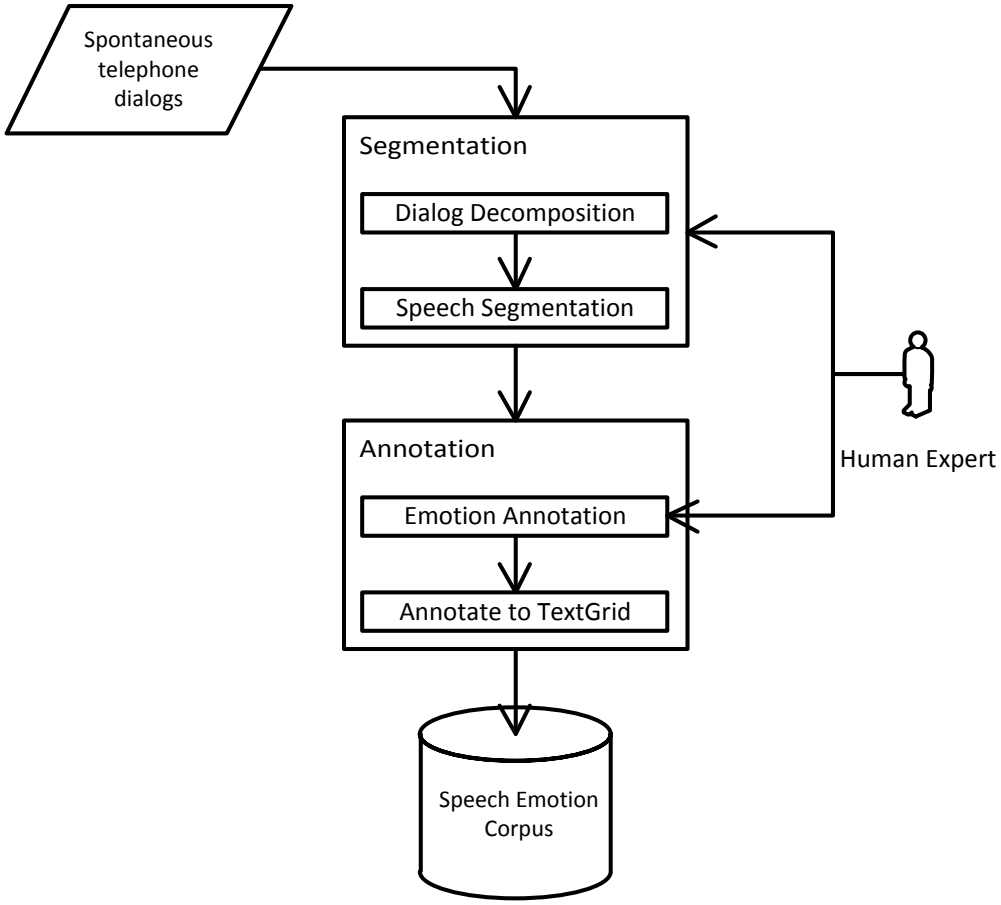

Figure 4.2: Speech Emotion Corpus Preparation

The following subsequent sections discuss the process we went through to build the spontaneous speech emotion corpus.

4.1.1 Segmentation

The segmentation phase is composed of two major tasks called dialog decomposition and speech segmentation.

Dialog Decomposition

Common telephone conversation involves two parties. In order to get each individual's speech separately, speaker turns of each party in a dialog is manually decomposed using Audacity (a free, open source, cross-platform software for recording and editing sounds). We then got isolated speeches from 18 male and 17 female speakers.

Speech Segmentation

In speech signal processing, waveforms need to be segmented to shorter parts (i.e., variable or fixed length) and annotated. This is even more important for spontaneous speech emotion recognition as the emotional state of a speaker in spontaneous speech is rarely constant but changes overtime. Unless the speech waveform is sliced down to a shorter segment that potentially contain a single emotion, it is difficult to annotate it with a single emotion class that correctly reflects the emotional state of the speech. The decision on the length of the segmentation units for natural spontaneous speech is not as simple as acted speech emotion (i.e., in acted corpus the entire utterance is assumed to have a single emotion). In spontaneous speech, each segment of speech needs to have a constant emotional state to annotate the speech data. In other words, more than one emotional state can be reflected in a single speech and difficult to isolate the boundary where a particular emotion leaves and another emotion starts. The objective here is to find a speech unit containing a constant emotional state which in practice vary in length and difficult to identify in spontaneous communication.

In this thesis, the recorded speeches are segmented manually into variable length intermediate chunks ranging from 1 to 15 seconds long. Intermediate chunks level approach is used as it compromises between the possibility of multiple emotion occurrences within longer segments such as utterances or speaker turns, and the complexity as well as challenges of finding sufficient emotional information in shorter segments such as word level segments. We followed a top-down approach to segment the recorded speeches into chunks where they are short enough so that constant emotional state can be assumed while keeping them long enough for the human annotators to listen and label the speech with no difficulty. To do this, we first segmented the recorded speech

into speaker turns and then we further divided all turns into utterances. Since there is still a possibility for the emotional state to be changed within a single utterance, we further divided all utterances longer than 15 seconds into chunks using noticeable pauses in the utterance as a segmentation boundary. Finally we ended up with 1200 variable length intermediate chunks with length ranging between 1 to 15 seconds long.

4.1.2 Annotation

Annotating speeches with emotional state is a very difficult task, especially when dealing with natural spontaneous speeches. As discussed in Section 2.4, the number of primary emotions varies from six basic emotions to the larger lists of emotional state classes. At least 60 emotional state classes are required to define emotions that occur with a reasonable frequency in everyday life. However, the more the number of classes used, the less the classification performance will be. This means that statistically reliable recognition can only be achieved with smaller number of emotional state classes. Due to this reason, we adopted the mapping onto cover class approach in order to reduce the number of emotional state classes as well as to combine less frequently appearing emotional states together onto a broader cover class to balance the distribution of the emotional states in the natural speech emotion corpus. Moreover, though it is difficult to avoid the application of down-sampling and up-sampling of the data to get evenly distributed data set across the emotional classes, cover class approach helps to minimize the data loss that can occur due to down-sampling. We also tried to reduce up-sampling of less frequent classes as no additional new information about the less frequent classes can be gained as the data are just duplicated to level with the other classes.

The annotation is done at the chunk level (intermediate unit between word and utterance or speaker turn). That is, each manually segmented speech data are labeled with emotional state classes. The annotation is carried out by 3 experienced senior psychologists having a counseling background with a good understanding of human feelings and emotional state reflections.

The annotators listened to each chunk independently to identify the emotional state of the speech data. Each chunk of speech is then annotated separately by the human experts. The experts had the freedom to annotate the chunks with any emotional state classes (even other than the 6 basic emotions) as restricting the choice of classes would not be an effective solution to the real-world problem. Hence, this open class annotation approach

gave the flexibility in annotating the speeches with emotional state classes other than the primary emotions that may appear frequently in a day to day conversations.

As a result, the experts annotated the 1,200 speech chunks with 29 emotional states. The complete list of the emotional state classes are listed below.

Table 4-1: The 29 Emotional States classes

Anger	Calm	Excitement	Hurt	Relief
Annoyed	Compassion	Fear	Insecurity	Sadness
Anxious	Despair	Happy	Panic	Satisfied
Arrogance	Disturbed	Hatred	Passion	Scared
Ashamed	Eager	Hope	Proud	Stressed
Broken	Empathy	Humiliation	Relaxed	

Once we organized all the emotional state labels used for the annotation, the mapping of the emotional classes onto broader classes is carried out. For the purpose of reducing the number of emotional state classes, we first applied a tree-structured list of emotions. This approach is a catalogue of emotions grouped into primary, secondary and tertiary emotions as shown in Table 4-2 [30].

Table 4-2: Inventory of Emotions

Primary Emotion	**Secondary Emotion**	**Tertiary emotion**
Love	Affection	Adoration • Fondness • Liking • Attractiveness • Caring • Tenderness • Compassion • Sentimentality
	Lust/Sexual desire	Desire • Passion • Infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement • Bliss • Gaiety • Glee • Jolliness • Joviality • Joy • Delight • Enjoyment • Gladness • Happiness • Jubilation • Elation • Satisfaction • Ecstasy • Euphoria

Primary Emotion	Secondary Emotion	Tertiary emotion
	Zest	Enthusiasm • Zeal • Excitement • Thrill • Exhilaration
	Contentment	Pleasure
	Pride	Triumph
	Optimism	Eagerness • Hope
	Enthrallment	Enthrallment • Rapture
	Relief	Relief
Surprise	Surprise	Amazement • Astonishment
Anger	Irritability	Aggravation • Agitation • Annoyance • Grouchy • Grumpy • Crosspatch
	Exasperation	Frustration
	Rage	Anger • Outrage • Fury • Wrath • Hostility • Ferocity • Bitter • Hatred • Scorn • Spite • Vengefulness • Dislike • Resentment
	Disgust	Revulsion • Contempt • Loathing
	Envy	Jealousy
	Torment	Torment
Sadness	Suffering	Agony • Anguish • Hurt
	Sadness	Depression • Despair • Gloom • Glumness • Unhappy • Grief • Sorrow • Woe • Misery • Melancholy
	Disappointment	Dismay • Displeasure
	Shame	Guilt • Regret • Remorse
	Neglect	Alienation • Defeatism • Dejection • Embarrassment • Homesickness • Humiliation • Insecurity • Insult • Isolation • Loneliness • Rejection
	Sympathy	Pity • Mono no aware • Sympathy

Primary Emotion	Secondary Emotion	Tertiary emotion
Fear	Horror	Alarm • Shock • Fear • Fright • Horror • Terror • Panic • Hysteria • Mortification
	Nervousness	Anxiety • Suspense • Uneasiness • Apprehension (fear) • Worry • Distress • Dread

As a result, we substituted the 29 originally annotated labels with their corresponding cover classes to reduce to 6 emotional state classes. Then a majority voting technique (agreement between at least two annotators) is applied for selecting the perceived emotional state class for each speech chunk. At this point, we obtained 1,200 speech chunks having the 6 emotional state labels (Anger, Joy, Fear, Love, Sadness and Surprise) as shown in Table 4-3. We also got 109 speech data annotated differently by the annotators. These chunks with conflicting labels are excluded from the training dataset but kept as a separate test set for future experiment.

Table 4-3: The Distribution of Emotion Classes after Mapping onto Primary Emotions

Emotional state classes	Frequency in the corpus
Anger	258
Joy	360
Fear	283
Love	19
Sadness	269
Surprise	11
Total	*1200*

Though mapping of the 29 emotional states to primary emotions significantly reduce the number of classes, dealing with a 6-class problem is still difficult due to the limited size of speech corpus. Moreover, the unbalanced distribution of emotional classes over the speech corpus can significantly affect the performance of the model. Therefore, due to

their low frequency in the corpus, the other two categories, Love and Surprise, have to be either omitted or further mapped onto a broader cover classes. We chose to keep them because discarding less frequently observed emotions is not a good approach for real application. Thus we decided to map Joy, Love and Surprise emotions onto a broader cover class called Positive. Finally, we are left with 4 emotional state classes containing Anger, Fear, Sadness and Positive. The number of speech chunks that belong to each emotion category is shown Table 4-4.

Table 4-4: The Distribution of the Final 4 Classes after Mapping onto Cover Classes

Emotional state classes	**Frequency in the corpus**
Angry	258
Fear	283
Sadness	269
Positive	390
Total	*1200*

Sample list of wav files along with emotion annotation is presented in Annex A.

The final task, mainly labeling of each speech chunks with its corresponding emotional state class, is carried out using Praat. For each annotated speech chunk, we feed the corresponding label into TextGrid file. The TextGrid object is generated automatically having the name of the input speech file and with ".TextGrid" file extension. A single tier TextGrid object with a tier name called "Emotion" is created for each audio file. Single tier TextGrid is selected because each audio file is assumed to contain a single emotional state class and the label is applied on the entire chunk.

As shown in Figure 4.3, we created a new empty TextGrid from a sound object named "02fc0006A.wav". In this way the time domain of the TextGrid will automatically be aligned with that of the sound file. The waveform and spectrogram of the speech chunk is displayed in the first two upper part of the TextGrid object window and the third section is a single tier containing the emotional state that represents the whole chunk (in this case, the speech chunk is labeled with Anger). The text window at the top also shows the label

of the selected interval tier. The bottom section displays the visible part and total duration of the speech chunk respectively.

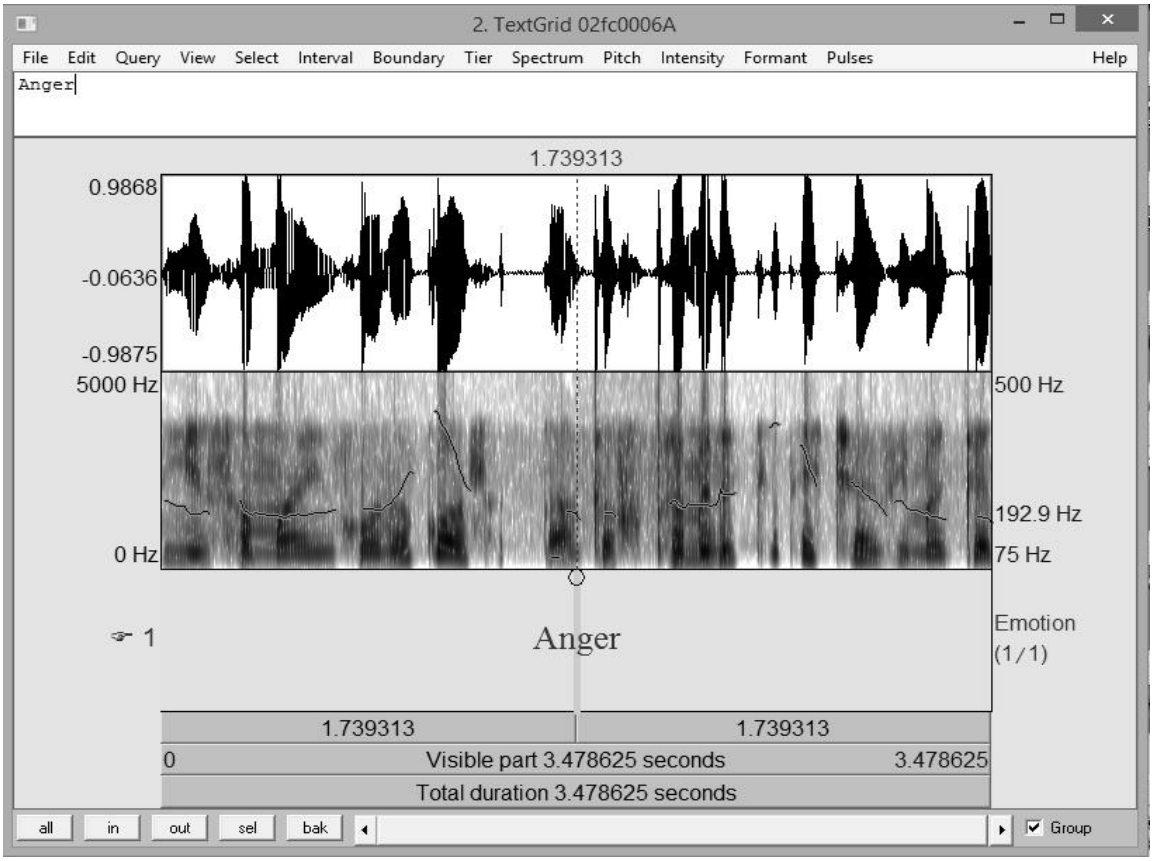

Figure 4.3: Sample Speech Waveform and Spectrogram along with its Annotation

4.2 Speech Emotion Recognition Model

As stated in Section 2.4, the main problem in the emotion recognition field is the lack of agreed upon benchmark database, feature extraction and selection or classification methods. This makes the performance comparison among different databases or methods very difficult. The ultimate objective here is to empirically select a classifier with a reasonably high classification performance and a feature set that best represents the emotional state of the speech signal.

We first carried out different preprocessing tasks to prepare our corpus in such a way that is suitable for speech signal analysis. We then extracted as many speech features as possible from each speech data and further investigated the extracted features to select only the most relevant features (i.e., a feature vector that potentially convey emotional information). Finally, each feature vector is mapped onto a potential emotional state class in the classification module. In order to conduct classification, the classifier is trained

using the training dataset containing a selected feature vector. Besides, the test dataset is used to test the performance of the classification model. The general structure of the emotion identification model is shown in Figure 4.4.

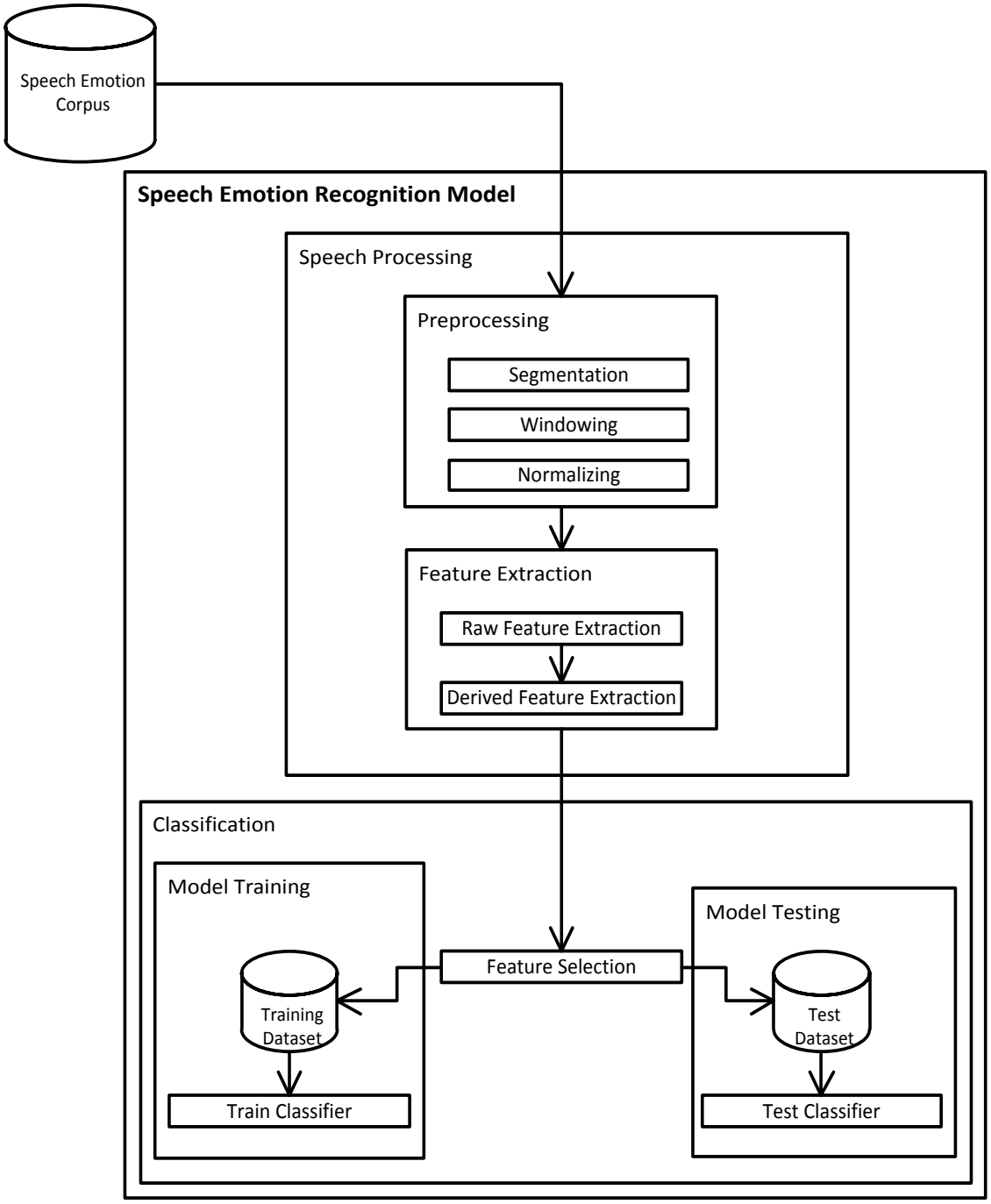

Figure 4.4: Speech Emotion Recognition Model

4.2.1 Preprocessing

The telephone dialogs obtained from the call center were recorded by Sony ICD-UX recording device. The waveforms were recorded in MP3 format with bit rate of 128kbps

at a sampling rate of 44.1 kHz, stereo channel. The first step we took is transforming those recorded speech signals into a common representation or uniform standard format.

We conducted preprocessing on each speech chunk before applying any speech signal analysis techniques. These speech chunks are preprocessed in order to reduce the effort of the successive processing steps as well as to increase the later classification performance. The preprocessing tasks including the channel and format conversion, down-sampling as well as segmentation and normalization are carried out on each chunk using Audacity. As a result, each speech chunk is converted to a mono channel sound and down-sampled to 16 kHz to comply with the standard sampling rate in speech processing. The preprocessed chunks are then saved as '.wav' files.

Finally, in order to reduce speaker and recording variability, we conducted intensity and pitch normalization on each speech chunk.

Intensity normalization: the speech chunks are scaled such that the average Root Mean Square (RMS) energy of the emotional speech database are the same for each speaker. Intensity normalization is separately applied for each speech chunk in the corpus before the analysis process. The goal of this normalization is to compensate for different recording settings in the corpus.

Pitch normalization; the pitch contour is normalized for each subject. The average pitch across speakers is estimated. Then, the average pitch value for the speech emotion corpus is estimated for each speaker. The goal is to minimize speaker dependency while keeping the relevant emotional information in the speech.

4.2.2 Feature Extraction

The feature extraction component of our model maps each speech waveform into a set of speech features called feature vector. We only used the acoustic characteristic of speech signal to recognize the emotional state of a speaker. In other words, we focused on how the words are spoken rather than the actual semantics of the speech. These acoustic features can be divided into prosodic features, spectral features, and voice quality features. Recalling Chapter 2, there is no common agreement on the number and type of features that yield the best recognition performance. In this work, we have decided to extract as many features as possible and select the most relevant feature set out of them.

As discussed in Chapter 2, speech feature vectors can be categorized as short-time (segmental) and long-time (suprasegmental) information based on their temporal structure. It is also discussed that due to the dynamic nature of speech signal, it is common in speech processing to divide a speech signal into small segments called frames as within each frame the signal is considered to be approximately stationary. Short-time features are extracted over a frame, while the long-time information are computed over the entire length of the speech segment.

In this thesis, we used global statistic features because, long-time information identify emotions more efficiently than short-time features. First, each speech chunk is segmented into frames with overlapping windows. The size of the overlapping window is 25ms, and the step between two successive windows is 10ms. In other words, the first set is taken from the beginning of the speech chunk and the second set is 10ms to the right of the first set. This is repeated until the window covers the entire speech chunk. Then, the feature vector is extracted from each frame consisting of pitch, energy, MFCC, LPC, LFCC, and voice quality features. Finally, the statistical functional of these raw features such as mean, maximum, minimum, standard deviation, and median, are computed over time to obtain the global statistics features of the speech chunk.

Finally, a total of 170 acoustic features are calculated from each speech chunk. The Praat script used to extract these features is presented in Annex B. The total number of chunk-level acoustic features derived from each raw feature along with their corresponding categories are shown in Table 4-5.

Pitch: a pitch is the Fundamental Frequency (F0) of the quasi-periodic speech signal. The detection of this acoustic periodicity is done on the basis of autocorrelation method available in Praat. The normalized pitch contour is computed with the default pitch floor and pitch ceiling values (pitch floor = 75 Hz and pitch ceiling = 600 Hz) with a time step value of 10ms. Next, the statistical parameters of the raw pitch is calculated to obtain a total of 14 pitch features. In order to reduce gender differences, we normalized the pitch feature vector. Pitch mean, median, first and third quartile are normalized by minimum and maximum pitch for each chunk. The normalized mean is calculated using the following equation.

$$mean_{norm} = \frac{mean - min}{max - min}$$

Table 4-5: The Summarized List of Features and their Corresponding Categories

Raw Feature	Number of Derived Features	Category	Total Number of Features per Category
Intensity	11	Prosodic	29
Pitch	18		
Jitter	5	Voice Quality	18
Shimmer	6		
HNR	7		
F1	10	Spectral	123
F2	10		
F3	10		
F4	10		
F5	10		
Spectrum	12		
MFCC	18		
LPC	21		
LFCC	22		

Intensity: the values in the speech chunk are first squared, then convolved with a Gaussian analysis window. The effective duration of this analysis window is calculated in order to guarantee a periodic signal is analyzed as having a pitch-synchronous intensity ripple not greater than 0.00001dB.

$$v = \frac{3.2}{(minimum\_pitch)}$$

where v is the analysis window and $minimum\_pitch$ represents the minimum periodicity frequency in the signal.

We set the minimum pitch of the resulting intensity contour to 100 Hz with a time step of 0 second. This made the time step computed as one quarter of the effective window length.

Finally, the global statistics of the raw intensity is computed in the whole speech chunk to get a total of 11 energy features.

MFCC: For spectral features such as MFCC, each speech signal is blocked into N overlapping samples (N varies depending on the length of the signal) with a frame size of 25ms Hamming window sliding at 10ms. For each overlapping frame, the feature vectors were calculated and then the average vector for the entire speech signal was used as a final value. The importance of overlapping frames and windowing are discussed in Section 2.6. We then computed the spectral measure consisting of the norm of the absolute vector derivative of the first 12 coefficients for MFCC. In addition to the mean of each of the 12 coefficient, statistics based on MFCC such as mean, standard deviation, minimum and maximum are calculated across all frames in each speech chunk. Finally, we obtained 18-dimentional chunk-level MFCC features.

Linear Frequency Cepstrum Coefficients (LFCC): in order to represent cepstral coefficients on a linear frequency scale, we computed the first 16 cepstral coefficients as a function of time with constant sampling period of 25ms sliding 10ms at a time.

Linear Predictive Coding (LPC): similar to MFCC, we set the effective duration of each analysis window to 25ms with 10ms overlap between two consecutive analysis windows. We computed the first 16 linear prediction coefficients representing filter coefficients as a function of time.

Formant-based Features: the first 5 formants of the speech signal are tracked over time in order to model the change in the vocal tract shape. We performed a short-term spectral analysis, approximating the spectrum of each analysis frame by a number of formants. The speech signal is resampled to a sampling frequency of twice the value of Maximum formant and pre-emphasis is applied. For each analysis window, a Gaussian-like window is applied and the LPC coefficients are computed with the Burg algorithm. In this work, a maximum formant of 5500 with a window length of 25ms and time steps of 10ms are used to extract the formant frequencies (F1, F2, F3, F4, and F5). The global statistics of formant contour across all frames is then calculated in order to get a total of 50 formant features at chunk-level.

Voice Quality Features: measure the cycle-to-cycle variation of the period length (reciprocal of the fundamental frequency) and the peak amplitude, respectively. In our experiment, HNR, the measure of the degree of periodicity of a sound, as well as jitter and

shimmer of the glottal pulses is calculated for the whole speech chunk. The algorithm performs an acoustic periodicity detection on the basis of a forward cross-correlation analysis. The complete list of extracted features are presented in Annex C.

4.2.3 Feature Selection

After the feature extraction stage, each speech chunk is represented by a feature vector consisting 170 features. However, dealing with a large number of features is rather complex and time consuming. In other words, the complexity of the classification model increases with increase in dimensionality of feature space. Besides, some features might convey redundant, irrelevant, or contradicted information which potentially reduce the performance of the classification model.

As discussed in Section 2.6, the selection of speech features appears to be application and/or data dependent. The main objective of feature selection here is to choose the most significant feature sets in spontaneous speech data for building the speech emotion recognition model.

For feature selection task, Weka is used to select the most relevant speech feature set that potentially convey emotional state information. In this thesis work, we extracted as many acoustic features as possible and let the generic attribute selection algorithms implemented in Weka to find out the most relevant feature set for our 4-class problem. We combined the 3 generic attribute evaluators in Weka (GainRatio, InfoGain, OneR) and performed an attribute ranking. Thus, the average of the 3 attribute evaluator with Ranking search method is applied to reduce the original 170 speech features to 33 best acoustic features. As a result, we obtained a feature set composed of 4 prosodic, 28 spectral and 1 Voice Quality features for Multilayer perceptron classifier. The selected acoustic features are listed in Table 4-6.

Once the best feature set is selected, a Praat script is used to extract only the selected features for MLPNN classifiers. This greatly enhance the extraction process for the testing and ultimately for the online emotion identification task.

Table 4-6: List of Selected Features for Multilayer Perceptron Classifier

Derived Features	Raw Feature	Category
Pitch Mean Pitch_medianQ Pitch_lowerQ Period_mean	Pitch	Prosody
HNR_min	HNR	Voice Quality
F1_mean F1_medianQ F1_lowerQ	F1	Spectral
F3_mean F3_medianQ F3_lowerQ	F3	
F4_mean F4_medianQ F4_lowerQ	F4	
Spectral_COG Spectrum_sd	Spectrum	
Ltas_freqMax	LTAS	
mfcc1_mean mfcc6_mean mfcc8_mean mfcc11_mean MFCC_max MFCC_sd	MFCC	
lpc1_mean lpc4_mean Lpc5_mean	LPC	

Derived Features	Raw Feature	Category
LFCC_meanEnergy lfcc2_mean lfcc7_mean lfcc12_mean LFCC_mean LFCC_max LFCC_sum	LFCC	

4.2.4 Classification

After the feature extraction and selection, each speech chunk is represented by a feature vector. The task of the emotion classification is to map the acoustic features to the perceived emotional state classes. The main objective here is that the selection of the best classification algorithm that is suitable for the kind of speech corpus as well as the type of feature vector used, in order to attain a maximum recognition performance. This is because, the choice of the classifier appears to be dependent on the kind of dataset being used (acted, derived, or spontaneous), the type feature vectors selected (global or local), the number of features (the whole feature or reduced set), and so on.

In this thesis, we formed 7 different datasets comprising of the selected feature vector as well as all the possible combinations of feature types to train and test MLPNNs classification algorithm. Hence, we carried out the classification with 7 different sets of inputs. We used the same dataset to train and evaluate the classifier using 10-fold cross validation technique. The MLPNN algorithm consists of two steps:

Forward pass: the predicted outputs are calculated corresponding to the given inputs.

Backward pass: partial derivatives of the cost function with respect to the different parameters are propagated back through the network.

The classifier takes the extracted relevant feature vectors as an input and then outputs the recognized emotional state class of the speech chunk for which the input features were extracted. In other words, the input layer takes the selected feature vectors for each speech chunk and the number of nodes in the output layer correspond to the 4 emotional classes to be recognized. The hidden layer uses a sigmoid transform function as the activation

function. The parameter of the sigmoid activation function is automatically adjusted during the training procedure. The number of hidden nodes are calculated follows:

$$neuron = \frac{attrib + class}{2}$$

where $neuron$ is the number of computation nodes, $attrib$ is the number of feature vectors to be used as an input to the classifier, and $class$ represents the number of emotional state classes.

Figure 4.5 shows the input, hidden and output layers of neural network.

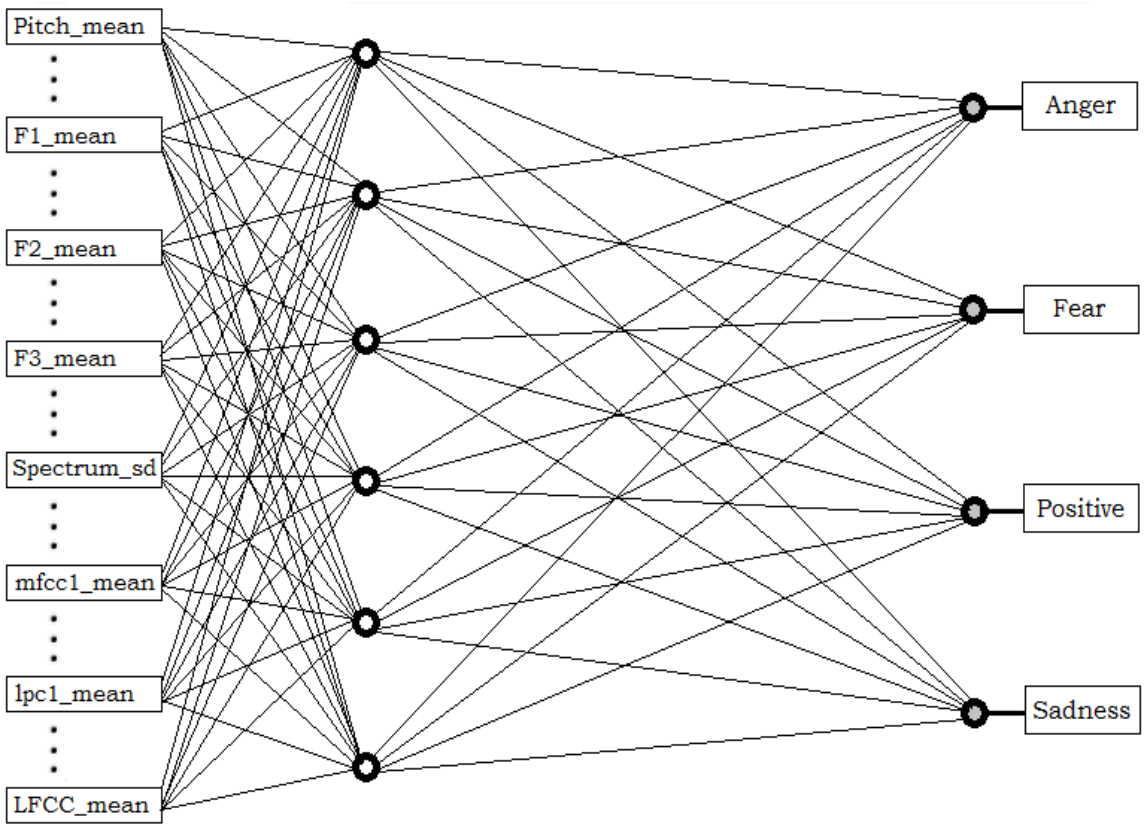

Figure 4.5: Model of a MLPNN

The first basic element, a set of connecting links from different input features consisting of the 33 combined feature vector. The second basic elements, the adder that sums the input signals and the activation function f, comprise the hidden layer limit the amplitude of the output layers representing emotion classes containing Anger, Fear, Positive and Sadness.

4.3 Online Emotion Identification

So far in this thesis work, we focused on the offline automatic classification of emotional state of the speech signal, where both the training and testing are done using a recorded speech corpus. One of the objectives of this thesis work is to identify the emotional state of the speaker while he/she is speaking. Thus, the next step will be capturing the speech online and analyze it to recognize the emotional state of the speaker in near real-time based on the constructed model. In this section, the task of analyzing and responding to emotional state in near real-time while a user speaks will be discussed.

Both the offline and online recognitions share similar characteristics such as speech segmentation, feature extraction and classification phases. However, online recognition introduces additional challenges to the automatic emotion recognition that need additional efforts. The first challenge is the trade-off between response time and capturing meaningful emotion. That is, high response time such as word-level recognition can respond fast however, identifying the emotion for each and every word may not be relevant. Even if it is relevant, it is computationally expensive. In addition, several subsequent words may carry the same emotion and redundant result can be observed. This is a critical factor for online speech emotion recognition as the system needs to respond with the recognized emotional information of the speaker in near real-time. The following are the main approaches adopted to achieve near real-time response.

- **Fixed length segmentation:** breaking down the continuously incoming speech signal into meaningful chunks delays the overall response time as it requires further linguistic knowledge such as pause detection or word recognition. However, fixed length segmentation does not require such knowledge and significantly improve the response time. The decision on the duration of the segment can be left to the user so that the user can define the value based on the requirement. For instance, in the case of conversations where stable emotional state is observed (i.e., if similar state for adjacent units observed frequently) the user can increase the duration. Whereas if the emotional state of the speaker changes frequently within a short period of time, the user can decrease the duration accordingly.
- **Feature Extraction and Selection:** features are extracted from the entire intermediate speech chunks automatically without any knowledge of speech units

such as word or utterance. In addition, optimal feature selection approach enables to reduce the dimensionality of feature vectors and thus improve the response time.

- **Classification:** different classifiers behave differently based on the kind of speech corpus used, types and natures of the speech features, and others. Though simple classifiers such as Naïve Bayes run faster than complex classifiers, their simplicity come with the cost of accuracy. Though complex classifiers train slowly, we still preferred to use MLPNN classifier in a similar way as we did for online recognition. This is because the training task is carried out offline where time is not considered as a critical factor.

Fast segmentation of the continuously incoming speech signal and instant application of the recognition process on each segment is very important for online emotion recognition. Fixed length segments may not contain semantically meaningful speech as word or utterance based segmentations. However, this task requires no further speech recognition knowledge. Thus, it improves the response time and makes it ideal for online speech emotion classification.

For online emotion recognition, a continuous speech is broken down into fixed length non-overlapping speech segments. Therefore, the first segment recoding, r_1 runs for x seconds (x depends on the user input) and stops. The next recording r_2 starts immediately to capture the next x seconds and this will continue till the end of the speech r_n. Meanwhile, the moment the first recording stops, the analysis on speech segment starts immediately. This includes the feature extraction, the application of the speech emotion model against the feature vectors and return the emotional state of the current segment. This process continues repeatedly that is, whenever we capture the i^{th} recording (r_i), the system analyzes the $(i-1)^{th}$ segment against the speech emotion recognition model to identify and display the emotional state reflected in that segment. Note that there is no need to perform feature selection this time as it is already done offline. The overall process of the online emotion recognition is illustrated in Figure 4.6.

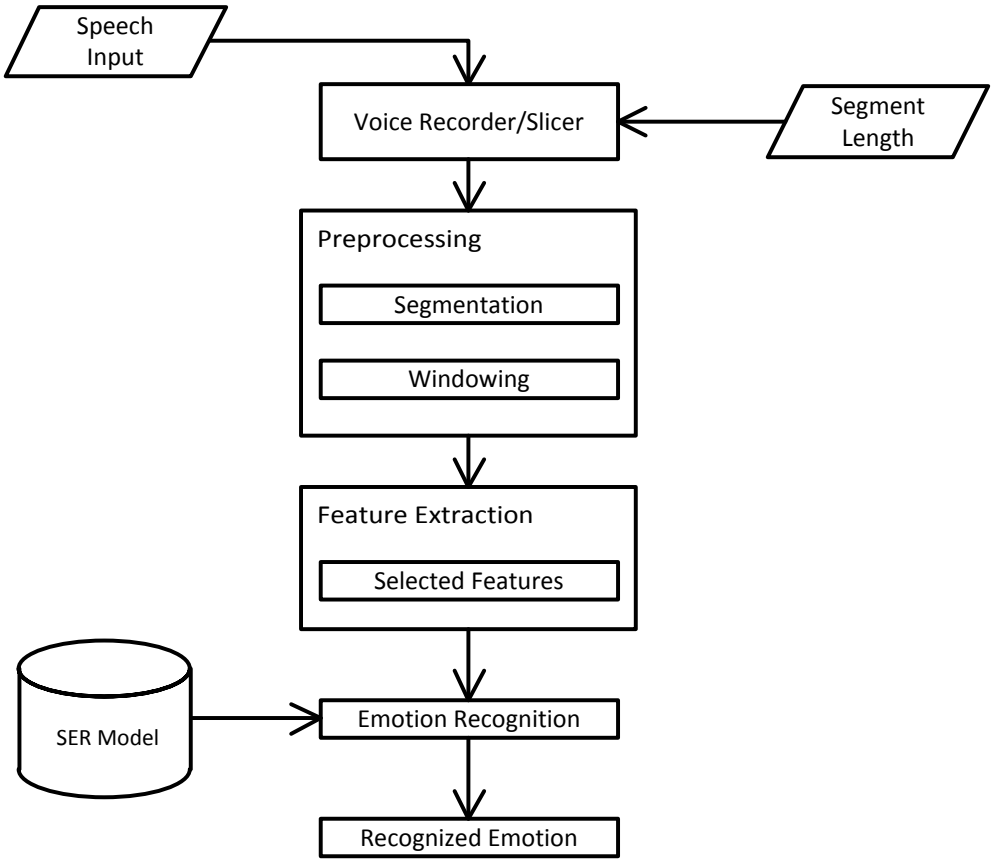

Figure 4.6: Online Emotion Identification

Chapter 5: Experiment

Chapter 4 focused on the design of spontaneous speech emotion recognition model that is capable of performing emotion recognition in near real-time. In this Chapter, the experimental setup, tools and methodologies used will be discussed in detail. The rest of this Chapter is organized into two major sections. Section 5.1, briefly demonstrates the design of a prototype application for the proposed spontaneous speech emotion recognition model. This section discusses the tools and programming language used in building the prototype and the rationale behind the selection as well as the main components of the prototype. Next, the experimental procedure and the results obtained will be discussed in Section 5.2. Finally, Section 5.3 will discuss the evaluation and performance analysis of our proposed model.

5.1 Prototype Design

The prototype for the online emotion recognition is developed to demonstrate the applicability of the proposed model. In this section, the proposed solution will be illustrated by means of a demo application where a continuous speech input is segmented into fixed length chunks and analyzed for the emotional state of each speech chunk in near real-time. The application displays the recognized emotion for each speech chunk as soon as the recording for the specified duration is completed. Our speech emotion model can take any speech as an input, segment it into fixed length chunk and predict the emotional state for each speech segment.

5.1.1 Tools and Programming Language

In order to justify the proposed speech emotion recognition model as well as the identification of emotion from spontaneous communication, we developed a prototype using Java Programming Language. Java is selected as a development environment as it offers flexible APIs suitable for this work such as Java Sound API to work with sound files and Java Weka API to interface with Weka.

Besides, Praat scripting language is used for speech signal processing such as preprocessing and feature extraction. Weka machine learning tool along with its API for Java is also used to carry out the classification task. The rationale behind the selection of these tools is that they are freely available yet powerful and widely used tools on different researches. Moreover, they are easy to use due to their graphical user interface.

5.1.2 Experimental Setup

The experimental setting such as specification of computer hardware, operating system and other software we used to build the prototype application as well as to carry out the experiment is shown in Table 5-1.

Table 5-1: Hardware and Software Specifications

Hardware and Software		**Version**
Operating System	Windows 8.1 Enterprise	64 bit
Application Software	Praat	6.0.14
	Weka	3.7.12
	JRE	1.8
	Java Eclipse IDE	4.5
Hardware	Memory	4 GB
	CPU	Intel
	Speed	Core i3
	Hard disk	500 GB
	Sound	High Definition Audio Device
	Recording Device	Integrated Microphone
	Playback Device	Integrated Speaker

5.1.3 Components of the Prototype Design

The prototype accepts a continuous speech and slices down to user specified interval (10 seconds by default) and saves it in a wave file. As soon as it finishes the recording of the first segment, it calls Praat to execute the preprocessing and feature extraction script against the saved speech segment. The resulting feature vectors then will be saved in *.arff* file. Next, it interfaces with Weka to call the classifier and apply the emotion recognition model on the *.arff* file. Finally, the recognized emotion will be displayed back to the user the moment the recognition process is done for that specific segment of speech. This

process continues every time the next speech segment is recorded and saved in the specified location until it is interrupted by the user.

In order to accomplish the aforementioned tasks, a prototype application is developed using Java Development Kit. The source code is presented in Annex D. The prototype is composed of four components: Interface, Recorder, Praat and SER. The following consecutive paragraphs discuss each component in brief.

- **Interface:** enables the user to interact to the emotion recognition system. The interface provides the option to set the window size (i.e., the length of speech segment to be analyzed for emotion recognition at a time), to start and stop the recognition process, as well as to display the recognized emotion for each speech interval. The user can start the recognition with the default analysis window size (i.e., 10 seconds). For example, a sample screenshot in Figure 5.1 shows an online recognition with the default window size of 10 seconds. The result of the recognition indicates that the system returned "Positive" for the first 30 seconds of the speech and from 30 to 70 seconds "Fear" emotional state is identified. Then the identifier start detecting "Sadness" emotion from the continuous speech in real-time. In addition, the user can also perform offline emotion identification using prerecorded speech file of any length by clicking the "Browse" button and locating the file. The window size setting holds true for offline recognition as well. If required, the user can select the "Settings" tab to modify window size as depicted in Figure 5.2.

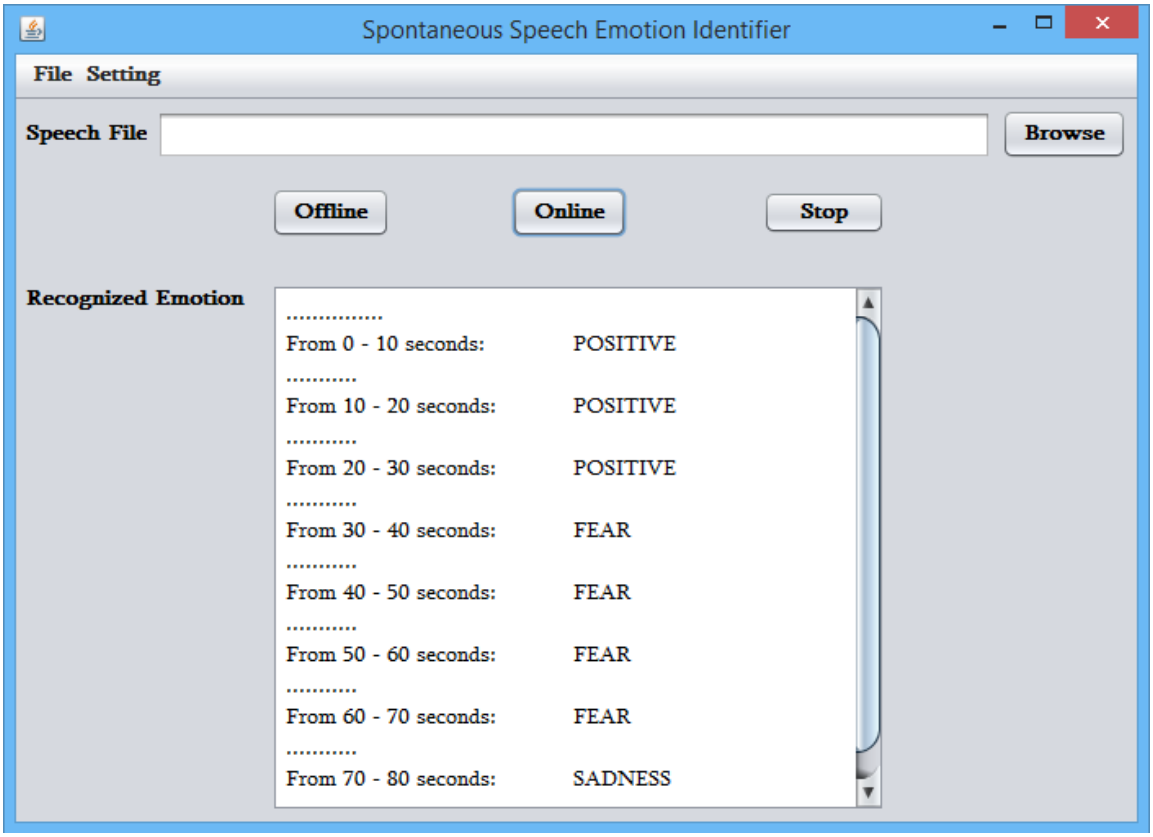

Figure 5.1: Online Speech Emotion Recognition System Interface

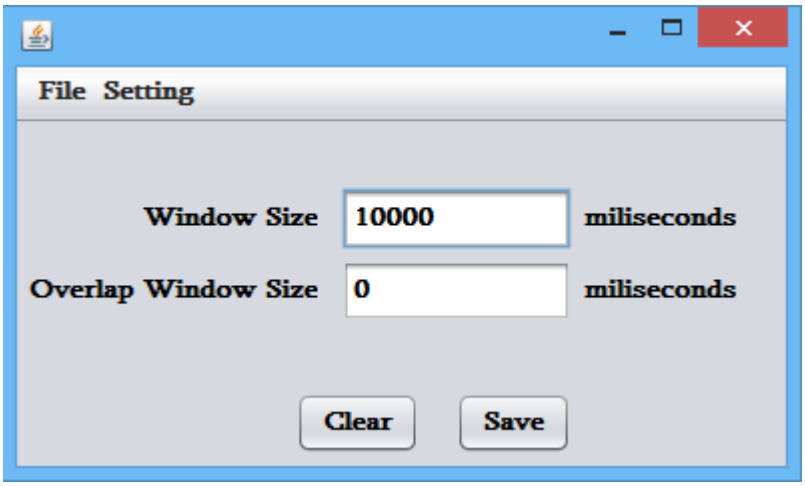

Figure 5.2: Online Speech Emotion Recognition System Setting Interface

- **Recorder:** is responsible for accepting the speech stream and segment it into a fixed length speech segment for analysis. This is accomplished by recording the continuous speech in every 10 seconds by default, and save the captured speech segments in sequences of wav files separately (such as, RAudio0.wav, RAudio1.wav, RAudio3.wav and so on) till the recording is interrupted by the user.

In other words, this component is responsible for slicing or segmenting the continuous spontaneous speech into the specified interval.

- **PraatFeatureExtraction:** waits the completion of the sliced wave file recording and run a saved Praat script against it the moment it is saved in the predefined location. The Praat script performs preprocessing and extract the selected feature vectors of the speech chunk and then saves in *.arff* file keeping the name of the input wave file.
- **SER:** waits till the RunPraat class finishes the extraction of feature vectors for the next *"wav"* file and takes the resulting *.arff* file to conduct the emotion recognition against it using the proposed speech emotion model. It then displays the recognized emotion back to the user.

The source code for the Java classes that are used to develop the prototype of the emotion identification is presented in Annex D.

5.2 Results

Experiments are conducted to evaluate the performance of the proposed model. In this section, the results of our experiments on identifying the emotional state classes in spontaneous speech is discussed. The experiments are conducted on the call center speech corpus consisting of 1200 segmented chunks. The chunk level features are classified using MLPNN as described in Chapter 4. In the first set of experiments, we focused on evaluating the effectiveness of all possible combinations of the three major categories of acoustic features (prosodic, spectral, voice quality features). As a result, the ultimate acoustic feature for natural spontaneous speech emotion recognition is used to build our speech emotion model. In addition to the evaluation of speech features, experiments using MLPNNs as a classifier are also conducted to evaluate the performance of the proposed speech emotion model. The evaluation methods used to evaluate the given classification task are presented next.

Recognition Rate: the total Recognition Rate (RR) is defined as the percentage of correctly classified samples (i.e., how many correct decisions the classifier made).

$$RR := \frac{1}{N}\sum_{k=1}^{K}(n_{kk} \cdot 100\%)$$

where k is the number of classes, n_{kk} is the number of times a class is correctly classified, N is the total size of the sample set.

5.2.1 Feature Evaluation

In this section, we evaluated the classification results of all the possible combinations of the three feature categories: prosodic features (i.e., pitch and energy features), spectral features (Formant, MFCC, LPC and LFCC) and voice quality features (jitter, shimmer and HNR) one by one. As a result, a total of seven evaluation scenarios containing seven unique combinations are obtained from the three feature categories (i.e., $2^3 - 1 = 7$). The list of unique combinations along with the results indicating the average of correctly classified instances for each scenario is depicted in Table 5-2 (The binary values under each category indicate their presence in the combination where, 1 is "Exists" and 0 is "Does not exist"). For each evaluation scenario, we performed feature selection independently to find out the best feature set as well as to reduce the dimension of the feature vector. Moreover, in the scenario where all the three categories are used, the number of features appeared from each category are also compared to evaluate the importance of each category for emotion identification assuming that the more features selected from a certain category, the more important that category is.

Table 5-2: Classification Results for the Seven Combinations of Features

Prosodic	Spectral	Voice Quality	Avg. Recognition Rate (%)
0	0	1	56.6
0	1	0	72.8
0	1	1	70.5
1	0	0	54.0
1	0	1	57.9
1	1	0	62.8
1	1	1	72.4

The result indicates that among the experiments done on all the possible combinations, three candidate feature sets achieve a promising recognition rate. Spectral feature, when used alone, gives the highest recognition rate of 72.8%. Another promising result of 70.5%

is observed by the aggregate feature set comprising spectral and voice quality features. Finally, the experiment involving all the three feature types (combined features) with the application of feature selection yields 72.4%. In addition to its high recognition rate, the existence of spectral features in each of the candidate feature vector shows that spectral features carry more relevant emotional cues.

In terms of recognition, no significant change is observed between the combined features and the spectral features. This implies that both can be taken as the best feature sets for spontaneous speech emotion identification. However, when we compare the size of each feature vector, the combined feature vector with 33 features clearly outperforms the spectral feature vector having 123 features. We conclude that the combined feature set is more suitable in terms of simplicity, manageability and minimized execution time.

Due to this reason, further experiments and comparisons are carried out in order to identify which spectral feature specifically contribute more for spontaneous speech emotion recognition. The results obtained from the 4 spectral features Formants, MFCC, LFCC and LPC along with a combined spectral feature set containing the two spectral features that exhibit the highest result is shown in Table 5-3.

Table 5-3: Classification Results for Spectral Features Types

Spectral Feature	**Avg. Recognition Rate (%)**
Formants	54.63
MFCC	61.45
LPC	51.66
LFCC	61.08
MFCC and LFCC	64.08

The result indicates that MFCC and LFCC exhibit relatively higher classification performance than Formants and LPC spectral features do. A combined spectral feature set containing MFCC and LFCC even provides a better recognition rate of 64.08%. This indicates that MFCC and LFCC convey relatively higher emotional information and are relevant for speech emotion identification from spontaneous communication.

5.2.2 Classifier Evaluation

The training of MLPNN model has been carried out with all the possible combinations taken from the three major categories of acoustic features involving prosodic, spectral and voice quality speech features.

Unlike acted emotional corpus, the size of natural speech corpus should be big enough to train the classifier. However, in reality this is difficult due to the reasons mentioned in the previous sections. Hence, we applied a 10-fold cross validation technique to increase the reliability of the result. In 10-fold cross validation, the data is split into 3 mutually exclusive subsets for training, validation and testing. We used 80% of the total dataset for training, 10% for validation and the remaining 10% is used for testing. The training repeats itself 10 times and use different one-tenth subsets of the data for testing and take the mean accuracy. The validation data is used in training to prevent over-fitting. The training, test and validation datasets are mutually exclusive in each run.

As stated in the previous section among the 3 candidate feature vectors, we picked the reduced set of combined features due to its optimal performance in terms of recognition rate and dimension. The evaluation of MLPNN classifier trained with the 33 combined features in order to classify the dataset containing 1200 instances of speech chunks is shown in Table 5-4.

The percentage of correctly classified instance representing the recognition rate of the classifier is measured using a 10-fold cross validation and the result indicates that 72.4% instances are correctly classified while 27.6% instances are incorrectly classified.

Table 5-4: Summary of Correctly and Incorrectly Classified Instances

	No. of instances	**% of instances**
Correctly classified instances	869	72.4%
Incorrectly classified instances	331	27.6%
Total number of instances	*1200*	*100%*

The confusion matrix displaying the classification results after using MLPNN classifier with the selected 33 feature vectors is shown in Table 5-5. In this matrix the accuracy and the false alarm rate are shown. The diagonal represents the accuracy for the respective

emotional classes and the columns representing the emotions (instances in the determined class). The rows represents the emotions recognized in the actual class. The errors or misclassified rate is shown outside the diagonal with non-zero values.

Table 5-5: Confusion Matrix for the Selected 33 features

Classified as →		Output Emotion			
		Anger	Fear	Positive	Sadness
Input Emotion	Anger	**179**	5	50	23
	Fear	3	**233**	29	18
	Positive	28	29	**298**	35
	Sadness	16	21	38	**164**

Performance is measured for all classes as absolute accuracy, computed from average precision and recall for MLPNN classifier. Unlike RR, Recall and Precision are specific for a single class. A class-level average accuracy in terms of true and false positive, precision and recall is shown in Table 5-6.

Table 5-6: Detailed Accuracy by Emotion Class

Emotion Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**
Anger	0.503	0.052	0.627	0.503
Fear	0.823	0.070	0.809	0.823
Positive	0.764	0.172	0.718	0.764
Sadness	0.686	0.092	0.683	0.686
Weighted Avg.	*0.724*	*0.109*	*0.721*	*0.724*

The result indicates that the MLPNN classification achieves the average precision of 0.721 and average recall of 0.724 for Anger, Fear, Sadness and Positive emotional state identification using the combined feature vector. The highest accuracy is achieved for Fear with precision (0.809) and recall (0.823), while relatively less accuracy is observed for Anger with precision (0.627) and recall (0.503).

5.2.3 Online Classification Evaluation

The previous sections focused on a systematic evaluation of the offline classification for automatic identification of spontaneous speech emotion. Fixed-time segmentation and classification tasks are done on prerecorded spontaneous speeches where time is not a factor.

In this section, a basic set of testing procedure was followed for the purpose of evaluating the performance of the online identification of emotion from spontaneous speech. To evaluate the performance of our online speech emotion recognition system, a group discussion consisting of 4 human subjects (2 male and 2 female) was made. None of the human subjects participated in the training of the model. The online recognition started the recognition and the result of every 10 seconds of the speech is displayed. Meanwhile, every 10 seconds of the speech are recorded and saved in a separate file along with the recognized emotion for evaluation. We used the built-in recording and playback device of the laptop running the prototype application.

After a couple of minutes, we stopped the automatic recognition and let the evaluators listen the recorded speech segments and annotate separately. We took 10 segments from each speaker randomly to get a total of 40 segments for evaluation. The evaluators had to choose between the 4 emotional state classes we considered in this thesis work. Once done, we compared the observed emotion by the human evaluators with the online recognizer. For each speaker, the total agreement between the system and the two evaluators is presented in Table 5-7.

Table 5-7: The Total Agreement between SER and Human Evaluators

	SER vs Evaluator1	**SER vs Evaluator2**	**Average**
F_Speaker1	40%	70%	55%
F_Speaker2	60%	50%	55%
M_Speaker3	80%	60%	70%
M_Speaker4	50%	70%	60%
Total Agreement in %	*57.5%*	*62.5%*	*60%*

The result shows that the system achieves 60% agreement with the human experts in online recognition. This result is even better than most of the offline speech emotion models reviewed in this thesis work.

In addition we measured the response time of the system. The processing time to capture the speech input plus the recognition latency (time taken to analyze and produce an output). The response time is approximated as:

$$Response\ time = window\ size + 3\ seconds$$

where $window\ size$ is the duration of the speech (in seconds) we need to recognize at a time and 3 $seconds$ represents the recognition latency.

5.3 Discussion

Automatic identification of speech emotion has been evaluated using spectral, prosodic and voice quality features, all modeled by MLPNN on intermediate chunk level. A speech emotion corpus collected from spontaneous call center dialogs was used to train the model. The experimental results show that spectral features carry useful information for emotion detection from spontaneous speeches.

In this thesis work, a model for identification of emotional state from spontaneous communication is presented. We discussed the numerous potential applications of such a model for social and technological sectors. We then proceeded to examine the features that can characterize the emotional information conveyed in natural speech. A variety of prosodic spectral and voice quality features that extracted from human speech are used to represent each speech chunk. These features include statistics relating to the pitch, intensity, formants, MFCC, LPC, LFCC, HNR as well as jitter and shimmer of speech signal. Feature selection techniques are applied in order to reduce the dimension of the feature vector as well as to obtain the optimal speech features that potentially carry relevant emotional cues. We also presented a neural network classification algorithm called MLPNN to develop a classification model for the purpose of automatic identification of speech emotion from spontaneous communication. Only the selected features extracted from each speech chunks are used as inputs to the classifier. Due to the limited size of our spontaneous speech corpus, 10-fold cross validation technique is used to evaluate the model. We built 6 more models using the possible combinations of prosodic, spectral and voice quality features. A performance comparison is made among the models to investigate the relevance of acoustic feature types in spontaneous emotion identification.

Finally, we presented the results obtained through the use of MLPNN on training and testing the spontaneous speech emotion corpus. The emotion recognition accuracy of these experiments allow us to explain which feature type in general and which specific features in particular carry the most relevant emotional cues in spontaneous communication.

Though the result obtained from the entire 123 spectral speech features is slightly better than the one with the reduced set of combined features, using high-dimensional speech feature vector can significantly affect the response time and thus not ideal for real-time application of the model.

In general, using the spectral feature vector as an input to our classification algorithms achieves the highest recognition accuracy for natural emotions in spontaneous communication. On the other hand, prosodic and quality features exhibit lower recognition rate. Further investigation on spectral features demonstrated that MFCC and LFCC carry relatively high emotional cues in comparison to formant and LPC features.

We observed that formant and LPC features do not carry much emotional information. Since formants are used to model the resonance frequencies (and shape) of the vocal tract, we can assume that unlike acted emotions, which are exaggerated, natural emotions do not significantly affect the vocal tract shape as format features reflects the vocal resonance in speech production.

Regarding the class-wise classification, the results show that the recognition rate for Anger is lower than the rest of emotional classes. We also observed that 25% of Anger is misclassified as Positive emotion. The high degree of confusion between Anger and Positive emotion may indicates the difficulty of distinguishing the two emotional states in natural communication.

In general, the good performance of MLPNN for natural spontaneous emotion identification indicates that neural network is ideal for classification problems with high-dimensional and noisy data. Though MLPNN trains slow and requires lots of training data, the effect on the performance of our recognition process is minimal as the training phase is done offline. Although it is impossible to compare recognition accuracy of our model with other studies because of the different dataset used, our proposed model achieved a promising recognition performance.

Chapter 6: Conclusions, Recommendations and Future Work

A Multilayer Perceptron Neural Network (MLPNN) based speech emotion model that makes use of the three major acoustic feature categories of speech for representation of natural emotions is proposed in this thesis work. The performance of each category are compared. The results of the experiments show that average accuracy of 72.8% is achieved in classifying the 4 emotion classes. In this Chapter, we summarize the activities done in the course of this thesis by drawing conclusions and indicating future works. The remainder of this Chapter is organized as follows. In Section 6.1, conclusions will be drawn and contribution of the work will be stated. Finally, some recommendations will be forwarded and potential future work will be suggested in Section 6.2.

6.1 Conclusions

In this thesis, MLPNN based emotion recognition model is developed and demonstrated in near real-time identification of emotion from spontaneous speech.

For this, we first tried to explore the applications and basic challenges that are associated with emotion recognition in Chapter 1. In Chapter 2, intensive assessments are done on relevant literatures regarding speech, speech emotion and their recognition technologies and components. In Chapter 3, we tried to look existing attempts on emotion identification and their constraints in order to address the limitations in the emotion recognition research domain. We grouped the related works into three major categories in terms of the emotion corpus used in each research work in order to explore its impact on the observed result. We also discussed that authentic speech corpus is an important component in speech processing in general and speech emotion detection in particular. Though it is difficult to capture, natural emotion corpus is the ideal input in order to train a model that is capable of identifying the emotional state of a speaker. Comparisons on the classification approaches also assessed in order to get a good insight on strengths and weaknesses of various classifiers for emotion recognition.

Next, telephone dialogs collected from a call center are used to build an authentic speech emotion corpus containing 1200 intermediate chunks. A mapping onto cover class approach along with a majority voting technique is used to decide the label of each segment after the annotation is done separately by 3 professional psychologists. Then, a total of 170 acoustic features comprising prosodic, spectral and voice quality are then extracted at a chunk-level and comparisons among them is carried out in order to identify which feature

type conveys more emotional information in spontaneous communication. In order to compare their performance, a MLPNN classifier is trained with 7 different input datasets consisting of all possible combinations of the three basic acoustic feature types. Feature selection is also done to identify the most relevant features in relation to spontaneous speeches as well as to reduce the dimension of the feature vector so as to achieve real-time emotion recognition. Moreover, further experiments are done on the feature type that exhibited the highest performance in order to find out which specific feature contribute the most for spontaneous speech emotion identification.

The result indicates that the model with a combined feature vector consisting of the selected 33 features as indicators of emotional content in spontaneous communication. Though a comparative result is observed using spectral features, the high dimensionality of the feature may affect the performance the recognition and thus not suitable for real-time application of spontaneous emotion identification. The comparison we made among the 4 spectral features shows that MFCC and LFCC carry more emotional information than Formants and LPC.

From this we can conclude that spectral features play a significant role in naturel emotion identification. Moreover, MFCC and LFCC features are the most relevant spectral features. We also conclude that training MLPNN classifier with combined feature vector serves as a feasible approach for near real-time identification of spontaneous speech emotion from spontaneous communication.

Finally, our objectives listed in Chapter 1 involve developing a speech emotion model that takes acoustic features as inputs so as to achieve speech emotion identification from spontaneous communication. Tasks such as extraction, comparison and selection of an optimal feature vector are done in order to build the model. The proposed approach is then implemented using a prototype application in order to demonstrate the applicability of the model. A promising result is observed meeting the goals of the thesis. In general, the thesis work contributes a speech emotion recognition model that is capable of identifying speech emotions from spontaneous communication using the acoustic characteristics of speech. The specific contributions the thesis include:

- High-level emotion information that need to be captured to represent natural speech communications for the purpose of mapping onto low-level information, classification and modeling.

- A comparison among different acoustic feature types in order to identify their relation with the emotional states of Amharic speakers.
- A combined speech feature set relevant to speech emotion in natural spontaneous communication.
- An optimal approach that is suitable for identifying emotion in near real-time and a prototype to demonstrate the applicability of the proposed model.

6.2 Recommendation and Future Works

One of the difficulties to conduct this experiment is the lack of a concrete approach to describe emotion. This limitation significantly affects the annotation task as natural emotions are difficult to understand and thus affect the overall recognition task.

On the other hand, environmental conditions, such poor audio recorder, overlapping speeches and background noise influence the input features available to the system.

An accurate classification of the emotional state of a speaker can enhance other speech processing systems such as speech translation, speech recognition, and speech synthesis by providing a more naturalistic way of communication.

Finally, we would like to recommend the following for future work:

- Comparative study of other sophisticated classifiers (such as ANN, SVM, HMM) with more training and testing data.
- Enhancing the recognition performance by applying a robust and fast noise reduction algorithm in order to reduce background noises common to most natural speech communications as they interfere with the recognition accuracy.
- The performance of the spontaneous speech emotion identification can further be enhanced by adding more natural speech sample to the speech emotion corpus.
- To help improve spontaneous speech emotion recognition, multi-modal emotional cues such as facial expression, gesture and linguistic features could be applied in order to take advantage of the emotional contents of words and physical reactions.

References

[1] P. Sehgal and R. Kumar Jain, “Speech Processing,” *International Journal of Engineering Sciences and Emerging Technologies* 2, Vol. 5, pp. 83-87, 2013.

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech Emotion Recognition using Hidden Markov Models,” in *Proceedings of Speech Communication*, Vol. 41, No. 4, pp. 603-623, 2003.

[3] M. Sigmund, “Information Mining from Speech Signal,” *Recent Advances in Signal Processing*, pp. 297-319, 2009.

[4] J. Tao, S. Pan, M. Yang, Y. Li, K. Mu, and J. Che, “Utterance Independent Bimodal Emotion Recognition in Spontaneous Communication,” *EURASIP Journal on Advances in Signal Processing*, No. 1, pp. 1-11, 2011.

[5] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-based Speech Emotion Recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Los Alamitos, CA, USA, 2003.

[6] Z. Callejas, D. Griol, and R. López-Cózar, “Predicting User Mental States in Spoken Dialogue Systems,” *EURASIP Journal on Advances in Signal Processing*, pp. 1-21, 2011.

[7] B. Schuller, M. Lang, and G. Rigoll, “Automatic Emotion Recognition by the Speech Signal,” *Institute for Human-Machine-Communication, Technical University of Munich*, 2002.

[8] J. Epps, R. Cowie, S. Narayanan, B. Schuller, and J. Tao, “Emotion and Mental State Recognition from Speech,” *EURASIP Journal on Advances in Signal Processing* 2012, No. 1, pp. 1-2, 2012.

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion Recognition in Human–Computer Interaction,” in *Proceedings of the IEEE Signal Process.* Mag. 18, pp. 32–80, 2001.

[10] S. Scherer, H. Hofmann, M. Lampmann, M. Pfeil, S. Rhinow, F. Schwenker, and G. Palm, “Emotion Recognition from Speech: Stress Experiment,” in *Proceedings of the LREC*, 2008.

[11] O. Pierre-Yves, “The Production and Recognition of Emotions in Speech: Features and Algorithms,” *International Journal of Human-Computer Studies* 59, No. 1, pp. 157-183, 2003.

[12] R. Cowie and R. R. Cornelius, “Describing the Emotional States that are Expressed in Speech,” in *Proceedings of Speech Communication,* Vol. 40 (1–2), pp. 5–32, 2003.

[13] E. Neiberg and K. Laskowski, “Emotion Recognition in Spontaneous Speech using GMMs,” in *Proceedings of INTERSPEECH*, pp. 809-812, 2006.

[14] 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 *Proceedings of the ICASSP 2004*, Vol. 1, pp. 577–580, 2004.

[15] C. Peter and R. Beale, “Affect and Emotion in Human–Computer Interaction: From Theory to Applications”, Vol. 4868, Springer-Verlag, New York, 2008.

[16] S. Ramakrishnan, *Recognition of Emotion from Speech: A Review*, InTech, 2012.

[17] S. Steidl, “Automatic Classification of Emotion Related User States in Spontaneous Children's Speech,” Unpublished PhD Dissertation, Erlangen, Germany: University of Erlangen-Nuremberg, 2009.

[18] Husain Seid and B. Gambäck, “A Speaker Independent Continuous Speech Recognizer for Amharic,” in *Proceedings of INTERSPEECH*, pp. 3349-3352, 2005.

[19] Solomon Teferra Abate and W. Menzel, “Automatic Speech Recognition for an Under-Resourced Language - Amharic,” in *Proceedings of INTERSPEECH*, pp. 1541-1544, 2007.

[20] Solomon Teferra Abate, Martha Yifiru Tachbelie, and W. Menzel, “Amharic Speech Recognition: Past, Present and Future,” in *Proceedings of the 16th International Conference of Ethiopian Studies*, Vol. 4, pp. 1391-1401, 2009.

[21] Sebsibe Hailemariam and K. Prahallad, “Extraction of Linguistic Information with the Aid of Acoustic Data to Build Speech Systems,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, Honolulu, USA, Vol. 4, 2007.

[22] P. R. Kleinginna and A. M. Kleinginna, “A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition,” *Motivation Emotion*, Vol. 5, No. 4, pp. 345–379, 1981.

[23] J. Cahn, “The Generation of Affect in Synthesized Speech,” *J. Am. Voice Input/ Output Soc. 8*, pp. 1–19, 1990.

[24] C. Williams and K. Stevens, “Vocal Correlates of Emotional States,” *Speech Evaluation in Psychiatry*, Grune and Stratton, pp. 189–220, 1981.

[25] L. Caponetti, C. A. Buscicchio, and G. Castellano, “Biologically Inspired Emotion Recognition from Speech,” *EURASIP Journal on Advances in Signal Processing*, pp. 1-10, 2011.

[26] J. Liscombe, “Prosody and Speaker State: Paralinguistic, Pragmatics, and Proficiency,” Unpublished PhD Thesis, Columbia University, 2007.

[27] R. Cowie and R. R. Cornelius, "Describing the Emotional States that are Expressed in Speech," in *Proceedings of Speech Communication*, pp. 1-28, 2002.

[28] J. O'Connor and G. Arnold, *Intonation of Colloquial English*, Second ed., Longman, London, UK, 1973.

[29] M. Schubiger, "English Intonation: Its Form and Function," Niemeyer, Tubingen, Germany, 1958.

[30] W. Parrott, *Emotions in Social Psychology*, Psychology Press, Philadelphia, 2001.

[31] M. E. Ayadi, M. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, Vol. 44, Issue 3, pp. 572-587, 2011.

[32] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "How to Find Trouble in Communication," Speech Communication, Vol. 40, pp. 117-143, 2003.

[33] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine," in *Proceedings of INTERSPEECH*, pp. 14-18, 2014.

[34] R. B. Lanjewar and D. S. Chaudhari, "Speech Emotion Recognition: A Review," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Vol. 2, Issue 4, pp. 68-71, 2013.

[35] T. Vogt and E. Andr´e, "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," in *Proceedings of International Conference on Multimedia & Expo*, Amsterdam, The Netherlands, 2005.

[36] S. Zhang, X. Zhao, and B. Lei, "Speech Emotion Recognition using an Enhanced Kernel Isomap for Human-Robot Interaction," *International Journal of Advanced Robotic Systems*, Vol. 10, pp. 114-120, 2013.

[37] K. Dai, H. J. Fell, and J. MacAuslan, "Recognizing Emotion in Speech Using Neural Networks," *Telehealth and Assistive Technologies*, Vol. 31, pp. 38-43, 2008.

[38] L. Rabiner and R. Schafe, *Digital Processing of Speech Signals*, First ed., Pearson Education, 1978.

[39] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features and Methods," *Speech Commun.* 48 (9) pp. 1162–1181, 2006.

[40] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model-based Speech Emotion Recognition," In *International Conference on Multimedia and Expo (ICME)*, Vol. 1, pp. 401–404, 2003.

[41] T. Vogt, E. Andr´e, and J. Wanger, "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization," in *Affect and Emotion in Human-Computer Interaction*, pp. 75-91. Springer Berlin, Heidelberg, 2008.

[42] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in Real-life Emotion Annotation and Machine Learning based Detection," *Neural Networks 18(4)*, pp. 407–422, 2005.

[43] C. Gobl and A. N. Chasaide, "The Role of Voice Quality in Communicating Emotion, Mood and Attitude," *Speech Commun.* 40 (1–2) pp. 189–212, 2003.

[44] D. D. Joshi and M. B. Zalte, "Speech Emotion Recognition: A Review," *International Journal of Electronics and Communication Engineering*, Vol. 4, pp. 34-37, 2013.

[45] M. Rahman, F. Khan, and A. Bhuiyan, "Continuous Bangla Speech Segmentation, Classification and Feature Extraction," *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 2, pp. 67-75, 2012.

[46] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis,* Mach. Intell. 22 (1), pp. 4–37, 2000.

[47] L. Fu, X. Mao, and L. Chen, "Speaker Independent Emotion Recognition based on SVM/HMMs Fusion System," in *International Conference on Audio, Language and Image Processing*, ICALIP 2008, pp. 61–65, 2008.

[48] J. Wagner, T. Vogt, and E. Andr´e, "A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech," In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, Lisbon, Portugal, pp. 114–125, 2007.

[49] M. S. Unluturk, K. Oguz, and C. Atay, "Emotion Recognition using Neural Networks," in *Proceedings of 10th WSEAS International Conference on Neural Networks*, Prague, Czech Republic, pp. 82-85, 2009.

[50] J. Novakovic, M. Minica, and A. Veljovic, "Classification Accuracy of Neural Networks with PCA in Emotion Recognition," *Theory and Applications of Mathematics & Computer Science,* Vol. 1, No. 1, pp. 11-16, 2011.

[51] K. B. Khanchandani and M. A. Hussain, "Emotion Recognition using Multilayer Perceptron and Generalized Feedforward Neural Network," *Journal of Scientific and Industrial Research,* Vol. 68, No. 5, pp. 367-371, 2009.

[52] M. M. A. Mia, S. K. Biswas, M. C. Urmi, and A. Siddique, "An Algorithm for Training Multilayer Perceptron (MLP) for Image Reconstruction Using Neural Network without Overfitting," *International Journal of Scientific & Technology Research*, Vol. 4, pp. 271-275, 2015.

[53] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Emotion Recognition in Spontaneous Speech using Hidden Markov Models," in *Proceedings of INTERSPEECH*, pp. 809-812, 2001.

[54] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy, "Amplitude Modulation Features for Emotion Recognition from Speech," In *Proceedings of INTERSPEECH*, pp. 2420-2424, 2013.

[55] D. Ververidis and C. Kotropoulos, "Emotional Speech Classification using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm," in *IEEE International Conference on Multimedia and Expo, 2005*. ICME 2005, pp. 1500–1503, July 2005.

[56] M. Grimm and K. Kroschel, *Emotion Estimation in Speech Using a 3D Emotion Space Concept, Robust Speech Recognition and Understanding*, Michael Grimm and Kristian Kroschel (ED), 2007.

[57] L. Vidrascu and L. Devillers, "Detection of Real-life Emotions in Call Centers," in *Proceedings of INTERSPEECH*, pp. 1841-1844, 2005.

[58] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech," in *Proceedings of the ICASSP 2007*, Vol. 4, pp. 941–944, 2007.

Annexes

Annex A - Sample Labeled Speech

Chunk	Annotator	Annotation	Map onto Cover Class	MV
01fu0005.wav	A1	Unhappy	Sadness	Sadness
	A2	Depression	Sadness	
	A3	Sadness	Sadness	
01mc0022.wav	A1	Humiliation	Sadness	Anger
	A2	Hatred	Anger	
	A3	Dislike	Anger	
02fu0006.wav	A1	Insecurity	Sadness	Fear
	A2	Nervous	Fear	
	A3	Fear	Fear	
02mu0203.wav	A1	Scared	Fear	Fear
	A2	Fear	Fear	
	A3	Guilt	Sadness	
03fc0164.wav	A1	Happy	Joy	Anger
	A2	Agitation	Anger	
	A3	Frustration	Anger	
04mu0206.wav	A1	Despair	Sadness	Sadness
	A2	Fear	Fear	
	A3	Shame	Sadness	
05fc0115.wav	A1	Relief	Joy	Positive
	A2	Pleasure	Joy	
	A3	Passion	Love	

Chunk	Annotator	Annotation	Map onto Cover Class	MV
08fu0002.wav	A1	Care	Joy	Positive
	A2	Compassion	Love	
	A3	Satisfaction	Sadness	
10fc0015.wav	A1	Anger	Anger	Anger
	A2	Depression	Sadness	
	A3	Dislike	Anger	
	A1	Panic	Fear	Fear
	A2	Hope	Anger	
	A3	Anxious	Fear	
10mu0162S.wav	A1	Hurt	Sadness	Sadness
	A2	Fear	Fear	
	A3	Regret	Sadness	
12mu0061.wav	A1	Anxiety	Fear	Fear
	A2	Worry	Fear	
	A3	Hate	Anger	
14mc0001.wav	A1	Dislike	Anger	Anger
	A2	Anger	Anger	
	A3	Depression	Sadness	
15fc0013.wav	A1	Disgust	Anger	Anger
	A2	Anger	Anger	
	A3	Annoying	Anger	
16mu0001.wav	A1	Hope	Joy	Positive
	A2	Proud	Joy	
	A3	Furious	Anger	

Annex B - Sample Praat Script for Feature Extraction

```
form var
sentence directory
sentence filename
sentence textfile
endform

base_file_name$ = ""

if fileReadable (textfile$) = 1
     filedelete 'textfile$'
endif

sep$ = newline$

# Write-out the header

fileappend 'textfile$' @relation Emostate'sep$'@attribute
Pitch_mean numeric'sep$'@attribute Pitch_medianQ
numeric'sep$'@attribute Pitch_lowerQ numeric'sep$'@attribute
Period_mean numeric'sep$'@attribute HNR_min
numeric'sep$'@attribute F1_mean numeric'sep$'@attribute
F1_medianQ numeric'sep$'@attribute F1_lowerQ
numeric'sep$'@attribute F3_mean numeric'sep$'@attribute
F3_medianQ numeric'sep$'@attribute F3_lowerQ
numeric'sep$'@attribute F4_mean numeric'sep$'@attribute
F4_medianQ numeric'sep$'@attribute F4_lowerQ
numeric'sep$'@attribute Spectral_COG numeric'sep$'@attribute
Spectrum_sd numeric'sep$'@attribute Ltas_freqMax
numeric'sep$'@attribute mfcc1_mean numeric'sep$'@attribute
mfcc6_mean numeric'sep$'@attribute mfcc8_mean
numeric'sep$'@attribute mfcc11_mean numeric'sep$'@attribute
MFCC_max numeric'sep$'@attribute MFCC_sd
numeric'sep$'@attribute lpc1_mean numeric'sep$'@attribute
lpc4_mean numeric'sep$'@attribute lpc5_mean
numeric'sep$'@attribute LFCC_meanEnergy numeric'sep$'@attribute
lfcc2_mean numeric'sep$'@attribute lfcc7_mean
numeric'sep$'@attribute lfcc12_mean numeric'sep$'@attribute
LFCC_mean numeric'sep$'@attribute LFCC_max
numeric'sep$'@attribute LFCC_sum numeric'sep$'@attribute
Emotion$ {Anger, Fear, Positive, Sadness}'sep$'@data'sep$'

time_step = 0.01
no_formants = 5
window_length = 0.025

# Select All time range
onset = 0
offset = 0

     Read from file... 'directory$'/'filename$'
     soundname$ = selected$ ("Sound")

# Intensity Normalization (85dB)
```

```
################################
      # Select wave file and extract intensity tier
      select Sound 'soundname$'

      To Intensity... 100 0

      Formula... self+(85-self)
      Down to IntensityTier
      select IntensityTier 'soundname$'
      plus Sound 'soundname$'
      Multiply

      # Cleaning
      select Intensity 'soundname$'
      Remove
      select IntensityTier 'soundname$'
      Remove
      select Sound 'soundname$'
      Remove
      select Sound 'soundname$'_int
      Rename... 'soundname$'

# Formants (burg)
#########################

      # Select wave file and extract a formant tier
      select Sound 'soundname$'
      To Formant (burg)... time_step no_formants 5500
window_length 50

# Pitch
########################
      # Select wave file and extract pitch tier
      select Sound 'soundname$'
      To Pitch... time_step 75 600

      ### Adjust Pitch Range ###
      q1 = Get quantile... 0 0 0.25 Hertz
      q3 = Get quantile... 0 0 0.75 Hertz

      pitchFloor = floor(0.65*q1)
      pitchCealing = ceiling(1.5*q3)

      select Pitch 'soundname$'
      Remove

      select Sound 'soundname$'
      To Pitch... time_step pitchFloor pitchCealing

# PointProcess for the pulses
############################
      # Select wave file plus pitch tier and extract
PointProcess for the pulses
      select Sound 'soundname$'
      plus Pitch 'soundname$'
      To PointProcess (cc)
```

```
# Harmonicity
############################
      # Select wave file and extract Harmonicity object
      select Sound 'soundname$'
      To Harmonicity (cc)... time_step 75 0.1 1.0

# Spectrum (fft) and Ltas (1-to-1)
##################################
      # Select wave file and extract a Spectrum object
      select Sound 'soundname$'
      To Spectrum... yes
      To Ltas (1-to-1)

# MFCC
########################
      # Select wave file and extract a MFCC object
      select Sound 'soundname$'
      To MFCC... 12 window_length time_step 100 100 0

# LPC (burg)
########################
      # Select wave file and extract LPC
      select Sound 'soundname$'
      To LPC (burg)... 16 window_length time_step 50.0

# LFCC
########################
      # Select wave file and extract LFCC
      select LPC 'soundname$'
      To LFCC... 16

      # Calculate the acoustic properties for each emotion
interval

      # Calculates formant (F1 - F5) values
      select Formant 'soundname$'
      f1_mean = Get mean... 1 onset offset Hertz
      f1_medianQ = Get quantile... 1 onset offset Hertz 0.50
      f1_lowerQ = Get quantile... 1 onset offset Hertz 0.25

      f3_mean = Get mean... 3 onset offset Hertz
      f3_medianQ = Get quantile... 3 onset offset Hertz 0.50
      f3_lowerQ = Get quantile... 3 onset offset Hertz 0.25

      f4_mean = Get mean... 4 onset offset Hertz
      f4_medianQ = Get quantile... 4 onset offset Hertz 0.50
      f4_lowerQ = Get quantile... 4 onset offset Hertz 0.25

      # Calculates F0 pitch values
      select Pitch 'soundname$'

      f0 = Get mean... onset offset Hertz
      f0_medianQ = Get quantile... onset offset 0.50 Hertz
      f0_lowerQ = Get quantile... onset offset 0.25 Hertz
```

```
      # Calculates Pulse values from the selected PointProcess
file
      select PointProcess 'soundname$'_'soundname$'

      period_mean = Get mean period... onset offset 0.0001 0.02
1.3

      # Calculates the Harmonics-to-Noise Ratio values in dB
      select Harmonicity 'soundname$'
      hnr_min = Get minimum... onset offset Parabolic

      # Calculates the complex Spectrum values
      select Spectrum 'soundname$'
      spec_cog = Get centre of gravity... 2.0
      spec_sd = Get standard deviation... 2.0

      # Calculates the Spectral Slope
       select Ltas 'soundname$'

      ltas_freqMax = Get frequency of maximum... 0 0 Parabolic

      # Calculates mean values for each MFCC C1-12 across the
MFCC Matrix

      select MFCC 'soundname$'
      To Matrix

      length = Get number of columns
      sum1 = 0
      sum6 = 0
      sum8 = 0
      sum11 = 0

      # Extract the values for each row
      for c_value from 1 to 'length'
            c1'c_value' = Get value in cell... 1 'c_value'
            sum1 += c1'c_value'
            c6'c_value' = Get value in cell... 6 'c_value'
            sum6 += c6'c_value'
            c8'c_value' = Get value in cell... 8 'c_value'
            sum8 += c8'c_value'
            c11'c_value' = Get value in cell... 11 'c_value'
            sum11 += c11'c_value'
      endfor

      mean1 = 'sum1' / 'length'
      mean6 = 'sum6' / 'length'
      mean8 = 'sum8' / 'length'
      mean11 = 'sum11' / 'length'

      # Calculates MFCC max and sd values from the MFCC Matrix
and remove the Matrix

      mfcc_max = Get maximum
      mfcc_sd = Get standard deviation... onset offset 0.0 12.0
```

```
      # Calculates mean values for each LPC C1-16 across the LPC
Matrix

      select LPC 'soundname$'
      Down to Matrix (lpc)
      select Matrix 'soundname$'_lpc

      sum1 = 0
      sum4 = 0
      sum5 = 0

      # Extract the values for each row
      for lpc_value from 1 to 'length'
            lpc1'lpc_value' = Get value in cell... 1 'lpc_value'
            sum1 += lpc1'lpc_value'

            lpc4'lpc_value' = Get value in cell... 4 'lpc_value'
            sum4 += lpc4'lpc_value'

            lpc5'lpc_value' = Get value in cell... 5 'lpc_value'
            sum5 += lpc5'lpc_value'
      endfor

      meanLPC1 = 'sum1' / 'length'
      meanLPC4 = 'sum4' / 'length'
      meanLPC5 = 'sum5' / 'length'

      # calculates mean values for LFCC C0 Energy across the
MFCC frames
      select LFCC 'soundname$'
      frame_len = Get number of frames

      lfcc0_sum = 0
      # Extract the LFCC0 values for each frames
      for f_num from 1 to 'frame_len'
            lfcc0'f_num' = Get c0 value in frame... 'f_num'
            lfcc0_sum += lfcc0'f_num'
      endfor

      mean_lfcc0 = 'lfcc0_sum' / 'frame_len'

      # Calculates mean values for each LFCC C2 and C8 across
the MFCC Matrix

      select LFCC 'soundname$'
      To Matrix
      sum2 = 0
      sum7 = 0
      sum12 = 0

      for lfcc_value from 1 to 'length'
            lfcc2'lfcc_value' = Get value in cell... 2
'lfcc_value'
            sum2 += lfcc2'lfcc_value'
            lfcc7'lfcc_value' = Get value in cell... 7
'lfcc_value'
```

```
           sum7 += lfcc7'lfcc_value'
           lfcc12'lfcc_value' = Get value in cell... 12
'lfcc_value'
           sum12 += lfcc12'lfcc_value'
     endfor

     meanLFCC2 = 'sum2' / 'length'
     meanLFCC7 = 'sum2' / 'length'
     meanLFCC12 = 'sum12' / 'length'

     lfcc_mean = Get mean... onset offset 0.0 16.0
     lfcc_max = Get maximum
     lfcc_sum = Get sum

     if f0 = undefined
           f0 = 0
        endif
     if f0_medianQ = undefined
           f0_medianQ = 0
     endif
     if f0_lowerQ = undefined
           f0_lowerQ = 0
     endif
     if period_mean = undefined
           period_mean = 0
     endif
     if hnr_min = undefined
           hnr_min = 0
     endif

                fileappend 'textfile$'
'f0','f0_medianQ','f0_lowerQ','period_mean','hnr_min','f1_mean'
,'f1_medianQ','f1_lowerQ','f3_mean','f3_medianQ','f3_lowerQ','f
4_mean','f4_medianQ','f4_lowerQ','spec_cog','spec_sd','ltas_fre
qMax','mean1','mean6','mean8','mean11','mfcc_max','mfcc_sd','me
anLPC1','meanLPC4','meanLPC5','mean_lfcc0','meanLFCC2','meanLFC
C7','meanLFCC12','lfcc_mean','lfcc_max','lfcc_sum',?
                fileappend 'textfile$' 'newline$'

# clean up

select all
Remove
```

Annex C - The Complete List of Extracted Acoustic Features

Prosody		
Pitch_mean	Pitch_upperQ	Intensity_max
Pitch_max	Pitch_interquartileRange	IntensityTimeOfMax
Pitch_TimeOfMax	Pitch_sd	Intensity_min
Pitch_minPre	F0_meanAbSlope	Intensity_TimeOfMin
Pitch_minPost	Pulse	Intensity_medianQ
Pitch_min	Period	Intensity_lowerQ
Pitch_TimeOfMin	Period_mean	Intensity_upperQ
Pitch_range	Period_sd	Intensity_interquartileRange
Pitch_medianQ	Intensity_dB	Intensity_sd
Pitch_lowerQ	Intensity_mean	
Voice Quality		
Jitter_local	Shimmer_localDB	HNR_min
Jitter_localAbsol	Shimmer_APQ3	HNR_timeMin
Jitter_rap	Shimmer_APQ5	HNR_max
Jitter_PPQ5	Shimmer_APQ11	HNR_timeMax
Jitter_DDP	Shimmer_DDA	HNR_range
Shimmer_local	HNR_mean	HNR_sd
Spectral		
F1_mean	F5_max	lpc3_mean
F1_max	F5_TimeOfMax	lpc4_mean
F1_TimeOfMax	F5_min	lpc5_mean
F1_min	F5_TimeOfMin	lpc6_mean
F1_TimeOfMin	F5_medianQ	lpc7_mean

F1_medianQ	F5_lowerQ	lpc8_mean
F1_lowerQ	F5_upperQ	lpc9_mean
F1_upperQ	F5_interquartileRange	lpc10_mean
F1_interquartileRange	F5_sd	lpc11_mean
F1_sd	Spectral_COG	lpc12_mean
F2_mean	Spectral_skewness	lpc13_mean
F2_max	Spectral_kurtosis	lpc14_mean
F2_TimeOfMax	Spectrum_sd	lpc15_mean
F2_min	Ltas_mean	lpc16_mean
F2_TimeOfMin	Ltas_min	LPC_mean
F2_medianQ	Ltas_freqMin	LPC_max
F2_lowerQ	Ltas_max	LPC_min
F2_upperQ	Ltas_freqMax	LPC_sum
F2_interquartileRange	Ltas_range	LPC_sd
F2_sd	Ltas_sd	LFCC_meanEnergy
F3_mean	Spectral_slope	lfcc1_mean
F3_max	MFCC_meanEnergy	lfcc2_mean
F3_TimeOfMax	mfcc1_mean	lfcc3_mean
F3_min	mfcc2_mean	lfcc4_mean
F3_TimeOfMin	mfcc3_mean	lfcc5_mean
F3_medianQ	mfcc4_mean	lfcc6_mean
F3_lowerQ	mfcc5_mean	lfcc7_mean
F3_upperQ	mfcc6_mean	lfcc8_mean
F3_interquartileRange	mfcc7_mean	lfcc9_mean
F3_sd	mfcc8_mean	lfcc10_mean
F4_mean	mfcc9_mean	lfcc11_mean

F4_max	mfcc10_mean	lfcc12_mean
F4_TimeOfMax	mfcc11_mean	lfcc13_mean
F4_min	mfcc12_mean	lfcc14_mean
F4_TimeOfMin	MFCC_mean	lfcc15_mean
F4_medianQ	MFCC_max	lfcc16_mean
F4_lowerQ	MFCC_min	LFCC_mean
F4_upperQ	MFCC_sum	LFCC_max
F4_interquartileRange	MFCC_sd	LFCC_min
F4_sd	lpc1_mean	LFCC_sum
F5_mean	lpc2_mean	LFCC_sd

Annex D - Source Code for Online Speech Emotion Identification

```
package emotion;

import java.awt.Font;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.nio.file.Path;
import java.util.Scanner;
import java.util.logging.Level;
import java.util.logging.Logger;

import javax.swing.GroupLayout;
import javax.swing.GroupLayout.Alignment;
import javax.swing.JButton;
import javax.swing.JFileChooser;
import javax.swing.JFrame;
import javax.swing.JLabel;
import javax.swing.JScrollPane;
import javax.swing.JTextArea;
import javax.swing.LayoutStyle.ComponentPlacement;
import javax.swing.ScrollPaneConstants;

public class SER extends JFrame {

      /**
       * Creates new form SER
```

```
       */
      public static String folder;
      public int j, l, k, y, z;
      public int v, v2;
      public boolean flag = true;
      public File wavFile, fFile, wavFile2;
      public Path wavPath, fPath, wavPath2;
      public String wavString, wavString1, wavString2;
      private volatile boolean canceled = false;
      private volatile boolean running = true;

      public SER() {
            initComponents();
      }

      public void initComponents() {

            jLabelFile = new javax.swing.JLabel();
            jTextField1 = new javax.swing.JTextField();
            jLabelFile.setLabelFor(jTextField1);
            jButtonBrowse = new javax.swing.JButton();
            jMenuBar1 = new javax.swing.JMenuBar();
            jMenuFile = new javax.swing.JMenu();
            jMenuItemNew = new javax.swing.JMenuItem();
            jMenuItemExit = new javax.swing.JMenuItem();
            jMenuSetting = new javax.swing.JMenu();


      setDefaultCloseOperation(javax.swing.WindowConstants.EXIT
_ON_CLOSE);

            jLabelFile.setFont(new java.awt.Font("Sylfaen", 1,
14)); // NOI18N
            jLabelFile.setText("Speech File");

            jTextField1.setFont(new Font("Sylfaen", Font.PLAIN,
13)); // NOI18N

            jButtonBrowse.setFont(new java.awt.Font("Sylfaen",
1, 14)); // NOI18N
            jButtonBrowse.setText("Browse");
            jButtonBrowse.addActionListener(new
java.awt.event.ActionListener() {
                  public void
actionPerformed(java.awt.event.ActionEvent evt) {
                        jButtonBrowseActionPerformed(evt);
                  }
            });

            jMenuBar1.setFont(new java.awt.Font("Sylfaen", 1,
14)); // NOI18N

            jMenuFile.setText("File");
            jMenuFile.setFont(new java.awt.Font("Sylfaen", 1,
14)); // NOI18N
```

```
            jMenuItemNew.setFont(new java.awt.Font("Sylfaen", 1,
14)); // NOI18N
            jMenuItemNew.setText("New");
            jMenuItemNew.addMouseListener(new
java.awt.event.MouseAdapter() {
                public void
mouseClicked(java.awt.event.MouseEvent evt) {
                    jMenuItemNewMouseClicked(evt);
                }
            });
            jMenuItemNew.addActionListener(new
java.awt.event.ActionListener() {
                public void
actionPerformed(java.awt.event.ActionEvent evt) {
                    jMenuItemNewActionPerformed(evt);
                }
            });
            jMenuFile.add(jMenuItemNew);

            jMenuItemExit.setFont(new java.awt.Font("Sylfaen",
1, 14)); // NOI18N
            jMenuItemExit.setText("Exit");
            jMenuItemExit.addActionListener(new
java.awt.event.ActionListener() {
                public void
actionPerformed(java.awt.event.ActionEvent evt) {
                    jMenuItemExitActionPerformed(evt);
                }
            });
            jMenuFile.add(jMenuItemExit);

            jMenuBar1.add(jMenuFile);

            jMenuSetting.setText("Setting");
            jMenuSetting.setFont(new java.awt.Font("Sylfaen", 1,
14)); // NOI18N
            jMenuSetting.addMouseListener(new
java.awt.event.MouseAdapter() {
                public void
mouseClicked(java.awt.event.MouseEvent evt) {
                    jMenuSettingMouseClicked(evt);
                }
            });
            jMenuSetting.addActionListener(new
java.awt.event.ActionListener() {
                public void
actionPerformed(java.awt.event.ActionEvent evt) {
                    jMenuSettingActionPerformed(evt);
                }
            });
            jMenuBar1.add(jMenuSetting);

            setJMenuBar(jMenuBar1);
            jButtonOffline = new javax.swing.JButton();
```

```
            jButtonOffline.setToolTipText("");

            jButtonOffline.setFont(new java.awt.Font("Sylfaen",
1, 14)); // NOI18N
            jButtonOffline.setText("Offline");
            jButtonOffline.addActionListener(new
java.awt.event.ActionListener() {
                  public void
actionPerformed(java.awt.event.ActionEvent evt) {
                        jButtonOfflineActionPerformed(evt);
                  }
            });

            JButton buttonStop = new javax.swing.JButton();
            buttonStop.addActionListener(new
java.awt.event.ActionListener() {
                  public void
actionPerformed(java.awt.event.ActionEvent evt) {
                        buttonStopActionPerformed(evt);
                  }
            });
            buttonStop.setText("Stop");
            buttonStop.setFont(new Font("Sylfaen", Font.BOLD,
14));

            jButtonOnline = new javax.swing.JButton();

            jButtonOnline.setFont(new java.awt.Font("Sylfaen",
1, 14)); // NOI18N
            jButtonOnline.setText("Online");
            jButtonOnline.addActionListener(new
java.awt.event.ActionListener() {
                  public void
actionPerformed(java.awt.event.ActionEvent evt) {
                        jButtonOnlineActionPerformed(evt);
                  }
            });

            label = new JLabel();
            label.setText("Recognized Emotion");
            label.setFont(new Font("Sylfaen", Font.BOLD, 14));

            scrollPane = new JScrollPane();
            scrollPane.setAutoscrolls(true);
            scrollPane.setFocusCycleRoot(true);

      scrollPane.setVerticalScrollBarPolicy(ScrollPaneConstants
.VERTICAL_SCROLLBAR_ALWAYS);

            javax.swing.GroupLayout layout = new
javax.swing.GroupLayout(getContentPane());

      layout.setHorizontalGroup(layout.createParallelGroup(Alig
nment.LEADING)
```

```
      .addGroup(layout.createSequentialGroup().addContainerGap(
)

      .addGroup(layout.createParallelGroup(Alignment.LEADING)

      .addGroup(layout.createSequentialGroup().addComponent(jLa
belFile)

      .addPreferredGap(ComponentPlacement.RELATED)

      .addComponent(jTextField1, GroupLayout.DEFAULT_SIZE, 511,
Short.MAX_VALUE))

      .addGroup(layout.createSequentialGroup()

      .addComponent(label, GroupLayout.PREFERRED_SIZE, 144,
GroupLayout.PREFERRED_SIZE)

      .addPreferredGap(ComponentPlacement.RELATED)

      .addGroup(layout.createParallelGroup(Alignment.LEADING)

      .addGroup(layout.createSequentialGroup().addComponent(jBu
ttonOffline).addGap(74)

      .addComponent(jButtonOnline)

      .addPreferredGap(ComponentPlacement.RELATED, 75,
Short.MAX_VALUE)

      .addComponent(buttonStop, GroupLayout.PREFERRED_SIZE, 75,

                  GroupLayout.PREFERRED_SIZE)

      .addGap(65))

      .addComponent(scrollPane, GroupLayout.PREFERRED_SIZE,
376,

      GroupLayout.PREFERRED_SIZE))))

      .addPreferredGap(ComponentPlacement.RELATED).addComponent
(jButtonBrowse).addContainerGap()));

      layout.setVerticalGroup(layout.createParallelGroup(Alignm
ent.LEADING)

      .addGroup(layout.createSequentialGroup().addContainerGap(
)

      .addGroup(layout.createParallelGroup(Alignment.BASELINE).
addComponent(jLabelFile)
```

```
      .addComponent(jTextField1, GroupLayout.PREFERRED_SIZE,
GroupLayout.DEFAULT_SIZE,

      GroupLayout.PREFERRED_SIZE)

      .addComponent(jButtonBrowse))

      .addPreferredGap(ComponentPlacement.RELATED, 17,
Short.MAX_VALUE)
                                    .addGroup(

      layout.createParallelGroup(Alignment.BASELINE).addCompone
nt(jButtonOffline)

      .addComponent(buttonStop, GroupLayout.PREFERRED_SIZE, 27,

      GroupLayout.PREFERRED_SIZE)

      .addComponent(jButtonOnline))
                                    .addGap(28)

      .addGroup(layout.createParallelGroup(Alignment.LEADING)

      .addComponent(label, GroupLayout.PREFERRED_SIZE, 19,
GroupLayout.PREFERRED_SIZE)

      .addComponent(scrollPane, GroupLayout.PREFERRED_SIZE,
323, GroupLayout.PREFERRED_SIZE))
                        .addContainerGap()));

            textArea = new JTextArea();
            textArea.setEditable(false);
            scrollPane.setViewportView(textArea);
            textArea.setFont(new Font("Sylfaen", Font.PLAIN,
14));
            getContentPane().setLayout(layout);

            pack();
      }// </editor-fold>//GEN-END:initComponents

      public void cancel() {
            canceled = false;
      }

      public boolean isCanceled() {
            return canceled;
      }

      private void
buttonStopActionPerformed(java.awt.event.ActionEvent evt) {//
GEN-FIRST:event_jMenuItem2ActionPerformed

            canceled = true;
            this.dispose();
```

```
            SER.main(null);

      }// GEN-LAST:event_jMenuItem2ActionPerformed

      private void
jButtonBrowseActionPerformed(java.awt.event.ActionEvent evt)
{// GEN-FIRST:event_jButtonBrowseActionPerformed
            // TODO add your handling code here:
            JFileChooser jf = new JFileChooser();
            jf.setCurrentDirectory(new File(""));

            int r = jf.showOpenDialog(this);
            if (r == JFileChooser.APPROVE_OPTION) {

      jTextField1.setText(jf.getSelectedFile().toString());
                  wavFile = jf.getSelectedFile();
                  fFile = jf.getCurrentDirectory();
                  wavPath = wavFile.toPath();
                  fPath = fFile.toPath();
                  wavString = wavFile.getName();
                  textArea.setText("");
            }
      }// GEN-LAST:event_jButtonBrowseActionPerformed

      private void
jButtonOnlineActionPerformed(java.awt.event.ActionEvent evt)
{// GEN-FIRST:event_jButtonOnlineActionPerformed

            try {
                  Scanner s = new Scanner(new
File("D:\\ImplementER\\SER\\setting.txt"));
                  v = s.nextInt();
                  s.close();
                  final long st = System.currentTimeMillis();

                  // run feature extractor alone
                  new java.util.Timer().scheduleAtFixedRate(new
java.util.TimerTask() {
                        public void run() {
                              if (canceled == false) {
                                    try {
                                          RunPraat rPraat = new
RunPraat();

      rPraat.extractionTask(folder, l, v);
                                          l++;
                                    } catch (Exception e) {
                                          System.out.println("End
of Praat");
                                    }
                              }
                        }
                  }, v + 2000, v + 2000);
```

```
                  new java.util.Timer().scheduleAtFixedRate(new
java.util.TimerTask() {

                        public void run() {

                              if (canceled == false) {
                                    try {
                                          final ClassifyNewSpeech
classify = new ClassifyNewSpeech();

      classify.classificationTask(folder, k, v);
                                          textArea.append("\nFrom "
+ (k * v / 1000) + " - " + ((k * v / 1000) + (v / 1000))
                                                      + "
seconds:\t" + classify.getEmotion().toUpperCase() + "\n");
                                          k++;
                                    } catch (Exception e) {
                                          textArea.append(".");
                                          // k++;
                                    }
                              }
                        }
                  }, 0, 1000);

                  new java.util.Timer().scheduleAtFixedRate(new
java.util.TimerTask() {

                        public void run() {
                              if (canceled == false) {
                                    try {
                                          final Recorder rec = new
Recorder();

                                          // creates a new thread
that waits for a specified of
                                          // time before stopping
                                          Thread stopperThread =
new Thread(new Runnable() {
                                                public void run() {
                                                      try {

      Thread.sleep(v);
                                                      } catch
(InterruptedException ex) {

      ex.printStackTrace();
                                                      }
                                                      rec.stopRec();
                                                }
                                          });
                                          stopperThread.start();

                                          // start recording
                                          rec.startRec(folder, j);
                                          j++;
```

```
                                    } catch (Exception e) {
                                         e.printStackTrace();
                                    }
                               }
                          }
                     }, 0, v + 1000);

          } catch (FileNotFoundException ex) {

     Logger.getLogger(SER.class.getName()).log(Level.SEVERE,
null, ex);
          } catch (IOException ex) {

     Logger.getLogger(SER.class.getName()).log(Level.SEVERE,
null, ex);
          }
     }// GEN-LAST:event_jButtonOnlineActionPerformed

     private void
jButtonOfflineActionPerformed(java.awt.event.ActionEvent evt)
{// GEN-FIRST:event_jButtonOfflineActionPerformed

          try {
               Scanner s = new Scanner(new
File("D:\\ImplementER\\SER\\setting.txt"));
               while (s.hasNextInt()) {
                    v = s.nextInt();
               }
               s.close();
               final RunPraat rPraat = new RunPraat();

               new java.util.Timer().schedule(new
java.util.TimerTask() {

                    public void run() {
                         try {
                              rPraat.extractionTask(folder,
fPath, wavFile, y, v);
                              y++;
                         } catch (Exception e) {
                              textArea.setText("Browse
speech file for Offline Recognition\n");
                         }
                    }
               }, 0);

               final ClassifyNewSpeech classify = new
ClassifyNewSpeech();

               new java.util.Timer().scheduleAtFixedRate(new
java.util.TimerTask() {

                    public void run() {
                         try (BufferedReader br = new
BufferedReader(
```

```
                                        new FileReader(folder +
"\\ResultForAudio" + z + ".arff"))) {


     classify.classificationTask(folder, z, v);
                                textArea.append("From " + (z *
v / 1000) + " - " + ((z * v / 1000) + (v / 1000)) + "
seconds:\t"
                                                    +
classify.getEmotion().toUpperCase() + "\n");
                                z++;
                            } catch (Exception e) {
                                textArea.append("End of
File");
                                //cancel();
                            }
                        }
                }, v, v);

            } catch (Exception e) {

            }
     }// GEN-LAST:event_jButtonOfflineActionPerformed

     private void
jMenuItemNewMouseClicked(java.awt.event.MouseEvent evt) {//
GEN-FIRST:event_jMenuItemNewMouseClicked
     }// GEN-LAST:event_jMenuItemNewMouseClicked

     private void
jMenuItemNewActionPerformed(java.awt.event.ActionEvent evt) {//
GEN-FIRST:event_jMenuItemNewActionPerformed
     }// GEN-LAST:event_jMenuItemNewActionPerformed

     private void
jMenuItemExitActionPerformed(java.awt.event.ActionEvent evt)
{// GEN-FIRST:event_jMenuItemExitActionPerformed
           System.exit(0);
     }// GEN-LAST:event_jMenuItemExitActionPerformed

     private void
jMenuSettingActionPerformed(java.awt.event.ActionEvent evt) {//
GEN-FIRST:event_jMenuSettingActionPerformed
     }// GEN-LAST:event_jMenuSettingActionPerformed

     private void
jMenuSettingMouseClicked(java.awt.event.MouseEvent evt) {//
GEN-FIRST:event_jMenuSettingMouseClicked
           this.setVisible(false);
           new Setting().setVisible(true);
     }// GEN-LAST:event_jMenuSettingMouseClicked

     /**
      * Launch the application.
      */
```

```
      public static void main(String args[]) {
            File f = new
File("D:\\ImplementER\\SER\\SegmentedSpeech\\" +
System.currentTimeMillis());
            f.mkdir();
            folder = f.getPath();
            try {
                  for (javax.swing.UIManager.LookAndFeelInfo info
: javax.swing.UIManager.getInstalledLookAndFeels()) {
                        if ("Nimbus".equals(info.getName())) {

      javax.swing.UIManager.setLookAndFeel(info.getClassName())
;
                              break;
                        }
                  }
            } catch (ClassNotFoundException ex) {

      java.util.logging.Logger.getLogger(SER.class.getName()).l
og(java.util.logging.Level.SEVERE, null, ex);
            } catch (InstantiationException ex) {

      java.util.logging.Logger.getLogger(SER.class.getName()).l
og(java.util.logging.Level.SEVERE, null, ex);
            } catch (IllegalAccessException ex) {

      java.util.logging.Logger.getLogger(SER.class.getName()).l
og(java.util.logging.Level.SEVERE, null, ex);
            } catch (javax.swing.UnsupportedLookAndFeelException
ex) {

      java.util.logging.Logger.getLogger(SER.class.getName()).l
og(java.util.logging.Level.SEVERE, null, ex);
            }
            // </editor-fold>

            javax.swing.SwingUtilities.invokeLater(new
Runnable() {
                  public void run() {
                        new SER().setVisible(true);
                        ;
                  }
            });
      }

      private javax.swing.JButton jButtonBrowse;
      private javax.swing.JButton jButtonOnline;
      private javax.swing.JButton jButtonOffline;
      private javax.swing.JLabel jLabelFile;
      private javax.swing.JMenu jMenuFile;
      private JLabel label;
      private JScrollPane scrollPane;
      private JTextArea textArea;
}
```

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____