

**REAL-TIME ANGER DETECTION IN ARABIC
SPEECH DIALOGS**

BY

ASHRAF ABDULHAMID AL-SHEIKH KHALIL

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

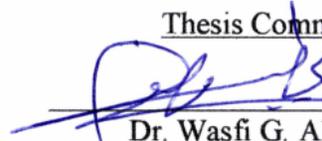
December, 2011

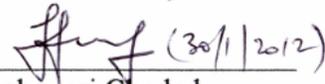
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

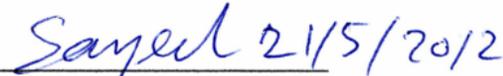
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Ashraf AbdulHamid AlSheikh Khalil** under the direction of his thesis advisors and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**

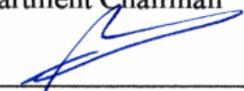
Thesis Committee

 14/5/2012
Dr. Wasfi G. Al-Khatib
(Thesis Advisor)

 (30/1/2012)
Dr. Lahouari Cheded
(Member)

 21/5/2012
Dr. El-Sayed El-Alfy
(Member)


Dr. Adel Ahmed
Department Chairman


Dr. Salam A. Zummo
Dean of Graduate Studies

22/5/12
Date



Affectionately dedicated

To my mother:

For her love and prayers

*To the memory of my father who passed on a love of
reading and respect for education*

To my beloved wife:

For her encouragement and endless support

To my children Jana, Ibrahim and Fares:

For their love

ACKNOWLEDGMENT

In the name of Allah, the Beneficent, Most Merciful

I am grateful to Almighty ALLAH by whom all good deeds are completed. I would like to express my deep appreciation to King Fahd University of Petroleum & Minerals for its support and assistance. I wish to extend my gratitude to my thesis advisor Dr. Wasfi Al-Khatib for his continuous support, patience, and much needed encouragement. I am also thankful to my thesis committee members Dr. Cheded Lahouari and Dr. El-Sayed M. El-Alfy for their time and useful comments. May Allah reward you all.

I wish to thank SAHARA NET, Internet service provider in Saudi Arabia, for their great support by providing me with the required data for the experiment, which added a real value to the thesis.

My appreciation goes to all colleagues and friends for their suggestions and encouragement during the writing of this thesis. Special Thanks to my colleagues Mohamed Yahya, Mohammad Amro, Mohamed Al-Qadi and Emad Jaha.

A very sincere appreciation is due to my mother, my wife and my children for their prayers, encouragement and continuous support throughout the writing of this thesis.

In particular, I am really grateful to my wife for her patience and for the care she extended to our children throughout the writing of this thesis. I would also like to use this opportunity to pray to Allah that He may rest the soul of my father in perfect peace under His boundless mercy and that He may also grant him a permanent place in Jannat-ul-Firdaus.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vii
THESIS ABSTRACT.....	viii
CHAPTER 1 INTRODUCTION.....	1
1.1. Introduction	1
1.2. Thesis Objectives.....	3
1.3. Thesis Organization.....	4
CHAPTER 2 LITERATURE SURVEY.....	5
2.1. Inter-Cultural vs. Cross-Cultural Emotions.....	5
2.2. Emotion Set Categorization.....	6
2.3. Emotions Data Sources.....	8
2.4. Online versus Offline Recognition.....	10
2.5. Emotion Recognition Applications	12
CHAPTER 3 REAL-TIME ANGER DETECTION FRAMEWORK FOR ARABIC SPEECH DIALOGS	14
3.1. Feature Selection and Extraction.....	14
3.1.1. Pitch.....	15
3.1.2. Energy.....	16
3.1.3. Formants	17
3.1.4. Mel-Frequency Cepstral Coefficients.....	17
3.2. Classification of Emotions	18
3.2.1. Support Vector Machine (SVM)	19
3.2.2. Probabilistic Neural Networks	20
3.2.3. Decision Tree Forests	22
CHAPTER 4 OFFLINE SYSTEM DEVELOPMENT	25

4.1. Emotion Set Selection	25
4.2. Corpus Development	29
4.2.1. ISP Customer Service Records.....	29
4.2.2. TV Debate Show	29
4.2.3. An Angry Customer's Call.....	30
4.3. Emotion Survey	35
4.3.1. Survey Results	36
4.4. Experimental Results of Classification	37
4.5. Handling Missing Values	41
4.6. Merging of Classes	45
4.7. Best Segment Length.....	47
CHAPTER 5 REAL-TIME SYSTEM DEVELOPMENT.....	49
5.1. Voiced Parts Segmentation Methodology	49
5.2. Feature Selection	51
CHAPTER 6 ANGER DETECTION PROTOTYPE SYSTEM	
DEVELOPMENT	56
6.1. Voice Recording Module	57
6.2. Online Segmentation Module.....	59
6.3. Online Feature Extraction.....	60
6.4. Real-Time Classification	60
6.5. Visual Presentation Module	61
CHAPTER 7 CONCLUSIONS AND FUTURE WORK.....	62
7.1. Main Contributions.....	65
7.2. Future Work.....	66
REFERENCES	67
APPENDIX A LIST OF SELECTED FEATURES	72
VITAE	81

LIST OF TABLES

TABLE	Page
Table 1 Distribution of perceived speaker affect in the natural corpus (NATURAL) ...	26
Table 2 Statistical properties of emotions prosodic features compared to neutral	28
Table 3 The content of the Speaker and Emotion tiers in dialog annotation	33
Table 4 Score function results	36
Table 5 Unclear vs. Duration.....	37
Table 6 Data included in both context-aware group and context-free group	38
Table 7 Number of clips included in training set and testing set	39
Table 8 Accuracy percentage for each classifier on both training subsets	40
Table 9 Attributes with missing values and low classification importance.	42
Table 10 Replaced new values for features attributes that may have null values	43
Table 11 Accuracy percentage of four classifiers after handling missing values	44
Table 12 Comparison of average accuracy for handling missing values	44
Table 13 Confusion matrix for the decision tree forest.	46
Table 14 Misclassification rate of three classifiers on the two groups.....	48
Table 15 Extraction times for various features of a 24.87 seconds long segment	52
Table 16 Classification of each feature subsets on context-aware data group	53

LIST OF FIGURES

FIGURE	Page
Figure 1: A two-dimensional emotion space with a valence and an arousal axes	7
Figure 2: Types of databases used for emotion recognition and their difficulty.....	9
Figure 3: Development Framework for Anger Detection in Arabic Speech Dialogs	15
Figure 4: Two-dimensional plotted data.....	19
Figure 5: All PNN networks have four layers.	21
Figure 6: Example of Single Decision Tree.	23
Figure 7: Example pitch contours for anger and neutral utterances	27
Figure 8: Label are given for each interval into two tiers of annotation	31
Figure 9: General format of annotation of each interval in the Speaker tier	32
Figure 10: Different speakers turns in one emotion period in TextGrid	32
Figure 11: Names of split .wav files from long sound file based on annotation.....	34
Figure 12: Survey on sound files.....	35
Figure 13: Accuracy vs. duration	48
Figure 14: Segmentation of long voiced part based on small break.....	50
Figure 15: Segmentation for long voiced part.....	51
Figure 16: Performance of the three classifiers on various feature sets	55
Figure 17: Anger detection prototype system development.....	56
Figure 18: Enabling Sound Mix	58
Figure 19: Recording Module GUI	59
Figure 20: Visual presentation of predicted emotion classes	61

THESIS ABSTRACT

Name: Ashraf AbdulHamid AlSheikh Khalil

Title: Real-Time Anger Detection in Arabic Speech Dialogs

Major Field: Computer Science

Date of Degree: December 2011

Anger is potentially the most important human emotion to be detected in human-human dialogs, such as those found in call-centers and other similar fields. It measures the satisfaction, or lack of it, of a speaker directly from his or her voice. Recently, many software applications were built as a result of anger detection research work. In this thesis, we design a real-time framework to detect anger from spontaneous Arabic conversations. We construct a well-annotated corpus for anger and neutral emotion states from real-world Arabic speech dialogs for our experiments. We use a hybrid segmentation method of sentence length and fixed-length splitting techniques. The classification is based on acoustic sound features that are more appropriate for anger detection. Many acoustic features will be studied such as the fundamental frequency (f_0), formants, energy and mel-frequency cepstral coefficients (MFCCs). Several classifiers are evaluated, and the experimental results show that support vector machine classifiers can yield 82.4% detection rate which is the best accuracy result as compared to other classifiers; in addition to being fast for online recognition. Moreover, a real-time software application is developed based on our proposed framework.

ملخص الرسالة

الاسم: أشرف عبدالحميد الشيخ خليل

عنوان الرسالة: الكشف الآني للغضب في المحادثات الصوتية العربية

التخصص: علوم الحاسب

تاريخ التخرج: ديسمبر 2011

تعد مشاعر الغضب من المشاعر المهمة والتي يجب الكشف عنها في المحادثات بين الناس في العديد من المجالات؛ على سبيل المثال في شركات الاتصالات والقطاعات المماثلة لقياس مدى رضا العميل عن الخدمات المقدمة من خلال نبرة الصوت. وفي الآونة الأخيرة ظهرت عدة تطبيقات للحاسب الآلي استناداً على نتائج بحثية في مجال كشف الغضب. في هذه الرسالة، قمنا بتصميم إطار آني لاكتشاف مشاعر الغضب في المحادثات العفوية باللغة العربية. كما بنينا مكنز بتصنيف جيد للمشاعر المختلفة في محادثات عربية حقيقية لاستخدامها في تجاربنا العملية. واستخدمنا عملية تقطيع تعتمد على خليط من تقنيتين مختلفتين وهما حساب نهايات الجمل والحساب المحدد بزمن ثابت. ويعتمد التصنيف على استخدام سمات سمعية ملائمة للكشف عن مشاعر الغضب. درسنا عدة سمات سمعية مثل التردد الأساسي ومدة المحادثة ومعاملات MFCC. وقد أظهرت نتائج التجارب التي أجريناها قدرة "آلة متجه القرار" مقارنة بطرق أخرى على تمييز مشاعر الغضب بنسبة تعرف تصل إلى 82.4%، كما أن عملية التصنيف لا تأخذ وقتاً طويلاً مما يلائم التطبيق الآني لها. وبالإضافة إلى ما سبق، فقد طورنا نظام برمجي آني بالاعتماد على تلك المنظومة المقترحة.

CHAPTER 1

INTRODUCTION

1.1. INTRODUCTION

Emotional speech detection, based on offline previously recorded and annotated databases, has received a lot of attention from Academia [1], [2]. However, there is growing need for real-time emotion detection systems that are based on online analysis of the speech signal. For example, in human-human dialogs, emotion recognition systems can provide users with improved services by being adaptive to their emotion. A clear example of an application area is customer service centers, which have become an important part of today's business sector. Using a computerized system application to report the customer's basic emotions in real-time to the supervisors will make it faster and easier to monitor the quality of customer service personnel and will, therefore, aid in handling tough and uneasy situations. Real-time emotion detection is considered to be more challenging and less predictable than offline detection for the following reasons: First, it imposes a constraint on the available time period to carry out the detection process. This may negatively affect its accuracy. Second, real-time emotion detection applications run on real-life emotions which are harder than acted emotions, with respect to collection, segmentation, labeling and noise processing of the data [3] [4].

In this thesis, we study the automatic real-time recognition of emotions in the Arabic culture. There are several reasons for restricting our domain to the Arabic culture. According to Ethnologue¹: “Arabic is ranked as the fourth most widely spoken language in the world.” It is being spoken by nearly 300 million people around the world and is considered of interest to more than one Billion Muslims world-wide. Despite the renewed interest in Arabic computing research since 2001, there has been little research work reported on emotion recognition, according to the authors' best knowledge. In addition, such research work is of a great importance to the business sector, where applications related to enhanced customer service call center support are needed to live up to the fierce competition in the Middle Eastern markets. Note that our emphasis in this study is on the Arabic culture, rather than its language. Our proposed framework is not based on Arabic speech recognition, but on the prosodic features of emotion in the Arabic culture, such as pitch, energy, formants and Mel-frequency cepstral coefficients (MFCCs). In addition, we restrict the set of emotions to be detected, to active anger² and neutral emotions because of the growing need of their automatic detection in dialogs, such as those in customer care call centers.

¹ <http://www.ethnologue.com/>

² Psychologists define anger as a response to the perception of threat due to a physical conflict, injustice, negligence, humiliation, or betrayal. The state of anger includes emotional states such as tense, alarmed, angry, afraid, annoyed, mad, and so on [36].

1.2. THESIS OBJECTIVES

The main objective of this thesis is to develop a framework for building a *real-time anger detection system* based on acoustic sound features extracted from spontaneous Arabic dialogs. Such dialogs may occur in meetings, customer-service and emergency-hotline calls.

According to our best knowledge, no work has been done in the Arabic language on real-time speech classification, even the one based on prosodic features. So our resulting framework will be of a pioneering nature and realistic, and may be a good start for other researchers who have interest in Arabic language processing. In order to achieve this goal, the following steps are performed:

- 1) Building a spontaneous and well-annotated Arabic emotion corpus. This real-world database will be used to train our automatic detection system.
- 2) Creating a machine learning model system to identify anger and neutral emotions using a set of prosodic features.
- 3) Developing, testing and implementing an application to detect anger and neutral emotions from speech signals in real-time using various classification techniques.
- 4) Demonstration of a working model that embodies objectives 2 and 3

1.3. THESIS ORGANIZATION

In this thesis, the problem of anger detection is studied in order to develop a real-time system that will automatically detect anger from the Arabic speech. The thesis is organized as follows:

- Chapter 2 presents a brief literature survey on the problem at hand. The literature will be reviewed and categorized based on cultural and cross-cultural studies, anger specific versus set of emotion states recognition, and types of data sources. Moreover, challenges on real-time processing when compared to offline process are addressed.
- Chapter 3 describes the main stages of a generic framework.
- Chapter 4 shows how we build our corpus and how we elaborate on the process of data acquisition, segmentation, feature extraction and classification.
- Chapter 5 elaborates the real-time system development.
- Chapter 6 explores the prototype system modules.
- Chapter 7 concludes the thesis and highlights main contributions, findings and future work.

Finally, a detailed list of all the features used in our research with proper definition and organization is presented in Appendix A.

CHAPTER 2

LITERATURE SURVEY

In the past few years, a lot of research has been conducted in the field of emotional speech recognition on different sets of premises and environments. This literature survey is classified into four categories. The first category elaborates on inter-cultural versus cross-cultural studies. The second category highlights various types of emotions. The third category includes acted versus spontaneous emotions. The fourth category discusses the challenges of online processing when compared to offline methods. The last category lists some of emotion recognition applications.

2.1. INTER-CULTURAL VS. CROSS-CULTURAL EMOTIONS

One type of studies focuses on a specific culture to detect basic emotions through its native speakers, such as English [5], Indian [6], German [7], Japanese [8] and Farsi [9].

Recently, some researchers compared results of detecting emotions in more than one language [2], [10], [11], [12]. The number of languages used in the comparison depends on the interest of the researchers, and the availability of audio databases for each language that were built using real-world speech corpora, or by recording acted emotions by native speakers. Recognition of emotions in the Arabic language has been carried out in recent years as part of cross-cultural emotion studies. In [11] and [12], emotions of Arabic language speech are compared with those in other languages. The recognition rate for

Arabic language tended to be lower for most emotions relative to the other languages, because encoders of these non-Arabic databases tended to have a wider acting experience (e.g., training or experience in theatrical productions of some nature) than their Arabic counterparts [2].

2.2. EMOTION SET CATEGORIZATION

In 1872, the seminal work “The Expression of Emotion in Man and Animal” by Charles Darwin discussed a large number of expressions of emotional states [13]. It includes not only states commonly accepted as being emotional, such as joy and anger, but also states such as patience and sulkiness. In his discussion, Darwin did not identify specific emotions to be basic emotions. Later, Paul Ekman and his colleagues studied the facial expression of emotion and suggested that there are six basic emotional states: happiness, sadness, fear, anger, disgust, and surprise [14]. Thus, a set of these basic emotions were targeted for automatic recognition. In [2], [10], [15], [16], [17], [18], [19], [20], [21], the authors address all these emotions or a subset of them in their studies.

In [22] Basic emotions are distributed into two main groups: *problem group* and *no-problem group*. The anger emotion states belong to the *problem group*. Other emotions, e.g. happiness, surprise and neutral emotions are in the *no-problem group*. It is claimed that this way of classification gives good result.

An alternative way of classifying emotion states is by placing them within a dimensional model of two affective dimensions Arousal and Valence [23]. The dimension of valence

ranges from highly positive to highly negative whereas the dimension of arousal ranges from calming or soothing to exciting or agitating in a way that different classification methods can be applied to each specific case as shown in Figure 1. Thus, emotion states can be mapped onto four quadrants positive/high, positive/low, negative/high and negative/low.

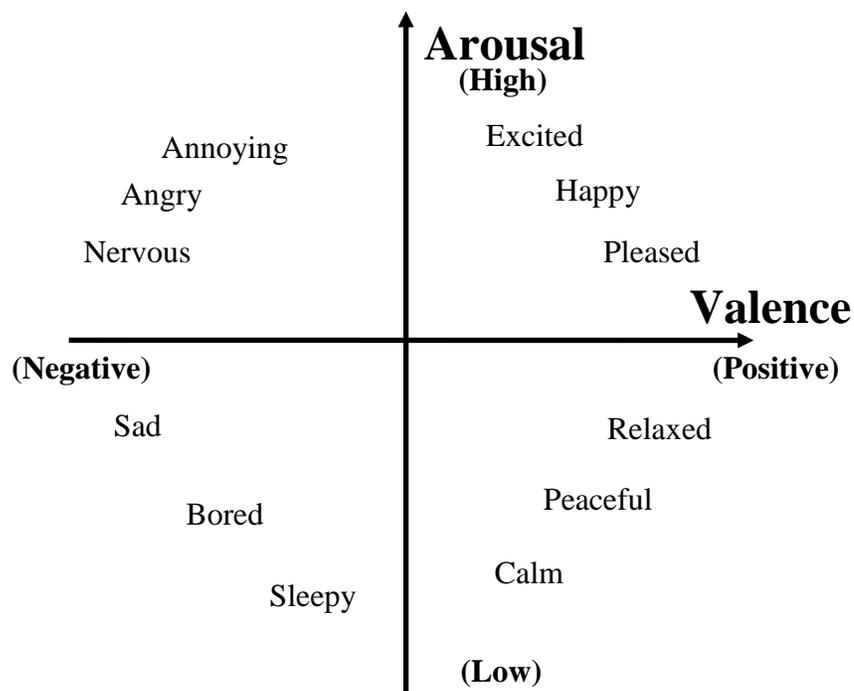


Figure 1: A two-dimensional emotion space with a valence and an arousal axes

On the other hand, some researchers based their studies on a single emotion (for example, anger). In the International Hand Book of Anger [24], anger is recognized as one of the best identified emotions from sound alone, and it is usually associated with changes in fundamental frequency, intensity, and speech rate. With the notable exception of [25], it is

argued that the fundamental frequency (what we typically hear as the pitch of the voice) increases during anger, and anger expressions tend to be more intense (i.e., louder to listeners) and faster. Further, when listeners are forced to choose between several alternative emotions to the one that they have just heard, all studies show that anger is identified at levels far above chance [24]. Furthermore, some studies related to anger recognition and their possible uses by upper management were carried out in the last few years [26] [19]. The objective of these studies is to ensure customer satisfaction of their services. For example, in call-centers, an early detection of anger response from a caller allows an automatic system to forward the caller to a human agent or trying to react with appropriate dialog strategies [10].

2.3. EMOTIONS DATA SOURCES

The lack of real-world emotion speech databases is considered a major challenge for this kind of studies. Audio digital records of spontaneous speeches which are available at governmental or business units are mostly confidential and are not available for public use due to ethical and moral issues, let alone its negative advertising aspect. Nevertheless, researchers use either speech databases or recorded emotions through actors as a solution to overcome this obstacle. For instance, emotion speech corpora in different languages, which were published for speech processing applications needs, have been used in [5], [8], [10], [16], [17], [18], [19], [27], [28]. Another way of finding the required resources is proposed in [5], where the authors built their own database using a semi-professional

female actress. The actress was asked to read the same sentences with four emotional states: anger, happiness, neutral, and sadness.

Another type of databases contains elicited emotions which are neither real nor simulated [29]. They are induced to the participants by different modalities, like stories, movies, virtual reality or role playing. The last type is databases of spontaneous speech which contain real emotions. Figure 2 demonstrates various types of emotion sources and the level of difficulty in recognition.

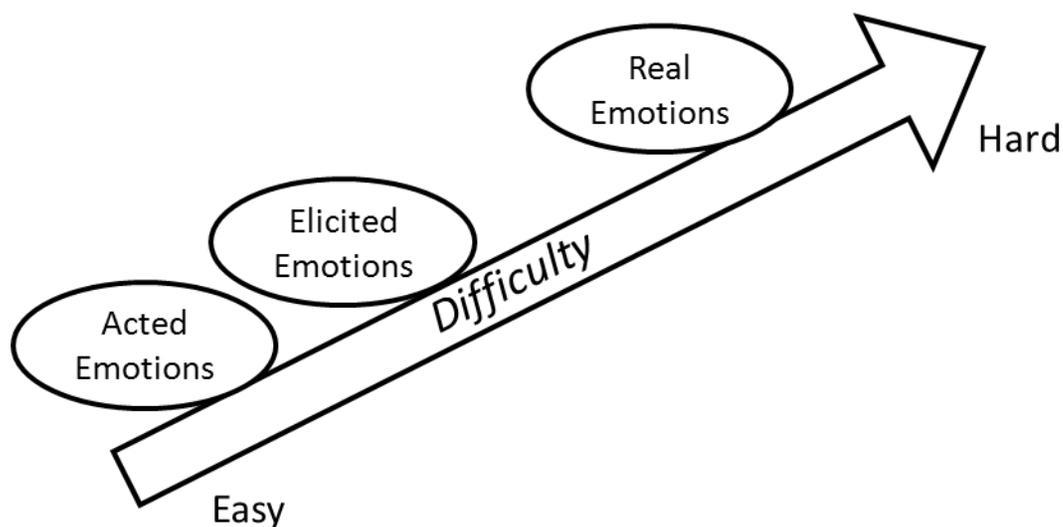


Figure 2: Types of databases used for emotion recognition and their difficulty

In an acted corpus, professional actors are able to imitate similar physiological responses as in real emotional states. Therefore, some people argue that these emotions are better suited for examination and training than real-life emotions because they are pure, that is

free from noises and other uncontrollable influences [4]. However, it is disputed as to how far this can be true. Acted emotions usually comprise only basic and very pronounced emotions whereas real-life emotions are often weak and shadowed, and contain a mixture of several emotions. In conclusion, earlier research shows that spontaneous emotion is much more subtle than that from actors and hence much more difficult to detect and predict [22].

2.4. ONLINE VERSUS OFFLINE RECOGNITION

Online recognition is a necessary prerequisite for the realization of human-computer interfaces that analyze and respond to the user's emotions while he or she is interacting with an application [30].

Offline experiments were conducted to find suitable audio segmentation, feature extraction and classification algorithms [4]. "Suitable" means in this context that they should be fast and at the same time give results as correct as possible.

The criteria used to identify the system as online or offline are: (i) whether the system responds to the speaker's emotional state during or after the interaction, (ii) whether explicit feedback to emotions is given or not, (iii) whether the system reacts itself emotionally in some way or not, and (iv) whether emotion processing is critical or not, that is if wrong processing impairs the application's aims or not.

An early work at MIT Media labs [31] involved recognition using physiological data collected from an actress over many weeks. This work also included heart rate information along with vocal expression recognition. The achieved recognition rates were around 81% for eight emotions in offline recognition. This system was also adjusted to run online but with less accuracy rate (by -8% in comparison with offline version).

In [32], several areas can potentially aid interaction with automated emotion recognition systems in real time. An example is given on how a teacher, human or computer, could tutor more effectively by accurately monitoring the student's affective state.

An experiment was done in [10] using an anger detection component for voice portals. The authors indicated that they achieved enhancements by using an adapted features set, adding cepstral analysis, and using SVM instead of GMMs. They concluded also that duration measures do not play an important role for voice portal systems.

Several previous studies have focused on improving customer relations in call centers [26]. They proposed a real-time affect classification system based on features extracted from the analysis of acoustic speech signal. They considered only two emotions, anger and neutral, and used neural networks in classification. Their system was designed to be speaker- and text-independent and is to be deployed in a call center environment to assist in handling customer inquiries.

On the other hand, this system was not able to adapt to the voice profiles of new speakers without awkward and lengthy offline retraining. As a result, this is improved in [20] to

enable correction of misclassified speech samples without retraining the model on the entire dataset. This methodology is known as incremental learning of the model and allows the system to adapt to new speakers and their emotional profiles over time. The classifier used was a Support Vector Machine (SVM) and it is applied to telephone-based spoken affect classification. A database of 391 natural speech samples depicting anger and neutral speech was collected from 11 speakers. Using this data and features shown to correlate speech with emotional states, an SVM-based classification model was trained. Forward selection was employed on the feature space in an attempt to prune redundant or destructive dimensions. The resulting model offers a mean classification rate of 88.45% for the two-class problem. The results were also compared with an Artificial Neural Network (ANN) designed under the same circumstances.

2.5. EMOTION RECOGNITION APPLICATIONS

Emotion recognition from speech is a relatively new research area with wide applications such as call centers. In [33] they investigate real-life emotion detection across three French corpora collected from call centers. (Anger and Neutral) emotional states were examined in different contexts (service complaints, stock exchange and medical emergency). By mixing the three corpora during training phase they improved the anger detection system. The recognition rate was 73.6% with best 25 selected features and 73.5% with 384 features.

For intelligent household robot, Xu Huahu et al [34] applied speech emotion recognition methods presented by integrated HMM/SOFMNN (Self-Organizing Feature Mapping Wavelet Neural Network) model to the robot platform. They performed categorized recognition of 5 speech emotions as represented by joy, grief, anger, fear and surprise by means of algorithm with the combination of HMM and SOFMNN hybrid models. Accordingly robot can recognize emotional information as contained in the human speech signals for friendly interaction with human beings.

Furthermore, emotion recognition systems could be used in military care systems. In a Japanese study titled "Usage of Emotion Recognition in Military Health Care" [35], authors investigate voice data which were collected from the personnel of military medical corps participating in a special stressful. They used "Sensibility technology ST Emotion" (AGI Japan Inc.) for emotion voice analysis system which determines emotional elements (such as anger, joy, sorrow, and calmness). It also measures feeling of excitement. The results showed that the techniques of emotion recognition may be used for screening of mental status in military situation.

CHAPTER 3

REAL-TIME ANGER DETECTION FRAMEWORK FOR ARABIC SPEECH DIALOGS

In order to develop a prototype system that detects anger, the framework shown in Figure 3 has been followed. This framework consists of two main stages. The first stage is an offline system development which is responsible for building and training a recognition system that can detect anger and neutral emotions using prosodic features. The second stage is the deployment of the constructed system in a real-time environment to detect anger and neutral emotions in Arabic language speeches.

In this chapter, we will discuss the offline development stage which is very similar to a traditional pattern recognition system. It consists of two main steps: feature selection and extraction, and feature classification. The key task of this stage is to choose an appropriate and effective classification model which can provide an accurately predicted result.

3.1. FEATURE SELECTION AND EXTRACTION

In order to be able to classify the sound, we need to extract useful features that will lead to proper classification. The acoustic features that we consider in this study include pitch (absolute and normalized), intensity/energy (absolute and normalized), formants, harmonicity, 12 MFCCs, jitter and shimmer. For pitch, energy and formants, we calculate

the median, mean, standard deviation, minimum and maximum. The average speaking rate is calculated as the inverse of the average length of the voiced part of the utterance. In the following, we will briefly describe the various features that we adopted, and a detailed list will be included in Appendix A.

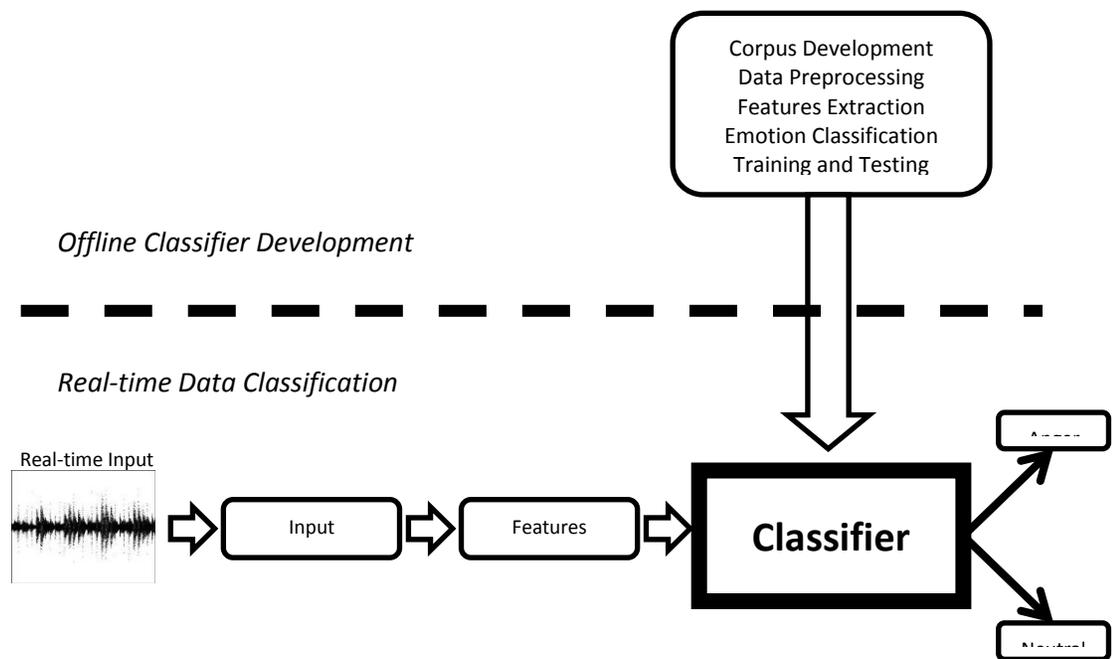


Figure 3: Development Framework for Anger Detection in Arabic Speech Dialogs

3.1.1. Pitch

Pitch is a very sensitive factor which responds to the auditory sense. It is also called fundamental frequency f_0 , and refers to the inverse of the periodic time of a wave pulse generated by air compressed through the glottis from the lungs [36]. Pitch is a very obvious property of speech. Also for non-experts, it is often erroneously considered to be

most important for emotion perception. Pitch does definitely have some importance for emotions, but it is probably not as significant as typically assumed. Generally, a rise in pitch is an indicator for higher arousal.

Pitch can be calculated from the time or the frequency domain. In the time domain, the pitch can, for example, be calculated from the zero-crossings rate. This method is however more suited for musical pitch detection. Obviously, pitch does not exist for the unvoiced parts of the speech signal.

3.1.2. Energy

Energy is the strength of a sound as perceived by the human ear. It is hard to measure directly. Energy can be calculated from the spectrum resulting from the Fourier transformation of the original signal. However, it differs from loudness in that all existing noises add to the signal energy while the ear perceives the loudness of speech as just that.

The energy curve depends on many factors, such as phonemes, speaking style, utterance type (e. g. declarative, interrogative, and exclamatory). It also depends on the affective state of the speaker. High energy roughly correlates with high arousal, and also variations of the energy curve give hints on the speaker's emotion [4].

3.1.3. Formants

Formants are local maxima in the frequency spectrum caused by resonance during speech production [37]. The fundamental frequency f_0 (a.k.a. pitch) is usually the global maximum whereas the formants f_1, f_2, f_3, \dots are the following local maxima in order. Especially for vowels, the distribution of the formants is a very important characteristic. For instance, the first formant is low, while it is high in f_2 and f_3 . The degree of significance of formants for emotions is often considered to be low. Formant position is influenced by the physiological state of a speaker, which is in turn influenced by emotions. Consequently, there exists a connection between formants and emotions, though it may not be obvious or unambiguous [4].

3.1.4. Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) are a parametric representation of the speech signal that is commonly used in automatic speech recognition [38]. They have been proved to be successful for other purposes as well, e.g. speaker identification and emotion recognition. MFCCs are calculated by applying a Mel-scale filter bank to the Fourier transform of a windowed signal. Subsequently, a DCT (discrete-cosine transform) transforms the logarithmic spectrum into a cepstrum. The MFCCs are then computed from the amplitudes of the cepstrum. Usually, only the first 12 coefficients are used [39]. Through the mapping onto the Mel-scale, which is an adaptation of the Hertz-scale for

frequency to the human sense of hearing, MFCCs enable a signal representation that is closer to human perception.

MFCCs filter out pitch and other influences in speech that are not linguistically relevant. Hence, they are very suitable for speech recognition. Though this should make them useless for emotion recognition, they turned out to yield very good results in several approaches [4].

3.2. CLASSIFICATION OF EMOTIONS

The role of a classifier is to assign class labels to objects. Classification should be fast and robust, as well. If no online retraining or adaptation of the classifier is necessary, feature selection and training of the classifier do not need to run in real-time. In our application, the data files are labeled into different emotion classes. We will divide our labeled files which were generated from the feature selection and extraction step into two major sets: a training set and a testing set, with 7:3 split ratio.

In our experiments, we used for classification a robust software tool called DTREG [40]. We choose four types of the available classifiers in DTREG: Support Vector Machine (SVM), Probabilistic Neural Networks (PNN), Decision Tree Forest, and Single Decision Tree. In the following subsections, we briefly review these classifiers.

3.2.1. Support Vector Machine (SVM)

Support Vector Machines (SVMs) have been successfully used in solving many pattern recognition problems. They were originally proposed to carry out binary classification, i.e. to classify data into two categories. They do so by constructing an N -dimensional hyper plane that separates the training data into two classes. As an example [40], consider a two dimensional classification problem, as shown in Figure 4. Input feature vectors from two categories have been plotted on the two dimensional plane, where each point in one category is represented by a circle, and points from the other category are represented by squares. The objective of SVM is to find a hyper-plane, i.e. a line in our example, which separates between the two categories. The data points from each category that constrain the space in which the separating line can reside in are called the support vectors. Figure 4 shows the support vectors for each class.

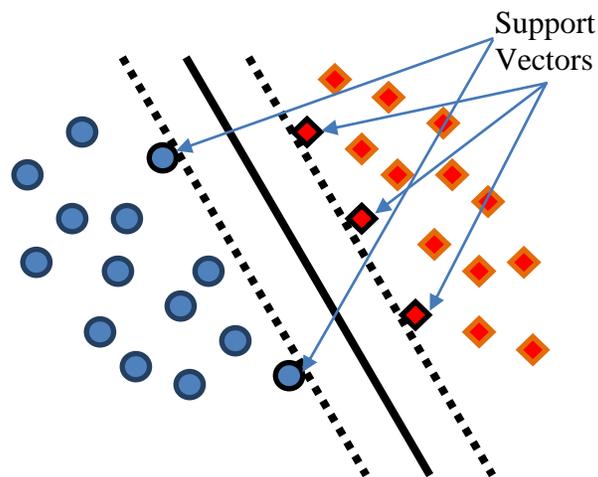


Figure 4: Two-dimensional plotted data

When the data is separable, there exist one or more hyper-planes (possibly infinite number of them) that separate the classes completely. If the data is separable by a non-linear region, SVM analysis finds a mapping, through a function called a kernel function, to a higher dimensional space, say with dimension N , in which an $N - 1$ dimensional hyper-plane separates the data. When the data is inseparable, a hyper-plane is sought where the error in training data is minimized.

3.2.2. Probabilistic Neural Networks

Probabilistic neural networks (PNNs) were originally proposed by Donald F. Specht in 1990 [41]. Instead of classifying new data directly from the training samples, probabilistic neural networks approximate the probability density function of the distribution of samples. A PNN consists of four layers [40]: an input layer, a pattern layer, a summation layer and an output layer, as shown in Figure 5.

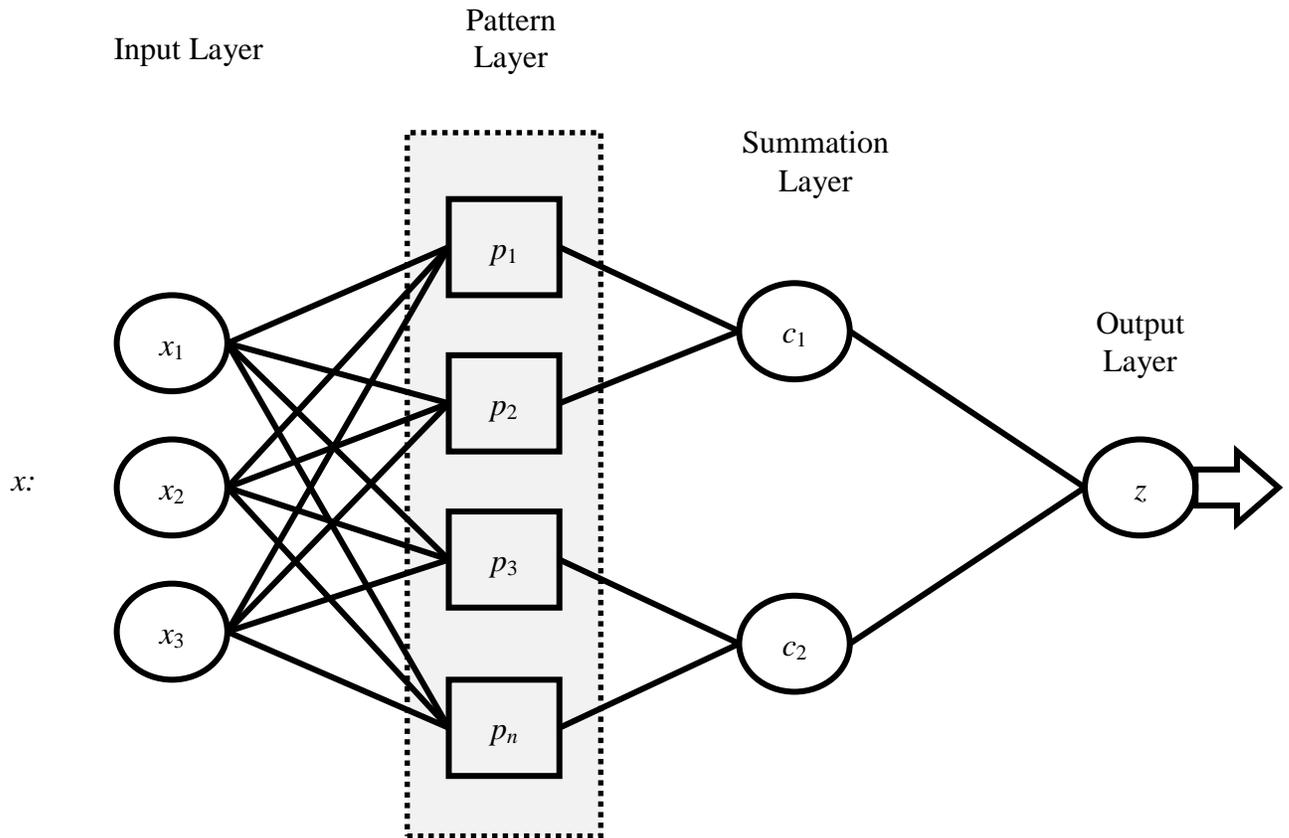


Figure 5: All PNN networks have four layers.

1. Input layer: For each input feature of a specific data point, one neuron is allocated. Values of the input values are normalized before being fed into each neuron of the pattern layer.
2. Pattern layer: For each training sample, a neuron is allocated. The neuron stores the values of the corresponding feature of the sample and its target value. When presented

with the x vector of input values from the input layer, it computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function using the sigma value(s). The resulting value is passed to the neurons in the summation layer.

3. Summation layer: For each classification class in the output, one neuron is allocated. The summation neurons add the values for the class they represent. The resulting weighted values from neurons belonging to a certain class in the pattern layer are fed into the corresponding summation neuron only.
4. Decision layer: It compares the weighted votes for each class coming from the summation layer and uses the largest vote to predict the target class.

Since PNN networks approach Bayes optimal classification, they can generate accurate predicted target class probability scores, provided they have enough training samples. In addition, they are relatively insensitive to outliers. However, they do require more memory space to store the model [40].

3.2.3. Decision Tree Forests

A decision tree forest extends single decision trees by constructing an ensemble of trees whose predictions are combined to make the overall prediction for the forest. A single decision tree [40] is a binary tree in which each node represents a set of samples, where the root node contains all samples. An example of a decision tree is shown in Figure 6.

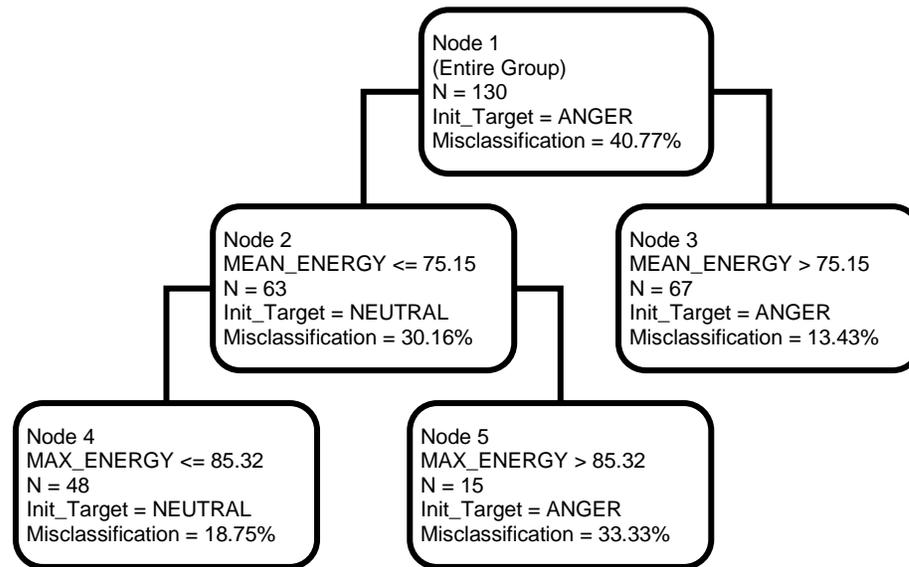


Figure 6: Example of Single Decision Tree.

A decision tree is constructed by dividing a node into two child nodes. The split partitions the samples in the parent node between its two children on one of the features. For example, Figure 6 splits the root node based on the Mean_Energy feature. Therefore, Node 3 contains all samples in which the Mean_Energy value is greater than 75.15. The same split process may continue for subsequent nodes, which is called recursive partitioning. The objective of this partitioning is to reach leaf nodes where the error in classification is the least. The process of classification proceeds as follows: Starting from the root node, go the left child or the right child based on the feature value of the splitting node. This procedure continues until a leaf node is reached, where the class shown in the leaf will be assigned to the sample.

One disadvantage of single decision trees is the high classification error. Decision tree forests, however, have shown much better performance. They were developed by Breiman [42]. In this approach, a decision tree forest grows a number of independent trees, in parallel. Among the many features of decision tree forests are that they are as easy to create as single decision trees, they can handle large numbers of features, and can handle missing values, and they do not suffer from the problem of overfitting [40]. On the other hand, they lack the ability to being visualized like single decision trees, and hence they act as a black box like neural networks.

Assume we have a dataset consisting of N samples. The algorithm for building decision tree forests is as follows [40]:

1. Take a random sample from the dataset with replacement, called “bagging”.
2. Using the samples selected in step 1, construct a maximum size decision tree without pruning. During building the tree, choose, randomly, a subset of features to be chosen to build the tree.
3. Repeat Steps 1 and 2 a large number of times constructing a forest of trees.

Classification of a sample is done as follows: Run the sample on each developed tree in the tree forest, until reaching a leaf node and use the category of the leaf node one vote for it. The category with the largest number of votes will be assigned as the class of the sample.

CHAPTER 4

OFFLINE SYSTEM DEVELOPMENT

The purpose of this chapter is to describe the offline development part of the framework.

The main steps in this process include:

- **Emotion Set Selection:** Choosing a set of emotion states to be included in the study.
- **Corpus Development:** The data acquisition, annotation and segmentation process.
- **Offline Features Extraction:** Selection of prosodic features.
- **Classification of Emotions and Testing:** Building an offline classifier for the selected feature vector.

4.1. EMOTION SET SELECTION

Since real-time emotion detection is tightly related to applications developed for customer support call centers, it is important to determine which set of emotions to detect. To achieve this, we need to know the type of emotions that frequently occur in call centers. In [20], they studied a natural corpus from call centers which we refer to it as NATURAL. This dataset consists mainly of neutral speech, with the second largest category made of angry speech as shown in Table 1. Because of the low distributions of sadness, fear,

disgust, and surprise, it can be assumed that their probabilities of occurring in applications requiring anger detection (such as call centers) are quite low. Call centers mainly concerned about customer satisfaction which can be monitored by measuring their customer's expressed emotions during customer service calls. Anger emotion could be considered a clearer indicator than happiness which rarely occurs when compared to anger emotion. Therefore, we opted to consider only anger and neutral emotional states and discarded data collected under the other emotional states (happiness, sadness, surprise, fear, disgust), which were either scarce or not available.

Table 1 Distribution of perceived speaker affect in the natural corpus (NATURAL)

Emotion class	Number of conversations (%)
Neutral	93.3
Anger	3.1
Happiness	1.8
Sadness	0.1
Surprise, Fear, Disgust	0.0

Another reason for limiting the study to two emotion states (Anger and Neutral) only is the difficulty of collecting real-life speech samples from our local region. We tried to contact local companies which provide phone call services to their customers. Call centers, Internet Service Providers and hotels are examples of these companies. We sent them an official request letter endorsed by King Fahd University of Petroleum and Minerals with proper guarantee that their confidential customer service recorded calls will

be used only for academic research. Despite that, we received answer from only one local Internet Service Provider, SAHARA NET. The details of this dataset will be described in Section 4.2.

In addition to the previous reasons of why we have chosen the anger emotion state to be detected, the clear characteristics of anger emotion state is expected to enhance the accuracy of recognition. Figure 7 shows the pitch contours of two example utterances from the NATURAL dataset. It can be seen that the angry sample has downward slopes and a greater range. On the other hand, the neutral sample has a monotonous contour with a shallow range [20].

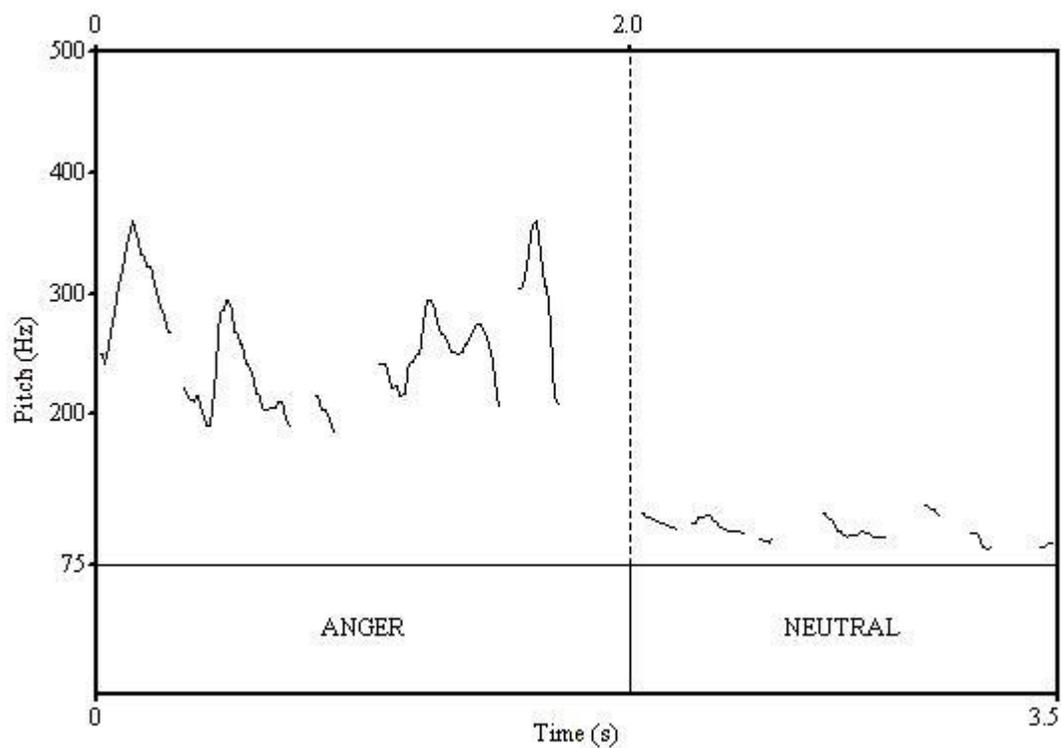


Figure 7: Example pitch contours for anger and neutral utterances

Table 2 lists the statistical properties of prosodic features (F0, Energy, speaking rate and Formants) for six basic emotions, with neutral being the reference point for comparison.

In Addition to prosodic features, in [25] the following additional acoustic features for Anger are listed:

- High volume
- High volume variance
- Fast rate of utterance
- Short time of speech duration
- Short time of pauses duration
- Small number of pauses

Table 2 Statistical properties of prosodic features for selected emotions compared to the neutral emotional state

Emotion	F0 mean	F0 range	Energy	Speaking Rate	Formants
Angry	Increased	Wider	Increased	High	F1 mean increased; F2 mean higher or lower; F3 mean higher
Happiness	Increased	Wider	Increased	High	F1 mean decreased; F1 bandwidth Increased
Sadness	Decreased	Narrower	Decreased	Low	F1 mean increased; F1 bandwidth decreased; F2 mean lower
Surprise	Normal or increased	Wider	-	Normal	-
Disgust	Decreased	Wider or narrower	Decreased or normal	Higher	F1 mean increased; F1 bandwidth decreased; F2 mean lower
Fear	Increased or decreased	Wider or narrower	Normal	High or low	F1 mean increased; F1 bandwidth decreased; F2 mean lower

4.2. CORPUS DEVELOPMENT

Pattern recognition systems require a huge number of data samples. No Arabic language corpus for anger detection is available, to the best of our knowledge. Therefore, it was a challenge for us to build our own Arabic emotion corpus. The following is a detailed description of our data sources:

4.2.1. ISP Customer Service Records

As mentioned above, we faced a difficulty in collecting real-life Arabic speech samples from our local industries. Only one local Internet Service Provider, SAHARA NET, responded positively to our request despite all our effort. They provided us with 16 customer service recordings in digital format ranging in duration from less than one minute to four minutes long. They were digitized in 16kbps/16kHz/mono MP3 files format. Although these recordings provide realistic data, unfortunately they do not include anger emotion states for neither the customer nor the agent. All given recordings were handled in a proper way with customers being seemingly satisfied. Therefore, we had to look for another source of recordings in order to collect enough anger emotion state data.

4.2.2. TV Debate Show

This source is publicly available on the Internet. It represents a set of recorded episodes of a live Arabic political debate show which is called “The Opposite Direction”. This is one

of Al-Jazeera Satellite Channel programs similar to CNN's Crossfire. There are several reasons for choosing this TV program for this research. First, more than 120 45-minute episodes of this program are available on the Internet. Second, the program hosts Arabian speaker guests from different Arab countries and from different genders providing a good variety of political subjects. Finally and most importantly, the anger emotion state exists in all episodes because of the nature of the conversational debates held in this program.

We selected a set of these episodes carefully for our research. We then proceeded with the segmentation process because of the long episode duration. The number of samples extracted from the selected episodes is more than 400 samples for anger emotion state. These samples vary from one second to 9 seconds with different speaker's Arabic cultures and gender.

4.2.3. An Angry Customer's Call

Since speaking in a TV show is still different from speaking through a telephone line, we kept looking for angry call data on the Internet. We found only one such call published in YouTube. The recording features a customer who was very angry while speaking with the agent. The call duration was four minutes long. We were able to extract 45 sample data which reflect anger states from this clip.

Since sentences in spontaneous speech do not have a fixed length of time and they could be extracted from different environments, there is therefore a need to split them into smaller parts with accurate labeling. For our purpose, we partitioned the digitized audio

into voiced and unvoiced parts. Unvoiced parts represent pauses or silences of the speaker. Unvoiced parts may not necessarily contain vocal words as they may, in addition, contain music, tones, humming or any other sound around the speaker environment.

In order to accomplish this, we used a software tool called Praat [43]. It is capable of annotating various intervals of the sound signals assigning text or labels to them. Two layers of annotation, called Tiers, were used: *Speaker tier* and *Emotion tier*; as shown in Figure 8.

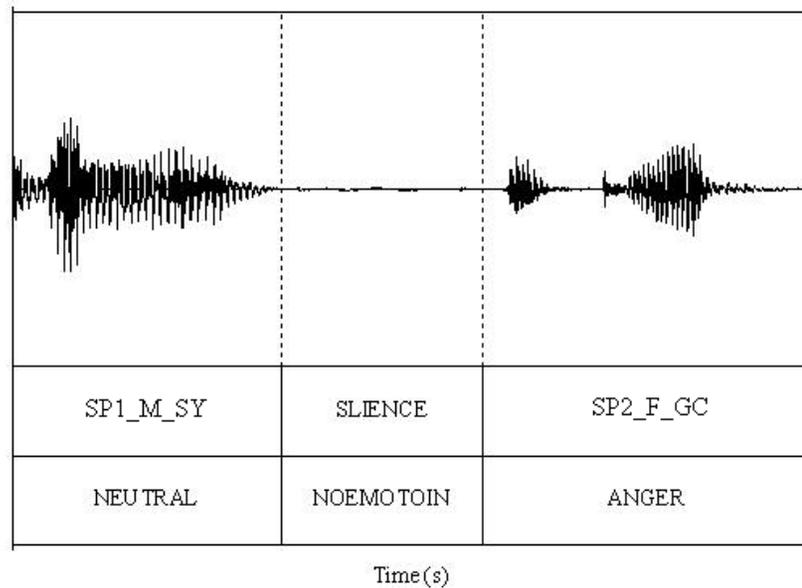


Figure 8: Labels are given for each interval into two tiers (levels) of annotation, Speaker tier and Emotion Tier

Each voiced interval in the speaker tier was given annotation that defines its content, which consists of three parts, the speaker ID, his/her gender and his/her nationality. If

there is more than one speaker talking at the same period of time, every one of them is included in the annotation. Figure 9 shows the general format of the labeling.

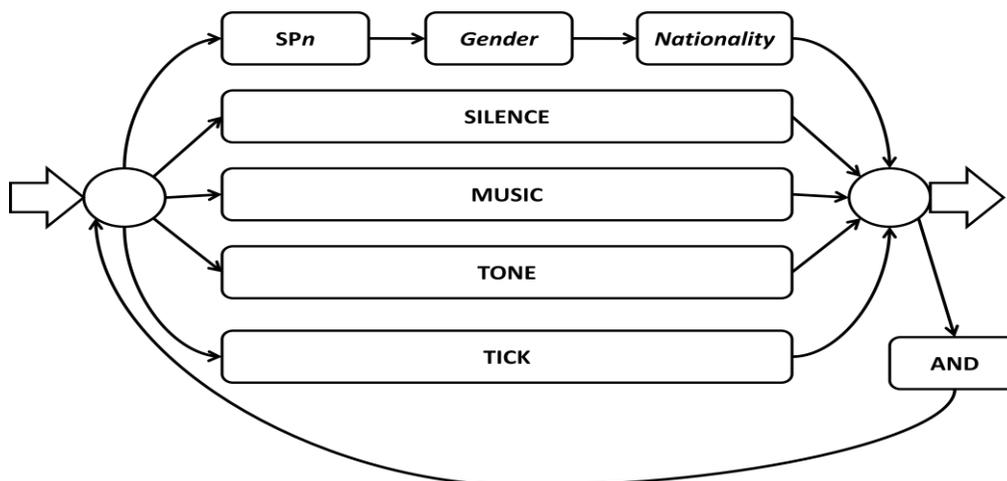


Figure 9: General format of annotation of each interval in the Speaker tier

The Emotion tier is assigned an initial labeling which describes the emotion. The emotion is chosen according to the labeler opinion based on the dialog context. The period of an emotion is independent from the period of the Speaker tier as shown in Figure 10. This can be useful for extracting speaker-independent emotions from the Emotion tier only.

SP1_M_GC	SP2_F_SY	SP1_M_CG	SP2_F_SY	SP1_M_GC	Speaker Tier
NEUTRAL					Emotion Tier

Figure 10: Different speakers turns in one emotion period in TextGrid

Contents of the Emotion tier are varied, containing real emotions or no-emotion. Anger and Neutral are the only emotion states in our experiments for a speaker. On the other hand, silence periods, music, tones, noise and ticks are labeled NoEmotion in the emotion tier. Table 3 shows annotation content of the Speaker tier and the Emotion tier in our emotion corpus.

Table 3 The content of the Speaker and Emotion tiers in dialog annotation

Speaker tier content	Emotion tier content
$Speaker_i_Gender_Nationality$ where $i > 0$ for one speaker voiced parts	NEUTRAL or ANGER
$Speaker_i_Gender_Nationality_AND_Speaker_j_Gender_Nationality$ where $i, j > 0$ and $i < j$ for more than one speaker voiced parts	
SILENCE	NO EMOTION
MUSIC	
LONG_TONE	
SHORT_TONE	
TICK	
NOISE	

After completion of the of annotation process, we wrote a script to split the long sound file into smaller segments as separated files with their label names as file names based on our annotation. Combinations of labels from both tiers in the same period of time were used as the name of the split file. For example, SP1_M_SY-NEUTRAL.wav and SP2_F_GC-ANGER.wav, represent the sound signals of the first male speaker who is Syrian with the segment being identified as neutral, and the second speaker who is a female from an Arabian Gulf country in anger emotion state, respectively, as show in Figure 11.

SP1_M_SY	SILENCE	SP2_F_GC	SP1_M_SY	Speaker Tier
NEUTRAL	NO EMOTION	ANGER		Emotion Tier
SP1_M_SY-NEUTRAL.WAV	SILENCE-NOEMOTION.WAV	SP2_F_GC-ANGER.WAV	SP1_M_SY-ANGER.WAV	Split File

Figure 11: Names of split .wav files from long sound file based on annotation

4.3. EMOTION SURVEY

In order to verify the accuracy of labeling, as it maybe subjective, a listening test of the initial emotional utterances was carried out with the help of groups of an odd number of randomly-selected volunteer human judges. Each group consisted of one female and two male judges. The listeners have not been made aware of the initially-assigned emotion. All sound files were published on a website, and each listener played these sound files one by one and is asked to select at most one emotional choice (anger, neutral, others and unclear), as shown in Figure 12. These choices were saved in a database associated with each sound file.

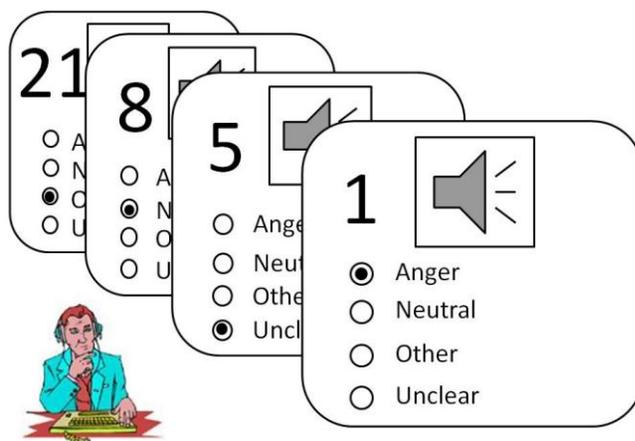


Figure 12: Survey on sound files.

4.3.1. Survey Results

A total number of 484 clips were divided into 10 sets, four of which contained 49 clips and the rest contained 48 clips. Each set was given to a group of participants, consisting of two males and one female Arabic native speakers. Each clip has been initially assigned one of two emotion states, either anger or neutral. The web survey included four options to describe each sound clip “Neutral”, “Anger”, “Other” and “Unclear”. Table 4 shows the results of the listeners’ judgments as compared to the initially-labeled emotion by the annotator based on dialogs context. It is worth mentioning that less than 15% of the responses included “Unclear” and “Other” has high scores, indicating that the data chosen was mostly clean to recognize.

Table 4 Score function results

Initially Labeled	Total	Score 3/3	Score 2/3	Total Match	Score 1/3	Score 0/3	Total Mismatch
Neutral	244	114 47%	71 29%	185 76%	42 17%	17 7%	59 24%
Anger	240	74 30%	78 33%	152 63%	88 37%	88 37%	88 37%
Other	-	5	17	22	101	361	462
Unclear	-	14	34	48	99	337	436

The results show that there is a level of disagreement between the initial labeling and the participants’ choices. In this case we will consider the initial labels as the ground truth, since the annotator has based his judgment on the dialog context listening to previous and/or consequent clips when needed. As a result of this, we carried out the classification

experiments using two sets of data. The first set consists of all data with initial labeling and will be called context-aware, whereas the second set excludes all clips having high score disagreement and will be called context-free.

Table 5 shows the comparison of all clips with a non-zero score function of “Unclear” against their duration.

Table 5 Unclear vs. Duration

Unclear option	Number of clips	Minimum Duration in seconds	Maximum Duration in seconds	Average Duration in seconds
Low score	99	0.32	17.85	1.94
Intermediate score	34	0.46	2.04	1.21
High Score	14	0.29	1.77	0.84

The purpose of this comparison is to help us to set a threshold value for the segmentation process, in order to split the long voiced parts into smaller recognizable parts that are suitable for building an efficient real-time classification framework.

4.4. EXPERIMENTAL RESULTS OF CLASSIFICATION

Based on the data corpus we described in Section 4.2, we studied the performance of the four classification techniques mentioned earlier. We carried out two sets of experiments

depending on the dataset used for training and testing. The first set includes all the clips that were generated, and labeled using our manual annotation that took the contextual knowledge of the clip into account. We call this set the "context-aware" dataset. The second set includes a subset of the clips generated that were annotated by the groups who participated in the web survey, which had non-conflicting labeling with our own labeling of the data. We call this set the "context-free" dataset. Table 6 shows the method by which data was included or excluded in the context-free dataset based on the high score labeling of the survey and our manual label. The reason for doing so is to test the influence of the knowledge of context in small segments on the proper classification of data.

Table 6 Data included in both context-aware group and context-free group

Corpus Data	Initial label	High score label	Included in "context-aware" group	Included in "context-free" group
SP1_F_GC	NEUTRAL	NEUTRAL	YES	YES
SP3_M_JO	ANGER	NEUTRAL	YES	NO
SP1_F_SY	ANGER	ANGER	YES	YES
SP2_M_AR	ANGER	NEUTRAL	YES	NO

As a result, a total of 2003 clips were included in the context-aware dataset and a total of 1802 clips were included the context-free dataset. These clips were distributed between the training and the testing sets as shown in Table 7, where 70% were included in the training set and 30% were included in the testing set.

Table 7 Number of clips included in training set and testing set

Class	Total (100%)		Training (70%)		Testing (30%)	
	Context-aware	Context-free	Context-aware	Context-free	Context-aware	Context-free
ANGER	524	417	367	292	157	125
LONG_TONE	41	41	29	29	12	12
MUSIC	34	34	24	24	10	10
NEUTRAL	1006	935	704	655	302	280
NOISE	56	56	39	39	17	17
OVERLAP_ANGER	163	146	114	102	49	44
OVERLAP_NEUTRAL	155	149	109	104	46	45
SHORT_TONE	6	6	4	4	2	2
TICK	18	18	13	13	5	5
Total	2003	1802	1403	1262	600	540

In order to develop the offline pattern recognition system, we initially selected support vector machines (SVM), probabilistic neural networks (PNN), single decision trees and decision tree forests to carry out the classification. We carried out data segmentation based on the pauses, where a segment's end point is a pause. However, we noticed that some of these segments have lengths greater than 7 seconds without any pause detected. Processing audio data of such lengths increases the processing time of the system considerably, especially that we will be applying these techniques on a real-time system. Therefore, we chose to segment the data based on pauses as long as the segment duration does not exceed three seconds. If the voiced segment exceeds three seconds, we only consider the first three seconds for this segment and make the rest part of the next segment. Then, we carry out the feature extraction process on these segments. We

extracted all features listed in Appendix A. Then, we carried out the classification process, where the classes consisted of ANGER, NEUTRAL, OVERLAP_ANGER, OVERLAP_NEUTRAL, MUSIC, NOISE, TICK, SHORT_TONE and LONG_TONE. The OVERLAP_ANGER and OVERLAP_NEUTRAL classes refer to clips containing more than one voice talking simultaneously. This usually indicates a conflict among the speakers (people are talking, but not listening to each other), especially in call center conversations. To better understand the impact that context-aware data may have on the classification accuracy, this process was applied first on context-aware data, and then on context-free data. Table 8 shows the resulting accuracy of the classification for both datasets.

Table 8 Accuracy percentage for each classifier on both training subsets

Class	Classifiers accuracy %							
	SVM		PNN		Decision Tree Forest		Single Tree	
	Context Aware	Context Free	Context Aware	Context Free	Context Aware	Context Free	Context Aware	Context Free
ANGER	75.8	77.08	74.52	77.08	75.16	72	64.97	59.2
LONG_TONE	100	100	100	100	100	100	100	100
MUSIC	90	71.43	100	100	100	100	60	70
NEUTRAL	88.74	87.63	83.44	77.96	89.74	91.07	82.78	84.29
NOISE	76.47	44.44	64.71	55.56	70.59	76.47	47.06	23.53
OVERLAP_ANGER	73.47	70	69.39	82.5	61.22	64.44	53.06	68.89
OVERLAP_NEUTRAL	69.57	58.62	71.74	65.52	56.52	63.64	50	59.09
SHORT_TONE	50	50	50	50	50	50	0	0
TICK	20	0	0	0	20	20	20	20
Overall Accuracy	81.83	79.37	78.33	77.29	80.17	81.3	71.33	72.41

We observe from these results that SVM has the highest accuracy rate among all classes of context-aware datasets, whereas context-free dataset has the highest accuracy rate on Decision Tree Forest. SVM on “Anger” class tops the other classifiers in both context-aware and context-free datasets. The results also show that with the exception of few cases, the difference in accuracy between context-aware data and context-free data is minimal. Hence, the choice of prosody to carry out anger detection is justified, as compared with the more computationally intensive speech recognition and understanding techniques.

4.5. HANDLING MISSING VALUES

Missing values occurred mainly due to the failure of the feature extraction process to return values within the defined range of these features. For example, pitch values range from 75Hz to 600Hz. Whenever the feature extractor cannot assign a certain segment a pitch value in that range, it will assign it a null value. As a result, some classifiers handle these null values by either replacing them with uninformed calculated values, such as the median of all the dataset, or drop the whole feature vector from the training process. This will either decrease the number of feature vectors available to build the classifier, or increase the error in misclassification due to the incorrect choice of calculated values. Hence, the classification accuracy will be negatively affected. To avoid this problem, we used two approaches. The first one was to drop the feature and not extracted if it proved to be of low importance to the classification process. Seven features were handled using this

approach as shown in Table 9. The second approach was to replace the null values with an informed value that is more reasonable than the other calculated values. For example, pitch values that were marked as null were replaced with zero, whereas the third formant null feature was replaced by 5500 which represents a value greater than the maximum value in the third formant dataset. Nineteen features were handled in this way, as shown in Table 10.

Table 9 Attributes with missing values and low classification importance.

Feature attribute	Number of missing values out of 1403	Classifier importance %		
		SVM	PNN	Decision Tree Forest
jitter_loc	25	0	0.082	0
jitter_loc_abs	25	0	0.116	0
jitter_rap	26	0	0.061	0
jitter_ppq5	30	0	0.141	0
shimmer_loc_dB	26	0	0.828	0
shimmer_apq5	35	3.125	0.097	3.125
shimmer_apq11	64	6.250	1.017	3.125

Table 10 Replaced new values for features attributes that may have null values

Feature Attribute	New value	Feature Attribute	New value
Min_F0	0	minF1	0
Min_Stylfit_F0	0	maxF3	5500
Max_F0	0	medianF3	5500
Max_Stylfit_F0	0	meanF3	5500
Mean_F0	0	sdF3	0
Mean_Stylfit_F0	0	minF3	5500
Mean_Pitch_Slope	0	meanPeriod	0
medianPitch	0	sdPeriod	0
Standard Dev. Pitch	0	shimmer_loc	0
		shimmer_apq3	0

Subsequently, we created a new feature set by replacing all null values with the suggested new values to train our selected classifiers. Table 11 shows the results of the same experiments of Table 8 after handling the missing values. There is a notable improvement in the accuracies of neutral and anger emotions classification in the case of SVMs and PNNs. In the decision tree forest classifier, almost all accuracies were improved for all classes. However, the performance of the single tree classifier is low and inconsistent as can be observed from Table 8 and Table 11. Therefore, we will drop this classifier from our subsequent experimental work. Table 12 compares the overall performance of the SVM, PNN and decision tree forest classifiers on the datasets in the presence of missing values and after handling the missing values. We can observe that the small discrepancies between the context-aware and context-free accuracies were greatly reduced after handling missing values.

Table 11 Accuracy percentage of four classifiers after handling missing values

Category	Accuracy percentage %							
	SVM		PNN		Decision Tree Forest		Single Tree	
	Context aware	Context free	Context aware	Context free	Context aware	Context free	Context aware	Context free
ANGER	77.07	75.2	77.71	75.2	80.26	73.6	73.25	53.6
LONG_TONE	100	100	100	100	100	100	100	100
MUSIC	90	90	100	100	100	100	70	70
NEUTRAL	89.07	90.36	87.42	85.71	89.40	92.14	80.46	81.79
NOISE	52.94	70.59	64.71	76.47	70.59	82.35	41.18	52.94
OVERLAP_ANGER	67.35	68.89	69.39	73.33	63.27	57.78	57.14	46.67
OVERLAP_NEUTRAL	65.22	65.91	65.22	75	58.7	56.82	50	63.64
SHORT_TONE	50	50	50	50	50	50	50	0
TICK	40	0	20	20	20	20	40	60
Overall Accuracy	81	81.67	80.83	80.97	81.67	81.3	73	69.63

Table 12 Comparison of average accuracy results for handled and unhandled missing values

Classifiers	SVM		PNN		Decision Tree Forest	
	Context aware	Context free	Context aware	Context free	Context aware	Context free
Unhandled missing values	81.83	79.37	78.33	77.29	80.17	81.29
Handled Missing values	81	81.67	80.83	80.93	81.67	81.3

4.6. MERGING OF CLASSES

Another issue to consider here is the dependency that exists among the classes we defined. For example, ANGER and OVERLAP_ANGER refer to the same emotion, despite the difference in their context. Similarly, NEUTRAL and OVERLAP_NEUTRAL refer to the same emotion. In order to investigate this issue, we considered the confusion matrix of these classes and studied the source of inaccuracy in the classification process. Table 13 shows the results for the decision tree forest classifier on the context-aware dataset. We can observe that 11 samples out of the 18 misclassified samples of the OVERLAP_ANGER class have been “misclassified” as ANGER. Similarly, 15 samples out of the 19 misclassified samples from the OVERLAP_NEUTRAL class have been misclassified as NEUTRAL. This led us to further investigate those statistics by considering merging of the results of ANGER and OVERLAP_ANGER into one ANGER class, and the results of NEUTRAL and OVERLAP_NEUTRAL into one class in the confusion matrix.

Table 13 Confusion matrix for the decision tree forest.

Class	ANGER	LONG_TONE	MUSIC	NEUTRAL	NOISE	OVERLAP_ANGER	OVERLAP_NEUTRAL	SHORT_TONE	TICK	Total
ANGER	126			26		4	1			157
LONG_TONE		12								12
MUSIC			10							10
NEUTRAL	30			270			2			302
NOISE				3	12				2	17
OVERLAP_ANGER	11			7		31				49
OVERLAP_NEUTRAL	2			15		2	27			46
SHORT_TONE				1				1		2
TICK					3			1	1	5
Totals	169	12	10	322	15	37	30	2	3	600

These results indicate that there is no need to differentiate between the ANGER-OVERLAP_ANGER and the NEUTRAL-OVERLAP_NEUTRAL classes. They also indicate that there is minimal difference in the feature values among the different classes belonging to the same emotion. Since the remaining classes NOISE – TICK – MUSIC – LONG_TONE – SHORT_TONE have significantly fewer number of samples than ANGER and NEUTRAL emotion classes, which negatively affects their classification accuracies, we dropped them from the reported results in all subsequent results.

4.7. BEST SEGMENT LENGTH

Since one of our objectives is to carry out the classification in real time, it is important to determine the best segment length that has a high accuracy rate with low latency time. We achieve this by studying the relationship between the length of a segment and the accuracy of its classification in the decision tree forest classifier when applied to context-aware data with handling of missing values. The reason of this choice is because it has the highest classification accuracy among other classifiers. Figure 13 depicts the relation between the accuracy rate and the duration of the classified samples. Each line represents the accuracy of the classification when the overlap classes are not merged and when they are merged with its corresponding emotion classes, respectively, according to their duration from shortest to longest. It is evident that misclassification occurs in segments with short duration much more than its occurrence in segments with long duration. The accuracy starts to grow when the duration exceeds 1.75 seconds.

To determine the best duration rate, we got assistance from our reported manual classification results reported in

Table 5 of Section 4.3.1, which were obtained from using a group of judges. We noticed that the average duration of segments that had the highest unclear percentage voting was 0.84 second, with minimum and maximum durations of 0.29 and 1.77, respectively. The lowest unclear percentage voting occurred with segments with average duration of 1.94 seconds, with minimum and maximum durations of 0.32 and 17.85, respectively. These

results corroborate our decision to choose two seconds as the duration that can be utilized for real-time classification of anger and neutral emotions. Hence, any voiced segments with greater durations will be split into two or more segments of maximum duration of 2 seconds each. Table 14 shows the results of 10-fold cross validation on chosen classifiers after segmenting our samples to a maximum of 2 seconds each. The results show that, the SVM classifier outperformed in general the decision tree forest and the probabilistic neural network classifiers in context-free data and had a similar accuracy of decision tree forest classifier on context-aware data.

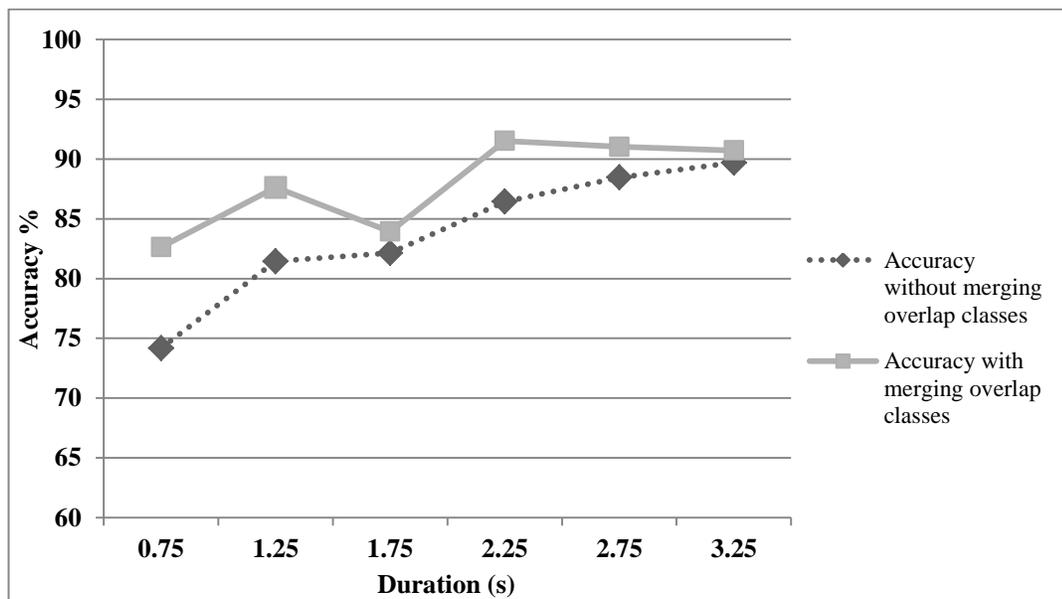


Figure 13: Accuracy vs. duration

Table 14 Misclassification rate of three classifiers on context-aware and context-free groups

Datasets	SVM			PNN			Decision Tree Forest		
	Anger	Neutral	Overall	Anger	Neutral	Overall	Anger	Neutral	Overall
Context-aware	82.4	89.39	86.8	80.4	89.86	86.35	82	89.63	86.8
Context-free	81.64	91.41	88.06	79.71	90.91	87.07	79.71	91.41	87.4

CHAPTER 5

REAL-TIME SYSTEM DEVELOPMENT

Detection of anger in real time involves the following challenge: There is a constraint on the time available to carry out feature extraction and classification. This implies the likely need for omitting some features for the sake of reducing the processing time. This constrains the classifier to be used.

5.1. VOICED PARTS SEGMENTATION METHODOLOGY

A number of different techniques have been proposed in the literature to carry out speech segmentation for various available emotion speech corpora. Some of them segment data using a fixed time duration ranging between 0.5 second and 2 seconds. Other techniques have segmented data at the word, word in context and utterance levels, regardless of the time duration [4], [22]. Vogt, et al. proposed the use of pause detection in order to segment the data [22]. A pause can be detected by measuring the short-time energy on a frame by frame basis. They concluded that the longer the duration of the segment is, the higher the accuracy of emotion detection. However, this does not suit our situation since the duration of such segments can be longer than seven seconds, considerably increasing the feature extraction time, and hence reducing its ability to carry out real-time classification. In order to overcome this, we employed the following approach: We seek a pause, within two seconds from the beginning of the utterance. We chose 65 dB as an

energy threshold to determine a pause with a minimum duration of 0.6 second based on pause definition in a research done on question detection in Arabic audio monologues using prosodic features [44]. The noise in real-life records increases energy threshold from 59 dB to 65 dB. If such a pause is found, the utterance is segmented at this point. If not, we look for the last pause with less than 0.6 second duration before reaching the end of the current two-second utterance in order to determine its end, as shown in Figure 14. Otherwise, if we again fail to identify such a pause, the two-second utterance will be extracted as one segment and a pause is searched for within the next two seconds, in the same way; see Figure 15.

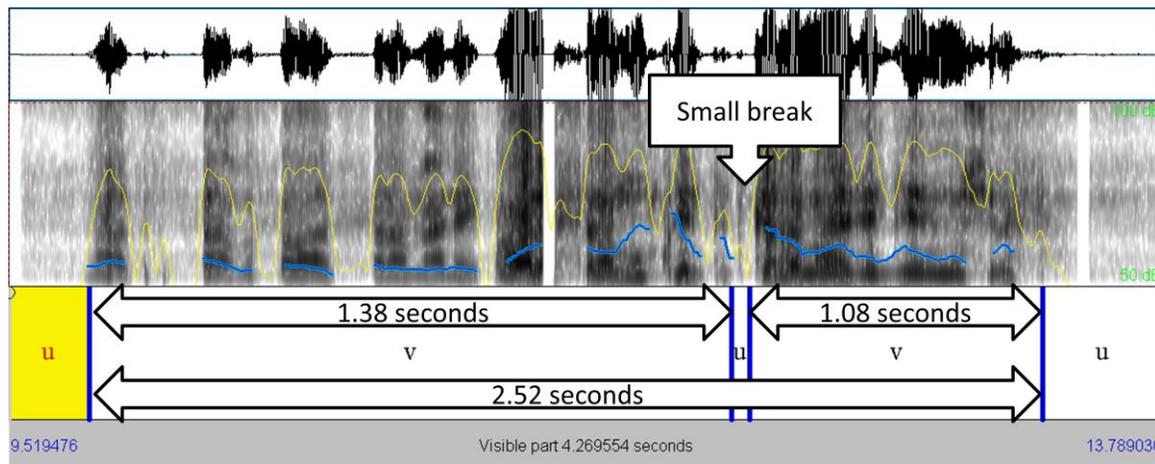


Figure 14: Segmentation of long voiced part based on small break

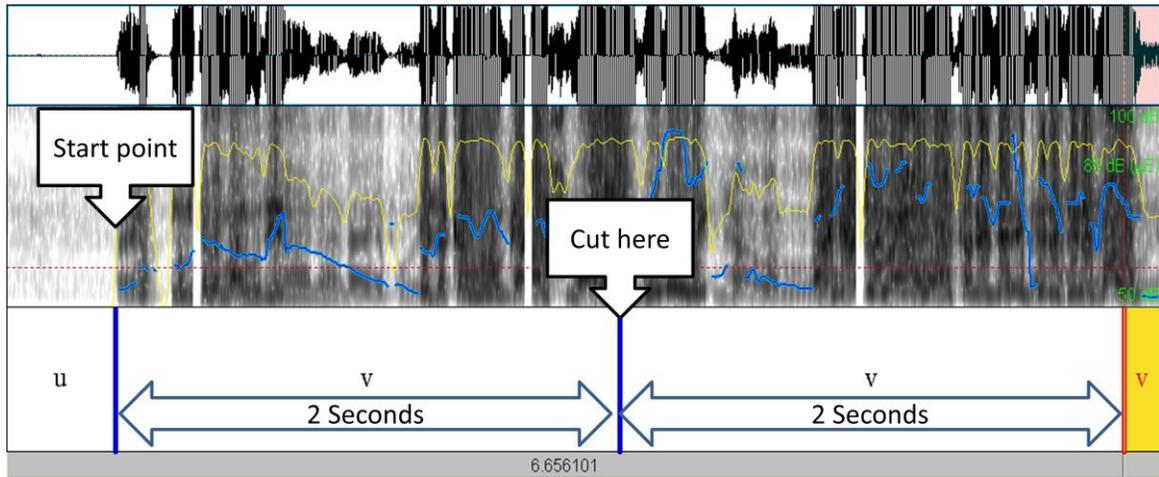


Figure 15: Segmentation for long voiced part

5.2. FEATURE SELECTION

Selected acoustic features from Appendix A will be extracted from each voiced part of the speaker's speech signals that have been previously segmented. Since the objective is to carry out emotion detection in real time, we need to consider a set of features that costs the least in terms of processing time, but at the same time performs reasonably well in the classification process.

We first measured the time required to extract the various features. We did that on a core 2 duo 2.53GHz CPU with 3GB RAM. To have a clear picture of the difference in extraction time, we chose an audio segment of length 24.87 seconds. Table 15 shows the duration in seconds for the extraction of formants, pitch, MFCC, energy, and all features. We found that formants and pitch require the most CPU time, whereas MFCC and energy features require the least.

Table 15 Extraction times for various features of a 24.87 seconds long segment

Feature Calculation Process	Time/seconds
Formants	6.46
Pitch	5.04
MFCCs	1.76
Energy	1.51
All Features Extraction	14.77

This variation in extraction time suggests that it may be necessary to sacrifice some classification accuracy by reducing the number of extracted features in exchange of achieving real-time performance. Therefore, we need to study the performance of various subsets of these features in order to choose the subset with the most suitable (accuracy, extraction time) pair. In order to do that, we applied the SVM, PNN and Decision Tree forest classifiers on all subsets of the feature set containing the four features. Table 16 lists the classification results of each subset, when ran on the context-aware data.

Table 16 Classification of each feature subsets for the three classifiers on context-aware data group

Features Subsets	SVM			PNN			Decision Tree Forest		
	Anger	Neutral	Overall	Anger	Neutral	Overall	Anger	Neutral	Overall
Pitch	60.00	90.80	79.38	61.20	89.39	78.93	63.60	85.38	77.30
Energy	66.00	86.56	78.93	68.00	86.32	79.53	62.80	83.97	76.11
Formants	56.80	88.21	76.56	60.00	86.09	76.41	60.00	87.74	77.45
MFCCs	68.80	87.50	80.56	68.00	87.97	80.56	63.20	87.50	78.49
Pitch – Energy	74.40	90.09	84.27	73.60	88.92	83.23	72.80	89.62	83.38
Pitch – Formants	74.80	88.92	83.68	69.60	90.33	82.64	73.20	90.09	83.83
Pitch – MFCCs	73.60	91.75	85.02	72.80	89.39	83.23	74.00	88.44	83.09
Energy – Formants	69.20	87.74	80.86	76.40	86.09	82.49	75.60	87.50	83.09
Energy – MFCCs	77.20	88.68	84.42	77.20	88.92	84.57	72.80	88.68	82.79
Formants – MFCCs	74.80	89.15	83.83	68.40	87.50	80.42	65.20	89.62	80.56
Pitch - Energy – Formants	81.20	89.62	86.50	79.60	89.86	86.05	80.40	88.92	85.76
Pitch - Energy – MFCCs	80.40	88.92	85.76	80.40	89.15	85.91	78.00	89.86	85.46
Pitch - Formants – MFCCs	76.80	89.15	84.57	77.20	89.86	85.16	73.60	89.62	83.68
Energy - Formants – MFCCs	76.80	87.26	83.38	75.60	87.26	82.94	76.40	87.97	83.68
Pitch - Energy - Formants – MFCCs	82.40	90.09	87.24	79.60	88.92	85.46	80.80	89.62	86.35
All Features	82.40	89.39	86.80	80.40	89.86	86.35	82.00	89.62	86.80

We observe from this table that the highest accuracy rates for detecting anger were achieved using more than two features. In particular, pitch, energy, formants and MFCC features achieved an anger detection accuracy rate of 82.4%. However, feature extraction in these case take considerably more time than using a lesser number of features. Considering feature extraction time as a crucial factor, we can better explain matters using Figure 16, where the accuracy vs. feature extraction time has been represented as a bar graph of Table 16. It is clear that the highest bar at the first quarter of time axis represent the best candidates for a good balance between accuracy and feature extraction time. In particular, Energy and MFCC features using either SVM or PNN classifiers, performed best in detecting anger in the first quarter in 3.27 seconds with 77.2% accuracy. It is worth to mention that in case of a single feature, which has low latency extraction time, MFCCs with PNN as a classifier achieved 68.8% accuracy in 1.76 seconds extraction time.

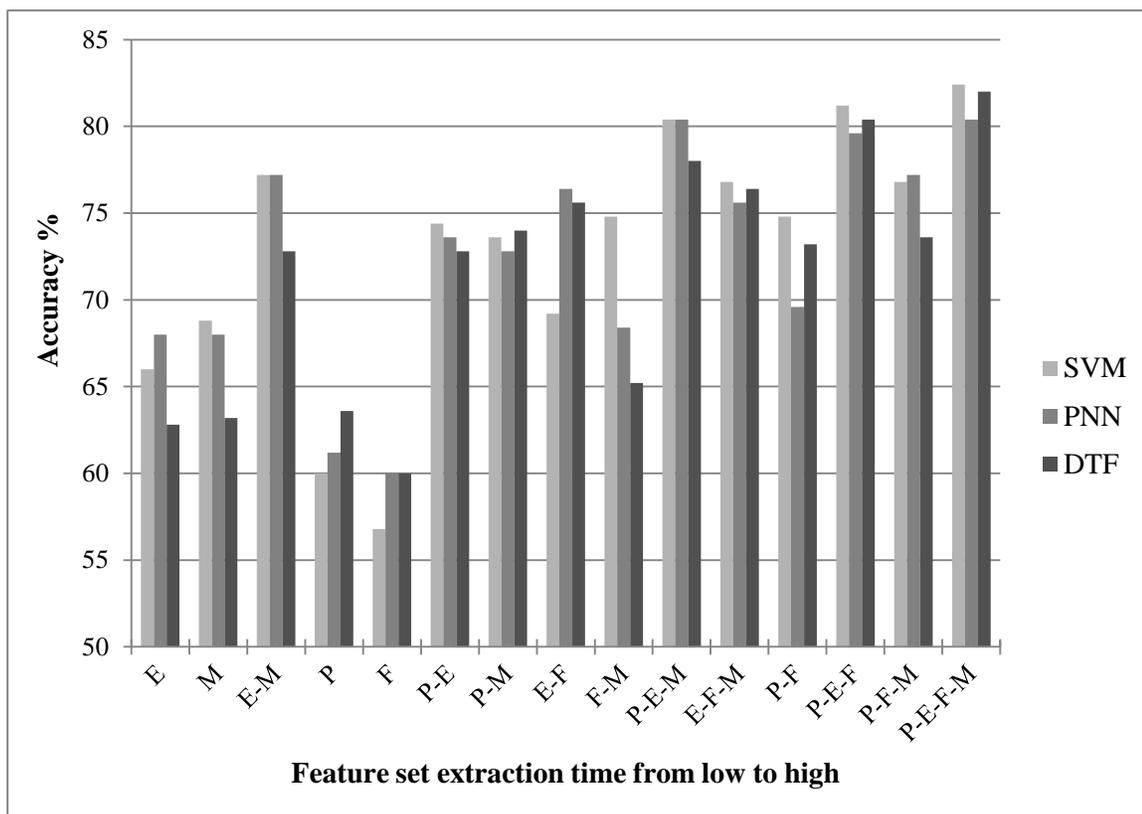


Figure 16: Performance of the three classifiers on various feature sets in terms of accuracy and feature extraction time. E) Energy, F) Formants, M) MFCCs and P) Pitch

CHAPTER 6

ANGER DETECTION PROTOTYPE SYSTEM DEVELOPMENT

We developed a prototype for a real-time anger detection system consisting of four modules: Voice recording module, segmentation and feature extraction module, classification module and output visualization module, as shown in Figure 17. We will elaborate on each of these modules in the subsequent sections

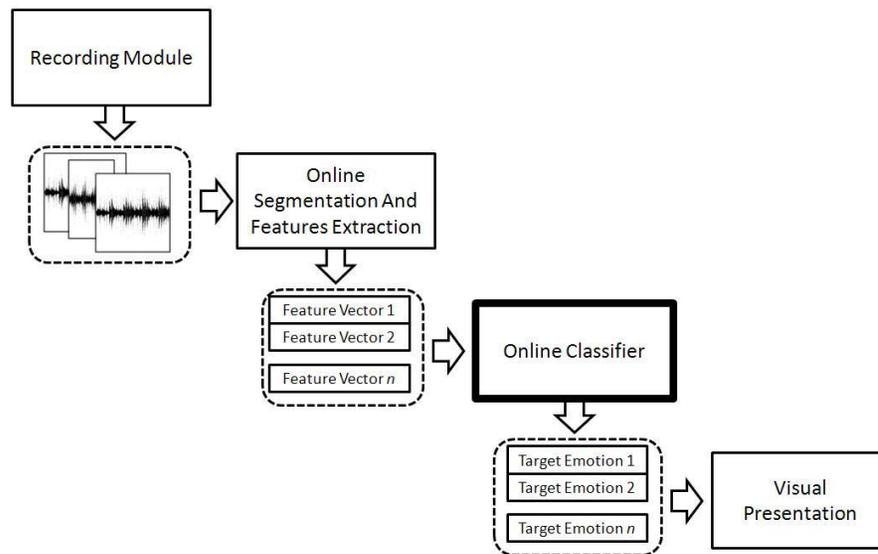


Figure 17: Anger detection prototype system development.

6.1. VOICE RECORDING MODULE

For real-time speech processing, the technique of streaming the sound data into the application should be taken into consideration. In practice, when the magnitude of a speech signal changes due to the magnitude of a speaker's voice, the distance between the speaker and the microphone or the characteristics of the microphone, the recognition performance deteriorates [45] [46].

Besides the quality of the input data, reality of the speech sound is also the main factor for testing online anger detection frameworks. Some researchers test their proposed systems by simulating anger conversations or neutral dialogs themselves or with the help of professional actors. This can be accepted by their systems but it does not reflect real-world data.

Therefore, we use a function of the device sound card used to control audio in MS Windows XP which is called **Sound Mix**³. When enabled (see Figure 18), Stereo Mix loops sound back through the sound card. In other words, it will take anything you hear and send it back to your microphone.

³ **Stereo Mix** can also be called **What U Hear** or **Wave Out** depending upon your sound card

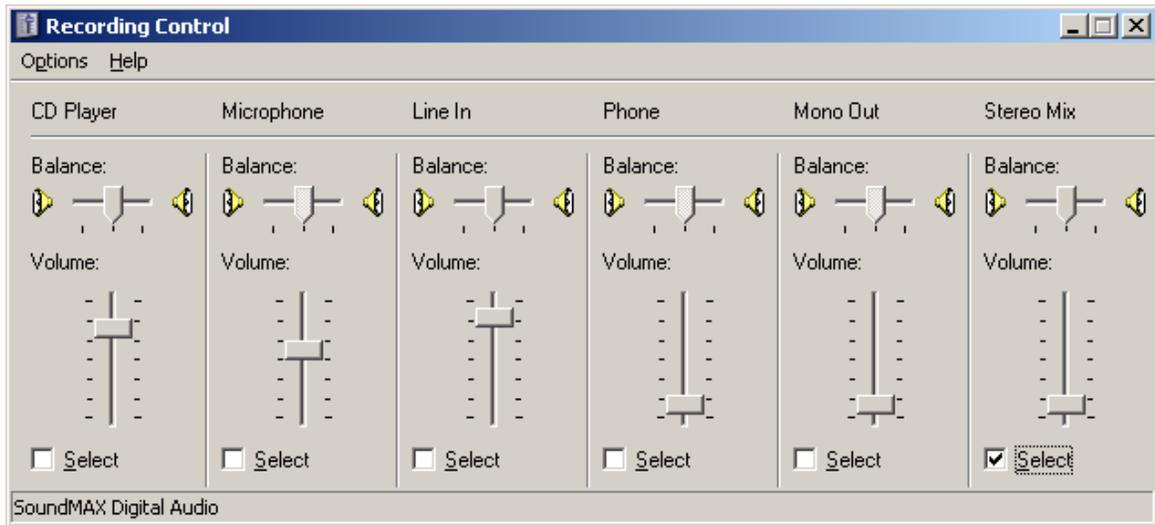


Figure 18: Enabling Sound Mix

This utility is very helpful for our framework. Instead of “1-2-3 hello hello” testing phrases, we can play any real speech sound file in our hard drive or play any clip from the Internet, such as clips from YouTube. This adds more value and credibility to our results and it can be tested while developing the system and before publishing it.

We developed a graphical user interface (GUI) model for recording sound files. Thus, we use `mciSendString` windows API function to send command strings to a sound device, as shown in Figure 19.

Sequence of Open, Record and Save command strings are sent every one second to simulate real-time speech recording. The result of this process is a sound file which grows in size and which is used as input for segmentation and feature extraction.

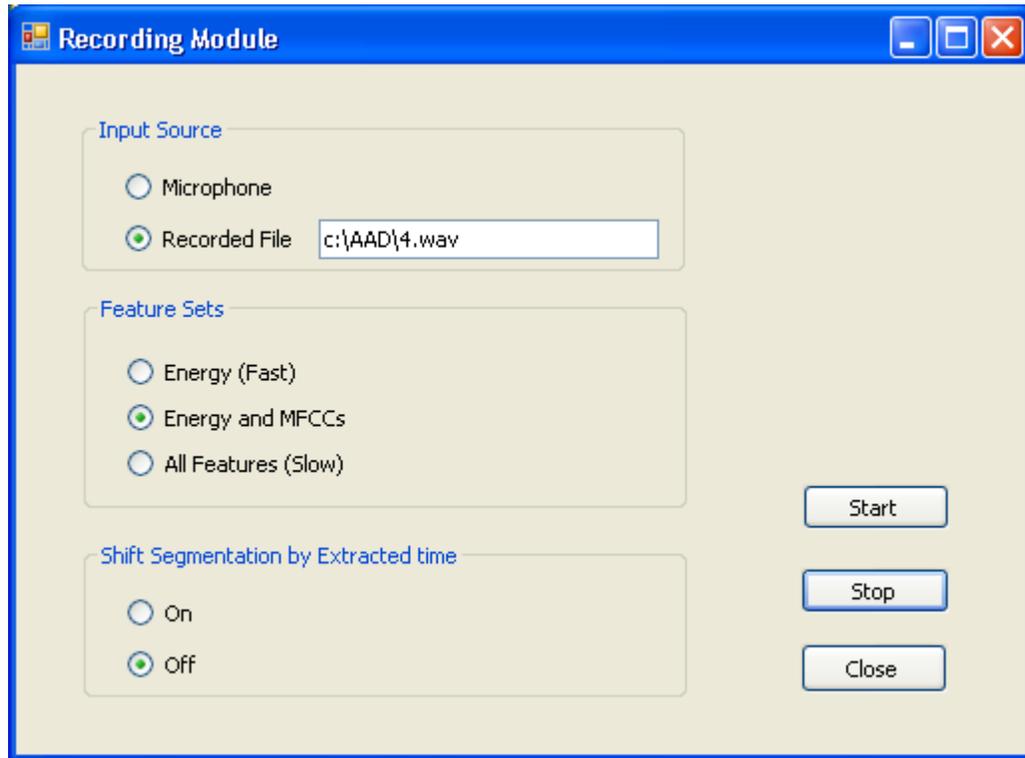


Figure 19: Recording Module GUI

6.2. ONLINE SEGMENTATION MODULE

Speech segmentation of two-seconds maximum are used for each voiced part as discussed in Section 3.1.4.4. Different techniques may be useful to detect voiced parts and unvoiced parts. Some of these techniques use one or more sound features, such as pitch and energy. Using more features leads to more time consumption. So, we developed our segmentation model based on the single feature of energy. When this feature falls below 65 dB within 0.6 second, then we consider the current part as an unvoiced part, otherwise, it will be

treated as a voiced part. A Praat script is developed to extract voiced parts and unvoiced parts as files based on examining the energy feature and time duration.

6.3. ONLINE FEATURE EXTRACTION

In this stage, a developed Praat script is continuously running to look for a new incoming part of the sound file from the segmentation process. When the new part of the sound file is found and is readable, then the script opens that file for feature extraction. Only voiced parts are extracted and unvoiced parts whose energy level falls below the “silence” threshold will be excluded to speed up the real-time detection process.

Acoustic sound features from the set listed in Appendix A are extracted from the voiced parts and compiled in a file of comma-separated vectors (CSV). Each file is time stamped, i.e. is given a sequence number to preserve the time of the feature extraction. The process of extraction ends when the recording program sends an end-of-process file.

6.4. REAL-TIME CLASSIFICATION

After the feature extraction process is complete, the feature vector needs only to be passed through the classification model. The time needed to classify the utterance is negligible and sufficiently fast, given that no modification to the model is required.

DTREG generates C codes of classifier models. The number of rows generated for the Decision Tree Forest classifier is more than 404,000 lines of codes. We rewrote this code

into a visual basic application. The application runs in an infinite loop waiting for the feature vector file to be generated by the feature extraction process in a CSV file format for only voiced parts.

6.5. VISUAL PRESENTATION MODULE

The next step is to visually present the main findings of developed real-time anger detection system, namely the duration (in seconds) of the analysis segment, and a different color reflect the emotion classified with its starting and ending time. The classifier application stops when the recording program sends an end-of-recoding file, as shown in Figure 20.

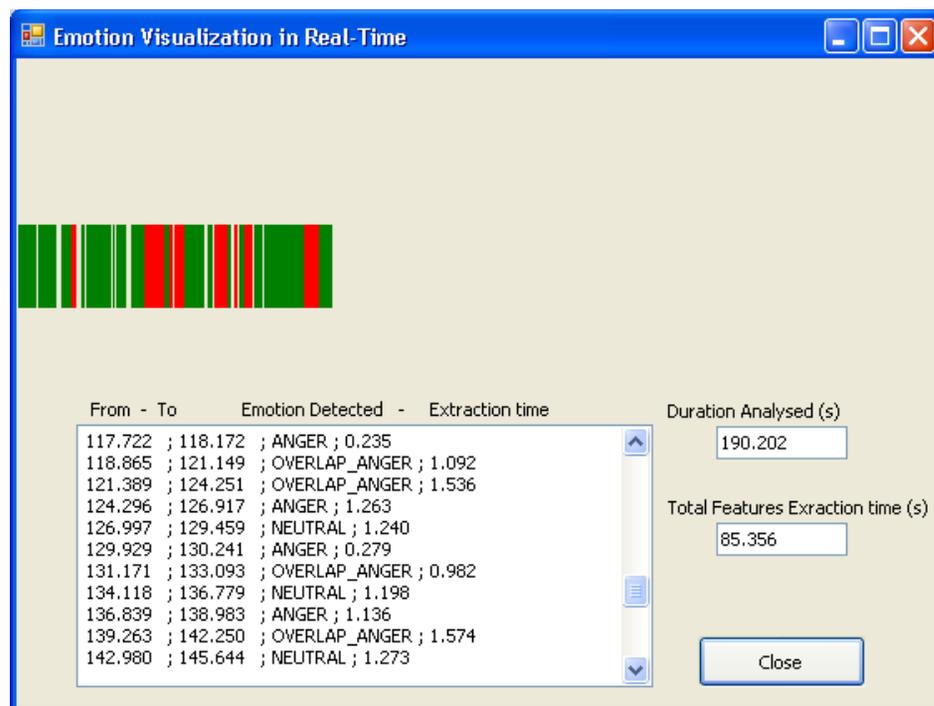


Figure 20: Visual presentation of predicted emotion classes

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this thesis, an automatic framework for real-time anger detection from Arabic dialogs was designed and developed. Various emotion detection techniques were studied, and it was determined that anger emotion state with neutral is a preferred practical class model. There are two reasons for this: first, the lack of an Arabic emotion corpus which limits our choices and second, the importance of anger and neutral emotion to be detected in real-life business applications motivated our choice of the two emotions.

To overcome the first obstacle, we build our own spontaneous anger/neutral corpus obtained from native Arabic speakers, by collecting data from call-center customer services records, and from TV Arabic debate show. These sound files were annotated and segmented in smaller part based on their emotion contents. We found that in real-life recorded speech there are voiced segments other than anger/neutral emotions such as music and noises which should be considered and classified.

Each emotion content of a segment clip in our corpus is presented to group of voters for judgment. This score function shows that initial emotion label which is given by the annotator is enough for classification and is sufficiently reliable. This may be so because the annotator has background knowledge of the emotion context in the dialog of each

segment. Voters could not recognize emotion in a speech segment of duration less than one second.

On the other hand, the key task of the segmentation stage is to choose an appropriate segmentation method which can provide result for emotion recognition that is accurately and efficiently predicted in a short delay time. Our study revealed that in real-life speech, pause detection is not enough for online processing because we found some voiced segments where it take a long time to detect the endpoint pause, and which subsequently delays the system. So we add another approach which force the framework to split long voiced parts to smaller parts, each two-second long, and where no end point pause is to be detected. The two-second was period selected based on a statistical analysis of misclassification compared to time duration.

A set of main acoustic feature were extracted from only voiced segments while silences were not processed to speed up the real-time framework. The long process of feature extraction led us to reduce the number of features to be used. So we removed features which have low importance for the classifier. Unfortunately, the key features for classification are the relatively time consuming ones and so we kept them for better classification with acceptable latency.

Missing values generated in feature extraction process affect the classifiers' performance comparison, because some classifiers accept null values while other classifiers either remove totally vector of features which contain a null value or replace it by medians, which is not realistic.

According to this observation, we handle the missing values problem by either removing low-importance features which may contain missing values, or by replace them by appropriate values which were elicited from a statistical study, such as a value greater than the maximum value for that feature in our dataset. So, filling missing values will be done online after feature extraction and before passing values to the classifier to ensure that the classifier will reliably analyze it.

We observed from our study of the set of classifiers fed with a combination of features that the highest accuracy rates for detecting anger were achieved using more than two features. However, feature extraction in these cases take considerably more time than using a fewer features. So, for a good trade-off between accuracy and feature extraction time, Energy and MFCC features using the SVM classifier, performed best in detecting anger, followed by the Energy feature alone, which takes less time with the Decision Tree Forest as a classifier.

An Application program was developed based on our research findings and experimentation results. A selected classifier model is implemented to work online. The running framework can be tested either by inputting speech to it through a microphone or by feeding it with real-life recorded dialogs.

7.1. MAIN CONTRIBUTIONS

1) Creating an Anger/Neutral emotion corpus from Arabic spontaneous speech, which was totally inexistent and unavailable to researchers interesting in the field of Arabic emotion recognition. The created corpus is well annotated; each sample is labeled by emotion content and type of the speaker according to his or her gender and nationality. This corpus provides real-life sound speech. So, it is useful for other Arabic Language Processing researches and experimentation.

2) Cross-cultural emotion recognition research has taken place in the last few years. These studies either compare results by doing lab experiments by simulating emotions in different languages or by using available corpora. Another type of comparison may involve going through the available literature to compare different cultures' important sound features such as speaking rates and pause durations. No research studies on Arabic emotion recognition are available to the best of our knowledge. So, publishing such study on Arabic cultural emotion will add value to, and stimulate the current academic research in the field of cross-cultural emotion recognition.

3) Designing a real-time anger detector product framework, which could be applied in local area business sectors, such as call-centers?

7.2. FUTURE WORK

1) There was a trial in our experiments to detect overlap speech in a conversation, which highly indicates a problem or sometime angry talk. This could be done by defining the main features for this type of speech. But the low number of collected samples of OVERLAP_ANGER class makes us to postpone it to later time.

2) Building a more comprehensive corpus from data centers is a challenge (many refuse to cooperate)

3) Investigating speech-word spotting in speech to detect anger more accurately than prosodic features and compare the results

REFERENCES

- [1] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293- 303, 2005.
- [2] M. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages", *Journal of Phonetics*, vol. 37, pp. 417-435, 2009.
- [3] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies", *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119-131, 2010.
- [4] T. Vogt, "Real-time automatic emotion recognition from speech", Ph.D. Dissertation, Faculty of Engineering, University of Bielefeld, 2010
- [5] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan and I. Media, "Emotion recognition based on phoneme classes", in *International Conference on Spoken Language Processing (ICSLP 2004)*, Jeju Island, 2004.
- [6] S. Koolagudi, S. Nandy, and K. Rao, "Spectral features for emotion classification", in *IEEE International Advance Computing Conference, IACC 09*, 2009.
- [7] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge", in *10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009*, 2009.
- [8] T. Stroemhaug, "Detection of emotional speech", N.Y., 2007.
- [9] D. Gharavian and S. Ahadi, "Recognition of emotional speech and speech emotion in Farsi", in *Proceedings of International Symposium on Chinese Spoken Language Processing*, 2006.
- [10] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", *IEEE Transactions on Audio, Speech,*

and Language Processing, vol. 17, no. 4, pp. 582-596, 2009.

- [11] M. Pell, S. Kotz, L. Monetta, and S. Paulmann, "Recognizing emotions in a foreign language", *Journal of Nonverbal Behavior*, pp. 107-120, 2009.
- [12] M. Pell and V. Skorup, "Implicit processing of emotional prosody in a foreign versus native language", *Speech Communication*, vol. 50, pp. 519-530, 2008.
- [13] C. Darwin, "The expression of the emotions in man and animals", *John Murray*, London, 1872.
- [14] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124-129, 1971.
- [15] R. Stibbard, "Automated extraction of ToBI annotation data from the reading/leads emotional speech corpus", in *Proceedings of the ISCA Workshop on Speech and Emotion*, 2000.
- [16] V. Petrushin, "Emotion in speech: recognition and application to call centers", in *Artificial Neural Networks in Engineering (ANNIE '99)*, 1999.
- [17] L. T. Bosch, "Emotions: what is possible in the ASR framework", in *International Speech Communication Association Tutorial and Research Workshop on Speech and Emotion (ISCA)*, 2000.
- [18] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech", in *International Conf. on Spoken Language Processing (ICSLP '96)*, 1996.
- [19] F. Archetti, G. Arosio, E. Fersini, and E. Messina, "Audio-based emotion recognition for advanced automatic retrieval in judicial domain", in *1st Int. Conf. on ICT Solutions for Justice*, Greece, 2007.
- [20] D. Morrison, R. Wang, and L. De Silva, "Ensemble methods for spoken emotion recognition in call-centres", *Speech Communication*, vol. 49, pp. 98-112, 2007.
- [21] F. Zotter, "Emotional speech", Technical Report, Signal Processing and Speech Communication Laboratory (SPSC Lab), Graz University of Technology, 2003. Available at

http://www2.spsc.tugraz.at/www-archive/AdvancedSignalProcessing/SpeechSynthesis.new/zotter_report.pdf

- [22] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME 2005)*, 2005.
- [23] R. Tato, R. Santos, R. Kompe, and J. Pardo, "Emotional space improves emotion recognition", in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver-Colorado, USA, 2002.
- [24] M. Potegal, G. Stemmler, and C. Spielberger, *"International handbook of anger"*: Springer, 2010.
- [25] C. Sobin and M. Alpert, "Emotion in speech: the acoustic attributes of fear, anger, sadness, and joy", *Journal of Psycholinguistic Research*, vol. 28, no. 4, pp. 347–365, 1999.
- [26] D. Morrison, R. Wang, L. De Silva, and W. Xu, "Real-time spoken affect classification and its application in call-centres", in *Proceedings 3rd Int. Conf. Information Technology and Applications (ICITA)*, 2005.
- [27] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs", in *Proceedings of INTERSPEECH'2006*, 2006.
- [28] R. Shah and M. Hewlett, "Emotion detection from speech", 2007. Class Project, Stanford University. Available at:
<http://www.stanford.edu/class/cs229/proj2007/ShahHewlett%20-%20Emotion%20Detection%20from%20Speech.pdf>
- [29] I. Chiriacescu, "Automatic emotion analysis based on speech", MSc Thesis, Delft University of Technology, 2009.
- [30] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech a review of the literature and recommendations for practical realisation", in *Affect and Emotion in Human-Computer Interaction*, 2008.
- [31] E. Vyzas and R. Picard, "Offline and online recognition of emotion expression from

- physiological data", in *Workshop on Emotion-Based Agent Architectures, 3rd International Conference on Autonomous Agents*, Seattle, WA, 1999.
- [32] R. el Kaliouby, R. Picard, and S. Baron-Cohen, "Affective computing and autism", *Annals of the New York Academy of Sciences*. vol. 1093, pp. 228-248, 2006.
- [33] C. Chastagnol and L. Devillers, "Analysis of anger across several agent-customer interactions in French call centers", in *Proceedings ICASSP*, 2011.
- [34] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot", in *Artificial Intelligence and Computational Intelligence and Computational Intelligence (AICI)*, 2010.
- [35] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mitsuyoshi and M. Shimura, "Usage of emotion recognition in military health care", in *Defense Science Research Conference and Expo (DSR)*, 2011.
- [36] J. Rong, G. Li, and Y. Chen, "Acoustic feature selection for automatic emotion recognition from speech", *Information Processing and Management*, vol. 45, no. 3, pp. 315–328, 2009.
- [37] J. Clark, C. Yallop, and J. Fletcher, *An introduction to phonetics and phonology*, 3rd ed. Malden, MA, USA: Blackwell Publishers, 2007.
- [38] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental", in *Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence*, Hyannis, MA, USA, 1976.
- [39] G. Saha and U. Yadhunandan, "Modified mel-frequency cepstral coefficient", in *Signal and Image Processing*, 2004.
- [40] DTREG, Software For Predictive Modeling and Forecasting. [Online]. <http://www.dtre.com>
- [41] D. Specht, "Probabilistic Neural Networks", *Neural Networks*, vol. 3, pp. 109-118, 1990.
- [42] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

- [43] Praat: doing phonetics by computer. [Online]. <http://www.praat.org>
- [44] O. Khan, "Detection of questions in arabic audio monologues using prosodic features", MSc Thesis, King Fahd University of Petroleum & Minerals, 2007.
- [45] B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl, "Towards more reality in the recognition of emotional speech", in *Proceedings of ICASSP 2007*, Honolulu, 2007, pp. 941-944.
- [46] E. Kim, "Robust emotion recognition feature, frequency range of meaningful signal", in *IEEE Workshop on Robot and Human Interactive Communication (ROMAN)*, 2005, pp. 667-671.

APPENDIX A

LIST OF SELECTED FEATURES

The feature script was generated according to the following principles:

- Start extracting features corresponding to the acoustic parameters that are known to be most important for speech emotion recognition studies and systems (i. e. pitch, intensity, etc.);
- Even if some feature look irrelevant, still add it to the script, the selected classifier will decide whether to keep it or not based on its importance to classification;
- Real-time processing forces us to select features which can be extracted in a reasonable time.

Features 1 - 9: Pitch

A Pitch object represents periodicity candidates as a function of time. It does not mind whether this periodicity refers to acoustics, perception, or vocal-cord vibration. It is sampled into a number of frames centered around equally spaced times.

Raw pitch values are calculated by Praat's autocorrelation method which analyzes the non-speech signals, which may require different standard settings of the parameters. These values are then smoothed and voiced/unvoiced regions determined and stored in TextGrid. Praat's pitch stylization function is used to stylize raw pitch values over

The algorithm performs acoustic periodicity detection on the basis of an accurate autocorrelation method, as described in Boersma (1993). This method is more accurate, noise-resistant, and robust, than methods based on cepstrum or combs, or the original autocorrelation methods.

Feature 1: Median Pitch

Feature 2: Mean Pitch

Feature 3: Standard deviation Pitch

Feature 4: Minimum Pitch

Feature 5: Maximum Pitch

Feature 6: Mean Stylized Pitch

Feature 7: Minimum Stylized Pitch

Feature 8: Maximum Stylized Pitch

Feature 9: Mean Absolute Pitch Slope

Features 10 - 17 : Intensity/Energy

An Intensity object represents an intensity contour at linearly-spaced time points, with values in dB SPL, i.e. dB relative to 2×10^{-5} Pascal, which is the normative auditory threshold for a 1000-Hz sine wave.

Since there is no intensity stylization function in Praat, intensity values are stylized by pitch stylization function. Stylization is performed over entire intensity contour, whereas in pitch case, this stylization only applies in voiced regions.

Feature 10: Median Intensity

Feature 11: Mean Intensity

Feature 12: Standard deviation Intensity

Feature 13: Minimum Intensity

Feature 14: Maximum Intensity

Feature 15: Mean Stylized Intensity

Feature 16: Minimum Stylized Intensity

Feature 17: Maximum Stylized Intensity

Features 18 - 32: Formants

A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several formants, each at a different frequency, roughly one in each 1000Hz band. Or, to put it differently, formants occur at roughly 1000Hz intervals. Each formant corresponds to a resonance in the vocal tract. We limit our study on the first three formants F1, F2 and F3:

Feature 18: Median F1

Feature 19: Mean F1

Feature 20: Standard deviation F1

Feature 21: Minimum F1

Feature 22: Maximum F1

Feature 23: Median F2

Feature 24: Mean F2

Feature 25: Standard deviation F2

Feature 26: Minimum F2

Feature 27: Maximum F2

Feature 28: Median F3

Feature 29: Mean F3

Feature 30: Standard deviation F3

Feature 31: Minimum F3

Feature 32: Maximum F3

Features 32 - 35: Pulses

Representing the glottal closures of the wave.

Feature 32: Number of pulses

Feature 33: Number of periods

Feature 34: Mean period

Feature 35: Standard deviation of period

Features 36 - 38: Voicing

Feature 36: Fraction of locally unvoiced frames

Feature 37: Number of voice breaks

Feature 38: Degree of voice breaks

Features 39 - 43: Jitter

Jitter is the deviation, in or displacement of, some aspect of the pulses in a high-frequency.

Feature 39: Jitter (local); the local jitter is defined as the relative mean absolute second-order difference of the point process (= the first-order difference of the interval process).

Feature 40: Jitter (local, absolute); the absolute local jitter (in seconds) is the mean absolute (non-negative) difference of consecutive intervals.

Feature 41: Jitter *Relative Average Perturbation* (rap); the average absolute difference between a period and the average of it and its two neighbors, divided by the average period.

Feature 42: Jitter *five-point Period Perturbation Quotient* (ppq5); the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period.

Feature 43: Jitter *Difference of Differences of Periods* (ddp); the relative mean absolute (i.e. non-negative) third-order difference of the point process

Features 44 – 49 :Shimmer

Shimmer is similar to jitter but based on the energy contour rather than the pitch contour.

Feature 44: Shimmer (local); the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

Feature 45: Shimmer (local, dB); the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.

Feature 46: Shimmer (apq3); the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude.

Feature 47: Shimmer (apq5); the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbors, divided by the average amplitude.

Feature 48: Shimmer (apq11); the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbors, divided by the average amplitude.

Feature 49: Shimmer (dda); the average absolute difference between consecutive differences between the amplitudes of consecutive periods.

Features 49 - 51: Harmonicity:

A harmonicity object represents the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR). Harmonicity is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is $10 \cdot \log_{10}(99/1) = 20$ dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise. Harmonicity can be

used as a measure for: The signal-to-noise ratio of anything that generates a periodic signal.

Feature 49: Mean autocorrelation: an acoustic periodicity detection on the basis of an accurate autocorrelation method.

Feature 50: Mean noise-to-harmonics ratio.

Feature 51: Mean harmonics-to-noise ratio.

Features 52 - 63: Mel Frequency Cepstral Coefficients:

Twelve coefficients are represented in frames with constant sampling period.

Feature 52: MFCC 1

Feature 53: MFCC 2

Feature 54: MFCC 3

Feature 55: MFCC 4

Feature 56: MFCC 5

Feature 57: MFCC 6

Feature 58: MFCC 7

Feature 59: MFCC 8

Feature 60: MFCC 9

Feature 61: MFCC 10

Feature 62: MFCC 11

Feature 63: MFCC 12

Feature 64: Speaking Rate

Speaking rate is defined as the number of spoken units (e.g., words/syllables) per unit of time (minute/second). It is calculated across continuous speech segments, which may include pauses, disruptions or dysfluency.

VITAE

PERSONAL INFORMATION		
Nationality	Palestinian	
Marital status	Married	
Date and Place of Birth	May 12, 1977 - Al-Ahsa, Saudi Arabia	
Language	Arabic (Native language) English (TOEFL CPT - Score 207/300)	
E-mail	ashrafmis@hotmail.com ashrafkh@kfu.edu.sa	
EDUCATION		
2012	King Fahd University of Petroleum and Minerals	Dhahran – Saudi Arabia
	Master of Science in Computer Science	GPA 3.06 out of 4
1998	Applied Science University	Amman – Jordan
	B.Sc. Degree in Computer Science	87.8% Excellent
EMPLOYMENT		
2003 - Present	Oracle Developer And Systems Analyst	
	Deanship of Information Technology – King Faisal University – AlAhsa	
1999 - 2003	IT Manager	
	Al-Moosa General Hospital – AlAhsa	
AWARDS RECEIVED		
	Jeraisy Computer & Comm. Services Company	
	Best employee in application systems department	
	Al-Moosa General Hospital	
	Best employee in IT department	
	Applied Science University	
	On the students honors list of the university	