ANALYZING ACOUSTIC MARKERS OF EMOTION IN ARABIC SPEECH

A DISSERTATION IS SUBMITTED TO THE UNIVERSITY OF MANCHESTER FOR THE DEGREE OF THE MASTER OF PHILOSOPHY (MPHIL) IN THE FACULTY OF SCIENCE AND ENGINEERING

2017

By

Muna Bin Othman

School of Computer Science

Supervisor: Prof. Allan Ramsay
<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract .......................................................... 9</td>
</tr>
<tr>
<td>Declaration ........................................................... 10</td>
</tr>
<tr>
<td>Copyright .............................................................. 11</td>
</tr>
<tr>
<td>Dedication ............................................................. 12</td>
</tr>
<tr>
<td>Acknowledgement ......................................................... 13</td>
</tr>
<tr>
<td>1. CHAPTER ONE : INTRODUCTION ...................................... 14</td>
</tr>
<tr>
<td>1.1 Introduction ....................................................... 14</td>
</tr>
<tr>
<td>1.2 The area of problem ................................................. 15</td>
</tr>
<tr>
<td>1.3 Research Aim ....................................................... 16</td>
</tr>
<tr>
<td>1.4 Research Tasks ..................................................... 16</td>
</tr>
<tr>
<td>1.5 Thesis Structure .................................................... 16</td>
</tr>
<tr>
<td>2. CHAPTER TWO : EMOTIONS AND SPEECH RECOGNITION ............ 18</td>
</tr>
<tr>
<td>2.1 Introduction ....................................................... 18</td>
</tr>
<tr>
<td>2.2 Speech ............................................................... 18</td>
</tr>
<tr>
<td>2.2.1 What is speech? .................................................. 18</td>
</tr>
<tr>
<td>2.2.2 Human Speech production ....................................... 20</td>
</tr>
<tr>
<td>2.3 Speech and emotion ................................................ 24</td>
</tr>
<tr>
<td>2.3.1 What are emotions ? ............................................. 24</td>
</tr>
<tr>
<td>2.3.2 Categories of emotion ........................................... 25</td>
</tr>
<tr>
<td>2.3.3 Emotion in natural speech ...................................... 28</td>
</tr>
<tr>
<td>2.3.4 The acoustic correlates of emotions in human speech ......... 29</td>
</tr>
<tr>
<td>Section</td>
</tr>
<tr>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>2.3.5 Cultural aspects of emotional speech</td>
</tr>
<tr>
<td>2.3.6 Applications for detecting emotion in human speech</td>
</tr>
<tr>
<td>2.4 Summary</td>
</tr>
<tr>
<td>3. CHAPTER THREE: ISSUES FOR ARABIC SPEECH PROCESSING</td>
</tr>
<tr>
<td>3.1 Introduction</td>
</tr>
<tr>
<td>3.2 Mapping written forms to phonetic sequences</td>
</tr>
<tr>
<td>3.2.1 Writing system, Semi-vowels, Diacritics</td>
</tr>
<tr>
<td>3.2.2 Arabic phonetic system (Arabic phonology)</td>
</tr>
<tr>
<td>3.2.3 Grapheme-to-phoneme</td>
</tr>
<tr>
<td>3.3 Prosodic Features</td>
</tr>
<tr>
<td>3.3.1 Intensity</td>
</tr>
<tr>
<td>3.3.2 Duration</td>
</tr>
<tr>
<td>3.3.3 Intonation</td>
</tr>
<tr>
<td>3.4 Dialects of Arabic</td>
</tr>
<tr>
<td>3.5 Summary</td>
</tr>
<tr>
<td>4. CHAPTER FOUR: CONSTRUCTION AND VALIDATION OF THE CORPUS</td>
</tr>
<tr>
<td>4.1 Introduction</td>
</tr>
<tr>
<td>4.2 Corpus design</td>
</tr>
<tr>
<td>4.2.1 Selection of actor</td>
</tr>
<tr>
<td>4.2.2 Selection of emotions</td>
</tr>
<tr>
<td>4.2.3 Design of material texts</td>
</tr>
<tr>
<td>4.2.4 Database collection</td>
</tr>
<tr>
<td>4.2.4.1 Recording process of data</td>
</tr>
<tr>
<td>4.2.4.2 Labeling the data</td>
</tr>
<tr>
<td>4.2.4.3 Evaluation of the data</td>
</tr>
</tbody>
</table>
List of Tables
Table 2.1: Emotion and speech parameters (Murray & Arnott, 1993) ........................................31
Table 3.1: Different shapes of Arabic letters with SAMPA and IPA symbols ................................39
Table 3.2: List of Arabic consonants with IPA symbols ..........................................................43
Table 3.3: Arabic vowel .............................................................................................................44
Table 3.4: Arabic diphthongs in different positions (Khalifa, Obaid, Naji, & Daoud, 2011) ......45
Table 4.1: Selected sentences ..................................................................................................64
Table 4.2: Labeling for emotions .............................................................................................68
Table 4.3: An example of listener response .............................................................................70
Table 4.4: Confusion matrix for the listeners’ responses ..........................................................71
Table 4.5: Confusion matrix for the listeners responses without neutral emotion ....................72
Table 4.6: Evaluation of happy emotion by listeners .................................................................73
Table 4.7: Evaluation of neutral emotion by listeners ...............................................................74
Table 4.8: Evaluation of sad emotion by listeners .....................................................................75
Table 4.9: Evaluation of anger emotion by listeners .................................................................76
Table 4.10: The expressive data recognised by subjective human evaluation .........................78
Table 4.11: The number of final expressive selected sentences ...............................................79
Table 6.1: dataset attribute information ..................................................................................108
Table 6.2: The expressive utterances used of our dataset ........................................................108
Table 6.3: The features are used in each experiment with different learning algorithms ..........117
Table 6.4: Simulation Result of six chosen Algorithms ............................................................122
Table 6.5: Error Rate Evaluation Parameters for classifiers .....................................................124
Table 6.6: Details of five datasets ............................................................................................126
Table 6.7: The features are used in each experiment with DecisionTable algorithms ..............126
Table 6.8: The result of using two features with DecisionTable algorithm ................................128
Table 6.9: The result of using one feature with DecisionTable algorithm ................................128
Table 6.10: Confusion matrix for DecisionTable obtained with 10 fold cross-validation with using all features ..............................................................................................................130
Table 6.11: The features are used in each experiment without Neutral Emotion (E00) ............132
Table 6.12: Confusion matrix of all utterances ..........................................................................135
Table 6.13: The result of six experiments with missing emotions ............................................136
Table 6.14: Confusion matrix of all utterances after adding the missing emotions ..................137
Table 6.15: The result of six experiments with adding of missing emotions .......................137
List of Figures

Figure 2.1: Waveform of sound ................................................................. 19
Figure 2.2: Speech organs ................................................................. 21
Figure 2.3: IPA vowel chart ................................................................. 22
Figure 2.4: A diagram of the human ear (Bosi & Goldberg, 2012) ............... 23
Figure 2.5: Plutchik’s Wheel of Emotions ............................................. 25
Figure 2.6: Circumplex Model of Emotions (Posner, Russell, & Peterson, 2005) .... 27
Figure 2.7: PANA Model (Watson & Tellegen, 1985) ................................ 27
Figure 3.1: Prosodic dependencies ....................................................... 48
Figure 4.1: Phonemes frequency .......................................................... 65
Figure 4.2: Sound recorder of Praat program ............................................ 67
Figure 5.1: The process of extracting the features .................................... 80
Figure 5.2: Two major processing stages in HTK toolkit (Sjölander, 2009) .......... 82
Figure 5.3: Data preparation of our Data ............................................... 84
Figure 5.4: Pronunciation dictionary file ............................................. 89
Figure 5.5: The task grammar designed for the first sentence ..................... 90
Figure 5.6: Wav_config file .................................................................. 92
Figure 5.7: Forced alignment (Young et al., 2002) .................................... 94
Figure 5.8: The files used in recognising the test data ............................... 94
Figure 5.9: Recognition output (recognized transcription_recout.mlf) .......... 96
Figure 5.10: A TextGrid window in Praat software .................................. 97
Figure 5.11: SCP_files ..................................................................... 98
Figure 5.12: TextGrid file of sample1.wav ........................................... 99
Figure 5.13: The extracted information ............................................. 100
Figure 5.14: CSV file with suitable names for recorded data .................... 101
Figure 5.15: First lines of the features database from CSV file ................. 102
Figure 6.1: The framework for the classification of emotional speech ........... 104
Figure 6.2: ARFF file of our dataset ................................................... 107
Figure 6.3: Decision tree algorithm ................................................... 112
Figure 6.4: Multilayer Perceptron ................................................... 115
Figure 6.5: Experiment Environment window .................................................................119
Figure 6.6: Preprocessing of E00 ..................................................................................131
Abstract

ANALYZING ACOUSTIC MARKERS OF EMOTION IN ARABIC SPEECH

Muna Bin Othman

A dissertation submitted to the University of Manchester for the degree of Master of Philosophy (MPHIL), 2017

This study aims to obtain detailed acoustic knowledge of how speech is modulated when a speaker’s emotion changes from neutral to certain emotional states based on measurements of acoustic parameters related to speech prosody.

This can be used effectively in many applications of synthesis or recognition systems of emotions in Arabic speech.

The common problems often faced by studying emotions in Arabic speech are explored, including the complexity of the phonetic rules and diacritic system of Arabic which makes Arabic speech harder to process than for most other languages, and a lack of freely available emotional corpora for Arabic speech.

The acoustic features of pitch, intensity and duration are extracted from a small corpus and then used to classify the four emotions: neutral, happy, sad and anger in Arabic speech.

A range of experiments are conducted in order to identify and investigate the key features of emotional speech in Arabic and to determine which one has the most effect on the emotional speech and to use them to classify the emotions. The results of the experiments are analyzed. The main findings are that using the combination of these features enhanced the performance and Anger was the most challenging to identify while other emotions were classified in different range of accuracy. Finally, several suggestions for future experiments in this area are discussed.
Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii.Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.
Dedication

To my mother Fatima.
Acknowledgement

All praise is due to Allah, who guided me to this. I would like to express my sincere gratitude to my supervisor; Prof. Allan Ramsay. I’m greatly indebted to his assistance, guidance and support.

I am very grateful for my mother, husband and my friends whom I consider as my sisters. Thank you for believing in me and supporting me.

Finally, I hope this thesis be a useful addition to the research activities of Arabic natural language processing

MUNA
CHAPTER ONE: INTRODUCTION

1.1 Introduction

Speech is the most important of the different forms of communication used in daily life. It occurs via the variations in air pressure that are caused by the articulatory system in the production of the human voice (Dutoit, 1997).

Speech also conveys certain information to listeners about the language being spoken, and about the gender and emotional state of speakers. This is achieved by the use of natural acoustic effects of the voices, which are very important in representing a speaker’s feelings and emotional state. In fact, people can convey and recognise emotions easily, even in a language which they do not understand, and adding the ability express emotion in synthetic speech would substantially improve its quality. This makes the emotions a part of speech and one of the core components in human communication.

Understanding how emotions are expressed in speech and embedding them as a component of speech processing can be useful for speech technology tools that perform specific tasks by making them more natural and effective.

In the field of speech synthesis, it is important to introduce emotional characteristics into synthetic speech in order to provide more freedom of expression from a machines to speak like humans, which is important to a growing variety of cases of the use of speech synthesis. There are many studies that have contributed towards the creation of expressive speech synthesis aimed at achieving a more natural result during human-computer interaction (Cahn, 1989),(Murray & Arnott, 1993).

Several target applications can be found, such as emotional greetings and reading applications for blind people. Speech recognition is another area which has received much attention from researchers attempting to enable computers to automatically recognise the emotional or physical state of humans from their speech.
It is particularly useful for enhancing the naturalness of speech-based human machine interaction (Schuller, Rigoll, & Lang, 2004), (Dellaert, Polzin, & Waibel, 1996), (Koolagudi, Maity, Vuppala, Chakrabarti, & Rao, 2009).

Applications of emotion recognition systems include tasks such as emotion/problem recognition in automatic dialog systems, emotion recognition for psychologic analysis and lie detection in forensic investigations (Zotter, 2003). In addition, information from customers’ behaviour in call centre conversations could be analysed (Lee & Narayanan, 2005) and information about the mental states of car drivers could be used by onboard driving systems to keep them alert while driving (Schuller et al., 2004).

There seem to be a number of parameters, such as pitch, speech rate etc., which can be examined and measured when looking for properties of speech which correlate with multiple emotions(Murray & Arnott, 1993),(Scherer, 2003), (Shami & Verhelst, 2007).

The aim of this research is to investigate how the emotional states affect Arabic speech and to extract the key features of selected emotional states from samples of recorded human speech.

1.2 The area of problem

Although the Arabic language is spoken throughout the Arab world and is known by millions of people as an official language in 24 different countries, and has a religious value for more than 1.6 billion Muslims worldwide, it has gained less attention in the field of speech systems in general, and of emotional speech corpora in particular, compared with other languages. It is noted that this may be due to limitations in natural language processing (NLP) in Arabic (El-Imam, 1989). In recent years there have been advances in Natural Language Processing (NLP) and Digital Signal Processing (DSP) which improved the quality and naturalness of available Arabic voices. All of that was in the area of processing of speech. Although, there has been a great increase in the demand for speech systems, and adding emotions to these systems and their applications has become one of the major trends in this field, to our knowledge there has been very little work on developing emotionally expressive voices for the language.
1.3 Research Aim

The main aim of this research is to find the acoustic correlates of emotion in spoken Arabic. This is important because people express their attitudes to what they are saying by changes in the characteristics of the sounds they are producing, and because if synthetic speech is generated without understanding how it expresses emotion then it may well express unintended emotions.

1.4 Research Tasks

In order to achieve our aim, there are a number of tasks that have to be carried out. These tasks can be related to the research aim as below:

*RT*: In order to find out what the acoustic correlates of emotions are, we have to do two things:

*RTA*, we have to start by gaining a good understanding of what emotions are by examining the psychological literature, previous studies, research papers, and library and internet searches on theories of emotion to illustrate how researchers have responded to what is assumed about these emotions. These studies also include the conclusions made by anthropologists, particularly from observations of cultural-specific emotions.

*RTB*, we need to obtain a database of Arabic speech which expresses different emotions (anger, sadness, happiness and neutral), marking them up for emotion, and then extracting from them the acoustic features, in an attempt to find, either by inspection or by machine learning, what the significant features are. This material will be discussed in chapter 5.

Existing speech corpora for Arabic do not contain a lot of emotional materials, so cannot easily be used. Due to the lack of Arabic speech corpora marked up for emotion, a decision was made to record a number of Arabic sentences spoken by a trained actor aiming to express different emotional states (anger, sadness, happiness and neutral state).

1.5 Thesis Structure.

The rest of the thesis will be organised as follows:

- **Chapter two** begins by firstly giving background information about the field of speech processing. The chapter starts with the basic definition of speech terms and the
mechanism of human speech production before taking a look into the process of speech processing in general with its applications in speech synthesis and speech recognition. 

**Secondly**, this chapter also explains the concept of emotions from different points of view and how they are expressed in speech, and discusses the existing emotional speech of some applications. **Lastly**, it discusses the cultural aspects of emotional speech and gives examples from corpora of emotional speech.

- **Chapter three** explains the relevant properties of the Arabic language. This chapter is divided into five sections. **Section 1** introduces the chapter. **Section 2**, *Mapping written forms to phonetic sequences*, introduces some characteristics of Arabic, such as the writing system, diacritics, semi-vowels, and the Arabic phonetic system (Arabic phonology). **Section 3** presents prosodic features (Intensity, duration and Intonation) of Arabic speech. **Section 4** discusses one of the challenges for Arabic text to speech, which is dialects of Arabic. The chapter ends with a summary in **Section 5**.

- The aim of **Chapter four** is to report on the design and validation of an Arabic emotional speech corpus. A brief introduction is presented in **Section 1**. **Section 2** discusses the different ways of creating a corpus of emotional speech and gives some reasons why the data were recorded by an actor, rather than finding a database of recordings of emotional speech. This chapter also describes the problems that could arise from making our own recordings and explains these risks. Explanations are given of which target emotions were selected, and how the actor and materials for recording were selected. The processing of the data for recording, labeling and evaluation by a number of native Arabic speakers are also presented. **Section 3** presents the summary of this chapter.

- **Chapter five**: *extracting and analysing the acoustic correlate features*. This chapter gives an overview of these prosodic features and explains how they were extracted.

- **Chapter six**: This chapter presents the set of experiments, along with their results, which were conducted by applying machine learning algorithms to find out which are the most important features for identifying emotions and also to discover which utterances and phonemes are more difficult to classify.

- **Chapter seven**: *conclusion and future work.*
CHAPTER TWO: EMOTIONS AND SPEECH RECOGNITION

2.1 Introduction

The aim of this work is to find the acoustic correlates of emotions in spoken Arabic as a basis for identifying the emotional content of Arabic speech. Therefore, it is worthwhile to go into a short overview to see how emotions are expressed in normal human speech. In fact, to understand how emotions are expressed in speech we have to understand the basics of what speech is and how human speech is produced; and we also have to know what emotions are. It is also good to know a little bit about some applications of detecting emotion in human speech such as in the areas of speech synthesis systems and emotional speech recognition.

2.2 Speech

2.2.1 What is speech?

Speech is the principal means of communication between people. It has evolved over many centuries to become the rich and elaborate language structure today.

Speech consists of the acoustic realisation of a sequence of units called morphemes (the smallest meaning-bearing unit in a language). The way speech encodes morphemes is by using elementary acoustic symbols known as phonemes where a phoneme is the smallest element of sound (Rowden, 1992) that can be used to distinguish between different messages (i.e. different morphemes). This contrasts with text, which uses graphical units known as characters for the same task.

However, speech signals convey more than spoken words, they also contain some information about the message content and meaning, about the nature of the transmission medium, and about the identity and condition of the speaker.
Physically, sound is any vibration (wave) travelling through the air or other medium which can be heard when it reaches a person's ear. Figure 2.1 shows the waveform of a simple sound.

![Figure 2.1: Waveform of sound](image)

A pure note corresponds to a sound wave where the pressure along the signal varies as simple sinusoidal wave. Such a wave involves displacement of each particle in the medium through which the sound is travelling around a mean position, where the displacement also varies sinusoidally with time. The time for a complete cycle of this motion is referred to as the period. The wavelength is, equivalently, the distance between two pressure peaks at an instant or the velocity of sound in the medium divided by the frequency.

All sound signals possess the attribute of intensity and many possess the properties of pitch and timbre.

- Intensity refers to the energy in the signal. Loudness refers to the listener’s experience of that energy. The link between the two is that perceived loudness is proportional to the log of energy, i.e. if one sound has four times the intensity of another it will be perceived as being twice as loud, if the first has nine times the energy of the second it will be perceive as being three times as loud.
- Pitch refers to how high or low a sound is. For simple sounds the pitch is determined by the frequency, with sounds with a low frequency sounding deeper than ones with a
high frequency. For speech, the perceived pitch is not straightforwardly linked to the frequency and is typically found by using auto-correlation.

- Timbre-tone or tone quality has been defined as the characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech (Trask, 1996). It depends primarily on the waveform of the sound wave. Timbre helps your ears distinguish one type of sound production from another.

### 2.2.2 Human Speech production

The human speech production system is an interesting and complex mechanism, in this subsection I will explain how speech is created by the human vocal system.

Sound production requires two things: Power/energy source and vibrating element; in human speech production, air is the power source that comes from the lungs while the vocal cords are the vibrating element. So, all sounds are produced by some movements of air which goes up the windpipe (trachea) and into the larynx and out of the body through the vocal tract (i.e. mouth or nose). These produced sounds are conveyed by the ear to the brain for processing.

There are a set of unique vocal characteristics which a speaker has to enable other listeners to recognize him from his voice, these characteristics are related to the physiology of the speaker (Hansen & Hasan, 2015; Rodman, 1999).

There are a set of components of the human speech system which play a major role in the speech production process which can be grouped into three systems: respiratory system; lungs generating air stream, phonatory system includes larynx and vocal folds and articulatory system; and vocal tract. The ways in which sounds are produced is through organs of speech (articulators), as shown in Figure 2.2.
Speech sounds are generated by forcing the air from the lungs to pass through the folds of the vocal cords that are present in the larynx. When the air flow increases, complex pressure differences are created by the glottis, leading to the vibration of vocal folds in a way resembling the reed in a wind instrument. The produced sounds are called ‘voiced’ sounds and the whole process is called phonation. The stress on the vocal folds determines the frequencies of the vowel sounds (Wheddon & Linggard, 1990).

The basic sounds of a language are called phonemes which are considered as a working definition of the perceptual unit of language and the manifestation of each phoneme depends on the word being spoken and the position of the phoneme within the word.

There are three aspects of studying phonemes as scientists of phonetics suggest:

- Acoustic phonetics deals with the physical properties (acoustic aspects) of the sounds of language.
- Auditory phonetics is concerned with how listeners perceive the sounds of language.
- Articulatory phonetics focuses on how the vocal tract or organs produce the sounds of language.

A typical speech utterance consists of a string of vowel and consonant phonemes whose temporal and spectral characteristics change with time.
Consonants in English are described according to three features: place of articulation; manner of articulation; whether the vocal cords vibrate in articulation, that is if they are voiced. Their formation depends upon an aggressive (outward-flowing) airstream initiating in the lungs. and can be produced as result of obstruction of airflow in the vocal tract.

English has five vowels in its alphabet: a, e, i, o, u. The IPA chart ("International Phonetic Alphabet Sounds," 2017) for vowels in English can be shown Figure 2.3:

![IPA vowel chart](image)

Figure 2.3: IPA vowel chart

Vowel articulation is described using four features: tongue height (high, mid, low), tongue backness (front, central, back), lip rounding (round, unround) and tenseness i.e. whether the tongue muscle is tense or lax. As an example; [i] as in meet is high front tense unrounded vowel while [æ] as in pat is low front lax unrounded.

All vowels in English are voiced and involve a continuous flow of air through the oral cavity (no obstruction of airflow that comes from the lungs in vocal tract). Vowel sounds change according to the shape of the vocal tract. Because the vowels do not have any point of articulation or manner of articulation like consonants, it is more difficult to provide an articulatory description of them than it is for consonants.

As mentioned, voiced sounds occur when the vocal cords vibrate when the sound is produced. In contrast to this, There is no vocal cord vibration when producing voiceless sounds.

Measurement of fundamental frequency can be used as a way to detect if sounds are voiced or not. The fundamental frequency is the frequency of vocal fold vibration. Calculating the energy
in a section of the signal (signal frame) is another way to identify voiced and unvoiced sounds where a voiced sound has more energy than an unvoiced sound.

Next, a brief explanation to how humans hear the sounds. Hearing is a complex process of picking up sound and attaching meaning to it.

The ear can be divided into three parts leading up to the brain: the outer ear, middle ear and the inner ear as shown in Figure 2.4.

![A diagram of the human ear (Bosi & Goldberg, 2012)](image)

Waves enter the outer ear (called the pinna or auricle) and they travel through the ear canal and make their way toward the middle ear. These waves reach the part of the middle ear called the eardrum (a piece of thin skin stretched tight), and cause the eardrum to vibrate. These sound vibrations are carried to the three tiny bones of the middle ear, which turn waves into vibrations and then delivers them to the inner ear. The vibrations in the inner ear go into the cochlea (a small, curled tube in the inner ear). When the sound vibrations hit the liquid in the cochlea, these vibrations cause the sensory hairs in the cochlea to move. The sound vibrations are transformed into nerve signals and delivered to the brain via the hearing nerve, also called the “eighth nerve”. The brain then interprets these electrical signals as sound (Bosi & Goldberg, 2012).
2.3 Speech and emotion

In this section, we highlight some of the different notions of emotion and ways of classifying emotions in emotion research. Furthermore, we will describe how emotion relates to natural speech and which features influence the recognition of emotion in speech. In addition, we introduce cultural aspects of emotional speech and some applications for detecting emotion in speech.

2.3.1 What are emotions?

Emotions are internal feeling we experience, they are used to communicate in everyday life and make speech more expressive and effective.

Researchers have not managed to agree on an objective definition of the term emotion. Different people use the word with a variety of contextual meanings so this term is considered uncertain and complex. This is a fundamental hurdle to overcome in proceeding with a scientific approach towards research in the area (Schröder & Cowie, 2006).

Psychologists describe emotions in general as mental states that arise spontaneously as a result of external or internal stimulants rather than through conscious effort. However, some researchers claim that emotion is dependent on the influence of religion and socio-cultural customs in addition to social phenomena and the conditioned behavior of the group which an individual inhabits.

People can recognize the underlying emotions that other people feel by the way they express them in speech and facial expressions (Koolagudi & Rao, 2012). In the following section, we will describe and give categories of emotion.
2.3.2 Categories of emotion

Before we can attempt to link acoustic features to emotions, we have to have a target set of emotions. Unfortunately, there is no uniformly agreed set of categories for emotions, with most categorising done on a relatively subjective basis. The reason for this is that there are a range of emotions that can be confusing, challenging and complicated to describe because there are so many different perspectives on them. There are, however, two commonly used approaches, called discrete and dimensional.

- In the discrete approach, a set of emotions called basic emotions which are universally experienced in all human cultures are identified, namely happiness, sadness, disgust, fear, surprise, and anger. These emotions can be found in human everyday communication and are easy to understand and recognise. However, other theories and new research continue to explore many different types of emotions and it has been found that there are far more basic emotions than previously believed. Plutchik (Plutchik, 1990) proposed that there are eight primary emotions joy, sadness, acceptance, disgust, fear, anger, surprise and anticipation. The emotions that he identified, which are the foundation for all others, are grouped into polar opposites as show in figure 2.5 where the cone’s radial dimension represents intensity – emotions intensify as they move from the outside to the center of the wheel. For example, a feeling of boredom can intensify to loathing if left unchecked.

Figure 2.5: Plutchik’s Wheel of Emotions
There are, however, numerous competing classifications, and it can be hard to compare descriptions of the psychological properties and the associated acoustic features of emotions when different classifications are used.

- In the dimensional approach, emotions are defined according to one or more dimensions. For example Ressel (Ressel, 1980) suggests that emotions are distributed in a two-dimensional space, containing arousal and valence. Arousal represents the degree of activity that is generated by a particular stimulus. It relates to the intensity of an emotional experience and may range from energised or alert to calm or drowsy (i.e. active or passive). Arousal may also be referred to as Activation. Valence represents the degree to which a particular stimulus is viewed as being pleasant or unpleasant (i.e. positive or negative). Valence may also be referred to as Evaluation.

Murray and Arnott (Murray & Arnott, 1993) and Stibbard (Stibbard, 2001), on the other hand, suggest that there are three dimensions rather than two: Strength (corresponds to attention – rejection), Valence (corresponds to positive – negative) and Activity (corresponds to sleep – tension), or another example are the dimensions of pleasure, arousal and power, e.g. Pereira (Pereira, 2000). Schlosberg (Schlosberg, 1954) named three dimensions of emotion: "pleasantness–unpleasantness", "attention–rejection" and "level of activation".

Several dimensional models of emotion have been developed, though there are just a few that remain as the dominant models currently accepted by most (C. & M; David C. Rubin; Rubin & Talarico, 2009). The two-dimensional models that are most prominent are the circumplex model and the Positive Activation – Negative Activation (PANA) model as shown in Figures 2.6 and 2.7.
In this thesis, four emotion labels from the discrete approach are used in our experiments, namely anger, happiness, neutral, sadness, as we will describe in Section 4.2.2. This set of emotions appears most frequently and is most used with computational approaches, due to its simplicity and familiarity.
2.3.3 Emotion in natural speech

Humans have different ways, like laughing, shouting, or crying, to express their feelings; they also use more subtle modulations of speech. Consequently, in expressing and understanding emotions, different types of sources need to be considered.

Emotional expression also relates to a wide range of somatic correlates, including heart rate, skin resistivity, temperature, pupil diameter and muscle activity. For example, when feeling anger the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is then loud, fast and enunciated with strong high frequency energy. Meanwhile, experiencing a bored or sad state arouses the parasympathetic nervous system, whereby we observe decreasing of the heart rate and blood pressure decrease and salivation increases, producing speech that is slow, low-pitched and with little high frequency energy (Breazeal, 2000).

There are some obstacles to research into speech and emotion that make the study of emotions in speech difficult. In particular some features, such as pitch, intensity and duration, can be used both to encode emotions and to distinguish between different words. It can be hard to see when a feature is being used as an emotion marker and when it is used for distinguishing between lexical items, which can cause problems when looking at their use for encoding emotions.

Additionally, the lack of a settled approach to describing emotion and of clear, theory-based hypotheses makes comparing different analyses difficult.

Understanding how emotions are marked by various aspects of the voice requires more intensive work and contributions from researchers in multiple fields such as psychology, acoustics, speech science, linguistics, medicine, engineering, and computer science (Juslin & Scherer, 2008). However, analysis of expressed emotion can be done at three different levels according to Juslin and Scherer: the physiological, the phonatory-articulatory and the acoustic. In physiological level, researchers describe nerve impulses or muscle innervations patterns of the major structures involved in the voice-production process, the phonatory-articulatory level includes describing the
position or movement of the major structures such as the vocal folds, while describing characteristics of the speech wave form emanating from the mouth as the acoustic level. The physiological and phonatory-articulatory levels can take into consideration most of the current methods for measurement and analysis of expressed emotion, but require specialised equipment as well as a high level of expertise. In contrast, acoustic measurement of voice cues, requiring basic training in voice physiology and acoustics but no special equipment, is perhaps the method that holds the greatest promise for interdisciplinary research on emotional speech.

In the next section, we present in brief some studies conducted in order to investigate the features (parameters) of the acoustic speech signal which carry emotional information.

2.3.4 The acoustic correlates of emotions in human speech

A number of studies with a detailed analysis of recorded human emotional speech have been carried out in order to examine specifically parameters in the speech signal that are affected by the speaker's emotional state.

Williams and et al.(C. E. Williams & Stevens, 1972) reported two principal reasons why it is important to study and discover speech attributes related to emotional states. Firstly, studies of the parameters of the speech signal that relate to the emotional state of a speaker contribute towards a general theory of speech performance. Secondly, analysis of these characteristics of utterances may help to monitor the physiological and emotional state of an individual and hence make it possible to respond to their needs in certain situations and it would be convenient if an indication of this state could be obtained through analysing these parameters.

However, while identifying these characteristics is important, currently, researchers are still attempting to delineate some of the acoustic correlates of the emotions of a speaker, and there is debate about which features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together.

The acoustic characteristics of emotions in the speech signal that reflect the emotional state of a speaker have been investigated by many researchers through a number of experiments using
computer-based sound manipulation techniques (Banse & Scherer, 1996; Stevens & Williams, 1969; C. Williams, Stevens, & Hecker, 1970). Their results generally agree on the speech correlates that come from physiological constraints and correspond to broad classes of basic emotions, but are unclear when one looks at the differences between the acoustic correlates of, for instance, fear and surprise or boredom and sadness. Certain emotional states are often correlated with particular physiological states (Picard, 1995) which in turn have quite mechanical and thus predictable effects on speech, especially on pitch, (fundamental frequency F0) timing and voice quality.

Their findings refer to the fact that physiological constraints have an effect on emotions. For example, when we get angry, heart rate, arterial tension and testosterone production increase, cortisol (the stress hormone) decreases, and the left hemisphere of the brain becomes more stimulated. These things by themselves will change how your voice sounds in that situation, while by contrast, there are also things you do deliberately in order to express emotions which might be rising intensity and pitch for example when you get angry. It is hence hard to know whether the change in the way the sounds are produced is a result of autonomic physiological changes associated with the emotion in question or whether it is a deliberate choice. In the long run this may not matter for emotion detection, since whatever the cause the change will be associated with the emotion.

All studies in the field point to the pitch (fundamental frequency) as the main vocal cue for emotion recognition. The other acoustic variables contributing to vocal emotion signaling are vocal energy, frequency spectral features, formants (usually only one or two first formants (F1, F2) are considered), and temporal features (speech rate and pausing) (Banse & Scherer, 1996). Some more recent studies have shown that voice quality, which refers specifically to those auditory qualities that arise from variations in the voice source signal (Gobl & Chasaide, 2000) and certain co-articulatory phenomena (Kienast & Sendlmeier, 2000), are also reasonably correlated with certain emotions. Overall however it is generally agreed that the most crucial aspects are those related to prosody: the pitch contour, the intensity contour and the timing of utterances.
One of the most often mentioned studies in the field of emotion recognition through speech signals is work done by Murray and Arnott (Murray & Arnott, 1993) who have summarised the best acoustic parameters for detecting basic emotions in terms of speech rate, pitch (fundamental frequency), intensity, duration. The acoustic characteristics for the different emotions are presented in the Table (Murray & Arnott, 1993). Another important research on the acoustic correlates of speaker states was accomplished by Pittam and Scherer (Pittam & Scherer, 1993).

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Fear</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech rate</td>
<td>Slightly faster</td>
<td>Faster or slower</td>
<td>Slightly slower</td>
<td>Much faster</td>
<td>Very much slower</td>
</tr>
<tr>
<td>Pitch average</td>
<td>Very much higher</td>
<td>Much higher</td>
<td>Slightly slower</td>
<td>Very much higher</td>
<td>Very much lower</td>
</tr>
<tr>
<td>Pitch range</td>
<td>Much wider</td>
<td>Much wider</td>
<td>Slightly narrower</td>
<td>Much wider</td>
<td>Slightly wider</td>
</tr>
<tr>
<td>Intensity</td>
<td>Higher</td>
<td>Higher</td>
<td>Lower</td>
<td>Much wider</td>
<td>Slightly wider</td>
</tr>
<tr>
<td>Voice quality</td>
<td>Breathy, chest tone</td>
<td>Breathy, blaring</td>
<td>Resonant</td>
<td>Irregular voicing</td>
<td>Grumbled chest tone</td>
</tr>
<tr>
<td>Pitch changes</td>
<td>Abrupt, on stressed syllables</td>
<td>Smooth, upward inflections</td>
<td>downward inflections</td>
<td>Normal</td>
<td>Wide downward terminal inflections</td>
</tr>
<tr>
<td>Articulation</td>
<td>Tense</td>
<td>Normal</td>
<td>Slurring</td>
<td>Precise</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 2.1: Emotion and speech parameters (Murray & Arnott, 1993)

In summary, acoustic parameters, such as F0, intensity and duration, which have seemingly clear relationships with the most prominent perceptual characteristics of speech (i.e. pitch, loudness and speech rate) have received the most attention. These parameters also tend to be the easiest to analyze.

However, decoding emotions in speech is a complex process that is influenced by the cultural, social, and intellectual characteristics of subjects (Petrushin, 1999). The next section introduces the cultural aspects of emotional speech.

### 2.3.5 Cultural aspects of emotional speech

Many transcultural studies emphasise how language and culture play an important role in the way vocal emotions are encoded and recognised in speech and other communication channels (Elfenbein & Ambady, 2002). Many researchers have suggested that any theory relating to emotion should take account of linguistic and cultural factors.
They also showed evidence that the way specific effects are expressed acoustically is very similar across different cultures (Scherer, Banse, & Wallbott, 2001; Thompson & Balkwill, 2006). The latter finding suggests that the vocal expression of emotion is, to some extent, innate (Hammerschmidt & Jürgens, 2007), so a listener can identify an emotionally spoken sentence correctly even if it was not spoken in his or her mother language (Beier & Zautra, 1972; Kramer, 1964; Scherer et al., 2001)

Some researchers such as Ekman and et al. (Ekman & Friesen, 1971; Izard, 2001) reported that expressions of basic emotion such as joy, anger, disgust, sadness, and fear hold certain properties which allow them to be recognised independently regardless of culture and learning, this would arise from the fact that at least some emotions are directly linked to physiological changes: the changes in hormone levels that arise when person is in a dangerous situation will predate the evolution of language, so if those changes result in things like a raised heartbeat that then affect how quickly someone do things then this will almost certainly be uniform across cultures.

In contrast, others claim that there are differences in how emotions are expressed across human cultures depending on types of expressing emotions and study group (Beaupré & Hess, 2005; Ekman et al., 1987; Elfenbein, Beaupré, Lévesque, & Hess, 2007; Matsumoto, 1993).

Pell et al in their study (Pell, Monetta, Paulmann, & Kotz, 2009) also emphasise that the ability to understand expressed emotions in speech is partly independent of linguistic ability, although this ability is also shaped by linguistic and cultural variables. They investigate whether basic expressed emotions can be identified pan-culturally from a speaker’s voice and the face, regardless of the culture or linguistic ability of the individual. They compared how Argentine Spanish speakers (61 participants) recognise basic emotions from pseudo-utterances (“nonsense speech”) produced in their native language and in three foreign languages (English, German, Arabic). Results indicated that vocal expressions of basic emotions could be recognised from vocal cues in each language at accuracy levels exceeding chance.

They found also from the perspective of language that the Arabic task was the most difficult for the majority of participants. On the level of recognising emotions, disgust was the hardest emotion to identify by participants, followed by joy and fear.
Another study conducted by Pell et al. (Pell, Paulmann, Dara, Alasseri, & Kotz, 2009) with four different languages (English, German, Hindi, and Arabic) supported the conclusion that vocal expressions of basic emotion in speech are largely unaffected by language or linguistic similarity. They conducted their study of vocal expressions of six emotions (anger, disgust, fear, sadness, happiness, pleasant surprise) and neutral expressions which were elicited from four native speakers of each language (English, German, Hindi, and Arabic) and the recordings were judged for their perceived emotional meaning by a group of native listeners in each language condition. The results of analysing acoustic patterns and emotion recognition across each language show that all emotions could be recognised strictly from vocal cues in each language at levels exceeding chance.

The most accurately recognised emotions were anger, sadness, and fear, irrespective of language. Also they found by analysing acoustic and discriminate function that relative pitch level and variability (fundamental frequency) is an important feature for signaling vocal emotions in all languages.

The debate concerning whether emotions are universal or culturally specific still goes on. However, it has already been suggested that emotions could be both universal and recognisable in all cultures.

2.3.6 Applications for detecting emotion in human speech

The main purpose of employing speech emotion recognition is to adapt the system’s response to allow a highly natural speech interaction by understanding the verbal and emotion content. The interaction between user and computer will be more meaningful if the system can recognise not just who said something and what they said but also how it is being said (Ingale & Chaudhari, 2012).

Speech emotion recognition is particularly useful for applications which require that the response of those systems are natural and depend on the detected emotion such as computer tutorial applications and in-car systems in which the emotional and mental state of the driver play an important role in her/his safety (Schuller et al., 2004).

Other applications include call centre and mobile communication (J. Ma, Jin, Yang, & Tsai, 2006). Speech emotion recognition has also been used by therapists as a diagnostic tool (France,
Shiavi, Silverman, Silverman, & Wilkes, 2000), and can be applied in a variety of situations to help people with the highest levels of suffering or need, such as in a nursing home. Speech emotion recognition can also be useful with real time applications such as verification of emotions to analyse whether requests are genuine or analysis of voice messages in the emergency services (Koolagudi & Rao, 2012).

El Ayadi et al (El Ayadi, Kamel, & Karray, 2011) summarise some of the reasons that make the task of speech emotion detection challenging:

- first, difficulties in determining the most significant and powerful feature when distinguishing between emotions is due to the direct effects of many of factors such as speaking rate and style, different speakers and sentences used.
- another challenge is how a certain emotion is expressed generally depends on the speaker, his or her culture and environment.
- A further difficulty is that it is not clear which emotion the automatic emotion recogniser will detect, especially in cases where that person may undergo a certain emotional state for many days (long-term emotion), so other emotions will be transient and will not last for more than a few minutes.

However, there are two different applications areas which require understanding how emotions are expressed in speech. There are (1) synthesis of emotional speech, and (2) emotion recognition. This section focuses on these areas in brief.

In synthesis of emotions in speech, employing the appropriate emotional expressivity in synthetic speech leads to the increased natural quality of these systems. This fact has motivated many researchers in this area to attempt to incorporate the expression of emotions into synthetic speech over more than a decade.

Marc (Marc, 2001) attempts to give an overview of studies that can be found in literature concerned with the expression of emotion in synthetic speech. These studies can be grouped according to the type of synthesis technique employed, such as formant synthesisers which have captured the attention of many researchers because of the high degree of control that they provide, as in (Burkhardt, 2001; Burkhardt & Sendlmeier, 2000; Cahn, 1989; I. R. Murray,
Emotion recognition systems is another an important research area today that attracts the attention of the research community increasingly towards understanding emotions and the idea of automatically identifying the emotional or physical state of a human from his or her voice. There are several applications where speech emotion recognition can be developed (Gadhe, Shaikh Nilofer, Waghmare, Shrishrimal, & Deshmukh, 2015). The recent literature on speech emotion recognition can be found in many studies such as in (Gadhe et al., 2015; Koolagudi & Rao, 2012) and (Ververidis & Kotropoulos, 2006).

Gadhe et al. (Gadhe et al., 2015) conducted a survey included in the present literature review of different features addressing the importance of choosing different classification techniques to be used for speech emotion recognition and some important issues in this area. They found that developing an emotion recognition system depends essentially on a well-designed database and on the selection of classifiers. They observed that the Gaussian Mixture Model is a more efficient classifier compared with others such as the Hidden Markov Model (HMM), the Bayes classifier, the Support Vector Machine (SVM), k-nearest neighbors (KNN), the Artificial Neural Network (ANN) and the Maximum Likelihood Bayesian classifier. The Gaussian Mixture Model is based on expectation-maximization algorithm or maximum a posterior (MAP) parameter estimation. Koolagudi and et al. in their study (Koolagudi & Rao, 2012) considered the issues related to emotional speech recognition such as different types of speech features and classification models used for recognition of emotions from speech. In addition, emotional speech corpora and some important issues in this area of speech emotion recognition are discussed. They concluded many facts such as a need for a larger emotional speech database with a reasonably large number of speakers and text prompts because the existing databases are used with a limited number of speakers so that most of the research produces results that cannot be generalized. Also the performance of emotion recognition can be enhanced by a consideration of linguistic contents and the features extracted from speech which have an effect on emotion expression.

Another study conducted by Ververidis et al. (Ververidis & Kotropoulos, 2006) provides an overview of emotional speech recognition from three facets: first, the available recent emotional
speech corpus, addressing a number of emotional states and speakers, the language, and the kind of speech; second, presenting the most frequent acoustic parameters used for recognising emotional speech and for assessing how the emotion affects them; a third reviewing appropriate techniques in order to classify speech into emotional states. Some of their findings include the fact that the most interesting features are pitch, formants, short-term energy, MFCCs and the cross-section areas, in addition, the Teager energy operator-based features which refers to the number of harmonics due to the non-linear airflow in the vocal tract that produces the speech signal and is considered a useful feature for emotion recognition (Teager & Teager, 1990)

2.4 Summary

This chapter has given an overview of speech and emotion, pointing out the acoustic correlates of emotions in human speech and taking a look at some cultural aspects of emotional speech, finally, some applications of detecting emotion in human speech are presented. In the next chapter we are going to give an overview of issues for Arabic speech processing
CHAPTER THREE : ISSUES FOR ARABIC SPEECH PROCESSING

3.1 Introduction

The study of the phonological system of Arabic is an important step in the development of speech systems such as text to speech (TTS) and Automatic Speech Recognition (ASR) systems.

This chapter investigates the issues related to Arabic speech processing. The chapter is structured as follows: Section 2 considers the properties of the Arabic language through a discussion of both its written and its phonological systems. An in-depth study is introduced for Arabic consonants, vowels, semi-vowels, diphthongs and syllables. Section 3 presents general information about the prosodic features of Arabic language, namely Intensity, duration and intonation, which form the basis of our work. Section 4 presents the problem of the variety of dialects as a challenge for Arabic language speech studies. A summary of this chapter is presented in Section 5.

3.2 Mapping written forms to phonetic sequences

In order to discuss how to get from written forms to phonetic sequences, we have first to understand the nature of the Arabic writing system and of the set of phonemes, and then we have to see how to map one to other.

3.2.1 Writing system, Semi-vowels, Diacritics

The Arabic writing system uses an alphabet with a set of 28 distinct characters which denote consonants and semi-consonants (see the next section) and three that represent long vowels, along with a set of marks called diacritics that are attached to these characters to represent short vowels and other phonetic events. These diacritics are often omitted, which can lead to significant problems when trying to interpret written forms (see the next section).
In Arabic, each consonant letter has a different shape depending on its location within the word as shown in Table 3.1, which also presents the correspondence between the written forms and the phonemes used in Arabic, using SAMPA (The Speech Assessment Methods Phonetic Alphabet) for the Arabic alphabet\(^1\) and IPA (the International Phonetic Alphabet) symbols\(^2\).

<table>
<thead>
<tr>
<th>ARABIC GRAPHEME</th>
<th>IN A WORD</th>
<th>PHONEMIC SYMBOL (In SAMPA)</th>
<th>PHONEMIC SYMBOL (In IPA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ا</td>
<td>ايا</td>
<td>/ʔ/</td>
<td>ئ</td>
</tr>
<tr>
<td>ب</td>
<td>بب</td>
<td>/b/</td>
<td>b</td>
</tr>
<tr>
<td>ت</td>
<td>تت</td>
<td>/t/</td>
<td>t</td>
</tr>
<tr>
<td>ث</td>
<td>ثث</td>
<td>/θ/</td>
<td>ظ</td>
</tr>
<tr>
<td>ج</td>
<td>حجه</td>
<td>/dʒ/</td>
<td>dʒ</td>
</tr>
<tr>
<td>ح</td>
<td>حجه</td>
<td>/h/</td>
<td>h</td>
</tr>
<tr>
<td>خ</td>
<td>خخ</td>
<td>/x/</td>
<td>x</td>
</tr>
<tr>
<td>د</td>
<td>ديد</td>
<td>/d/</td>
<td>d</td>
</tr>
<tr>
<td>ذ</td>
<td>ذذا</td>
<td>/ð/</td>
<td>ظ</td>
</tr>
<tr>
<td>ر</td>
<td>ربر</td>
<td>/r/</td>
<td>r</td>
</tr>
<tr>
<td>ز</td>
<td>زيز</td>
<td>/z/</td>
<td>z</td>
</tr>
<tr>
<td>س</td>
<td>سسس</td>
<td>/s/</td>
<td>s</td>
</tr>
<tr>
<td>ش</td>
<td>ششش</td>
<td>/ʃ/</td>
<td>ʃ</td>
</tr>
<tr>
<td>ص</td>
<td>صصص</td>
<td>/sˤ/</td>
<td>sˤ</td>
</tr>
<tr>
<td>ض</td>
<td>ضضض</td>
<td>/dˤ/</td>
<td>dˤ</td>
</tr>
<tr>
<td>ط</td>
<td>ططط</td>
<td>/tˤ/</td>
<td>tˤ</td>
</tr>
<tr>
<td>ظ</td>
<td>ظظظ</td>
<td>/dˤ/</td>
<td>dˤ</td>
</tr>
<tr>
<td>ع</td>
<td>ععع</td>
<td>/ʔʔ/ /ʔʔ/</td>
<td>ئ</td>
</tr>
</tbody>
</table>


Table 3.1: Different shapes of Arabic letters with SAMPA and IPA symbols

An Arabic word can be written with consonants and vowels in certain forms of classical Arabic writing which are used solely for text found in the Qur'an and classical Arabic literature, but in general, short vowels are not actually written in newspapers, magazines, books, and government documents in Modern Standard Arabic (MAS).

The vowels are represented by marks above or below the consonant letters called diacritics and are also known as the vocalisation or the voweling. These diacritics are summarised below:

- The six vowels are divided into three short vowels and three long vowels (M. Alghamdi, 2001; Alkhouli, 1990; Javed, 2013) as follows:
- Short vowels:
- "َ", is named in the diacritic system "fathah", this appears on top of a consonant in order to indicate the sound of the Arabic short vowel /a/.

- "و", is called "kasrah", which appears under a consonant to show that the consonant is vocalised by the sound of the Arabic short vowel /i/.

- "ُ", is called "dammah" in a diacritic system, which is like a comma, that appears on top of a consonant to indicate that the consonant is vocalised by the sound of the short vowel /u/.

- Long vowels which put the stress on a given vowel, these long vowels are also considered some sort of consonants, which are:

  - Alif, which presents both long vowel /a:/ and the consonant "hamzah" (glottal stop) /ʔ/.

  - Yaa, which presents both long vowel /i:/ and the consonant /y/.

  - Waw, which presents both long vowel /u:/ and the consonant /w/.

- There are another three symbols in the Arabic writing system which are called "Tanween" and are named as tanween dammah, tanween kasrah and tanween fathah. They appear as symbols "ٌ، ٍ، " respectively and can be shown on any consonant to indicate certain phonemic sequences. For example the following words: "بخيل" /baki:lun/ (Stingy), "طويل" /t'awu:lin/ (tall) and "كبيرا" /kabi:ran/ (big) have the tanween symbols,"ٌ، ٍ، ".

- The Shadda, is another symbol "ٰ" which in general appears on a consonant to indicate that the consonant is doubled or its pronunciation is repeated twice, for example, "قيم"/qajjim/ (Valuable) has the Shadda symbol on the letter "ي" /j/ that means the sound of the letter doubled. The Shadda also appears with diacritic marks, short vowels (َ، ُ، ُ) which are placed on top of the Shadda.
• Arabic uses another diacritic called sukun “ًّۡ” It appears on the top of consonant and is not followed by a vowel sound and indicates that the consonant is not provided with a vowel, so it is not necessary to write sukun in the text.

• Dagger Alif “ـ 
الخنجرية” is written as a short vertical line on top of a consonant. It indicates a long alif /a:/ sound, but the alif is normally not written. The dagger alif occurs in a few common words, such as: هنا /ha:Da/ (this) and لكن /la:kin/ (but).

• Hamza “ء 
و ۡء” and stand alone "ء " The Hamzah indicates a glottal stop accompanied by any of the diacritic (fatha, kasrah, dammah or sukun); e.g. أحمد (?hamd), سؤال, (question) and هدوء (quietness).

• Maddah “آ 
آ ۡآ” The Maddah is a diacritic like the shadow of a bird flying, which can appear only on top of an alif “أ” and indicates a glottal stop (Hamzah) followed by another alif representing the long /a:/; e.g. قراران /qur’a:n/ (the holy book).

As mentioned above, the diacritics that include short Arabic vowels, Shadda, Tanween etc., are represented using diacritic symbols in the written form which refer to vowel phonemes in the designated words. These diacritics are frequently omitted in most written Arabic text, leading to a lack of essential information about the meaning of the word. Additionally, identification of the function of the word in the sentence also makes the orthography difficult to understand, since a single written form may correspond to numerous underlying forms, for example, the diacritic in the top of the letter "r" in word مدرسة, if it was the short vowel "ا" fathah then the word will be مدرسة /madrasah/ (school), but if was the symbol “ُ” Shadda, then the word will be مدرسة /mudarrisah/ (teacher).

However, the reader is expected to understand the meaning of the text which requires enough knowledge of the language in order to infer vowels from the context along with other missing cues to enable him or her to pronounce the text correctly.

Consequently, one of the biggest challenges facing Arabic text preprocessing is that the text must be diacritized to be read, so that each character and its diacritic must be determined to avoid ambiguity.
The full diacritization of Arabic script provides useful information for pronunciation modeling and higher level processing of developing Arabic speech and natural processing language applications, since it is likely to reduce ambiguity in these tasks.

In synthesis of speech, it is essential for text-to-speech synthesis (TTS) applications that the correct vowel(s) are pronounced, because the meaning of the utterance will be altered completely by using an incorrect vowel (Ananthakrishnan, Narayanan, & Bangalore, 2005). In Automatic Speech Recognition (ASR) systems, the absence of diacritization of Arabic text leads to many similar word forms, and consequently, decreases predictability in the language model (Alotaibi & Hussain, 2010). In contrast, using vocalised text can improve the accuracy of speech recognition applications (Zitouni, Sorensen, & Sarikaya, 2006).

The problem of determining the correct vowelisation has been investigated by many researchers in different areas of Arabic speech processing (Elshafei, Al-Muhtaseb, & Al-Ghamdi, 2002); in (Al-Otaibi, 1988; Elshafei, Al-Muhtaseb, & Alghamdi, 2006; Kirchhoff et al., 2002; Nelken & Shieber, 2005; Vergyri & Kirchhoff, 2004) the focus was to develop techniques for automatic diacritization of Arabic script for applications for MSA.

There are two obvious approaches to solving the vowelisation problem: One approach is to implement a module for automatic vowelisation in order to infer the vowels (Elshafei et al., 2002). Generating vowels automatically requires integration of morphological, syntactical, and semantic information (Elshafei et al., 2002). Most applications rely on automatic Arabic diacritization as a pre-processing step to transcription (Rashwan, Al-Badrashiny, Attia, Abdou, & Rafea, 2011). Another approach is by using electronic lexicons that provide the vowelisation for an orthographic string (Tomokiyo, Black, & Lenzo, 2003).

However, in our work we avoided the inclusion of these levels of solutions by requiring that the text be diacritized from the outset.
### 3.2.2 Arabic phonetic system (Arabic phonology)

The phonetic system of Modern Standard Arabic (MSA) has basically 36 phonemes classified into two types, consonants and vowels.

Arabic uses different set of consonants and vowels from English and therefore we need to have a look at specific vowels and consonants of Arabic, these summarized in the following tables.

<table>
<thead>
<tr>
<th>Stops</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blabial</td>
<td>b /ب</td>
<td>d /د</td>
</tr>
<tr>
<td>Labiodental</td>
<td>d /د</td>
<td>t /ت</td>
</tr>
<tr>
<td>Interdental</td>
<td>d /د</td>
<td>k /ك</td>
</tr>
<tr>
<td>Alveodental</td>
<td>d /د</td>
<td>q /ق</td>
</tr>
<tr>
<td>Palatal</td>
<td>d /د</td>
<td>z /ز</td>
</tr>
<tr>
<td>Uvular</td>
<td>d /د</td>
<td>? /؟</td>
</tr>
<tr>
<td>Pharyngeal</td>
<td>d /د</td>
<td></td>
</tr>
<tr>
<td>Laryngeal</td>
<td>d /د</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fricative</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blabial</td>
<td>D /د</td>
<td>t /ت</td>
</tr>
<tr>
<td>Labiodental</td>
<td>D /د</td>
<td>t /ت</td>
</tr>
<tr>
<td>Interdental</td>
<td>D /د</td>
<td>t /ت</td>
</tr>
<tr>
<td>Alveodental</td>
<td>D /د</td>
<td>t /ت</td>
</tr>
<tr>
<td>Palatal</td>
<td>D /د</td>
<td></td>
</tr>
<tr>
<td>Uvular</td>
<td>D /د</td>
<td></td>
</tr>
<tr>
<td>Pharyngeal</td>
<td>D /د</td>
<td></td>
</tr>
<tr>
<td>Laryngeal</td>
<td>D /د</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nasal</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blabial</td>
<td>m /م</td>
<td>n /ن</td>
</tr>
<tr>
<td>Labiodental</td>
<td>m /م</td>
<td>n /ن</td>
</tr>
<tr>
<td>Interdental</td>
<td>m /م</td>
<td>n /ن</td>
</tr>
<tr>
<td>Alveodental</td>
<td>m /م</td>
<td>n /ن</td>
</tr>
<tr>
<td>Palatal</td>
<td>m /م</td>
<td></td>
</tr>
<tr>
<td>Uvular</td>
<td>m /م</td>
<td></td>
</tr>
<tr>
<td>Pharyngeal</td>
<td>m /م</td>
<td></td>
</tr>
<tr>
<td>Laryngeal</td>
<td>m /م</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trill</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blabial</td>
<td>R /ر</td>
</tr>
<tr>
<td>Labiodental</td>
<td>R /ر</td>
</tr>
<tr>
<td>Interdental</td>
<td>R /ر</td>
</tr>
<tr>
<td>Alveodental</td>
<td>R /ر</td>
</tr>
<tr>
<td>Palatal</td>
<td>R /ر</td>
</tr>
<tr>
<td>Uvular</td>
<td>R /ر</td>
</tr>
<tr>
<td>Pharyngeal</td>
<td>R /ر</td>
</tr>
<tr>
<td>Laryngeal</td>
<td>R /ر</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lateral</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blabial</td>
<td>l /ل</td>
</tr>
<tr>
<td>Labiodental</td>
<td>l /ل</td>
</tr>
<tr>
<td>Interdental</td>
<td>l /ل</td>
</tr>
<tr>
<td>Alveodental</td>
<td>l /ل</td>
</tr>
<tr>
<td>Palatal</td>
<td>l /ل</td>
</tr>
<tr>
<td>Uvular</td>
<td>l /ل</td>
</tr>
<tr>
<td>Pharyngeal</td>
<td>l /ل</td>
</tr>
<tr>
<td>Laryngeal</td>
<td>l /ل</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semi vowels</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blabial</td>
<td>w /و</td>
</tr>
<tr>
<td>Labiodental</td>
<td>w /و</td>
</tr>
<tr>
<td>Interdental</td>
<td>w /و</td>
</tr>
<tr>
<td>Alveodental</td>
<td>w /و</td>
</tr>
<tr>
<td>Palatal</td>
<td>w /و</td>
</tr>
<tr>
<td>Uvular</td>
<td>w /و</td>
</tr>
<tr>
<td>Pharyngeal</td>
<td>w /و</td>
</tr>
<tr>
<td>Laryngeal</td>
<td>w /و</td>
</tr>
</tbody>
</table>

**Table 3.2: List of Arabic consonants with IPA symbols**

Arabic is particularly rich in uvular, pharyngeal, and pharyngealised ("emphatic") sounds. Arabic phonemes can be classified to two distinctive classes, one is pharyngeal and another is emphatic phonemes (Ahmed, 1991; Alkhouli, 1990; Omar, 1991).

The emphatic sounds are generally considered to be /s /س, /d /د, /t /ت, and /z /ز (Sabir & Alsaeed, 2014). These sounds are also referred to by linguists as (Alhuruf AlmuTbaqa) "الحروف المطبقة", covered letters (Alsabaan, Alsharhan, Ramsay, & Ahmad) because it is produced physiologically with retraction of the root of the tongue and raising of the back of the tongue towards the velum (Ferrat & Guerti, 2013).
In Arabic, there are six vowels as mentioned before, these vowels are always voiced sounds and they are distinguished from consonants in that the passage through which the air travels is never so narrow as to obstruct the free flow of the airstream. There are four features for vowel articulation which can be used to distinguish between different vowels as shown in Table 3.3: tongue height (high, mid, and low), the part of the tongue involved (front, central and back), lip rounding (rounded and unrounded), and tenseness or laxness of the vocal tract (tense and lax). These features are essential or principal qualities and are most commonly found in the world’s languages (Maddieson & Disner, 1984).

<table>
<thead>
<tr>
<th>Tongue position/height</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High or closed</td>
<td>/i/</td>
<td>/i:/</td>
<td>/u/</td>
</tr>
<tr>
<td>Low or open</td>
<td>/a/</td>
<td>/a:/</td>
<td>/u:/</td>
</tr>
<tr>
<td>High or closed</td>
<td>/u/</td>
<td>/u:/</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Arabic vowel

However, the spectral properties of long and short vowels are similar, but durations of long vowels are longer than the short version (M. Alghamdi, 2001). This can lead to a different meaning of the word, for example, two words: 

يِمل/jamil/ (become bored) with short vowel /i/ and يِمِل/jami:l/ (to lean) with long vowel /i:/, differ only in the vowel length, but they have two different meanings.

Two Semi-vowels: the letters (و، يـ ) /w/ , /j/ which can be seen in different positions within the word (the beginning, middle or the end) have two functions. They can be realised as the consonant sounds /w/, /j/, typically if they occur between two other vowels are at the start of a word with a following vowel; and they can be realised as vowels, either by themselves as long versions of the vowels /u/ or /i/ or in combinations with other vowels, forming what are called Diphthongs.

- "أـ" /aw/~ The glide begins from the vowels /a/ to the consonant /w/ (pronounced like the "ow" in "power"). As in /nawm/ نوم "sleep".

44
"أي" /aj/- in which the glide /y/ is preceded by the short vowel /a/ generates a diphthong /ay/ (pounced like English word "eye") as in the word /bajt/ ("بيت" "house") (Amer). Table 3.4 shows some examples of diphthongs:

<table>
<thead>
<tr>
<th>Orthography</th>
<th>IPA</th>
<th>Examples</th>
<th>Front</th>
<th>Meaning</th>
<th>Central</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>وِْ</td>
<td>[ov]</td>
<td>فوز</td>
<td></td>
<td>Winning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>وِْ+ accentuated</td>
<td>[o'v]</td>
<td>ضوء</td>
<td></td>
<td>Light</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يَٰ</td>
<td>[ei]</td>
<td>بيت</td>
<td></td>
<td>House</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يَٰ+ accentuated</td>
<td>[e'i]</td>
<td>خير</td>
<td></td>
<td>Good</td>
<td></td>
<td></td>
</tr>
<tr>
<td>وَّا</td>
<td>[ov]</td>
<td>أوراق</td>
<td>Papers</td>
<td>They went away</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يْ ا</td>
<td>[ei]</td>
<td>ابعث</td>
<td>Ripened</td>
<td>رأيت</td>
<td>I saw</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4: Arabic diphthongs in different positions (Khalifa, Obaid, Naji, & Daoud, 2011)

There are numerous contextual phonetic variations that can occur within Arabic sounds, such as replacing, in the pronunciation, some consonants with other consonants. We will not be dealing with such variations in our experiments, and we are therefore not going to discuss them in detail here.

3.2.3 Grapheme-to-phoneme

Grapheme-to-phoneme conversion is the process of converting any sentence of the text (from the orthographic transcription) in Arabic to its phonetic pronunciation by simple one-to-one mapping. Grapheme-to-Phoneme also known as Letter-to-sound conversion. For example, the phonetic pronunciation of Arabic word 'كاتب' is /ka:lib/.

This process is considered an easy task to achieve in Arabic compared to languages which use the Roman alphabet, because Arabic has a regular spelling system which can be modeled with specific rules.

The grapheme-phoneme transcription is considered an important prerequisite and an essential task for applications involving speech recognition and/or speech synthesis. In the context of
speech synthesis, having a tool that generates correct Grapheme-to-phoneme is a fundamental step leads to a good performance of TTS (Polyakova & Bonafonte, 2005).

In general, transformation of a written Arabic text into a spoken utterance requires knowledge of the Arabic language for the realisation of an automatic phonetisation system and needs a work of analysis, comprehension and an appropriate language model.

So several works have been realised in the area of text-to-speech synthesis of phonetisation of the Arabic text, we mention the works realised by (Ahmed, 1991); (Al-Ghamdi, Al-Muhtasib, & Elshafei, 2004); (El-Imam, 1989, 2004; Saidane, Zrigui, & Ahmed, 2005; Zemirli, 2006); (Imedjdouben & Houacine, 2014; Khalifa, Obaid, Naji, & Daoud, 2011).

And in the area of speech recognition as in (M. Ali, Elshafei, Al-Ghamdi, & Al-Muhtaseb, 2009), El-Imam (2004) noted that three approaches have been used for transformation of a written text into a spoken text. They are: dictionary-based methods, rule-based methods and the relatively newer trained data-driven methods, which are classified into three classes: the pronunciation by analogy (PbA), statistical methods (based on stochastic theory and nearest neighbour), and methods based on neural networks (Bigi, Péri, & Bertrand, 2012).

The first two of these methods will be discussed in the next sections.

- Dictionary or Lexicon-based methods.

Words that do not follow phonemic transcription rules are transcribed by this lexicon which contains phonological information to represent the phonemes of words directly (Alotaibi, 2012).

This lexicon or dictionary contains morphemes and their pronunciation and might contain other syntactic and semantic information. Some errors during transcription can be generated by this method. So we have to use some rules in order to treat these errors.

A phonetic dictionary of Arabic must take into account inflected forms of words, resulting in a huge number of entries. However, all derived forms abide by the Arabic spelling rules. In addition, the simplicity of the Arabic spelling system makes using a rule-based transcription system with a dictionary of exceptional words a possible solution to the Arabic letter-to-sound
conversion problem (El-Imam, 2004). Speed, flexibility and simplicity can be improved by using this approach.

- Phonemic Transcription Rule-based methods.

The words which are not treated by the dictionary are transcribed by using a base of phonemic transcription rules. In this approach, a set of rules of phonemic transcription which rely on expert linguistic and phonetic knowledge are used. These rules are supported by a lexicon of exceptions when the rules are not applicable, this lexicon contains a list of special words or phrases and abbreviations that require special pronunciation attributes and the phonemic representation of the words (Andersen et al., 1996).

The main advantage of this approach is the ability to model the linguistic knowledge of human beings by a set of rules that can be incorporated in an expert system (Tebbi, Hamadouche, & Azzoune).

However, rule-based systems have difficulties to change some of the rules without introducing unwanted side effects. Furthermore, employing such systems to new tasks in other languages is really time consuming and requires a lot of experience of phonetic knowledge (Andersen et al., 1996).

Many of these pronunciation rules can be found in (Ahmed, 1991; M. Ali et al., 2009; Al-Ghamdi et al., 2004; Selim & Anbar, 1987), who describe the rules of Arabic phonology which is based on rewrite rules in order to map each grapheme to its phonemic representation.

Major challenges for a letter-to-sound conversion step raise some issues such as assimilation, deletion, insertion, emphasis, and sound variation which are discussed earlier in this chapter.
3.3 Prosodic Features

Prosody is often defined on two different levels (Chentir, Guerti, & Hirst, 2009):

- an abstract, phonological level (phrase, accent and tone structure),
- a physical phonetic level (fundamental frequency, intensity and duration).

Khan, Al-Khatib, & Lahouari (2007) divided prosodic features into acoustic prosodic features that can be gained from the audio signal directly and linguistic prosodic features that are obtained from the word or phone transcript of a conversation on a lexical information level.

Prosody is the set of features of speech output that includes pitch, timing, pausing, speech rate, and emphasis on words, etc. Without these features, speech would sound like reading of a list of words. So these features are very important and are critical to speech processing tasks such as automatic speech recognition and speech synthesis.

There are many factors that can have an effect on prosodic features of natural speech, which can make accurate modeling very difficult (Thakur, Chettri, & Shah, 2012). Figure 3.2 shows some of these factors.

![Figure 3.1: Prosodic dependencies](image)

Dialog act detection (Shriberg et al., 1998), (Fernandez & Picard, 2002); topic segmentation (Swerts, 1997); identifying the dialect of a speaker automatically with considerable accuracy
(Biadsy & Hirschberg, 2009); sentence boundary detection (Pfeiffer, 2001); emotion detection (Lee & Narayanan, 2005) etc., are examples of the many applications of speech that employ prosodic features to give the language its appearance and convey the meaning of the spoken language, in addition, providing the natural quality of the human voice.

Prosodic features are considered to be the most significant factor in emotional expression in speech (Kurematsu, 2008; Hashizawa, Takeda, Hamzah, & Ohyama, 2004) due to the fact that these prosodic features can be utilised to indicate the emotional state of the speaker, for example, expressing anger by increasing both loudness and pitch. In addition, prosodic features reflect whether the speaker is being ironic or sarcastic, also whether the utterance is a statement, a question, or a command.

So by analysing speech patterns of these features we can get emotional speech processing applications which are able to recognise the user's emotional state.

Nowadays, there are numerous computational studies aimed at prosodic modeling for emotional speech recognition, understanding and synthesis systems by using many techniques as listed by (Rajeswari & Uma, 2012).

As an example, (Jiang, Zhang, Shen, & Cai, 2005) investigate the importance of prosody modeling for emotional speech synthesis system by studying prosodic features and obtained results which show that the prosody features help to obtain meaningful results for some emotions. Also (Chabchoub & Cherif, 2012) introduce a new high quality Arabic speech synthesis technique by modeling the parameters of the fundamental frequency and duration for the MBROLA method. The result shows that syllables which produce reasonable natural quality speech and durational modeling are crucial for naturalness with a significant reduction in numbers of units of the total base developed.

Other studies have been conducted to discover the acoustic correlates of emotions in the field of expressive TTS, such as in (Bulut et al., 2005; Salvi, Tesser, Zovato, & Cosi, 2010; Shao, Wang, Han, & Liu, 2005) and their results proved that both prosodic (pitch, duration, intensity) and spectral parameters should be modified to achieve a successful emotion conversion.
In the area of speech recognition, there are a number of studies which have proposed some prosodic features for emotional speech. Devi, Srinivas, & Nandyala (2014), for instance, proposed speech features such as Energy and Pitch and used SVM for classification while features such as the Mel Frequency Cepstrum Coefficients (MFCC) was proposed in (Rawat & Mishra, 2015) and evaluated by using neural networks.

(Meddeb, Karray, & Alimi, 2014) proposed speech features such as Linear Prediction Cepstrum Coefficients (LPCC), Mel Frequency Cepstrum Coefficients (MFCC) in order to achieve Intelligent Remote Control for a TV programme based on emotion in Arabic speech,

Unfortunately, a reliable set of features for discriminating emotional states in speech has not yet been reached. It is an open question, because of difficulties of finding which features carry more information to describe the emotional content in speech and how to combine to gain a good recognition of emotion (Kuremastsu, 2008).

Among everything researchers have come up with, it looks as though intensity, duration and intonation are the most practically important and therefore I am going to say more about them in subsequent sections.

3.3.1 Intensity

The intensity of a sound, defined as power per unit area, is an acoustic property of the sound that can be physically measured. Loudness is a psychoacoustic property of sound, in that it represents our human perception of how “loud” or “soft” sounds of various intensities are. The relationship between intensity and loudness is rather complex, but it is monotonic -- if one sound has a greater intensity than another but is in all other respects similar to it then it will be perceived as being louder. Given the difficulties associated with measuring loudness, we will use intensity as a feature in the experiments below.

Stress reflects the degree of strength or emphasis with which a word, a syllable or phrase is spoken.
The stress plays a very important role in the process of automatic generation of the intonation in the standard Arabic language, where variation in stress may have an effect on the meaning of word. The stressed syllables may be characterized by increased loudness, longer duration or increased fundamental frequency (El-Bakry, Rashad, & Isma'il, 2011).

Determining if a syllable in Arabic is stressed or unstressed depends on two major factors as reported in (Abu-Salim, 1983), first, the internal structure of the syllable particularly 'weight', and second, the position of the stressed syllable and its relation to all other syllables in the same word. So from these two factors we can predict about the positions and levels of the stress.

There are three categories of stress that can be applied to syllables, words and to longer units of speech which are: primary (main) stress, secondary stress and tertiary (weak) stress. Al-Ani (Al-Ani, 1970) notes that every Arabic word contains a single stressed syllable, with the position of the stress within the word depending on the number, length and shape of the syllables in the word. Stress can be used in English as a free morpheme, in particular as a way of marking contrast or unexpectedness ("John stole my bike" suggests that while the speaker wasn't surprised that John stole something they were surprised that what he stole was their bike, "John stole my bike" suggests that they are surprised he stole anything). This use of stress does not occur in Arabic.

There are some rules which govern the place of stress in Arabic (Al-Ani, 1970). Applying these rules can help to enhance Arabic speech technology applications:

- If the word is made up of a sequence of short CV syllables, such as in: 'kataba' /CV-C V/ (he wrote), then the first syllable will receive the primary stress, and the rest of the syllables will get the weak stress.
- If a word includes a single long syllable, such as in the word: 'ka:tab' /CVV-CVC/ (writer), then this syllable will receive the primary stress, and the other syllables will receive the weak stress.
- If there are two or more long syllables, such as in the word: 'mustawda?a:tuhum' /CVC-CVC-CV-CVV-CV-CVC/ (their deposits), (for masculine plural), then the long one which is closest to the end of the word (but not the last one) will take the primary (main)
stress, and the long syllable which is closest to the beginning of the word (not the first one) will get the secondary stress; other syllables receive weak stress.

From these rules we can infer that determining the number of syllables correctly is essential for determining stress. However, there is disagreement about these rules that govern item stress in Arabic (Zaki, Rajouani, Luxey, & Najim, 2002).

Local stress can help the hearer to determine word boundaries. So there has been a number of studies about the importance and analysis of lexical stress in Arabic language such as (Halpern, 2009; Mitchell, 1993; Ryding, 2005); (Z. A. Ali & Ghani, 2014; Chentir et al., 2009) and (Zaki et al., 2002).

3.3.2 Duration

The duration of a sound can be defined as the actual time taken to read and produce the element contents of sound, which can be measured on either a spectrogram or a waveform in seconds or milliseconds. From a phonetic perspective, duration refers to the length of a linguistic unit. Duration can be defined also as the physical property that represents the measured duration of a speech sound from the articulatory and acoustic points perspective (Hassan, 1981).

The duration of certain phonemes depends on the contextual realization of these phonemes characterized both by the nature, size and the location of pauses and of word and syllable boundaries and the stress, etc.

Duration is significant in the Arabic language due to the fact that a difference in the length of a vowel or consonant makes a difference in meaning. Increasing the vowel duration for example may change the meaning of the word as in words: 'katala' /cv-cv-cv/ , (he killed) and 'ka:tala' /cvv-cv-cv/ , (he fought).

Arabic short vowels have similar spectral properties to those of their long vowel version but with longer durations, with long vowels lasting approximately twice as long as short one.
At the level of the syllables of an Arabic word, the durations of the consonants and the vowel within syllable types such as CVVC, CVCC and CVVCC are recognized to be longer than the other remaining types.

The duration of the short vowels is between 100 and 160 ms, the value could increase to 300 ms as (Al-Ani, 1970) has stated in his study, with the long vowels it is between200 and 350 ms.

The duration of consonants depends upon whether they occur initially, medially or finally. It also depends on other conditions, namely on the manner of articulation, voiced or unvoiced, and single or geminated. The duration of consonants varies from 40-375 ms.

The variation of sound duration for the Arabic language has been the subject of various studies, according to several factors: speakers, phonological phenomena, syllabic structures, consonantal geminating and Arabic dialects, etc. Among them, Rajouani, Ouadou, Nour, & Najim (1996) conduct their study on developing a modular text-to-speech system for Arabic by focusing on formalizing a set of rules to control the variations of the Arabic consonants duration in order to improve the acceptability of text-to-speech system.

In the same field of the synthesized speech, (Hifny & Rashwan, 2002) conducted experiments to predict the duration of Arabic phonemes by using an approach based on artificial neural networks (NN), The model achieved correlation coefficient accuracy of 0.91.

Khabet and et al (Khabet, Zemirli, & Mosteghanemi) also described in their article a method to develop a module to generate the sound duration of standard Arabic on continuous speech.

There are a number of studies concerning vowels in Arabic and addressing the duration of vowels in dialects of Arabic language, we state some of them, such as the study of (M. M. Alghamdi, 1998) which employed speakers of three Arabic dialects, Saudi, Sudanese and Egyptian speakers, to pronounce vowels. It has shown from the results that there was significant difference between these three dialects in the first formant frequencies, but in terms of duration, the vowels did not show any significant difference from one another. In addition, it was found that the behavior of the short vowels was less than half of the long vowels.
Another study on vowels in Arabic dialects has been addressed by Zawaydeh (1997) on Ammani–Jordanian Arabic and results were reported that the frequencies for the first formant in the low vowels in the uvularised environment are higher than first frequencies in the plain environment.

Lebanese Arabic has been studied by Khattab & Al-Tamimi (2008). It was shown from the results that there is no significant difference between the durational results for males and females.

Hassan (1981) studied vowel duration in Iraqi Spoken Arabic, aiming to investigate whether the factors governing the systematic variations of vowel duration are phonetic-universal or language-specific phenomena.

However, the vowel duration has the advantage of identifying the vowel and the intelligibility of speech (Ferguson & Kewley-Port, 2007).

### 3.3.3 Intonation

Intonation in many European languages as in Arabic refers to use of pitch melodies over an utterance to convey grammatical form and discourse-level meaning, e.g. questions, statements, contradiction, etc. The acoustical correlate of the intonation contour is the fundamental frequency (F0), as a function of time. The pitch of F0 ranges of female speakers is large compared to male speakers, between 160-400 Hz and between 80-160 Hz respectively.

The shapes of the intonation pattern depends on the sentence type (i.e., statements, commands, question, etc.)

Many works have studied the intonation for the Arabic language according to several factors: speakers, phonological phenomena, syllabic structures, consonantal geminating and Arabic dialects, etc., typically in order to make improvement of Arabic text to speech synthesis. Eldin & Rajouani (1999), in their study discovered the intonation characteristics of interrogatives in standard Arabic to suggest a model for the automatic processing of the intonation pattern in a
Text To Synthesis system. The results show that there is no need to use a syntactic analyzer or a lot of information about the utterance to get natural intonation.

Another study conducted by Rajouani, Loukili, & Najim (1997) described the development of an intonative model for Arabic language of text to speech synthesis.

Zemirli & Khabet (2007) discussed treating intonation under the aspect of the lexical stress with the level of the words in order to improve the naturalness of an Arabic Text To Speech Synthesis system (ARAVOICE)

The intonation is a different of varieties of colloquial Arabic, there are a number of studies on the intonation carried out in that level. For example, of colloquial Emirati Arabic in (Blodgett, Owens, & Rockwood, 2007), Lebanese Arabic in (Chahal, 1999) and (Hellmuth, 2005) with Cairene Arabic speakers, some results show the same pitch accent on every content word and another show the opposite.

### 3.4 Dialects of Arabic

Many varieties of Arabic are spoken across the Arab world, this variety of Arabic can differ in grammatical structure, in the lexicon and in the phonetic system.

Differences in speakers’ phonetic systems are what is generally denoted by the term 'accent'. Accent is the way somebody pronounces words, the prosody of their speech, etc. Dialect is a broader term that refers both to a person’s accent and the grammatical and lexical features of the way that person talks.

The wide variety of Arabic dialects reflects the ethnic and social diversity of its speakers and can differ within the same country.

This difference means that it can be a challenge, for example, for a Palestinian to understand and communicate with a Moroccan.
Arabic dialects include the Eastern dialects of Egypt. Sudanese is spoken in most areas of north-eastern Africa. The Gulf dialect is spoken in Saudi Arabia and parts of Yemen. The Levantine dialect is spoken in the region around Lebanon and Syria, while the Western dialects of North Africa of Maghreb Arabic are spoken in Saharan Africa, as well as Tunisia, Algeria, and Morocco.

The phonetic implementation of the MSA vowel system differs according to the accents associated with the different dialects (M. M. Alghamdi, 1998). For instance, the Levantine dialect has at least two extra types of diphthongs /aj/ and /aw/. Similarly, the Egyptian dialect has other extra vowels (M. Alghamdi, 2001). In addition, there are major phonological differences in recognition of specific phonemes, such as (ق ، ح ، ث ، ض).

Generally, understanding and modeling the accents associated with the various dialects of Arabic is a major challenge for current research on Arabic speech science and technology. In the speech synthesis field for example, (Tomokiyo et al., 2003) reported several reasons for considering dialects as a problem for speech synthesis, first deciding whether to consider modern standard Arabic (MSA) or the dialects to be generated. The second problem is that the limitation of the listener base who understand MSA or dialects has to be considered. MSA is widely understood only by communities with formal education, however, so its listener base is also limited.

In our work we deal with Modern Standard Arabic (MSA); although MSA is not every speaker's primary version of the language, most people can understand it, since it is the standard form of the language in formal settings and in broadcast media and hence at least comprehensible to most people, even if it is not widely used in informal everyday speech.

### 3.5 Summary

Arabic is a language which offers a number of challenges for speech researchers. In this chapter we have highlighted some of the characteristics of the Arabic language in spoken and written form which may affect speech research.
The major factor is diacritization, where vowels are represented partially and must be inferred. We discussed two obvious approaches to solving this problem, inferring the vowels automatically or using a dictionary. The correspondence between the written form of a word and its pronunciation in Arabic is another major factor that raises several issues. Here we focus on highlighting the Arabic phonetic features that make the language hard to process, describing some processes of phonological rules and explaining the designed letter-to-sound rules which can be used to improve many applications of Arabic speech, as discussed in subsection 3.2.2.2

Prosody features as an important factor of voice quality, which includes the stress, duration and intonation which might influence Arabic speech are discussed in Section 3.3. Moreover, we discussed the diversity of dialects of Arabic language which is considered problematic because they lack a fixed orthography; we also see major phonological differences in the realisation of specific phonemes, as explained in Section 3.4. However, there are other challenges for Arabic speech research such as the availability of databases of Arabic emotional speech, which are discussed in Chapter 4. It is worth saying that there are other challenges facing speech processing of Arabic which can be found in (Farghaly & Shaalan, 2009).

However, in our work, in order to overcome the problem of diacritization, we will suppose that the text is fully diacritized in modern standard Arabic.
CHAPTER FOUR: CONSTRUCTION AND VALIDATION OF THE CORPUS

4.1 Introduction

Emotional speech corpora are the most vital linguistic resources for studying emotional speech, with a range of applications including recognition and synthesizes of emotional speech, as well as for applications of other purposes such as customer-care centre and tutoring systems.

Compared to other languages, few corpora have been developed recently involving speech studies of Arabic language, exceptions being (M. Alghamdi et al., 2008), (Makhoul, Zawaydeh, Choi, & Stallard, 2005), (Harrag & Mohamadi, 2010), (Abushariah et al., 2012),(Selouani & Boudraa, 2010),(Col, LaRocca, & Chouairi, 2002).("ELRA - ELRA-S0219 : NEMLAR Broadcast News Speech Corpus," 2017). Most of these corpora have been designed for specific applications and in a specific dialect of Arabic, most of them are private and only available with a membership fee or exclusively reserved for subscribers, such as the speech corpora available from the linguistic data consortium (LDC) or European language resources association (ELRA). Moreover, these corpora are not marked up for expressive speech so cannot be exploited for the analysis of expressivity in Arabic language.

Therefore, there is a need for an emotional speech corpus to be available for Arabic in order to progress the study of emotion in Arabic speech. Unfortunately, collecting such data takes an enormous amount of time and effort which we are not in a position to give. Therefore, what we intend to do is gather sufficient data for the study of the relevant prosodic features, i.e., Pitch, Intensity and Duration. Achieving this goal requires the design of a specific database.

This chapter introduces the stages of the construction of an Arabic emotional speech corpus with which to conduct our experiments. It includes a set of sentences collected from Arabic news websites recorded by a professional actor and evaluated by Arabic-speaking listeners.
Firstly, a brief summary of the emotional corpus is introduced, including classification. Secondly, the construction stages for our corpus are described in subsections, discussing the preparation of data through choice of actors, types of emotions and text material, the process of data recording and evaluation, followed by the presentation of the evaluation results. Finally, a summary concludes this chapter.

4.2 Corpus design

Emotional speech corpora describe what type of records will be used in experimental speech processing of emotions. Such databases are mostly dependent on specific study goals and many considerations are usually taken into account in order to perform the recordings in specific environments and to distinct specifications.

Researchers (Ververidis & Kotropoulos, 2006), (Cullen, Vaughan, & Kousidis, 2008), (Scherer, 2003) have referred to different types of corpora which have been used for the study of emotions in speech. We list three main categories in this field which are:

- Elicited (Induced) emotional speech corpora.
- Acted (Simulated) emotional speech corpora.
- Spontaneous (Natural) emotional speech corpora.

Each type of emotional speech database has its particular advantages and drawbacks.

*Elicited emotional speech corpora* is speech in which the emotions are induced by putting a speaker under appropriate circumstances (Navas, Castelruiz, Luengo, Sánchez, & Hernáez, 2004). Sometimes these databases are recorded by asking the subjects to take part in verbal interaction with a computer whose speech responses are in turn controlled by the human being without the knowledge of the subjects (Batliner, Fischer, Huber, Spilker, & Nöth, 2000).

There are two approaches to inducing affective arousal. A direct way is the use of psychoactive drugs. An example of this method is found in (Helfrich, Standke, & Scherer, 1984) where their study compared the effects of antidepressant drugs on several vocal parameters with those of a placebo over several hours.
Another way is the indirect way favored by experimental psychologists because of the degree of control it affords. This is performed via stress induction through the completion of hard tasks under time pressure. However, the use of elicited emotion is potentially problematic. There are a number of drawbacks. One is the difficulty of generating similar emotional states in each of the individuals. Another disadvantage is the generally weaker results produced by these procedures.

**Acted emotional speech corpora** : is using professional trained people or radio actors to record and express emotional utterances in a conscious way.

Such a database can be collected from TV, soaps, plays, movies or serials clips. Generally, as (C. E. Williams & Stevens, 1972), (El Ayadi et al., 2011) point out, simulated emotions tend to be more clearly recognisable than real ones, so this is a widely used technique.

The main advantage of recording acted databases is that many aspects of the recording can be carefully and systematically controlled such as the phonetic and prosodic speech content (Le & Provost, 2013).

Problematic issues can arise from an extravagance of expressed emotion from actors during recording, partly due to lack of acting ability and partly due to personal reactions to different emotional situations. That can result in obtained data which is unrealistic in its ecological validity (Cullen et al., 2008). In addition, some researchers (Frijda, 1988), (Darwin, 1956) argue that there are physiological aspects to emotion inducing uncontrollable changes in speech (Johnstone, 1996) that reflect the underlying emotion. The presence of these uncontrollable changes may not be present in acted speech.

In summary, in using this method of recording data we must test our recordings by annotating and assessing them for various emotions in order to recognise which emotions are being expressed correctly by the actor.

However, there are many issues relating to corpora, such as size of database being insufficient for development, training, and testing of emotion systems (Douglas-Cowie, Campbell, Cowie, & Roach, 2003), the need to employ more professional speakers with different linguistic and cultural backgrounds; the recording of different sentences that would bring new phonemic and
phonetic contexts, and increasing the number of simulated expressions. All of these issues can be controlled for and taken into consideration when mapping out the goals of the research.

**Spontaneous emotional speech corpora**: is simply natural speech where all emotions are collected or recorded in a real time environment, such as cockpit recordings, live news reports, recording people on the street or on reality TV shows, conversations on TV, call center conversations (Lee & Narayanan, 2005), etc.

Although, this kind of corpus tends to be more natural than other types there are a number of serious drawbacks. Most importantly, it is uncontrolled and difficult to obtain and annotate. There are no ideal conditions for recording this type of corpus, a challenge which needs to be considered.

The resources mentioned are subject to bad recording quality, with a lack of consistency across broadcasts which leads to varying quality of voice samples depending on the nature of the program, whether it is recorded in a studio or outside in public spaces (as many reality television programs are). Additionally, some may contain noise added in the real time environment. Various other factors can affect the audio quality: noise from studio audiences, people talking across each other and environmental noise from outside broadcasts. The equipment used can also affect sound quality.

A major problem is the low levels of emotion expressed, meaning that huge quantities of material have to be marked up before obtaining anything of use.

However, collecting data from broadcast media to be used as acted or spontaneous speech is argued against by anthropological research, finding that the presence of the researcher and the equipment distorts the reality of the situation, causing people to feel constrained and to act unnaturally (Becker & Geer, 1957).

As mentioned above, due to the limitations of emotional corpora for Arabic speech compared to other languages, research on emotion in Arabic amounts to little or nothing. There are also difficulties in finding corpora containing suitable emotional texts for Arabic. This scarcity of relevant material from the media led to the present decision to use represented emotions for this study by recording a suitable number of sentences expressing emotions by a professional actor.
This has the advantage of allowing full control over the recording environment with respect to emotional content.

The next sections present the design procedure of an Arabic emotional speech corpus through the stages of selecting an actor to record texts expressing a major emotion, evaluating the recording through perceptual tests and using the results to assess and specify the selected sentences to be used to conduct our initial experiments.

4.2.1 Selection of actor

The first important task of constructing the emotional database is the selection of the actor, as the main factor responsible for producing the expressive speech which is to be as close to the desired emotion as possible, affecting the recordings and then the overall quality of our experiments.

There are a set of criteria for actor selection, primarily:

- Ability to express the selected emotions.
- Easy access to the speaker when needed.
- Ability to make spontaneous sounding recordings.

The researcher made the decision to record herself enacting the data for various reasons:

She meets the necessary conditions, having good experience of acting and of performing recordings. She has worked as a presenter of an Arabic radio station, presenting many types of programs which required acting such as news which requires for example the ability of delivering the content in a certain emotion to make the speech and the content more believable. And recording children’s stories, and she is therefore familiar with the recording environment.

Recording the data herself gives the researcher the advantage of a high level of attention to her work.

Despite the researcher being familiar with recording written material and with the environment of recording, as she has recorded Arabic speech in her research towards obtaining her Masters degree in the same area (Bin Othman, 2010) there nonetheless is a risk, because an actor who knows the nature of the research and its goals might be unconsciously biased about what she is
doing and might over-emphasise phonemes which she thinks are interesting so that the recordings may be influenced by these preconceptions.

While it would have been better to use recordings from several speakers, the fact that our subjects were able to recognise the emotion that was intended for the majority of the recordings suggests that the information that is required for expressing emotion was present in these recordings. Given the difficulty of collecting data from multiple speakers and annotating it with emotions, the fact that the intended emotions were detectable in our recordings gives us some confidence that the experiments are showing us something about how emotions are encoded, though in future work we would want to obtain recordings from more speakers.

### 4.2.2 Selection of emotions

The second step in creating an emotional speech corpus is the selection of emotions. One actor simulated the following three human emotions: sadness, happiness, anger, as well as a neutral state.

Although many categories of emotion can be discovered in speech generally, these three emotions are the most commonly recognised as basic emotions in the literature when attempts are made to classify them into meaningful groups (Izard, 2001; Plutchik, 1980; Parrott, 2001). These three emotions can be further divided into the set of emotions known as the “big six emotions” (P., Ekman, W. V. Friesen, & Ellsworth, 1982), (Douglas-Cowie et al., 2003): fear, anger, happiness, sadness, surprise, disgust, which have in turn been split into finer-grained categories, though there is less general agreement about the division beyond the big six. Given the difficulty in acting out a very fine-grained set of emotions, we chose to use the three that are most widely recognised as basic – in particular, it is extremely difficult to deliberately differentiate between anger, fear and disgust, and attempting to do so would have been likely to produce confusion for the listener.
4.2.3 Design of material texts

The third step in designing the corpus is selecting the set of sentences. Today, the newspaper is the most accessible source of modern Arabic. The BBC website is one of the most famous electronic newspapers of Arabic text for Arabic speakers. The dataset is an electronic archive for the newspaper of the year 2015, presented as windows HTML files. A set of MSA sentences of different lengths was selected randomly from the Arabic BBC News website. These sentences were the headlines of news reports. Table 4.1 shows these selected sentences.

<table>
<thead>
<tr>
<th>Sentence (Arabic)</th>
<th>Sentence (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 Dubay tuxa't'u litazwydi rijAli AlOIr't'Al' biOjhzapi t'ayarAnK mahmwl īp</td>
<td>Dubai plans to provide firefighters with Air Portable</td>
</tr>
<tr>
<td>S2 Rafa' darajapi AltOhubi limuw'Ajahapi Al'd'd'Abi AlduxAny fiy bikyn</td>
<td>Raise the degree of preparedness to meet the smog in Beikine</td>
</tr>
<tr>
<td>S3 Mu'tamaru AltGayuri AlmanAxy fiy P'Arys qad yakwnu nuq'tapa tahawl</td>
<td>Climate Change Conference Paris may be a turning point</td>
</tr>
<tr>
<td>S4 TaOmynu AlOnD'impi fiy SarikAti AlAntarnit Albrty'A'niyypa yax'd'aEu limur'Aja'yAtK s'Arimap</td>
<td>Insurance systems British Internet companies are subject to a strict review</td>
</tr>
<tr>
<td>S5 Tahfyzu Almux biAlkhorabA' yusA?íd'fu fiy Alta'Afy mina Alsaktapi AldimAGypa</td>
<td>Stimulating brain electricity helps recovery from stroke</td>
</tr>
<tr>
<td>S6 Riholpu fad'A'K Owvwhyp lltOkudi min wujwdi hayApK ʔITY Almaryx</td>
<td>European space flight to discover whether there is life on Mars</td>
</tr>
</tbody>
</table>

Table 4.1: Selected sentences

---

Table 4.1 shows the selected sentences, their meaning in English and their phonetic transcription. There are six sentences which contain from 5-9 words and they are diacritized in advance. The sentences contain most of the Arabic letters and phonemes, the letters are around 27 consonants with three short vowels (ُ /a/, ُ /u/, ُ /i/) and tanween Kasrah (ِ /en/), two long vowels (u:, i:), because our aim is to investigate which phonemes are the most significant.

Figure 4.1 shows the frequency of phonemes in these sentences.

Figure 4.1: Phonemes frequency

Here are some reasons why these six sentences are useful for the purpose of emotional speech recognition in Arabic.

- The dataset consists of real text of spoken Arabic, which requires a minimum of processing.

- These six sentences might be a suitable number for making a judgment about the emotion expressed, they cover all the sounds of Arabic and they are not intrinsically emotive and therefore it is possible to try reading them in different emotions and the listeners will be not swayed in their judgment. Also they are easy to speak immediately from memory, there is no need for a longer process of memorising or reading them off a paper, which may lead to a lecturing style.
An actor may experience some difficulty expressing emotion when the content of sentence refers to a certain emotion, for example, expressing sincere happiness when the text subject is death, especially when the speaker is not a professional actor. Because the meaning and content of sentences can play a role in the speaker's acting during recording, this can affect assessment of these sentences by the listener in the perceptual test phase. So most of these selected sentences are not intrinsically emotional and tend to be sentences taken from news websites which hold neutral emotion without any clues of emotional content. Therefore, the classification and recognition of these sentences would be on the basis of how the sound was rather than of the intrinsic emotional content.

4.2.4 Database collection

Recording of emotional speech data collections is definitely important for researchers interested in emotional speech analysis. In this section, the process of the creation of our own corpus will be presented as a subsection as following:

4.2.4.1 Recording process of data

The sentences are recorded in the Arabic language in the following manner. The first sentence was chosen and recorded directly to the computer four times in different expressive styles (sadness, happiness, anger and neutral). By portraying them emotionally using acted emotional states without interruption, the actor can repeat them many times until she is satisfied with the emotional tone of what she has simulated for the sentence in order to make the recording of the expression as realistic as possible. However, we have to be careful about recording the same sentence over several repetitions because that may cause the emotion to become dull and hence unrecognisable.

There are a number of problems with acted speech(Campbell, 2000):
Firstly, acting out emotions like anger or sadness while speaking sentences with unrelated content can sometimes be difficult and complex. It requires a lot of concentration and a certain theatrical capacity in the actor.

Secondly, overacting or exaggeration by the actor may result in unnatural expression, which differs from speech produced by a speaker when experiencing a genuine emotion (Scherer, Banse, Wallbott, & Goldbeck, 1991).

However, in an attempt to overcome the problem of exaggeration of expressing emotions, we use perceptual tests in order to evaluate and validate our recordings.

The data is recorded in a quiet room on a laptop by one female actor who is able to portray emotions in a convincing manner, four times over as mentioned (these amounted to 24 sentences (6 sentences x 4 emotions)).

A high-quality microphone was used to make the recordings. The distance between the speaker’s mouth and microphone was approximately 5-7 cm, and the Praat software program (Boersma & D, 2017) was used to control the recording. Figure 4.2 shows a screenshot of the sound recorder of the Praat program. A sampling frequency of 44100 KHz was used, 16 bits depth and mono channel.

![Figure 4.2: Sound recorder of Praat program](image-url)
4.2.4.2 Labeling the data

The recordings were saved in .wav format and labeled as the following: \textbf{SxxEyy}, where $S$ refers to sentence and $xx$ refers to the number of sentences and $E$ refers to emotion and $yy$ refers to the code of emotion as listed in the following table:

<table>
<thead>
<tr>
<th>Emotions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E00</td>
<td>Neutral</td>
</tr>
<tr>
<td>E01</td>
<td>Happiness</td>
</tr>
<tr>
<td>E02</td>
<td>Sadness</td>
</tr>
<tr>
<td>E03</td>
<td>Anger</td>
</tr>
</tbody>
</table>

Table 4.2: Labeling for emotions

For example: S00E00 means the first sentence with neutral emotion. Numbering of the six sentences goes from 00 up to 05.

The total number of files recorded is thus 24 files (1 speaker x 4 emotions x 6 sentences). The choice of which of the emotive sentences should be included in our study is evaluated in the next section.

4.2.4.3 Evaluation of the data

As mentioned in section 4.2.4.1 above, in order to record a certain emotion of emotive sentences, the actor could tend to over-act in order to record the data quickly and easily. Therefore, the recording has to be verified by carrying out a perception test to check whether human listeners can recognise the recorded emotion type and if it is close to the desired emotion; also if the actor is able to portray emotions convincingly. The procedure for this test was adopted from Meftah et.al (Meftah, Alotaibi, & Selouani, 2014).

The number of subjects selected amounted to 12 listeners (11 female and 1 male) who did not receive any training for the testing procedure due to the simplicity of the task. The subjects have these two important properties:

- They are fluent in spoken Arabic and have to listened to Arabic recordings.
- They do not have hearing-related health problems.
The subjects were told that the aim of the experiment was to recognise the emotion from the set of recordings and that they could listen to each utterance twice before they had to assign an emotional state.

A questionnaire was designed precisely to assess the correct emotion of these speech recordings in order to ensure data validation and establish an indication of how well the actor was expressing the intended emotion performed over the set of recordings. The questionnaire designed in paper format lists the recorded speech data in numbers with four different emotions. The listener had to play and listen to audio files and fill that part in by assigning the percentage of emotion they thought each sentence contained. The audio clips are not ordered as recorded but listed in a random way. However, we allowed them to repeat each file twice before they needed to make their decision. The questionnaire is presented in Appendix A.

The subjects were given 24 recordings of sentences to listen to and were asked to assign the correct emotion for each by choosing one or two emotions from a list of four possible emotions. For example, the listener could specify that a given sentence would comprise either 30% normal emotion and 70% sadness emotion, or 100% sadness. We allowed the subjects to choose two possible emotions and to give each of these a degree because we thought that there could be overlap between two emotions. Letting people choose two emotions, rather than one, would give us an indication of whether there is any degree of overlap between any two emotions.

The result of this questionnaire helps us to select the most appropriate sentences to be used in our experiments.

The reports on the results obtained and discussions on the findings are presented in the following section.
4.2.4.4 Results and analysis

The perception test results were analysed to determine the rate of emotion recognition. The results are shown in the following tables. Table 4.3 shows an example of the listener's response, including their percentage ratings.

<table>
<thead>
<tr>
<th>No. Utterance</th>
<th>Emotions No.</th>
<th>Neutral (%)</th>
<th>Happy (%)</th>
<th>Sadness (%)</th>
<th>Anger (%)</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:S00E02</td>
<td></td>
<td>80</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: An example of listener response

As mentioned before, and from Table 4.3, we can see that subjects were asked to express two views, for example, regarding the sentence S00 with sad emotion E02, the listener defines its expression of emotion as 80% neutral and 20% sad. Because neutral emotion is an absence of emotion, this choice means sentence S00 does not express much emotion, but what it does express is sadness. So if it had been intended as sad, this would actually be the right emotion, just not very much of it.

Four recordings, each intended to express one of the four emotions, were made of six distinct sentences. Twelve subjects each listened to these 24 recordings, i.e. each subject listened to six instances of a neutral recording, six instances of a sad recording, and so on, making 72 judgments of recordings of each intended class. Given that neutral means that no emotion was expressed, rather than being a label for a specific emotion, we treated cases where an annotator marked an utterance as being partly neutral and partly some concrete emotion as indicating that they felt that utterance was weakly expressive of the concrete emotion. In the tables below, cases where every annotator has given an utterance a mixture of neutral and the same concrete emotion, and the overall score for the concrete emotion was greater than 5, have been labelled as belonging to the concrete emotion rather than as neutral. We offered listeners the chance of assigning multiple emotions to a utterance to allow for situations where a single utterance was felt to express two differing concrete emotions, e.g. anger and sadness. It turned out that this very seldom happened, but that people used the possibility of assigning a mixture of a concrete
emotion and neutral, seemingly as a way of saying that the utterance expressed the concrete emotion, just not very strongly. We therefore assigned labels as follows:

- If the highest score was something other than neutral and was over five times as high as the score for any other concrete score then we accepted that.
- If everyone gave a mixture of neutral and the same other concrete emotion, and the overall score for the concrete emotion was reasonably high (the threshold for this was chosen manually) it seemed sensible to assume the utterance should be classified as belonging to the concrete emotion rather than neutral. In Table 4.6, for instance, sentence S2 has been assigned H because, whilst most of the annotators gave a mixture of H and N, the score for H, at 6.1, is higher than the score for N, at 5.7, and is much more than five times the score for A. S1, S3, S4, S5 and S6 have all been classified as H because no annotator has assigned any emotion other than H or N and the overall scores for H are all greater than 5.

This strategy allows us to cope with the unexpected way that annotators used a mixture of neutral and a concrete emotion to express the degree to which they felt that an utterance expressed that emotion: we accepted a concrete assignment if it had the highest score (including by comparison with neutral) or if every single person gave the same mixture. The overall result of the test is summarised in Table 4.4.

<table>
<thead>
<tr>
<th>Intended emotion</th>
<th>Assigned emotion</th>
<th>Neutral</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Anger</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>Neutral</td>
<td>70</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>97%</td>
</tr>
<tr>
<td>Sadness</td>
<td>2</td>
<td>64</td>
<td>4</td>
<td>2</td>
<td></td>
<td>89%</td>
</tr>
<tr>
<td>Happiness</td>
<td>5</td>
<td>0</td>
<td>66</td>
<td>1</td>
<td></td>
<td>92%</td>
</tr>
<tr>
<td>Anger</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>59</td>
<td></td>
<td>82%</td>
</tr>
</tbody>
</table>

Table 4.4: Confusion matrix for the listeners’ responses

Table 4.4 shows a confusion matrix of the subjective human evaluation of sentences, for example, the first row shows that 70 of the 72 utterances that were intended to be neutral were judged as neutral, 0 as sad, 2 as happy and 0 as angry. The recognition accuracy of neutral is approximately 97% which is the highest percentage of correct recognition. We can see that the
most easily recognisable category is neutral and then happy at 92% followed by sad at 89% and angry at 82% of utterances.

We also find that subjects sometimes show confusion in differentiating happy, sad and angry from neutral. However if we assumed the absence of a neutral choice, we can see that all three emotions would be identified correctly nearly all the time. Table 4.5 shows what subjects choose instead of neutral emotion.

<table>
<thead>
<tr>
<th>Intended emotion</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Anger</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sadness</td>
<td>64</td>
<td>4</td>
<td>2</td>
<td>96%</td>
</tr>
<tr>
<td>Happiness</td>
<td>0</td>
<td>66</td>
<td>1</td>
<td>99%</td>
</tr>
<tr>
<td>Anger</td>
<td>8</td>
<td>1</td>
<td>59</td>
<td>87%</td>
</tr>
</tbody>
</table>

Table 4.5: Confusion matrix for the listeners responses without neutral emotion

The table above presents the recognition rate of three emotions. It is clear that the happy expression was more recognisable than the others at 99%. The angry and sad expressions were not recognised as well, but the result is still higher than chance.

There are four tables of target emotions (Happy, Angry, Sad, and Neutral) which show the answers given by subjects for each simulated expression of our recordings. These tables allow us to decide which expressive utterance should be chosen in order to conduct our experiment of given expressions. Table 4.6 shows the results for the recordings which were intended to express happiness.
From this table we can see that almost all the subjects heard a degree of happiness in almost every utterance. The exception was listener number 3, who heard anger rather than happiness for sentence 2. There are five instances out of 72 where the subjects did not hear any emotion (referred to neutral emotion) when the actor tried to express the sound as happy. Meanwhile in the remaining 68 instances happiness, often mixed with a degree of neutrality, was heard.

As noted above, we have labelled the cases where there is no dissenting label and the overall score for H is greater than 5 as being H.

It is clear that there is little evidence of confusing it with sadness or anger in these sentences – people heard them all as being happy, just not terribly strongly so.
From confusion matrix Table 4.4, we found that all subjects were able to recognise neutral emotion, i.e. utterances which were not intended to carry any emotion, successfully for all sentences to the highest degree. In every case of a sentence that was intended to sound neutral the score was overwhelmingly in favor of that classification, though there are traces of other emotions. There is another choice of target emotion that can arise. For example, the subjects suggest that sentences (S1, S2, S5, S6) may have happy or sad emotions while one of subjects (L7) think that S3 may have also anger and some of subjects (L2, L3, L9, L11, L12) heard the sentences S4 as happy in addition to neutral. However, the occurrence of this was rare compared to the score of target emotions recognised correctly. Comparing Table 4.7 with Table 4.4, we can note that the neutral emotion was commonly confused by the listeners. Some of them think
that neutral sentences tend to indicate happiness; others selected sadness and anger, suggesting human evaluators experience difficulty in distinguishing neutral from the other three emotions.

<table>
<thead>
<tr>
<th>Target Emotion: SAD:E02</th>
<th>SUBJECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec. Of Sad</td>
<td>Emotions</td>
</tr>
<tr>
<td>S1</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>S2</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>S3</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>S4</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>S5</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>S6</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
</tbody>
</table>

Table 4.8: Evaluation of sad emotion by listeners

Here, the majority of subjects identified all the recordings as the target emotion, namely sadness. Again a number of utterances were marked as a mixture of sad and neutral, which indicates that the subjects felt that the target emotion was not expressed very strongly. Curiously subject L10 consistently assigned happiness rather than sadness as the main emotion to all these utterances: it is hard to see any reason for this.

There are three instances where subjects (L7, L10, L12) think that a sad sound could have a degree of anger emotion. Only the subject (L10) assigns sadness utterances in three cases of
utterances (S1, S3, S6) that could be completely happy and in two instances where subjects (L1, L4) said the sound of utterance S6 was completely neutral in emotion.

<table>
<thead>
<tr>
<th>Target Emotion: ANGER</th>
<th>Emotions</th>
<th>SUBJECTS</th>
<th>SUM</th>
<th>Rec. Emo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec. Of Anger</td>
<td></td>
<td>L1</td>
<td>L2</td>
<td>L3</td>
</tr>
<tr>
<td>S1</td>
<td>H</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.9</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>0.1</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>S2</td>
<td>H</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>0.2</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>S3</td>
<td>H</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.8</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>0.2</td>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td>S4</td>
<td>H</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>S5</td>
<td>H</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>1.0</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>S6</td>
<td>H</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 4.9: Evaluation of anger emotion by listeners

Again, in nearly every case the intended emotion was given the highest score out of the three main emotions, with an admixture of neutral indicating the degree to which the emotion was felt to be expressed. In a small number of cases sadness was mixed in with anger, and a few listeners assigned sadness rather than anger as the main emotion, though for every utterance the general assignment was anger.

From this table we can see that the anger emotion was recognised incorrectly in some cases, being confused with other emotions. A small number of cases (2) of subjects (L1, L9) heard anger as a happy expression with utterance S6.
In most cases, if the subjects misrecognized anger then they classified it as sadness but with a small number in (5) cases, with utterances S1, S4,S5 and S6. Overall the neutral sentences were the most easily recognisable by the listeners. In other cases listeners often assigned a mixture of neutral and the target emotion. This is to be expected, given that doing this can be used as a way of marking the intensity of the expressed emotion, and may even be taken as a positive sign that actor did not over-act during recordings.

From Tables 4.8 and 4.9, we can see that some individuals classified anger as sadness and sadness as anger, though even in these cases the majority vote was always for the target emotion. This suggests that people can find it difficult to distinguish the difference in spoken language between sadness and anger.

Psychologists suggest that emotions such as sadness can be detected beneath anger. Sadness is usually an uncomfortable experience as it makes people feel vulnerable and lacking in control. Due to this, people tend to go out of their way to avoid such feelings. One subconscious method is to shift into an angry mode.

We can conclude from the previous four tables that subjects were able to recognise the target emotion of recordings correctly with few mistakes. When we gave the subjects the freedom of choosing two emotions if they are unsure what the emotion is, they always choose a mixture of neutral and something else since neutral is absence of emotion, choosing neutral and another emotion is a way of expressing the degree to which the actor had expressed the target emotion.

Using the scoring algorithm outlined above, we obtained the results in Table 4.6-4.9, summarised in Table 4.10, which show that where our subjects heard any emotion at all they heard the one that the actor had intended to express. There are cases where they felt that the given emotion did not come across very clearly, but where there was a reasonable level of evidence that an utterance expressed an emotion (either because that emotion had the highest overall score or because the only other label assigned was neutral) the intended emotion was recognised.

Our goal is to perform human emotion classification of recorded speech through listening sessions in order to gain a set of emotive sentences. The following table shows that summaries of
the recordings were recognised and selected. We selected the ones with a clear majority vote for a single emotion other than neutral, since these cases clearly did express that emotion.

<table>
<thead>
<tr>
<th>Emotion Sentences</th>
<th>Happy</th>
<th>Neutral</th>
<th>Anger</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S5</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>S6</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Table 4.10: The expressive data recognised by subjective human evaluation

From Table 4.10, the sentence recordings S1, S2, S3 and S4 were recognised correctly all the time for all four emotions. Meanwhile, in recording S5 anger was classified incorrectly. In Table 4.9 we can see that subjects were confused between anger and sadness and it can be seen that subjects think that utterance S6 which was intended to express anger could also include a degree of sadness and happiness. There are two cases where subjects are completely convinced that the sound is sadness and not anger. For sadness in Table 4.8, the recording S6 was confused by the subjects as happiness, so we are not going to choose this recording.

For carrying out the next experiments, we only need the emotional speech data that can be recognised by most listeners. Given the results on Table 4.10, we omitted the ‘angry’ versions of S5 and S6 and the ‘sad’ version of S6 – if people did not recognise these emotions in these recordings then they must, by definition, fail to express them, and hence would be poor examples to use for training and testing a classifier. Thus only 21 sentences were selected from the original 24 sentences. These sentences are then put into an emotional database and can then be used as prosodic templates in the next experiments to extract information on acoustic parameters of phonemes.
Table 4.11 shows the number of the final expressive data to be selected.

<table>
<thead>
<tr>
<th>Desired Emotion</th>
<th>Number of selected sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>6</td>
</tr>
<tr>
<td>Sadness</td>
<td>5</td>
</tr>
<tr>
<td>Happiness</td>
<td>6</td>
</tr>
<tr>
<td>Anger</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total no:</strong></td>
<td><strong>21</strong></td>
</tr>
</tbody>
</table>

Table 4.11: The number of final expressive selected sentences

### 4.3 Summary

We have designed our own emotional speech corpus for Arabic containing six short sentences with four emotions each, namely Happiness, Sadness, Anger, and in addition Neutral emotion. These emotions are expressed by one female actor then a perceptual test involving a number of listeners is carried out in order to evaluate the degree to which the actor is able to express the desired emotions and to validate the recorded data for recognition of emotive elements.

The results of the evaluation test show that subjects are able to identify the emotions in target utterances. It can also be observed that subjects use the presence of neutral in the set of options as a way of expressing the confidence about the level of emotion the actor expressed. However, neutral utterances are needed to be compared with other emotive utterances.

The result also shows that people are good at recognising happy expressions, but less confident in recognising sad and angry expressions. Once the initial corpus had been created, analysed and evaluated the next step is to extract the basic prosodic features of these utterances, which are discussed in the next chapter.
CHAPTER FIVE: EXTRACTING AND ANALYSING THE RELEVANT PROSODIC FEATURES

5.1 Introduction

After selecting suitable sentences for use in our experiments, the next step is segmentation and labeling of our dataset, the purpose of which is to extract the prosodic features which are associated with four emotions (neutral state, happiness, sadness and anger) and which are involved in the actor's expression of speech.

This step is divided into two stages. The first stage is annotation, in which the recorded speech is transcribed and labeled, and the second is a process of extracting important prosodic information (pitch, duration and intensity) about emotional speech. Figure 5.1 shows this process.

5.2 The Annotation Process

In this stage, the recordings were segmented and labeled in order to identify the boundaries of the individual phonemes on the basis of the audio recordings and their orthographic transcriptions in the speech samples. This process is called forced alignment of phonemes or phonetic segmentation, a task which could be done manually or automatically. In the manual approach, we select the sound waves, listen to them, and mark them up along with spectrogram representations in order to decide where exactly to place the boundary of each phoneme. An
annotation tool, such as in Praat (Boersma & D, 2017) or ELAN (Rosenfelder, 2011), can be used for annotation of speech data manually.

This approach is considered the most accurate, but there is a major drawback represented by the time needed to concentrate on each chunk of phonemes from every single speech sample in the corpus. Some estimates put this at 800 times longer than real-time (Schiel, 2003).

We therefore decided to use an automatic method of carrying out this task. There is a set of algorithms for alignment, such as FST Aligner (Hazen, 2006), FST Aligner with disfluency model (Liu, Shriberg, & Stolcke, 2003), and Viterbi Alignment ("Viterbi algorithm - Wikipedia," 2017). We chose to use Viterbi alignment.

An HMM recogniser in the HTK (Sjölander, 2009) is used for the forced alignment; that is, a known sequence of phonemes is supplied to the Viterbi algorithm to segment and align the transcribed data. The main reasons for this choice are that we are already familiar with this toolkit for recognition of Arabic (Bin Othman, 2010), and it is well-known and widely used for the purpose of phonetic segmentation.

This approach requires many steps to prepare the data before using the automatic tool for the segmentation. In the next section, we will give a brief explanation of the HTK toolkit that will be used to align each utterance by using the Viterbi-based HVite tool in the HTK.

5.2.1 Overview of the HTK Toolkit

The name HTK refers to the “Hidden Markov Model Toolkit” developed by Cambridge University Engineering Department (CUED) in order to build and manipulate Hidden Markov Models (HMMs) (Sjölander, 2009).

The main function of the HTK is in speech recognition research and it involves two major processing stages which are shown in Figure 5.2:
**Training Phase:** This involves the use of a set of training tools to estimate the parameters of a prescribed set of HMMs. This is done by using training utterances and their associated transcriptions.

**Recognition Phase:** This involves using HTK recognition tools for the production of transcriptions of unknown speech utterances.

As mentioned before, the HTK toolkit is used to automatically segment and label the phonemes of each syllable to identify the time when each word in the transcription was spoken in the utterance. The information will be saved and converted to a suitable format (TextGrid) for acoustic analysis.

In our case, we model the speech input as a Hidden Markov Model (HMM) which takes a sequence of observed features from tokens (acoustic vectors) that can be used to decide which state the system is in, and give them some label (phones). This is done by modeling each label/phone as a sequence of “hidden” states.

We train the model to learn the characteristics of each state and the probability of being in a given state, given what the spectrum is doing at that point.

During training, observations are paired with labels so that transition probabilities between states and model vectors can be learned.

In order to identify the speech segments corresponding to individual phonemes, we train a recognition model using the full set of utterances and then use this to carry out forced alignment.
on each utterance in turn. Training such a model involves starting with 'flat start' models with uniform estimates of phoneme length and refining these to make better models and hence to produce better estimates of phoneme lengths. Whilst the training steps do include a forced alignment step, there are subsequent steps which make use of the results of this step, so that the best estimates of where the phoneme boundaries lie comes from the final trained model.

However, using HMMs of phones and the acoustics to determine where boundaries should be placed has some difficulties, including bad transcriptions, unknown words, overlap, noise, and code-switching.

In the following sections, we will explain the process for annotating our data.

5.2.2 Requirements of the HTK Toolkit

In order to conduct our experiments using the HTK, there are several important requirements for the proper functioning of the toolkit components. These prerequisites are briefly stated below:

1. Creating a transcription file for training data based on different aspects of the words.

2. Constructing phonetic transcriptions for the model that contain the pronunciation information of words derived from processing the transcription file.

3. Building the task grammar (grammar files) which consists of a set of variables defined by regular expressions, which depend on a finite state-based grammar. In our experiment, different grammar files were created called gram.txt.

4. A set of folders: This includes the training data folder (consisting of wave files - the actual recordings). In this project, our experiments will use a dataset of recording containing about 21 files. Also included in this set are the test data folder (consisting of the speech files of the test recogniser), and the HMM folders (containing definitions of HMMs). These three sets of folders must be arranged in parallel with specific names.
5.2.2.1 Data preparation

Before using the HTK toolkit with training data, a number of data preparation steps need to be completed to facilitate data processing by the toolkit.

For the purpose of training, the data were prepared as follows:

We used the corpus which was recorded before. This corpus includes 21 selected utterances recorded with four different emotions. Each recording is saved to a file with the same name as the utterance label. Figure 5.3 shows the procedure for the preprocessing of our data in order to obtain a suitable format of HTK toolkit for our data.

---

Figure 5.3: Data preparation of our Data
As shown in Figure 5.3, Arabic text in UTF-8 code is transliterated and converted to Roman script based on the Buckwalter transliteration as shown in the table 4.12 with some modifications made using a Python program called Transliterate.py (see Appendix B1). The result is a text file which will be modified and renamed as a ‘prompts’ file.

<table>
<thead>
<tr>
<th>Arabic letter</th>
<th>Transliteration in Buckwalter</th>
<th>Arabic letter</th>
<th>Transliteration in Buckwalter</th>
<th>Arabic letter</th>
<th>Transliteration in Buckwalter</th>
</tr>
</thead>
<tbody>
<tr>
<td>ء</td>
<td>o</td>
<td>ذ</td>
<td>m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>أ</td>
<td>١</td>
<td>ر</td>
<td>n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>أ</td>
<td>٢</td>
<td>ز</td>
<td>h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ا</td>
<td>&amp;</td>
<td>س</td>
<td>w</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ا</td>
<td>&lt;</td>
<td>ش</td>
<td>ي</td>
<td>Y **A</td>
<td></td>
</tr>
<tr>
<td>أ</td>
<td>}</td>
<td>ص</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ب</td>
<td>ب</td>
<td>ط</td>
<td>F ** an</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ت</td>
<td>ج</td>
<td>ح</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ج</td>
<td>ه</td>
<td>ٝ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ح</td>
<td>خ</td>
<td>ٝ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>د</td>
<td>د</td>
<td>ل</td>
<td>l</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.12: Buckwalter Arabic transliteration

After obtaining the prompts file which contains all our words in the recorded sentences, we will have a list of words that can be used to generate a pronunciation dictionary (lexicon) file. From this we automatically extract a phonetic dictionary (by using a Python program called

---


85
phonetic.py) to use when preparing the training data for the speech recogniser. The most challenging part of this involves handling semivowels (characters which are sometimes pronounced as vowels and sometimes as consonants) as described below. The advantage of doing this automatically is that it would make it easy to carry out experiments with more data in future work. Production of the pronunciation dictionary file will be explained in detail in the next section.

5.2.2.2 Generating the Pronunciation Dictionary (Lexicon)

The pronunciation dictionary essentially consists of a list of word entries with their respective phonetic transcriptions based on a given alphabet.

Due to the lack of existing freely available pronunciation dictionaries for Arabic and the use of a small amount of data in our study, we decided to generate the dictionary using pronunciation rules for Arabic.

Generating the pronunciation dictionary requires fully diacritised text. If the dictionary contained undiacritised words like “mn” or “bkyn” then we would have no chance of producing a phonetic transcription. We therefore start from the assumption that the text is fully diacritised.

Firstly, Arabic text will be converted to the Latin or Roman alphabet, and secondly this will be converted to phonetic letters using Buckwalter’s Transliteration. This is the most commonly used transliteration scheme for Arabic, developed at Xerox in the 1990s, and has been used in many publications.

There are some symbols or characters in the Buckwalter transliteration which have to be slightly modified in order to be easily readable and acceptable for some of the HTK Toolkit processes, such as $, >, <.


86
However, Arabic spelling is a phonemic system with one-to-one letter to sound correspondence; for example, phoneme names consisting of one or two letters are generally used. For instance ( ب ، ت ، م ) are mapped to /b/, /t/, /m/, and ( ظ ، ط ، ش ) are mapped to /zh/, /Th/, /sh/. However, there is no issue in mapping letters to sounds for most consonants.

We are not going to depend on a predefined set of phonemes because we think that converting the name of a phoneme to another different letter is more important in English than Arabic. In addition, it makes no difference for the HTK Toolkit if the letter is converted to the same or to some other different letter or phoneme. Thus in general simply using the Buckwalter transliteration of the text provides a reasonable set of phoneme names. The only exceptions are the semi-vowels, where we use "w" and "y" for cases where they are being used as consonants and "uu" and "ii" where they are being used as vowels. We thus have 44 phonemes, for which the HTK makes 44 individual models – the first 42 entries in Table 4.12 (the last two are diacritics that denote either the absence of a vowel in a position where one might be expected or the doubling of the previous sound, and hence do not correspond to any specific sounds), plus two for the distinct versions of the semi-vowels.

For vowels, as mentioned in section 3.2, there are two types, namely long and short vowels.

These vowels are also defined with long vowels, such as ( ٌِ ، ٍِ ، َِ ) being defined as /a:/, /u:/, /i:/ and short vowels such as ( ً، ٌ، َ ) being defined as /a/, /u/, /i/, these modified vowels and other consonants are shown in Table 4.12 after symbols **. There is, however, a small set of letters which cannot be mapped directly to one phoneme. Therefore, these exceptions must be taken into consideration when constructing the dictionary.

A Python program, phonetic.py (given in Appendix B2), is used for dealing with these special letters. The main issues relate to the semi-vowels, tā’marbūTa, shadda, hamzah, tanwīn and sukūn.

Next, the cases of these letters or diacritic will be explained:

- The diacritic tanwīnis defined such that ( ُِ ُِ ُِ ُِ ) are mapped to /an/, /un/, /in/.
- The sukūn diacritic is not mapped to any character because sukūn it is a symbol marking the lack of a vowel and hence represents the absence of a sound.
• The diacritic *shadda* is a symbol marking consonant doubling.

• We will at the same time fix the *tā’marbūTa* / (`) (p) , which is typically a feminine ending. It can only appear at the end of a word. If “p” is followed by a vowel it is pronounced “t”, otherwise it is pronounced “h”.

• *Hamzah*, hamzah has many different shapes or spellings in Arabic script depending on its vocalic context (Habash, Soudi, & Buckwalter, 2007). It can be a separate letter (‘) or can be combined with other letters: ئ , ئ , ئ. However all hamzahs are glottal stops and we want the phonetic transliteration to contain a single item which represents this glottal stop. We could use (’), as Buckwalter does, but this would be awkward because it would interact badly with Python string quotes, and would probably greatly confuse the HTK. We should therefore choose some other character so the symbol “O” will be used instead of (’). The various single characters, hamzah-under-alif(<), hamzah-over-alif(>) and for hamzah-on-waw(&) are thus “Oi”, “Oa” and “Ou” respectively; i.e. they are a glottal stop followed by one of the vowels, so we will transliterate them that way.

• There are two characters which sometimes behave as consonants and at other times as vowels, as mentioned section 3.2.2; these are *waw* and *ya*, referred to as semi-vowels. The purpose of these letters is to make specific short vowels sound longer or slightly stretched. For example, *yaa* is used to make the kasrah sound slightly longer, while *waw* is used to make the dammah sound slightly longer. The program manipulates the semivowels, applying a rule which says that, if a semivowel is followed by a vowel (i.e. by one of the actual vowels or by a semivowel) then the semivowel is a consonant, otherwise it is a vowel. The examples below illustrate this for “w” and “uu”:
  o In the word “taHawl” (which means “turning up”), “w” is preceded by a vowel, /a/, so “w” in this case is a consonant.
  o In “limuwAjahapi” and “TayarAnK” (“confrontation” and “flight” in English), “w” and “y” are preceded and followed by vowels. In these cases, “w” and “y” are clearly pronounced as consonants
  o In the word “maHmwlp” (“portable” in English), if we assume that we have a root “m?H?m?l+p” and a pattern “a-o-w”, we can then put the elements of the pattern into the slots in the root, “maHomwlp”, and delete the *sukun* because this just
represents silence, giving “maHmwlp”. The means that when “w” is a vowel we should just write it as such.

- In some cases, like in the word “litazwydi” (“to provide”), we get both of the semi-consonants together as “...wy...”. In these situations, the first one is always a consonant and the second is a vowel.
  - Finally, word final “Y”, for example in the word “ElY”, refers to the preposition “to”, which maps to “A”, which we will deal with in our program.

However, there are a lot of complex rules discussed earlier in Chapter 3, and because we are interesting about prosodic characteristic of vowels letters therefore applying most of these rules does not have much effect and is beyond the aim of our study.

Figure 5.4 shows a screenshot of the generated pronunciation dictionary (lexicon) file.

![Screenshot of pronunciation dictionary file](image)

**Figure 5.4: Pronunciation dictionary file**

### 5.2.2.3 The Task Grammar

Typical speech recognisers (e.g. HTK) are based on a set of Hidden Markov Models (HMMs), with each word having a distinct HMM. When an unknown word is detected, it is scored against all HMMs and the one with the maximum score is considered to be the recognised word.
Speech recognisers also use the lexicon (phoneme-word) to find the corresponding word in cases in which the same phoneme sequence occurs in multiple words, while the grammar (language structure) is used in contexts in which information is needed to narrow down the recognised word to fit the given grammatical construction. In ordinary usage, the grammar should cover what a user might say. However, because our concern is with using the HTK toolkit, not to recognise the sentences, but to determine the time positions of phone boundaries in speech, we therefore used the actual utterance of sentences as the grammar, enabling us to extract the timing information of individual phonemes. This process is termed *forced alignment*.

So “dictionary” and “grammar” files have to be created to be used with the HVite tool in HTK. The first file, the dictionary, is done and we will explain the second file in this section.

A task grammar is a language model that defines the list of words, along with their probability of occurrence in a given sequence, that are to be recognised by the system. This language is a high level grammar notation based on the extended Backus-Naur form (EBNF) used to specify recognition grammars. It consists of a set of variable definitions followed by a regular expression according to some syntactic rules describing the words to be recognised (Young et al., 2002).

The grammar file in our experiment consists of a list of the basic Arabic words which will be used in our sentences.

In the test step for the HTK Toolkit, six grammar files will be created for each sentence. These grammars are only used for recognition, and an example of the grammar file for the first sentence is given in Figure 5.5.

Figure 5.5: The task grammar designed for the first sentence.
As shown in Figure 5.5, the grammar file contains sets of predefined combinations of words which make up a sentence. Each word in a language model or grammar has an associated list of phonemes.

Here $SENT$ is a variable (referring to the sentence) the value of which is the expression appearing on the right hand side of the assignment. Because all words in the sentence will be used in the same order exactly, we do not need to use the symbol '|', which denotes optional items between words.

The start and end silences($SENT$-START and $SENT$-END) can be made optional by enclosing them within square brackets, which denote optional items, while angle brackets denote one or more repetitions. The system must know, via the dictionary file, the HMMs that correspond to the grammar variables.

After writing the grammar files for all our sentences, the HParse tool will be used to convert them automatically to a word net which will be used with the Viterbi of the HVite tool.

5.2.3 The Experiment and Preliminary Result from the HTK Toolkit

Once we have recorded the speech files, prepared the data and generated the pronunciation dictionary, we follow the steps of the HTK Toolkit for building an acoustic model in order to conduct the experiment.

There is a set of HTK commands that are executed using a DOS command window. This requires many steps, each containing multiple commands. In order to effectively conduct the experiment, a simple bash script was written to run the steps automatically to train the model. In addition, there are some scripts that can be used for assistance during the experiment.
5.2.3.1 Building and Training the Model

We will briefly explain the HTK steps which were applied in order to achieve our aims of finding the phone boundaries, extracting the required samples and investigating their phonetic properties.

Bringert (2005) summarises the steps involved in using the HTK to train a model as below:

1. Create phone level transcriptions of the training utterances by using the dictionary.

2. Parametrise the recorded data (the HCopy tool is used for decoding the audio data) using the settings from the HTK tutorial which are saved in a wav_config file, as shown in Figure 5.6, with mel frequency cepstral coefficients (MFCCs), 10 ms frames, a 26 channel filter bank, and an output of 12 MFCCs.

```
SOURCEFORMAT = WAV
TARGETKIND = MFCC_O_D
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
```

Figure 5.6: Wav_config file

3. Create simple prototype models, and re-estimate them using the data.

4. Use the current models for each word, select the closest pronunciation from the dictionary, and then re-estimate the models using the new data.

5. Fix and include the silence model for HMM models, a “sp” (short pause) silence model which refers to the types of short pauses that occur between words in normal speech, and then use a set of monophone models to build triphone models by copying these monophone models.
6. Re-estimate the triphone models, with the transcriptions converted to use triphones. These sets of triphone models are used for recognition.

7. Create tied-state triphone models, in order to make more robust models. This step requires decision tree clustering, where models are organised in a tree, the parameters for which are called questions. The decoder asks a question about the phone context and decides what model to use.

Of these steps, we omit step 4, since the dictionary that we are using contains a single pronunciation for each word and hence there is no point in carrying out this step, and we also omit the use of triphone models, since using triphones brings very little benefit with very small training sets. After training the model, the next step is to evaluate it. The next section explains this process.

5.2.3.2 Evaluating the Model

In order to evaluate the recogniser in the usual case, the grammar files will be used and some more utterances will be recorded. In our case there is no need to record more speech, as the current data which were used as training material will also be used as test material. This would not be the correct approach if we were trying to train the HTK as a recogniser, but this is not the case. We are only interested in determining the point in time when particular phonemes occur in a speech segment by achieving a forced alignment of the transcription of the audio speech segment and are not interested in speech recognition in itself. So the steps in this phase will be slightly different.

The HTK toolkit uses the forced aligner to extract the start and end times of phones from recordings, which is what we need in order to align our phonemes.

The HVite tool (with -a option) in the HTK Toolkit can be used in the compute forced alignments mode to select the best matching pronunciations. It computes a new network for each
input utterance using the word level transcriptions and a dictionary, and the output transcription will just contain the words and their boundaries (Young et al., 2002).

5.2.3.3 The Results of Automatic Speech Alignment in the HTK Toolkit

As mentioned in the previous section, time boundaries of phonemes can be obtained directly using the Viterbi alignment procedure in the HVite tool. This is done by searching time boundaries for known sequences of HMM models for phonemes.

Figure 5.7: Forced alignment (Young et al., 2002).

Figure 5.8 describes the files used in this stage.

Figure 5.8: The files used in recognising the test data
The HTK's HVite command is used as follows:

```
HVite -A -D -T 1 -H hmm9\macros -H hmm9\hmmdefs -C config -S test.txt -l "*" -m -i recout.mlf -w wnet -p 0.0 -s 5.0 lexicon.txt monophones1
```

*Config*: configuration file to allow word-internal expansion.

*Test.txt*: lists the names of the files (MFC).

*Recout.mlf*: transcription output.

*Wnet*: compiled grammar network.

*Monophones1*: the input data to be recognized, containing a list of all phonemes

The parameter, -m, is used to generate model level output transcriptions. The output label file will contain multiple levels. The first level will be the model number and the second will be the word name (not the output symbol).

This command uses the definition of the HMMs or *monophones* models stored in folder hmm9 to transform the input word level transcription *words.mlf* to the new phone level transcription *aligned.mlf*, using the pronunciations stored in the dictionary (*lexicon.txt*).

The output is the *recout.mlf* file which is the output recognition transcription file including all phonemes with timing information. Figure 5.9 shows this file.
Each line of an HTK label file contains start and end times denoting the start time and end time of the labeled segment in 100ns units, followed by the name of the phoneme, which is then optionally followed by a score which refers to a floating point confidence score (Young et al., 2002). It is worth mentioning that this segmentation provides us with more accurate phoneme boundaries than simply looking for voiced segments of a fixed length. It also provides scope for future work on looking at short and long, and stressed and unstressed, vowels to see whether different classes of vowel are more significant when looking at emotion.

The master label file, the recout.mlf file, will then be employed in the next step which uses the Praat software to extract information from speech.

5.2.4 Annotation with the Praat Program

In order to find what we want to measure or work with, we have to create a set of labels, or “TextGrids”. These labels can be used to identify individual words or speech sounds in the sound file.
The process of annotation is carried out by using Praat, a free computer software package for the scientific analysis of the feature patterns and acoustic parameters of speech. It was designed, and continues to be developed, by Paul Boersma and David Weenink of the University of Amsterdam. It can run on a wide range of operating systems (Boersma & D, 2017).

In Praat as in almost all existing aligners, annotation of a segment or label of a sound can be done manually in two steps. The first step involves listening to each chunk of sound, determining the sentence or word, annotating the phonemes by annotators, and then marking the boundaries on the audio and transcript files. The second step is the automatic alignment of each paired chunk (X. Ma, 2012). Figure 5.10 shows a TextGrid window with the annotation of sounds in a waveform.

![Figure 5.10: A TextGrid window in Praat software](image)

This TextGrid is the object type that handles annotation, as shown in Figure 5.10. From the figure, in which the TextGrid window has three parts. The first part, at the top, presents the waveform of the sound; the second part shows some analyses (in the figure, a spectrogram and a pitch contour) of the sound; the third part at the bottom shows the contents of the TextGrid in a set of tires. A tier is an ordered sequence of texts, each of which is connected to a point in time.
or to a stretch of time. In our case, each speech file will be annotated as two layers, words and phonemes.

However, doing this manually with all 21 wave files is a time consuming and labour intensive process, so instead of using the manual method, we used an open source script in the Perl language to automatically and easily generate TextGrid files for all 21 files.

The `mlf2praat.pl` file will be used, as mentioned above, to convert HTK master label file to Praat TextGrid files (Mark, 2017) by running the command below.

```
Perl mlf2praat scp_files.txt recout.mlf
```

`mlf2praat.pl` is a Perl file which will be used to generate Praat TextGrid files.

`recout.mlf` is an input HTK-format master label file for that command;

`scp_files` is a text file, in a format useful for HCopy; Figure 5.11 shows `SCP_files`.

![SCP_files](image)

Each line in `SCP_files` is assumed to contain one output filename (textgrid folder) and one input filename (input folder), with the output file extension being replaced with “TextGrid”, and the input file extension with “lab”. For example, if `SCP_files` contains the line “textgride/sample1.wav input/sample1.mfc”, then `mlf2praat` will look through `input/mlf` for an entry headed by “*/sample1.lab”; this entry will be converted to a TextGrid and written out `textgrid/sample1.TextGrid`. The output directory must already exist (Mark, 2017).
Thus, the result of this step is a set of TextGrid files; for example, Figure 5.12 shows one of these files.

![TextGrid file of sample1.wav](image)

Figure 5.12: TextGrid file of sample1.wav

### 5.3 Extraction of Prosodic Features

A speaker’s emotional speech holds a set of features. Exactly which features are the best for identifying emotions is still an open question. However, the changes in these features will result from a corresponding change in emotions. Therefore, extraction of these speech parameters which represent emotions is an important factor in emotional speech recognition and synthesis systems (Tawari & Trivedi, 2010).

What we do know is that features such as duration (measures the length of the signals), pitch (the fundamental frequency of audio signals) and intensity (measures the power of the signal) have been identified as important elements of emotional speech and are the most useful and widely used features (Levin & Lord, 1975; Nwe, Foo, & De Silva, 2003; Ververidis & Kotropoulos, 2006; C. E. Williams & Stevens, 1972).

Because of the importance of these features in identifying emotion and as the primary indicators for detecting an unknown speaker’s emotional state, prosodic features such as duration, pitch
and **intensity** are extracted and analyzed using speech analysis and the manipulation tool, Praat, and labeled chunks of phonemes.

By using Praat we are able to obtain all of the information about these features. We are using three modified Praat scripts which are available online (Mietta, 2017). These scripts look through two folders, the first folder being a wav file and the second a TextGrid folder (as obtained from the step above), open each sound and TextGrid pair, and extract the desired information about each labeled interval. The resulting alignments from running these scripts are a set of separate TextGrid files in text file format for word and phoneme segmentation. After obtaining these files we run a Python program, to merge and write these three text files to one tab-delimited text file (csv), which we can work with easily. The output is illustrated as a Screenshot in Figure 5.13.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>txt</td>
<td>phone</td>
<td>Max_pitch</td>
<td>Min_pitch</td>
<td>mean_pitch</td>
<td>Avg_int</td>
<td>Min_int</td>
<td>MinInt_T</td>
<td>Max_int</td>
<td>MaxInt_T</td>
<td>duration</td>
</tr>
<tr>
<td>2</td>
<td>test1</td>
<td>1sil</td>
<td>122.7953</td>
<td>88.23191</td>
<td>90.23813</td>
<td>48.42911</td>
<td>27.06482</td>
<td>1.023097</td>
<td>59.30372</td>
<td>1.138899</td>
<td>0.151306</td>
</tr>
<tr>
<td>3</td>
<td>test1</td>
<td>1d</td>
<td>228.7953</td>
<td>215.9163</td>
<td>221.9563</td>
<td>69.81574</td>
<td>65.35888</td>
<td>1.146989</td>
<td>72.34453</td>
<td>1.192531</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>test1</td>
<td>1uh</td>
<td>220.9484</td>
<td>203.296</td>
<td>205.5893</td>
<td>57.26231</td>
<td>47.34589</td>
<td>1.273771</td>
<td>67.73899</td>
<td>1.226989</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>test1</td>
<td>1ae</td>
<td>295.8111</td>
<td>220.9484</td>
<td>267.8161</td>
<td>73.28086</td>
<td>54.0934</td>
<td>1.282989</td>
<td>77.39527</td>
<td>1.535715</td>
<td>0.11</td>
</tr>
<tr>
<td>6</td>
<td>test1</td>
<td>1y</td>
<td>304.1233</td>
<td>295.8111</td>
<td>300.6223</td>
<td>69.60698</td>
<td>54.17837</td>
<td>1.602989</td>
<td>77.77809</td>
<td>1.446208</td>
<td>0.22</td>
</tr>
<tr>
<td>7</td>
<td>test1</td>
<td>1t</td>
<td>300.054</td>
<td>265.1076</td>
<td>259.5593</td>
<td>53.36012</td>
<td>26.79378</td>
<td>1.764651</td>
<td>51.48735</td>
<td>1.619095</td>
<td>0.19</td>
</tr>
<tr>
<td>8</td>
<td>test1</td>
<td>1uh</td>
<td>269.1076</td>
<td>265.1613</td>
<td>258.479</td>
<td>70.4258</td>
<td>60.59497</td>
<td>1.802989</td>
<td>74.27807</td>
<td>1.8336</td>
<td>0.05</td>
</tr>
<tr>
<td>9</td>
<td>test1</td>
<td>1kh</td>
<td>268.0217</td>
<td>216.3956</td>
<td>248.3327</td>
<td>59.47764</td>
<td>55.29274</td>
<td>1.946989</td>
<td>71.05994</td>
<td>1.659099</td>
<td>0.1</td>
</tr>
<tr>
<td>10</td>
<td>test1</td>
<td>1ae</td>
<td>359.9683</td>
<td>254.0841</td>
<td>326.5093</td>
<td>65.03314</td>
<td>53.33432</td>
<td>1.955026</td>
<td>70.24361</td>
<td>2.014372</td>
<td>0.08</td>
</tr>
<tr>
<td>11</td>
<td>test1</td>
<td>1th</td>
<td>359.1565</td>
<td>250.4603</td>
<td>354.073</td>
<td>56.20714</td>
<td>49.78202</td>
<td>2.106898</td>
<td>68.37596</td>
<td>2.034985</td>
<td>0.08</td>
</tr>
<tr>
<td>12</td>
<td>test1</td>
<td>1in</td>
<td>298.0718</td>
<td>269.1633</td>
<td>285.6931</td>
<td>63.0319</td>
<td>38.85539</td>
<td>2.123898</td>
<td>72.22561</td>
<td>2.169703</td>
<td>0.11</td>
</tr>
<tr>
<td>13</td>
<td>test1</td>
<td>1uh</td>
<td>291.1456</td>
<td>275.8559</td>
<td>283.8917</td>
<td>57.57924</td>
<td>53.79785</td>
<td>2.266989</td>
<td>64.59841</td>
<td>2.226989</td>
<td>0.05</td>
</tr>
<tr>
<td>14</td>
<td>test1</td>
<td>1l</td>
<td>298.8953</td>
<td>273.6754</td>
<td>287.5908</td>
<td>66.92467</td>
<td>47.40232</td>
<td>2.28954</td>
<td>75.73416</td>
<td>2.352114</td>
<td>0.16</td>
</tr>
<tr>
<td>15</td>
<td>test1</td>
<td>1l</td>
<td>296.3506</td>
<td>150.2242</td>
<td>218.1441</td>
<td>63.67538</td>
<td>49.7046</td>
<td>2.578389</td>
<td>73.33063</td>
<td>2.453897</td>
<td>0.15</td>
</tr>
<tr>
<td>16</td>
<td>test1</td>
<td>1t</td>
<td>183.4146</td>
<td>183.4146</td>
<td>183.4146</td>
<td>50.5886</td>
<td>45.58183</td>
<td>2.608945</td>
<td>60.77199</td>
<td>2.626989</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Figure 5.13: The extracted information

As shown in Figure 5.13, we have all of the features for our samples, the first column refers to the name of an audio file (type of expression); and the second column refers to the phones, while the remaining columns show the statistical values (mean, median, maximum, minimum) of the obtained of pitch, duration and intensity features.
The CSV file obtained above as an output is still not labeled. So we manipulate this file to have a suitable structure for our experiments, for example by changing the first column, which contains the names of files, by running a Python program to label suitable expressions. This code will read the name of the file and will add a label to the end of the CSV file corresponding to the third character of the filename. This will be done for every WAV file in the database and the results will be saved as a database of features. Each one of the WAV files from the database has the name of an audio file ending with a code for the emotions, from E00 to E03. For example, S00E00 is the first audio file that expresses neutral emotion for sentence number one, as shown in Figure 5.14.

![Figure 5.14: CSV file with suitable names for recorded data.](image)

Next, a Python program is written to modify the resulting file to a suitable format that is accessible for a machine learning algorithm. We run the Python program to achieve the following tasks:

- Extracting the desired columns in the CSV file and the target phones;
- Identifying the target class and putting this in the last column;
- Replacing the “undefined” values with a question mark (?), which indicates an unknown or missing value;
- Extracting acoustic properties of particular segments.
In our case, we think that the segments which are interesting to look at are voiced sounds, because voiced phonemes most clearly carry the important information about the acoustic features of emotional speech, especially pitch, but to a lesser extent, intensity and duration. Stop consonants in particular have low duration and intensity, and hence seem to be unlikely candidates for carrying information about emotion. There are many Arabic sounds which are voiced but we think that the most clearly voiced segments are vowels -- voiced stop consonants such as /b/ and /d/, for instance, have essentially zero duration and intensity, and consequently also have no detectable pitch. The result of the process of excluding all non-vowels from analysis is as follows:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>phone</td>
<td>mean_pitch(Hz)</td>
<td>Avg_Int(dB)</td>
<td>duration(Seconds)</td>
<td>txt</td>
</tr>
<tr>
<td>2</td>
<td>u</td>
<td>222.6344911</td>
<td>69.97020501</td>
<td>0.07</td>
<td>E00</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>267.8112983</td>
<td>71.26221966</td>
<td>0.12</td>
<td>E00</td>
</tr>
<tr>
<td>4</td>
<td>u</td>
<td>255.8462994</td>
<td>70.23353516</td>
<td>0.05</td>
<td>E00</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>339.2811178</td>
<td>62.07953332</td>
<td>0.14</td>
<td>E00</td>
</tr>
<tr>
<td>6</td>
<td>i</td>
<td>293.2418262</td>
<td>65.87538618</td>
<td>0.06</td>
<td>E00</td>
</tr>
<tr>
<td>7</td>
<td>u</td>
<td>286.243181</td>
<td>70.86519121</td>
<td>0.13</td>
<td>E00</td>
</tr>
<tr>
<td>8</td>
<td>i</td>
<td>187.9385911</td>
<td>64.03055132</td>
<td>0.09</td>
<td>E00</td>
</tr>
<tr>
<td>9</td>
<td>a</td>
<td>172.3591694</td>
<td>62.26751103</td>
<td>0.11</td>
<td>E00</td>
</tr>
<tr>
<td>10</td>
<td>i</td>
<td>174.9441026</td>
<td>65.74611551</td>
<td>0.1</td>
<td>E00</td>
</tr>
<tr>
<td>11</td>
<td>i</td>
<td>163.731945</td>
<td>62.5320484</td>
<td>0.08</td>
<td>E00</td>
</tr>
<tr>
<td>12</td>
<td>i</td>
<td>158.2614893</td>
<td>63.00755663</td>
<td>0.17</td>
<td>E00</td>
</tr>
<tr>
<td>13</td>
<td>i</td>
<td>?</td>
<td>45.06577122</td>
<td>0.03</td>
<td>E00</td>
</tr>
<tr>
<td>14</td>
<td>i</td>
<td>193.243337</td>
<td>63.83635924</td>
<td>0.1</td>
<td>E00</td>
</tr>
<tr>
<td>15</td>
<td>a</td>
<td>196.7982548</td>
<td>62.43242128</td>
<td>0.09</td>
<td>E00</td>
</tr>
</tbody>
</table>

Figure 5.15: First lines of the features database from CSV file.

The first column of Figure 5.15 shows that only three vowels (u, a and i ) have been selected after excluding all non-vowel sounds. Unlike English and many other languages, Arabic only has three vowels, with a short and long version of each as mentioned in Section 3.2.1, these vowels are associated with four emotions along with values of the three features, pitch, duration and intensity.
5.4 Summary

In this chapter we have described two processes for extraction of acoustic features of Arabic emotional speech; these are the annotation and extraction processes. We have briefly highlighted the theoretical background of the HTK Toolkit and the requirements for conducting our experiment for automatically obtaining the time domain phonemes for each utterance using adapted HMM models trained by the HTK Toolkit. We then extracted the acoustic properties, namely pitch, intensity and duration, of speech associated with four different emotions (neutral state, happiness, sadness and anger). All manipulation of extracted features was performed using scripts from the Praat software.

From the above investigation of the acoustic features of emotional speech, we would like to be able to classify the emotions based on these features.

In the following chapter, we will investigate how these features can be used for the classification of emotions in Arabic speech.
CHAPTER SIX : CLASSIFICATION.

6.1 Introduction

In the previous chapter some prosodic features are extracted using Praat as primary indicators of the speaker’s emotional state. These features are pitch, intensity and duration of three vowels in Arabic emotional speech which are then stored in a database where each feature is associated with its classification label, e.g. neutral, happy, sad or angry.

The classification simply consists of using a set of previous observations obtained in order to determine the category of a new observation or a sample. Many machine learning algorithms can be applied to the area of emotion classification, including K-nearest neighbors (KNN), hidden Markov model (HMM), Gaussian mixtures model (GMM), support vector machine (SVM) and artificial neural net (ANN). Figure 6.1 shows the framework for the emotion classification of speech.

As shown in Figure 6.1, the input to the system is recorded speech of different emotional classes which are transcribed and labeled, as mentioned in Chapter 5. The features subsequently extracted (pitch, intensity and duration) from recorded speech, as also mentioned in Chapter 5, and saved as numbers of values. These specific features will be exploited as input (training set) to the model of classification (classifier). The output of the classifier is a label of emotion.
classification of the emotional contents of which are unknown. There are four classes, namely neutral, happy, sad and angry and each label represents the corresponding emotion classification.

The structure of this chapter is as follows: the experiments used for classifying phonemes (vowels) in this research are presented in Section 6.2 with experimental results. Classifying utterances by conducting the experiment and its result is given in Section 6.3. Conclusions are delivered in Section 6.4.

The next section will present those experiments conducted in order to classify the phonemes for the dataset.

6.2 Classifying phonemes

Once the features are extracted, they are used to form the feature vector database. Each data sample in the database is an instance, i.e., the feature vector, and is used for classification.

In our experiments, open source software called Weka (Waikato Environment for Knowledge Analysis) is chosen to undertake this process, of which the code and executable are freely available and issued under license. Weka is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using JAVA. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualisation. It is also well-suited for developing new machine learning schemes (Frank, Hall, & Witten, 2016)

6.2.1 The experiments and Dataset description

The main objective of the experiments described here is primarily to investigate which of the features extracted are the most significant for identifying emotion by analysing different combinations of features at phonemic as well as utterance level. The Weka toolkit was investigated in order to identify the best algorithm for achieving this task.
To find out which features are the most significant, we first have to discover the optimum algorithm. By looking at the results of each Weka machine learning algorithm scheme, we note which one has the highest recognition accuracy. Once we have found which one is the most accurate we use it to identify the most significant features. We chose to use the Leave-One-Feature-Out (LOFO) technique. In our case, there are four features, Pitch, Duration, Intensity and Phone, and these are extracted in accordance with our aim, which is to provide more information on which features are more important for speech emotion identification.

We also investigated the effect of including or excluding neutral emotion on our result, because there will be some overlap between neutral emotion and other emotions as shown in the result of human judgment evaluation in Chapter 4. Recall that in some cases subjects annotated utterances as a mixture of neutral and some other emotion, which was presumably a way of indicating that the degree to which the utterance expressed the non-neutral emotion. We decided to conduct this experiment to investigate the possibility that if we dropped neutral emotion (E00) from our dataset, a more noticeable increase in accuracy could be gained than we first expected.

Therefore, this experiment will be divided into three parts. Experiment I presents the process for evaluating the effectiveness of different selected classification algorithms, Experiment II introduces the result of choosing different subsets of features on the accuracy of the performance of the classifier to be selected for investigation. Experiment III will present the result of including or excluding the neutral emotion (E00).

**Dataset description**

The ARFF file format is the data file normally used by Weka, which consists of special tags to indicate different elements in the data file such as attribute names, attribute types, attribute values and the data. Our dataset which uses ARFF File format is shown in Figure 6.2.
The file consists the following information:

- The name of the dataset

- A header which describes the attribute and its types, such as:
  - phone \{u, a, i\} : nominal
  - mean_pitch (Hz) : numeric
  - Avg_Int (dB) : numeric
  - duration (Seconds) : numeric

- Data section, containing one row for each wav file. The rows have the features of that sample separated by commas and in the last position there is the label indicating its emotion. An example of a row is: u, 222.634491, 69.970205, 0.07, E00 which refers to feature1; feature2; feature3; feature4; emotion label, respectively.

The dataset used in this experiment contains four features (Pitch, Intensity, Duration, Phone) with a target class of emotions called (txt). Target class (txt) with values \{E00, E01, E02, E03\} where E00 refers to Neutral emotion, E01 is Happiness, E02 is Sadness and E03 is Anger. as shown in Table 6.1.
### Table 6.1: dataset attribute information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type of Attribute</th>
<th>Possible values</th>
<th>Missing values</th>
<th>Mean</th>
<th>Std.Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>Nominal</td>
<td>i, u, a (Vowels)</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pitch</td>
<td>Numeric</td>
<td>Positive number</td>
<td>No</td>
<td>214.444</td>
<td>45.866</td>
</tr>
<tr>
<td>Intensity</td>
<td>Numeric</td>
<td>Positive number</td>
<td>No</td>
<td>63.004</td>
<td>7.127</td>
</tr>
<tr>
<td>Duration</td>
<td>Numeric</td>
<td>Positive number</td>
<td>No</td>
<td>0.067</td>
<td>0.039</td>
</tr>
<tr>
<td>Txt (target class)</td>
<td>Nominal</td>
<td>E00, E01, E02, E03</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The dataset contains 418 instances, as shown in Table 6.2, where each instance consists of a number of attributes representing details of prosodic information for one vowel in each expressed emotion.

### Table 6.2: The expressive utterances used of our dataset

<table>
<thead>
<tr>
<th>Emotions</th>
<th>E00</th>
<th>E01</th>
<th>E02</th>
<th>E03</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterances</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S00</td>
<td>***</td>
<td>✓</td>
<td>✓</td>
<td>***</td>
</tr>
<tr>
<td>S01</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S02</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S03</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S04</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S05</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>***</td>
</tr>
<tr>
<td>No. Instances</td>
<td>102</td>
<td>116</td>
<td>116</td>
<td>84</td>
</tr>
</tbody>
</table>

As shown in Table 6.2, six utterances S00, S01,..S05 where each recorded with four emotions (E00, E01, E02, E03) meaning that 24 utterances were recorded.

However, when we validated the data by asking human subjects to classify the various utterances, three utterances (S00:E00, S00:E03 and S05:E03) were not assigned the expected emotions. We therefore excluded these utterances from the experiments: if human subjects were unable to identify them as expressing the intended emotions then using them as training data would provide misleading clues to the classifier, and using them as test data would also be inappropriate.
6.2.2 Experiment I: Performance Evaluation of a set of classifiers

As mentioned before, the Weka Toolkit contains a number of different machine learning algorithms. Trying to use all these algorithms is very time consuming. So we have to choose one or two suitable algorithms rather than using all of them. However, it is worth working out which algorithm is best by looking at its various features in some detail. Finding the best machine learning algorithm raises the challenge of applied machine learning; however, Weka makes that task more easy by providing built-in options that enable us to run lots of algorithms on the same data. So once we have tested a set of algorithms from the Weka toolkit on our dataset, we can identify the one with the best performance and use this in our experiments. It is worth mentioning here that examining the reasons why one algorithm outperforms another is outside the scope of this thesis.

The process of identifying the best algorithm or classifier in general is as follows:

1. Design an experiment including the dataset and any data preparation as well as the test options, using 10-fold cross validation; this usually gives a good estimate of the accuracy of an algorithm: cross-validation, a standard evaluation technique, is a systematic way of running repeated percentage splits. Having 10 folds means 90% of full data is used for training (and 10% for testing) in each fold. Having more folds means that more of the data is used for training in each fold, thus increasing the accuracy of the trained model, but at the same time it means that it is necessary to train and test more models -- using 5 folds would involve only having 80% of the data for training, but you would only have to carry out five rounds of training and testing. With small amounts of data doing 10-fold cross validation is a reasonable compromise, since if you only have a small amount of data then you want to use as much of it as you can for training in each round, but training and testing will be reasonably fast and hence having 10 rounds will not take an excessive amount of time.

2. Conduct the experiment on all features of our dataset based on phoneme level along with selecting a set of different algorithms to be applied to our dataset.

3. Analyse the results and identify the best scoring achieved by algorithms in general.

4. Select one algorithm to be applied in the next experiment.
The algorithms that we use will be briefly explained in the following section in terms of how they work.

6.2.2.1 Classifiers Used

The classifiers in WEKA are categorized into different groups such as classifiers that extract rules, trees-based classifiers, classifiers that make use of Bayesian interface, and other miscellaneous classifiers.

Rule-based classification

In rule-based classification, a set of (if-then-else) rules are used for classification. An if-then rule is an expression of the form IF condition THEN conclusion. In the above expression the “IF”-part of a rule is known as the rule antecedent or precondition. And the “THEN”- part is the rule consequent. In the rule antecedent, the condition consists of one or more attribute tests that are logically ANDed.

The rule’s consequent contains a class prediction. If the condition in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuple.

For example, if we are predicting whether a student will get admission in ph.d or not, then R1 can be written as R1: (age = 25) ^ (post graduate = yes)) (get admission in PhD = yes).

Five variants are discussed in the following subsections

- DecisionTable: builds a decision table classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation. An option uses the nearest-neighbor method to determine the class for each instance that is not covered by a decision table entry, instead of the table’s global majority, based on the same set of features.

- JRip: implements Ripper algorithm, it is one of the basic and most popular algorithms which including heuristic global optimization of the rule set. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error. JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the
members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

- **OneR**: short for "One Rule", is a simple classification scheme based on the value of a single predictor, that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, we have to construct a frequency table for each predictor against the target. OneR provides a useful baseline -- any reasonable algorithm ought to beat OneR.

- **Part**: obtains rules from partial decision trees. The algorithm producing sets of rules called "decision lists". It builds the tree using C4.5’s heuristics with the same user-defined parameters as J4.8, it uses the separate-and-conquer strategy, where it builds a rule in that manner and removes the instances it covers, and continues creating rules recursively for the remaining instances (Frank & Witten, 1998).

- **ZeroR**: is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictive power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods (Witten, Frank, Hall, & Pal, 2016).

### Decision Tree Based Classification

A decision tree is a flowchart-like tree structure, it is one of the predictive modeling approaches used in statistics, data mining and machine learning. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making (Fathima, Manimegalai, & Hundewale, 2011).

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.

In these tree structures, each internal node denotes a test on an attribute, each branch represents a result of the test, and each leaf node holds a class label. The topmost node in a tree is the root node. We can simply obtain the rules related to the tree by traversing each leaf of the tree starting from the node. Figure 6.3 show the shape of a decision tree.
Most decision tree classifiers perform classification in two phases: tree-growing (or building) and tree-pruning. The tree building is done in a top-down manner. During this phase the tree is recursively partitioned till all the data items belong to the same class label. In the tree pruning phase the full grown tree is cut back to prevent over fitting and improve the accuracy of the tree in bottom up fashion.

There are many specific decision-tree algorithms. Notable ones include ID3, J48 (C4.5), C5, and REPTree (CART). As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy.

Decision trees are widely used as they are easy to interpret and are not restricted to functions that can be represented by rule "If-then-else" condition.

- **DecisionStump**: is basically a one level decision tree, designed for use with the boosting methods, which builds one-level binary decision trees for datasets with a categorical or numeric class, dealing with missing values by treating them as a separate value and extending a third branch from the stump. Trees built by RandomTree consider a given number of random features at each node, performing no pruning.

- **HoeffdingTree**: is an implementation of the incremental decision tree algorithm. It offers options to create splits based on information gain or the Gini index. Predictions at the leaves of the tree can be made by either majority class or naive Bayes models.
- **J48**: is an optimized implementation of the C4.5 decision tree learner or improved version of the C4.5 is an efficient method for estimation and classification of fuzzy data. The output given by J48 is the Decision tree (Quinlan, 2014).

- **LMT**: A classification model with an associated supervised training algorithm that combines logistic prediction and decision tree learning is logistic model tree (LMT) (Landwehr, Hall, & Frank, 2005). Logistic model trees use a decision tree that has linear regression models at its leaves to provide a section wise linear regression model.

- **RandomForest**: are defined as a group of un-pruned classification or regression trees, trained on bootstrap samples of the training data using random feature selection in the process of tree generation. After a large number of trees have been generated, each tree votes for the most popular class. These tree voting procedures are collectively defined as random forests. RF has excellent accuracy among current classifier algorithms. It also has an effective method for estimating missing data and it maintains accuracy when a large proportion of the data are missing.

- **RandomTree**: it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In standard tree every node is split using the best split among all variables. In a random forest, every node is split using the best among the subset of predictors randomly chosen at that node (Landwehr et al., 2005).

- **REPTree**: builds a decision or regression tree using information gain/variance reduction and prunes it using reduced-error pruning. Optimized for speed, it only sorts values for numeric attributes once. It deals with missing values by splitting instances into pieces, as C4.5 does.

**Bayesian based classification**

Bayesian classification is based upon Bayes’ probability rules and depends on likelihood functions. Four variants are discussed in following subsections.

- **Simple Bayes’ Classification**: bayesian networks are a powerful probabilistic representation. Bayesian algorithms predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. It learns from training data the
conditional probability of each attribute $A_i$ given the class label $C$. Classification is then done by applying Bayes’ rule (based on Bayes’ theorem) to compute the probability of $C$ given the particular instances of $A_1...A_n$ and then predicting the class with the highest posterior probability (Domingos & Pazzani, 1997). This Bayesian Network consists of two components. First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes or random variables represents the probabilistic dependencies among the corresponding random variables. Second component is a set of parameters that describe the conditional probability of each variable given its parents. The conditional dependencies in the graph are estimated by statistical and computational methods. Thus the BN combine the properties of computer science and statistics. Probabilistic models predict multiple hypotheses, weighted by their probabilities.

- **BayesNet**: Probabilistic graphical model that represents random variables and conditional dependencies in the form of a directed acyclic graph. A Simple Estimator algorithm has been used for finding conditional probability tables for Bayes net. A K2 search algorithm was used to search network structure.

- **Naive Bayes**: The Naive Bayes classifier is based on Bayes’ Theorem with independent assumptions between predictors. Naive Bayesian model is easy to build without complicated iterative parameter estimation. It analyzes all the attributes in the data individually, which means the value of a predictor ($X$) on a given ($C$) is independent of the values of other predictors. This assumption is called class conditional independence. In almost all cases this assumption is not valid, which has significant effects on the reliability of classifiers based on Bayes’ Theorem.

**Other miscellaneous classification**

- **MultilayerPerceptron** (MLP) is one of the most common neural network models, it is a class of feed forward artificial neural network that is trained using back propagation that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple
layers of nodes at least three layers in a directed graph with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. It can distinguish data that is not linearly separable (Van Der Malsburg, 1986).

The graphical representation of MLP is shown in Figure 6.4 with one input layer, one hidden layer and one output layer.

![Multilayer Perceptron](https://en.wikipedia.org/wiki/Multilayer_perceptron)

Figure 6.4: Multilayer Perceptron

MLPs are very flexible about incomplete, missing and noisy data. It can be updated with fresh data and implemented in parallel hardware. When an element of this algorithm fails, it can continue without any problem due to its parallel nature. They are useful in research for their ability to solve problems stochastically, which often allows approximate solutions for extremely complex problems. Also they can be used to create mathematical models by regression analysis (Mallios, Papageorgiou, & Samarinas, 2011).⁶

- Simple Logistics, It is a classifier used for building linear logistic regression models. LogitBoost with simple regression functions are base learners used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection (Landwehr et al., 2005).

---

• **SMO:** Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones (Keerthi, Shevade, Bhattacharyya, & Murthy, 2001). A single hidden layer neural network uses exactly the same form of model as an SVM.

• **k-Nearest Neighbor (IBk)** One of the simplest forms of classification algorithms is Nearest Neighbor implementations, the number of nearest neighbors (k) can be set manually, or determined automatically using cross-validation. Such learning schemes are depicted as statistical learning algorithms and are generated by simply storing the given data. For the classification to be performed a distance metric is chosen and any new data is compared against all-ready “memorized” data items. The new item is assigned to the class which is most common amongst its k nearest neighbors (Coomans & Massart, 1982).

• **K-Star:** K star is one of the nearest neighbour lazy learning classification method with generalized based on transformations. It provides a reliable approach to handle symbolic attributes, real valued attributes and missing values. Space required for the storage is very large as compared to other algorithms and it is generally slower in evaluating the result. Advantages of K-star: It is robust to noisy training data and it is more effective when applied on large data set.

The results of carrying out this experiment, which is based on these algorithms, are presented in the following section. It should be noted that our data is a mixture of features with numerical values and the categorical feature ‘phone’. All the classifiers we have used can cope with categorical data so long as it is so labelled in the training file.
### 6.2.2.2 Experimental results of Experiment I

Table 6.3 shows the results of the experiment aimed at finding the best classifier for our task.

![Table 6.3: The features are used in each experiment with different learning algorithms](image)

<table>
<thead>
<tr>
<th>Machine learning algorithms</th>
<th>The features are used Pitch, Intensity, Duration, Phone (Correctly Classified Instances %)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rules</strong></td>
<td></td>
</tr>
<tr>
<td>DecisionTable</td>
<td>43</td>
</tr>
<tr>
<td>JRip</td>
<td>39</td>
</tr>
<tr>
<td>OneR</td>
<td>30</td>
</tr>
<tr>
<td>PART</td>
<td>33</td>
</tr>
<tr>
<td>ZeroR</td>
<td>27</td>
</tr>
<tr>
<td><strong>Trees</strong></td>
<td></td>
</tr>
<tr>
<td>DecisionStump</td>
<td>40</td>
</tr>
<tr>
<td>HoeffdingTree</td>
<td>41</td>
</tr>
<tr>
<td>J48</td>
<td>37</td>
</tr>
<tr>
<td>LMT</td>
<td>40</td>
</tr>
<tr>
<td>RandomForest</td>
<td>34</td>
</tr>
<tr>
<td>RandomTree</td>
<td>33</td>
</tr>
<tr>
<td>PEPTree</td>
<td>41</td>
</tr>
<tr>
<td><strong>Bayes</strong></td>
<td></td>
</tr>
<tr>
<td>BayesNet</td>
<td>42</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>41</td>
</tr>
<tr>
<td><strong>Other miscellaneous classifiers</strong></td>
<td></td>
</tr>
<tr>
<td>Logistic</td>
<td>42</td>
</tr>
<tr>
<td>MultilayerPerceptron</td>
<td>39</td>
</tr>
<tr>
<td>SimpleLogistic</td>
<td>40</td>
</tr>
<tr>
<td>SMO</td>
<td>41</td>
</tr>
<tr>
<td>IBk</td>
<td>30</td>
</tr>
<tr>
<td>KStar</td>
<td>33</td>
</tr>
</tbody>
</table>
The first column shows a list of machine learning algorithms of the *Weka* toolkit from four groups which are Rules, Trees, Bayes and other miscellaneous classifiers. The second column shows the features used. In this first case, all features are used because in this phase we are interested in finding the best algorithm for our experiment. The values refer to the correctly classified instances of each experiment in percentage terms.

We make a comparison of all these algorithms to assess their performance; we use the default settings for all the algorithms analysed.

The distribution of phonemes to emotions in our dataset is 102 examples of Neutral, 116 of Happiness, 116 of Sadness and 84 of Anger. This means that simply classifying every sentence as Happy or Sad would score 27% (which is indeed the score returned by ZeroR).

Based on Table 6.3, we can clearly see that the performances accuracy of number of algorithms are around or just over 40%. It is worth noting that Random Forest and Random Tree, which are normally regarded as very effective algorithms, score poorly in these experiments. These algorithms perform well when there are large numbers of features, since it can be difficult to determine which features are the most significant. Given that we have a fairly small set of features, the advantages of these algorithms are lost when they are applied to our data.

The performance scores reported in the table above are based on measure of the accuracy of these classifiers, accuracy simply treats all examples the same and reports a percentage of correct responses and that would be fine when we are dealing with balanced (or approximately balanced) datasets.

Although the vast majority of research results and practical projects report that accuracy is the default metric. We think that the classification accuracy alone may be not enough information to make the decision of which classifier is the best for our experiments, especially given that we are dealing with unbalanced datasets. So we are going to use other measures in order to evaluate that numbers of classifiers and in the end choose one to carry out our experiments. In order to do that firstly, the classifiers with higher score from each group will be chosen from Table 6.3, secondly, these classifiers will be tested by a set of statistical measures in WEKA.
Approach followed:

- Five algorithms are considered for evaluation which had the higher score in their groups as listed in Table 6.3, they include:
  i. Rules. DecisionTable
  ii. Trees. HoeffdingTree
  iii. Tree. PEPTree
  iv. Bayes. BayesNet
  v. Functions. Logistic

- Weka tool is used for an experimental study, where Weka allows us to perform statistical tests through the Experiment Environment interface on the different performance measures in order to draw conclusions from the experiment. Figure 6.5 shows the Experiment Environment window of Weka.

![Experiment Environment window](image)

Figure 6.5: Experiment Environment window

Figure 6.5 gives analysis of all five algorithms experiment test, where the number of result lines available (Got 500 results) is shown in the Source panel. This experiment consisted of 100 runs, for 5 schemes, for 1 dataset, for a total of 500 result lines.
• 10 fold cross validation is used in this experimental study (as described earlier in section 6.2.2). The final accuracy of an algorithm will be the average of the 10 trails. There are 418 instances in our Dataset. Using 10 fold cross validation to evaluate the algorithms, each fold will comprise of 376 training samples and 42 test samples, 10 fold cross validation is repeated ten times and results are averaged.

• A corrected resampled paired t-test is performed via Weka analysis toolbar for claiming which of these applied classifiers are better, it will compare each algorithm in pairs and make some reasonable assumptions about the distribution of the results collected, such as that they are drawn from a Gaussian distribution. Also by taking into account both mean and variance of the differences between these two measures over several runs, it calculates a t value. Using this value and desired significance level (normally 5%), the probability that these two measurements are significantly different can be obtained by looking at result; there are some annotations such as “v” or “*” which appear next to any classifier results in the table which indicate that a specific result is statistically better (v) or worse (*) than the baseline scheme (in our case, DecisionTable) at that significance level specified (currently 0.05)

• Two performance parameters will be considered for experimental evaluation:
  
  i. Accuracy which is what we are interested in as a first pass. Also there are other options we can use to evaluate the performance of the classifiers such as TP rate, FP rate, Precision, Recall, F-Measure, Kappa statistic, these measurements are partitioned in numeric and percentage.

  ii. Error rate: there are some error measures which are used to find out how far way the predicted value is from actual known value, such as mean absolute error and root mean squared error which are in numeric value only. Also there are the relative absolute error and root relative squared error which are show in percentage for references and evaluation.

The next sections show these measurements and their results.
- **Accuracy Parameters.**

*Accuracy:* Accuracy defines percentage of correctly classified instances from the test set by the classifier. For accuracy measurement six parameters are considered:

- **TP Rate:** True Positive rate is proportion of examples which were classified as a class X among all examples which truly have the X class. This shows how much the part of the class is truly captured.

- **FP Rate:** False Positive rate is the proportion of the examples which were incorrectly identified and belong to another class.

- **Precision:** is the proportion of examples which truly have class X among all those which were classified as X. It is a measure of exactness. Positive predictive value is called as precision: \( PPV = \frac{TP}{TP+FP} \)

- **Recall:** is the proportion of examples which were classified as class X among all examples which truly have class X. It is a measure of completeness. Negative predictive value is called as recall: \( NPV = \frac{TP}{TP+FN} \)

- **F-measure**, is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test, in other words, It is combination of both precision and recall. F-score is calculated as following:
  \[
  F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
  \]

- **Kappa statistic:** is a measure of how closely the instances classified by the machine learning classifier matched the data labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy. Kappa result varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement.
Table 6.4 summarizes the results for each simulation calculated across all folds for a given classifier (the results for the collected folds for a classifier are more reliable than the individual results for each fold).

<table>
<thead>
<tr>
<th>Accuracy Parameters</th>
<th>Classifiers</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DecisionTable</td>
<td>HoeffdingTree</td>
<td>PEPTree</td>
<td>BayesNet</td>
<td>Logistic</td>
</tr>
<tr>
<td>Accuracy</td>
<td>43</td>
<td>40</td>
<td>40</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td>Std. deviations</td>
<td>6.64</td>
<td>6.83</td>
<td>7.17</td>
<td>6.59</td>
<td>5.51</td>
</tr>
<tr>
<td>TP rate</td>
<td>0.54</td>
<td>0.19 *</td>
<td>0.37 *</td>
<td>0.52</td>
<td>0.30 *</td>
</tr>
<tr>
<td>FP rate</td>
<td>0.33</td>
<td>0.11 *</td>
<td>0.23 *</td>
<td>0.32</td>
<td>0.21 *</td>
</tr>
<tr>
<td>Precision</td>
<td>0.35</td>
<td>0.37</td>
<td>0.36</td>
<td>0.34</td>
<td>0.32</td>
</tr>
<tr>
<td>Recall</td>
<td>0.54</td>
<td>0.19 *</td>
<td>0.37 *</td>
<td>0.52</td>
<td>0.30 *</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.416</td>
<td>0.244 *</td>
<td>0.340</td>
<td>0.403</td>
<td>0.304 *</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.23</td>
<td>0.20</td>
<td>0.18</td>
<td>0.22</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 6.4: Simulation Result of six chosen Algorithms

Based on the above results, we can observe that the highest accuracy is 43% for DecisionTable and BayesNet) and the minimum is 40% for (PEPTree) while in case of precision HoeffdingTree (0.37) performs best among all. A measurement system is considered valid if both (accuracy & precision) are accurate and precise. We can see also that there is no character “v” appears next to any algorithm. This means that the difference in the accuracy for these algorithms compared to DecisionTable is not statistically significant. We can also see that the accuracy for these algorithms compared to DecisionTable around or just over 40%, so we can say that these five algorithms do not achieved a statistically significantly result compared to the DecisionTable baseline.

Recall is the fraction of relevant instances that are retrieved. Precision and recall both are inversely proportional to each other. Both precision and recall are based on measurement of relevance. So in most cases performance of a classifier is measured in terms of both: from Table 6.4 it can be seen that Decision Table scores substantially better on F-measure, which is the

---

7The character "*" next to numbers indicates that to the performance of that classifier compared to the baseline classifier (DecisionTable) is statistically significantly different at significance level 5%.
standard compromise between precision and recall, and for that reason we use this algorithm in the later experiments.

• **Error Rate Evaluation Parameters.**

The error measures used to find out how far off the predicted value is from the actual known value. There are four parameters of error measurement:

- **MAE Mean absolute error**
- **RMSE Root mean squared error**
- **RAE Relative absolute error**
- **RRSE Root relative squared error**

- **Mean absolute error (MAE)** measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

- **Root mean squared error (RMSE)** is a quadratic scoring rule which measures the average magnitude of the error. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. Alternatively, the relative errors are also used.

Loss functions measure the error between and the predicted value. The most common loss functions are (Han, Pei, & Kamber, 2011):

- **Absolute error:** $|y_i - y'_i|$
- **Squared error:** $(y_i - y'_i)^2$

Based on the above methods, the error rate is the average loss over the test set.
Mean absolute error: \[
\frac{\sum_{i=1}^{d}|y_i - y'_i|}{d}
\]
Mean squared error: \[
\frac{\sum_{i=1}^{d}(y_i - y'_i)^2}{d}
\]

The mean squared error exaggerates the presence of outliers, while the mean absolute error does not. If we take the square root of the mean squared error, the resulting error measure is called the Root Mean Squared Error (RMSE).

The formula is defined as

Relative absolute error (RAE): \[
\frac{\sum_{i=1}^{d}|y_i - y'_i|}{\sum_{i=1}^{d}|y_i - \bar{y}|}
\]
Relative squared error (RSE): \[
\frac{\sum_{i=1}^{d}(y_i - y'_i)^2}{\sum_{i=1}^{d}(y_i - \bar{y})^2}
\]

Where \( \bar{y} \) is the mean value of the \( y'_i \)'s of the raining data, \( \bar{y} = \frac{\sum_{i=1}^{t} y_i}{d} \)

Table 6.5 shows simulation errors of each classifier.

<table>
<thead>
<tr>
<th>Error Rate Parameters</th>
<th>DecisionTable</th>
<th>HoeffdingTree</th>
<th>PEPTree</th>
<th>BayesNet</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>0.35</td>
<td>0.33 *</td>
<td>0.34</td>
<td>0.34</td>
<td>0.33 *</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.42</td>
<td>0.42</td>
<td>0.43</td>
<td>0.42</td>
<td>0.41 *</td>
</tr>
<tr>
<td>RAE %</td>
<td>93</td>
<td>89</td>
<td>91</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>RRSE%</td>
<td>97.0</td>
<td>97.3</td>
<td>99.5</td>
<td>97.6</td>
<td>95.7 *</td>
</tr>
</tbody>
</table>

Table 6.5: Error Rate Evaluation Parameters for classifiers

MAE : Mean absolute error  
RMSE: Root mean squared error  
RAE: Relative absolute error  
RRSE: Root relative squared error

An algorithm which has a lower error rate will be preferred as it has more powerful classification capability.

It is discovered from Table 6.5 that the highest error is found of MAE is found of DecisionTable at 0.35, The rest of them average around 0.338, which is not very different from DecisionTable rate. Similarly the root mean squared error for DecisionTable is very slightly higher than for some of the other classifiers.

However, although the scores for Hoeffding Tree and Logistic are marked as being significantly different from the baseline estimate and the scores for Decision Tree are not so marked, the
differences between the error rates for the various classifiers are extremely close and do not provide compelling evidence for choosing one rather than another.

Decision Tables are much faster to train, which is important when there are a large number of experiments to be carried out. Given that the difference in accuracy of the various classifiers is marginal, the speed of Decision Table training makes this the most attractive option for our experiments.

6.2.3 Experiment II: Exploring which feature is important

After selecting the best learning algorithm, we need to find out which are the most significant features in our dataset by applying that algorithm. In order to achieve this aim, we leave out one of the features each time and see how accurate the classifier is without that feature. So, the experiment will be repeated many times with a different number of features of our dataset each time. This is achieved by using different datasets which contain a different number of features in Weka.

A different dataset shown in Table 6.2 will be used with the selected algorithm to discover which feature at phone level is more important than others.

The next experiment aims to investigate what the most important feature is for assessing the emotional clues in speech.
6.2.3.1 Dataset Description of Experiment II

In this experiment, we generated five datasets from the original dataset for use in the Leave-One-Out experiments. The details of each dataset are shown in Table 6.6.

<table>
<thead>
<tr>
<th>No .EXP</th>
<th>Description of Datasets</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP #1</td>
<td>Using all features</td>
<td>5 (P, I, D, Ph, txt)</td>
</tr>
<tr>
<td>EXP #2</td>
<td>Dropping Pitch</td>
<td>4 (I, D, Ph, txt)</td>
</tr>
<tr>
<td>EXP #3</td>
<td>Dropping Intensity</td>
<td>4 (P, D, Ph, txt)</td>
</tr>
<tr>
<td>EXP #4</td>
<td>Dropping Duration</td>
<td>4 (P, I, Ph, txt)</td>
</tr>
<tr>
<td>EXP #5</td>
<td>Dropping Phone</td>
<td>4 (P, I, D, txt)</td>
</tr>
</tbody>
</table>

Table 6.6: Details of five datasets

The first column refers to the name of experiment, the second column describes each dataset is used for each experiment, third column describes the features will be used, where Symbols P, I, D, Ph and txt refer to features Pitch, Intensity, Duration, Phone and target class (txt) of dataset respectively. The first dataset is the original dataset which contains all the features, and the other datasets are generated from it by leaving out one of the features. The number of instances of each dataset is 418.

6.2.3.2 Experimental results of Experiment II

Table 6.7 shows the result of our experiments by using the DecisionTable algorithm.

<table>
<thead>
<tr>
<th>EXP.NO</th>
<th>Feature Name</th>
<th>DecisionTable algorithm (Correctly Classified Instances %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Pitch, intensity, duration, phone</td>
<td>43%</td>
</tr>
<tr>
<td>2.</td>
<td>intensity, duration, phone</td>
<td>38%</td>
</tr>
<tr>
<td>3.</td>
<td>Pitch, duration, phone</td>
<td>37%</td>
</tr>
<tr>
<td>4.</td>
<td>Pitch, intensity, phone</td>
<td>42%</td>
</tr>
<tr>
<td>5.</td>
<td>Pitch, intensity, duration</td>
<td>43%</td>
</tr>
</tbody>
</table>

Table 6.7: The features are used in each experiment with DecisionTable algorithms

The first column in Table 6.7 refers to the number of the experiment with a set of features saved in the dataset and used in each experiment in the second column.
The last column shows the performance scores of the classifier with these features.

Comparing the scores of each experiment, we can see that scores obtained using all features in our experiment differ from the other experiments conducted with leaving one feature out. However, with four classes we would expect a score of 0.25 if we simply made random choices and 0.33 if we used three classes. Also the purpose of this experiment is to show which feature is useful, and does not matter at this point whether that the classifier is not actually giving good performance.

The result of conducting Experiment 1 with all the features produces a better score at 43% compared with using three features which are (I, D, PH) at 38% in EXP 2, (P, D, PH) at 37% in EXP 3, (P, I, PH) at 42% in EXP 4. And it generates the same score compared with the features (P, I, D) in EXP 5. However, we observed that the combination of features that exclude pitch or intensity generate low scores compared with others, in particular, the lowest score is seen when intensity is absent, which indicates that intensity is the most useful feature. Also, in EXP 5, we can see from the result of the scores of the experiment in this table that the combination of features, namely (P, I, D), has the highest score compared with Exp 2, 3, 4. However, absence of the phone feature indicates that phone does not make much difference, and as a result is not as important as a feature. This result is expected due to the fact that, when expressing the emotion of anger for example, we change the intensity and pitch, and we do that no matter which phoneme (vowel, in our case) is being uttered. The identity of the individual phonemes is not, of course, by itself a likely indicator of emotion -- it would, for instance, be extremely surprising if /a/ were more common in words that express anger than in words that express happiness. However, it could happen that the ways in which the other features vary across different emotions depend on which phoneme is being uttered, e.g. that the pitch of /a/ is raised in angry utterances but that this does not happen with /i/. Given that the machine-learning algorithms we applied look at combinations of features, it seemed worthwhile to include the identity of the different phonemes to see whether this was in fact the case. The fact that leaving out the phone name from the training data has no effect suggests that this is not.

As mentioned above, a set of features which are (P, I, D) in EXP 5 has the same score as using all combination of features in EXP 1, so we are interested to see what happens if we drop one of
the features in this experiment. In other words, we are interested to see the accuracy of the algorithm with the following pairs of features: (pitch, intensity), (pitch, duration) and (intensity, duration), and then to see the result of conducting this experiment again with individual features in order to investigate which one is more important than others individually. Tables 6.8 and 6.9 show the results. We did not carry out experiments with phone and one other feature or phone by itself (i.e. four of the fifteen non-trivial combinations) because the initial round of experiments confirms the hypothesis that the phone does not contribute anything useful to the classification.

<table>
<thead>
<tr>
<th>EXP.NO</th>
<th>Feature Name</th>
<th>DecisionTable algorithm (Correctly Classified Instances %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Pitch, intensity</td>
<td>43 %</td>
</tr>
<tr>
<td>2.</td>
<td>Pitch, duration</td>
<td>37 %</td>
</tr>
<tr>
<td>3.</td>
<td>Intensity, duration</td>
<td>40 %</td>
</tr>
</tbody>
</table>

Table 6.8: The result of using two features with DecisionTable algorithm

<table>
<thead>
<tr>
<th>EXP.NO</th>
<th>Feature Name</th>
<th>DecisionTable algorithm (Correctly Classified Instances %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Pitch</td>
<td>37 %</td>
</tr>
<tr>
<td>2.</td>
<td>Duration</td>
<td>28 %</td>
</tr>
<tr>
<td>3.</td>
<td>Intensity</td>
<td>40 %</td>
</tr>
</tbody>
</table>

Table 6.9: The result of using one feature with DecisionTable algorithm

Table 6.8 shows that the pair of features (pitch and intensity) has the higher score with 43 %, followed by the pair of features (intensity and duration) at 40%.

By comparing this result with the obtained result from Table 6.7, we note that the combination of three features (pitch, intensity, duration) has the same classification accuracy as the use of two features (pitch and intensity) at 43% when the duration was dropping. This indicates that duration is less important than other features. In contrast, a lower score is obtained when the intensity feature was dropped, as shown when using two features (pitch and duration) at score 37% in Table 6.8.
We can observe that including an intensity feature improves performance rather than weakens it. Furthermore, this finding confirms the result of the previous experiments in Table 6.7. In addition, from Table 6.9 we can see that the duration has the lowest score at 28%. Because 28% is barely more than 25%, which is the score we would achieve if we made random choices, duration is therefore of no use as an indicator.

Comparing Table 6.8 with Table 6.9 shows that the performance of a classifier is the same whether including or excluding duration in these cases:

- Using two features (pitch and duration) or using pitch only of 37%
- Using two features (intensity and duration) or using intensity only of 40%.

This result supports our finding that duration is the least important feature.

Consequently, from our result we can deduce that the importance of features are ordered as following from highest to lowest: Intensity, Pitch and then Duration.

We can observe that there is no substantial difference in the performance scores when using a number of different features compared with the use of all the features, so we are going to include all these features in the next experiments.

Next, we use a confusion matrix table to show the number of predictions for each class compared to the number of instances that actually belong to each class. This is very useful to get an overview of the types of mistakes the algorithm made. This done with 10-fold cross validation for classification of the four different emotional states in our database. Consider the algorithm DecisionTable under the Rules group running on our dataset in Weka. For our dataset we obtain four classes and hence we have a 4x4 confusion matrix. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. From the confusion matrix we can see how the four emotions will be classified based on using all features of phonemes (vowels). Table 6.10 shows the result of this experiment.
From the confusion matrix table it can be observed that the highest precision and recall are achieved for the class “Sadness” at 50% and 61% respectively followed by the “Neutral” class, and the lowest are achieved for the class “Happiness”, while the class “Anger” is not recognised at all. The classes “Happiness”, “Sadness” and “Anger” are mixed with the class “Neutral” at all times and vice versa. The emotions which present less misclassification by using the 10 fold cross-validation approach are sadness and neutral emotion with 45 out of 116 instances and 49 out of 102 instances of misclassification respectively. Meanwhile the happiness emotion and anger emotion present 64 out of 116 instances and 84 out of 84 of instances of misclassification.

6.2.4 Experiment III : The effect of excluding neutral emotion

As shown in the confusion matrix Table 6.10, we found that the three emotions were mixed with “Neutral” class.

Neutral emotion in all cases was the second choice of prediction for all emotions. An utterance that was assigned a mixture of neutral and some other emotion can reasonably be interpreted as
being a weak example of the non-neutral emotion. We therefore decided to investigate the effect of leaving this class (E00) out of our data, on the assumption that in many cases examples which have been primarily classified as neutral are actually weak examples of other emotions.

6.2.4.1 Dataset Description of Experiment III

Experiment III will run with all instances of the neutral class (E00) removed and use of a different number of features retained as in Experiment II, to assess whether there are any substantial improvements in accuracy.

In Weka under the preprocess tab, we use the "RemoveWithValues" filter under "unsupervised-->attribute" part, in order to remove all instances of class E00. This a step will be applied for all datasets used in this experiment. Figure 6.6 shows screenshot of this process.

![Figure 6.6: Preprocessing of E00](image-url)
6.2.4.2 Experimental results of Experiment III

Table 6.11 shows the results of dropping neutral emotion (E00) from the target class (txt)

<table>
<thead>
<tr>
<th>No.EXP</th>
<th>The features</th>
<th>DecisionTable algorithm (Correctly Classified Instances %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Pitch, intensity, Duration, phone</td>
<td>51 %</td>
</tr>
<tr>
<td>2.</td>
<td>intensity, duration, phone</td>
<td>52 %</td>
</tr>
<tr>
<td>3.</td>
<td>Pitch, duration, phone</td>
<td>48 %</td>
</tr>
<tr>
<td>4.</td>
<td>Pitch, intensity, phone</td>
<td>51 %</td>
</tr>
<tr>
<td>5.</td>
<td>Pitch, intensity, duration</td>
<td>52 %</td>
</tr>
</tbody>
</table>

Table 6.11: The features are used in each experiment without Neutral Emotion (E00)

In comparing Tables 6.7 and 6.11, the accuracy of the experiment increases slightly when we drop all instances of the neutral emotion of target class.

From our observation, the percentage increase of the scores in all cases is between 8% and 14%; the aim of this experiment was to investigate whether leaving out E00, which is essentially the absence of any emotion, improves the performance of the classifier on examples where a genuine emotion is present. In other words, we want to see whether it is easier to assign the correct emotion to an utterance which is known to express one of the major emotions than to simultaneously decide whether it expresses any emotion and if so to determine which one. To assess this we need to consider what the expected effect of decreasing the number of classes would be. In particular, if we take into consideration the fact that using three classes is easier than four, the expectation of increase in performance should accommodate this fact and should be greater than would be predicted from random choices.

If we made random guesses with four classes (E00, E01, E02, E03) we would expect to score 0.25, but 0.33 with three classes. So to say that removing the need to decide whether some utterance expressed an emotion at all makes a bigger difference on our result, we have to expect the increase in the scores to be proportionally greater than with the original experiments. In other words, if the comparison of these scores is based on one of them obtained from three classes being more than 4/3 (1.33), then removing neutral has a greater effect on the result than otherwise. By comparing the obtained score from Tables 6.7 and 6.11 respectively, we have the
scores 0.43, 0.38, 0.37, 0.42, 0.43 for four classes with different sets of features, and these scores are 0.51, 0.52, 0.48, 0.51, 0.52 for three classes. For example, if we take the first scores of these two tables as the result of using all features and do a simple compute by dividing the larger number of classes by the smaller, i.e. (0.51/ 0.43) we note that the result is noticeably less than what we expected 1.333, so there does not seem to be a useful effect here in almost any of the cases.

As a conclusion, there was no substantial difference between the accuracy achieved by leaving the E00 class out and including it from our target class txt. Therefore, for further experiments we will include this class along with the other emotions.

6.2.5 Summary

To summarise this section, the first experiment was conducted to identify the best performance of classifiers for our experiments, from the Initial experimental results we found that most machine algorithms perform differently. So we decided to evaluate only five classification algorithms namely DecisionTable, HoeffdingTree, PEPTree, BayesNet and Logistic. These classifiers were experimentally evaluated based on number of accuracy and error rate using WEKA tool with 10 cross validation test option. From the result we decided to select the DecisionTable classifier to conduct our next experiments.

The main goal of this section was to study the effect of different combinations of three features and their contributions to the accuracy of the classifiers.

The observed result from the experiment shows that the Intensity feature is the most important because as we noticed if we drop the intensity feature we achieve a much worse score than if we include it. As regards the others, it does not seem to matter which are dropped; neither does dropping phone features make any difference. In summary, the importance of the three features is in the order of Intensity, Pitch and then Duration. It is, perhaps, unsurprising that the deleting the name of the phoneme as a feature makes little difference, but it was worth investigating whether this feature would interact with the others in significant ways.
Regarding the second experiment conducted to see the effect of leaving out the neutral emotion (E00) of our dataset on the performance of the classifier, the result shows that there is no real improvement.

6.3 Classifying utterances

The earlier experiments were aimed at classifying individual phonemes, but that what we are actually interested in is classifying utterances, and the obvious way to do this is by counting the classification of the phonemes that make up an utterance.

The main aim of this experiment is to examine the majority vote of the phonemes that make up specific utterances via the machine learning classifier.

This experiment uses the classifier developed in Section 6.2.2 to classify the phoneme that make up an utterance, and then simply counting their votes for the various accents. In Section 6.2.2.2 we showed that the classifier (DecisionTable), though unreliable, does do a reasonable job of assigning individual phonemes to emotions. The question to be investigated here is whether the aggregate accuracy when assigning complete utterances to emotions is higher than the individual accuracy.

6.3.1 Experimental results of the experiment

Table 6.12 shows the results obtained when we classify whole utterances by counting the classifications of the sets of phonemes that make up each utterance.
The result obtained from running the six experiments for the individual utterances (S00, S01, S02, S03, S04, S05) explained in appendix C over 21 utterances. For each of these we count the majority vote of the phonemes that make up the utterance and assign that as the emotion for the whole utterance.

From Table 6.12, we can see how the classifier classified these utterances according to their emotion.

The number that are classified correctly is 167 out of the total is 405. We can see also that two emotions, neutral and sad, were classified correctly.

The result shows that happy emotion was classified correctly 39 out of 121 instances and 82 out of 121 instances were misclassified. Sad emotion was classified correctly 63 out of 98 and misclassified 35 times as another emotion while anger was never classified correctly. As noted in Section 4.2.4.4, human subjects often find it hard to distinguish between anger and sadness, and our algorithms always classify anger as either neutral or sadness. This suggests that the features we have used behave very similarly for anger and sadness, and that distinguishing between these two requires further investigation. The overall classification success is 41%.

Table 6.12: Confusion matrix of all utterances

<table>
<thead>
<tr>
<th>REAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>65 17 20 0</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>53 39 29 0</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>23 12 63 0</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>38 13 33 0</td>
</tr>
<tr>
<td>sum</td>
<td>179 81 145 0</td>
</tr>
</tbody>
</table>
The result obtained from running the six experiments for the individual utterances (in appendix C1) are outlined in the following table for each utterance:

<table>
<thead>
<tr>
<th>Exp. no</th>
<th>Utterances</th>
<th>E00</th>
<th>E01</th>
<th>E02</th>
<th>E03</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Testing of S00</td>
<td>****</td>
<td>√</td>
<td>√</td>
<td>****</td>
</tr>
<tr>
<td>2.</td>
<td>Testing of S01</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td>3.</td>
<td>Testing of S02</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td>4.</td>
<td>Testing of S03</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>5.</td>
<td>Testing of S04</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td>6.</td>
<td>Testing of S05</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>****</td>
</tr>
</tbody>
</table>

Table 6.13: The result of six experiments with missing emotions

Table 6.13 shows the correct and incorrect classifying of each utterance, where (****) refers to missing emotion of that utterances, and √ refers to correct classification of classifier of emotion and × the inverse.

By looking at the emotions we can see that the anger emotion was classified incorrectly in all experiments, while the neutral emotion was classified correctly all the time. The sad emotion was also classified correctly except in one case. Happy emotion was classified incorrectly in most cases.

We notice from this experiment that E03 were misclassified almost every time, and we think the reason may be the missing examples or instances of this emotion in our training set. Also we missed E00 of utterance S00, so in this experiment, we are going to add an extra copy of these utterances which are underrepresented to our training data in order to make our training dataset balanced and repeat the experiment to find out if this is an effect of missing data.

In our case, there are a set of utterances missed in our dataset which are shown in Table 6.2 in Section 6.2.1, such as S00E00, S00E03 and S05E03, so we are going to use extra data from other utterances, for example from S01E00, S03E03 and S04E03 respectively. This is a way of giving balance to our data.
Then, we will run a set of six experiments again manually, and each test data will be the recordings of one sentence with their available emotions, and training data will be the recordings of other sentences with extra copies or examples. The result for individual utterance is shown in appendix C (C2), and the result for accounting the majority vote on all the phonemes that make up the utterance is shown in Table 6.14.

| Conf_Matrix | Decision Table | PREDICTED CLASS | | | |
|---|---|---|---|---|---|---|
| REAL | CLASS | Neutral (N) | Happiness (H) | Sadness (S) | Anger (A) |
| | | 54 | 44 | 26 | 37 |
| | | 17 | 39 | 10 | 12 |
| | | 38 | 24 | 79 | 33 |
| | | 12 | 9 | 1 | 2 |
| sum | | 161 | 78 | 174 | 24 |

Table 6.14: Confusion matrix of all utterances after adding the missing emotions

The results of Tables 6.12 and 6.14 show that overall accuracy is achieved by using the method of adding more instances, at 40%, which is slightly less than before.

Also the results of the remaining five experiments in Appendix C2 show that there is a decrease in accuracy after adding more instances of missing emotions in some cases, such as in S01 and S03; or show the same accuracy as other utterances. However, the next, Table 6.15, outlines the results of classifying the utterances of all experiments.

<table>
<thead>
<tr>
<th>Exp.no</th>
<th>Emotions</th>
<th>Utterances</th>
<th>E00</th>
<th>E01</th>
<th>E02</th>
<th>E03</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>H</td>
<td>S</td>
<td>A</td>
</tr>
<tr>
<td>1.</td>
<td>Testing of S00</td>
<td></td>
<td>****</td>
<td>x</td>
<td>x</td>
<td>****</td>
</tr>
<tr>
<td>2.</td>
<td>Testing of S01</td>
<td></td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>3.</td>
<td>Testing of S02</td>
<td></td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>4.</td>
<td>Testing of S03</td>
<td></td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>5.</td>
<td>Testing of S04</td>
<td></td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>6.</td>
<td>Testing of S05</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>****</td>
</tr>
</tbody>
</table>

Table 6.15: The result of six experiments with adding of missing emotions
These results show that adding more instances of missing emotions does not show any marked improvement of classification, and we think maybe there is not enough information to work on.

6.3.2 Summary

For classifying the emotions of utterances in an unknown test sample, the DecisionTable classifier has been chosen due to its proven efficiency in previous experiments.

The overall accuracy using DecisionTable classifier is 54%, made up of 83% for sad, 100% for neutral, 33% for happy and 0% for anger.

Also this section investigated the effect of missing instances of some utterances on the prediction accuracy of the classifier. The results show that adding more instances of missing emotions does not produce any improvement.

6.4 Conclusion

In this chapter, we have conducted a set of experiments based on phoneme and utterances levels.

First we tried to investigate the importance of three features extracted from Arabic emotional speech namely pitch, intensity and duration. We analysed the different combinations of these features by using the DecisionTable classifier in Weka. This classifier was selected as result of our experiment on various algorithms of Weka. The main findings show that the intensity feature was the most important while the duration feature was the least. Also we found that using the combination of three features produces differing recognition accuracy compared with using different combinations of two or one features. However, we concluded that prosodic features are more suitable and add meaningful information when the features are used together to classify emotions. Also we discovered that excluding neutral emotion has no substantial effect on the accuracy of classification.
Second, we investigate the accuracy of classification on the basis of different features extracted from utterances of different emotional states.

The overall accuracy using DecisionTable classifier was 54% and also the results show that there are two emotions classified correctly, sad and neutral at 83% and 100% for respectively. The other emotions (happy and anger) were misclassified.

We removed some sample recordings from the training data, on the grounds that our human subjects did not perceive them as expressing the target emotions. This left us with unbalanced training data, which could potentially have led to the over-represented classes dominating the classification. In order to overcome this we added extra copies of other recordings expressing these emotions, in order to produce a more balanced training set. Unfortunately this did not have much effect on the overall accuracy.
CHAPTER SEVEN : CONCLUSION

7.1 Conclusion

The work is divided into three main parts:

- Creation and evaluation of an Arabic emotional speech corpus.
- Extraction and analysis of relevant prosodic features of emotional speech in Arabic.
- Classification of Arabic emotion based on the details of features identified.

In the first part, we randomly select a set of six Arabic sentences from Arabic BBC news reports and recorded them by one speaker acting out the sentences in four different emotional states: happy, sad, angry and neutral. The dataset considered in this study consists of 24 Arabic emotional sentences evaluated by listeners recruited for the purpose, who found that 21 sentence were expressed successfully by the actor. These sentences were then labeled and prepared for the next step.

The second part comprises two stages: the first stage is annotation, where recorded speech is transcribed and labeled. An essential task here is to determine the time boundaries of phonemes using the HTK toolkit which employs the Viterbi alignment procedure provided by the HVite tool to align the phonemes. The second stage constitutes a process of extracting important prosodic information from these phonemes of emotional speech, namely pitch, intensity and duration, for classifying emotions. This is achieved by using the speech analysis and manipulation tool Praat to extract all the information of these features.

In the third part, we attempt to classify emotions based on extracted features of emotional speech. The Weka machine learning toolkit is used in this step. We attempt to evaluate a set of different machine learning algorithms in Weka on our dataset to pick one high performance algorithm. Analysis within default parameters shows that the set of algorithms performed well.
DecisionTable classifier is used for analysing the performance of emotion classification. Also in this phase, we run a set of experiments with different feature combinations and analyse the results obtained by using different combinations of features to identify the most important one. The conclusion is that on the basis of practical implementation, intensity is the most significant feature.

As mentioned above, a DecisionTable classifier is used to classify emotions based on feature values extracted from the speech samples of different emotional states at the level of phonemes and utterances. Each dataset is split into two separate files, one for training and another for testing data in order to achieve this task.

The results show that certain emotions seem more confusing than the others, for example, anger and happiness are more confusing than the other emotions, whereas neutral and sad emotions were classified correctly.

In particular, the anger emotion is not recognised, and the researcher presumes this may be due to a lack of sufficient examples of this emotion in the training set, suggesting the addition of further examples and resulting analysis. However, the observed result did not show any improvement in the overall emotion classification performance.

Finally, we can say that our small corpus seems to be adequate for determining which features are most important, and it suggests that it may be possible to develop a classifier based on these features.

7.2 Future work

Supplementary future work can be added to this study to seek improvement in many aspects including the following:

- This study achieves its results using a small sample size corpus; also we engage in training using very little data. Therefore, the results cannot be generalised to the whole
population. The work may in future be extended to encompass a larger corpus of Arabic emotional speech.

i. The corpus was recorded by only one female actor. Inclusion of new professional speakers from other Arabic countries could in future be implemented in order to diversify the speech corpus and provide a richer database.

ii. Recording of different sentences can provide a rich phonetic covering all syllables and sentence types and can address issues of pronunciation variations such as deletion and insertion etc.

iii. In our study, four states were considered: anger, happiness, sadness and a neutral state. We think it would be desirable to have additional expressive speech types such as fear, disgust, tiredness etc.

iv. Up until now, the perceptual test in the evaluation phase of our corpus was achieved by only a small number of subjects – 12 (11 female and one male) – therefore, we could employ a greater number of subjects with more of a gender balance which would increase the dataset size and could result in better accuracy in the corpus. For example, selecting more diverse subjects who do not speak Arabic or who have different linguistic and cultural backgrounds would be helpful in order to make comparisons between results from a wider range of people. Also it would be interesting to use other evaluation methods in order to give the evaluators more options and that would potentially enhance evaluation of expressive sentences.

v. We hope to explore a set of effective tools which can be used to generate high quality transcriptions as well as to evaluate and compare the output of labeled material (detecting errors in forced alignment). This is likely to yield positive results.

- In order to improve results and achieve better performance, we think it is necessary to select and extract more numerous and more diverse features, which would help acquire more emotional detail. For the current research, the features that have been extracted are pitch, intensity and duration. For future work, an insight into a different set of features, such as energy, Linear Prediction cepstrum coefficients (LPCC) or Mel Frequency cepstrum coefficients (MFCCs), could potentially improve the results.

- Other learning machine techniques can be used with an appropriate choice of parameters to gain a good improvement of accuracy in the classification.
References

Amer, W. An investigation into the differences between English and Arabic consonant and vowel sounds: a contrastive study with pedagogical implications. In.


Stevens, K., & Williams, C. (1969). On determining the emotional state of pilots during flight—An exploratory study(Pilot emotional state during stressful situations from tape recorded vocal utterances of air to ground radio communications using spectrographic analysis). Aerospace Medicine, 40, 1369-1372.


Appendix A: Questionnaire

THE EMOTION IN ARABIC SPEECH TASK STUDY

SCHOOL OF COMPUTER SCIENCE

The aim of this questionnaire is to help me testing and evaluating the Arabic emotional recorded sentences in my corpora. I would appreciate if you answer freely and as honest as possible. I would like to thank you for your help and participation.

You are going to listen to 24 audio files in Arabic language as listed in the folder, after each sentence please write down the appropriate emotion you think that was expressed by that sentence.

*note: we allowed for you to repeat each file twice only before deciding on the emotion type.*

Please Give percentage of the most appropriate emotion in the box below:

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Normal (%)</th>
<th>Happy (%)</th>
<th>Sadness (%)</th>
<th>Anger (%)</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO# files</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Thank you for your participation.*
Appendix B: Python programs

B1: Transliterate.py

---

```python
# this is a file of stuff for transliterating from Arabic to Roman

import codecs
import string
import os.path

def readAndWrite(ifileName, enc, ofileName):
    
    readAndWrite reads everything in ifileName, assuming that the encoding named by enc was used when that file was created, and writes out the character (using the UTF-16 encoding) and the number corresponding to that character. The point of this one is to find out what numbers correspond to the Arabic and Roman characters, so that we can decide how to map (numbers corresponding to) Arabic characters to (numbers corresponding to) Roman characters

    ifile = codecs.open(ifileName, encoding=enc)
    ofile = codecs.open(ofileName, "w", "UTF-8");
    characters = ifile.read()
    # this bit's a loop for looking at each character in turn
    for i in range(0, len(characters)):
        # get the next character
        c = characters[i]
        # get the number that encodes it
        x = ord(c)
        print x
        # print them out, each on a new line
        ofile.write(c)
        ofile.write(\r\n )
        ofile.write(str(x))
        ofile.write(\r\n )
    ofile.close()

def transliterateFile(ifileName, enc, ofileName):
    
    This one reads a file of, probably Arabic, characters and uses transliterate to swap them for Roman equivalents

    ifile = codecs.open(ifileName, encoding=enc)
    characters = ifile.read()
    numbers = []
    ofile = open(ofileName, 'w')
    for i in range(0, len(characters)):
        n = ord(characters[i])
        t = str(transliterate(n))
        if t=='~':  # to present ?symbol
            n = ord(characters[i-1])
            t = str(transliterate(n))
            ofile.write(t)
        else:
            ofile.write(t)
    ofile.close()
```

---
import os
def listing(dir):
    return os.listdir(dir)

def transliterateAllFiles(file, path):
    top = path+'/'+file
    print "scanning"+top
    try:
        st = os.lstat(top)
    except os.error:
        print top
        print "OK"
    if stat.S_ISDIR(st.st_mode):
        for name in listing(top):
            transliterateAllFiles(name, top)
    else:
        transliterateFile(top, "UTF-8", path+'\TRANSLITERATED\"'+file)

def transliterate(a):
    """
    a is the number corresponding to a character. We look through all
    the 2-tuples in the list of translations: if a is the first element of one of these then we want the
    second (which is supposed to be the number corresponding to the Roman equivalent)
    """
    translations = {1569:'\',1570:'|',1571:'O',1572:'W',1573:'I',1574:'}',1575:'A'
    ,1576:'b',1577:'p',1578:'t',1579:'v',1580:'j',1581:'H',1582:'x',1583:'d',1584:'*',1585:'r'
    ,1586:'z',1587:'s',1588:'$',1589:'S',1590:'D',1591:'T',1592:'Z',1593:'E',1594:'g'
    ,1600:' ','1601:'f',1602:'q',1603:'k',1604:'l',1605:'m',1606:'n',1607:'h',1608:'w',1609:'Y',1610:'y',1611:'F'
    ,1612:'N',1613:'K',1614:'a',1615:'u',1616:'i',1617:'~',1618:'o'}
    if a in translations:
        return translations[a]
    elif a <128:
        return chr(a)
    else:
        return ''

import os,string
def read(ifilename):
    ip = open(ifilename,"r")
    word  = ip.readlines()
    return word
    ip.close()
ifile=input ('Enter file name which you want to transliterate it to romanse file: ') 
transliterateFile(ifile+'trans.txt', "UTF-8", ifile+'trans.txt')
op=open(ifile+'trans.txt','w')
w=read(ifile+'trans.txt')
op.close()
#to have proper format which can be used as transcription in HTK
for index, item in enumerate(w):
    item = item.strip()
    if item!='':
        op.write("*/sample\"+str(index)+"' +item)
        op.write("\n")
op.close()
import os
import stat

    if not isinstance(wlist, str):
        wlist = [word.split("#")[0].strip() for word in open(wlist)]
    transcriptions = ["%s	[%s]\%	%s" % (w, phontrans(w), w) for w in sorted(wlist)]
    if lexicon:
        transcriptions += ["SENT-END\t[]\tsil", "SENT-START\t[][]\tsil"]
    with open(lexicon, 'w') as out:
        for x in sorted(transcriptions):
            out.write("%s\n"%x)
    else:
        return transcriptions

def phontrans(w0):
    w1 = ""
    for c in w0:
        if c == ":
            w1 += w1[-1]
        else:
            w1 += c

    allographs = {"Y":"A", "F":"an", "K":"in", "N":"un",
                  ":":"Oa", ":":"Oi", ":":"Ou", ":":"O", ":":"n", ":":"O"}
    for a in allographs:
        w1 = w1.replace(a, allographs[a])

    ""
    fixing the semivowels. If a semivowel is followed by a vowel
    (i.e. by one of the actual vowels or by a semivowel) then it is a
    consonant. We will denote the consonant version of a semivowel by
    its upper case equivalent, i.e. the vowel "w" is "W" and the
    consonant "w" is "W".

    We will at the same time fix tarmabutas. If "p" is followed by a
    vowel it is pronounced "t", otherwise it is pronounced "h".

    ""
    w2 = ""
    semivowels = set(["w", "y"])
    vowels = semivowels.union(set(["a", "i", "u", "A"]))
    sep ='
    for i, c in enumerate(w1):
        w2 +=sep
        if c in semivowels and i < len(w1)-1 and w1[i+1] in vowels:
            w2 += c.upper()
    elif c == "p":
        if i < len(w1)-1 and w1[i+1] in vowels:
            w2 += "t"
    else:
        w2 += "h"

    return w2
Appendix C: The result of experiments before and after adding the missing emotions

C1: The result of experiments before adding the missing emotions listed from 1 to 6 and shown in confusion matrix:

Experiment 1: S00

<table>
<thead>
<tr>
<th>Conf. Matrix</th>
<th>Decision Table</th>
<th>PREDICTED CLASS</th>
<th>REAL CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral (N)</td>
<td>N</td>
<td>H</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Happiness (H)</td>
<td>6</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Sadness (S)</td>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Anger (A)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td></td>
<td>0</td>
<td>55</td>
<td>57</td>
</tr>
</tbody>
</table>

Experiment 2: S01

<table>
<thead>
<tr>
<th>Conf. Matrix</th>
<th>Decision Table</th>
<th>PREDICTED CLASS</th>
<th>REAL CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral (N)</td>
<td>15</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Happiness (H)</td>
<td>9</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Sadness (S)</td>
<td>3</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Anger (A)</td>
<td>7</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Precision (%)</td>
<td></td>
<td>44</td>
<td>31</td>
<td>44</td>
</tr>
</tbody>
</table>
Experiment 3:S02

<table>
<thead>
<tr>
<th>Conf: Matrix DecisionTable</th>
<th>PREDICTED CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL CLASS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral (N)</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>35</td>
<td>71</td>
</tr>
</tbody>
</table>

Experiment 4:S03

<table>
<thead>
<tr>
<th>Conf: Matrix DecisionTable</th>
<th>PREDICTED CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL CLASS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral (N)</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>41</td>
<td>37</td>
</tr>
</tbody>
</table>
## Experiment 5: S04

<table>
<thead>
<tr>
<th>Conf. Matrix DecisionTable</th>
<th>PREDICTED CLASS</th>
<th>REAL CLASS</th>
<th>N</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>20</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>15</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>6</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>16</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>35</td>
<td>60</td>
<td>48</td>
<td></td>
<td>Acc= 41.34%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Experiment 6: S05

<table>
<thead>
<tr>
<th>Conf. Matrix DecisionTable</th>
<th>PREDICTED CLASS</th>
<th>REAL CLASS</th>
<th>N</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>6</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>61</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>61</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>40</td>
<td>65</td>
<td>65</td>
<td></td>
<td>Acc= 55.55%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

161
C2: The result of Experiments after adding the missing emotions listed from 1 to 6 and shown in confusion matrix:

Experiment 1:S00

<table>
<thead>
<tr>
<th>Conf:Matrix DecisionTable</th>
<th>PREDICTED CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL CLASS</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>Neutral ( N )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Happiness ( H )</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Sadness ( S )</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Anger ( A )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>0</td>
<td>56</td>
</tr>
</tbody>
</table>

Experiment 2:S01

<table>
<thead>
<tr>
<th>Conf:Matrix DecisionTable</th>
<th>PREDICTED CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL CLASS</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>Neutral ( N )</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Happiness ( H )</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Sadness ( S )</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Anger ( A )</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>39</td>
<td>36</td>
</tr>
</tbody>
</table>
### Experiment 3: S02

<table>
<thead>
<tr>
<th>Conf: Matrix</th>
<th>Decision Table</th>
<th>PREDICTED CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral (N)</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>Happiness (H)</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Sadness (S)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Anger (A)</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td></td>
<td>35</td>
<td>71</td>
</tr>
</tbody>
</table>

### Experiment 4: S03

<table>
<thead>
<tr>
<th>Conf: Matrix</th>
<th>Decision Table</th>
<th>PREDICTED CLASS</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral (N)</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>Happiness (H)</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Sadness (S)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Anger (A)</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Precision (%)</td>
<td></td>
<td>33</td>
<td>38</td>
</tr>
</tbody>
</table>
### Experiment 5: S04

<table>
<thead>
<tr>
<th>Conf: Matrix Decision Table</th>
<th>N</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL CLASS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral (N)</td>
<td>20</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>15</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>6</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>16</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>35</td>
<td>60</td>
<td>48</td>
<td>0</td>
<td>Acc= 41.34 %</td>
</tr>
</tbody>
</table>

### Experiment 6: S05

<table>
<thead>
<tr>
<th>Conf: Matrix Decision Table</th>
<th>N</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>RECALL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL CLASS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral (N)</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>Happiness (H)</td>
<td>6</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>Sadness (S)</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>Anger (A)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>40</td>
<td>71</td>
<td>60</td>
<td>0</td>
<td>Acc=55.55 %</td>
</tr>
</tbody>
</table>