ADDING EMOTIONS TO SYNTHESIZED MALAY SPEECH USING DIPHONE-BASED TEMPLATES

SYAHEERAH BINTI LEBAI LUTFI

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2007

ADDING EMOTIONS TO SYNTHESIZED MALAY SPEECH USING DIPHONE-BASED TEMPLATES

SYAHEERAH BINTI LEBAI LUTFI

DISSERTATION SUBMITTED IN FULFILMENT
OF THE PARTIAL REQUIREMENT
FOR THE DEGREE OF MASTER OF SOFTWARE ENGINEERING

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA,
KUALA LUMPUR, MALAYSIA

FEBRUARY 2007

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: SYAHEERAH BINTI LEBAI LUTFI

Registration/Metric No: WGC040001

Name of Degree: MASTER OF SOFTWARE ENGINEERING

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Adding Emotions To Synthesized Malay Speech Using Diphone-based Templates

Field of Study: Synthesized Speech/Human-Computer Interaction

I do solemnly and sincerely declare that:

2 I am the sole author/writer of this Work;
3 This Work is original;
4 Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
5 I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
6 I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
7 I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 25th January 2007

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

ACKNOWLEDGEMENT

In the Name of Allah, the Most Beneficent and the Most Merciful

A journey is easier when you travel together; therefore interdependence is certainly more valuable than dependence. When I first started this project, I was in the middle of nowhere and quite lost. There were sleepless nights resulting from anxiety and nervousness. Thankfully, throughout my journey in producing this Master's dissertation, I have been accompanied and supported by many people, those who helped me not to get lost. It is of utmost pleasure for me that I have now the opportunity to express my gratitude to all of them.

I am greatly indebted to my supervisor, *Assoc. Prof. Pn. Raja Noor Ainon Zabariah*, the Deputy Dean of Faculty of Computer Science and Information Technology, Universiti Malaya for providing her kind guidance throughout the development of this study. Her comments have been of greatest help at all times. She has also provided all the necessary resources such as a lab and the equipments I needed throughout this project. My fullest gratitude also goes to my second thesis advisor *Prof. Zuraidah Mohd Don* from the Faculty of Language and Linguistic. It was quite a struggle for me to gather information regarding the Malay culture, as there were limited literature, but Prof. Zuraidah has greatly helped by providing important guidance in linguistic and Malay culture as well as proofreading my chapters. My warm and sincerest thanks also go to *Dr. Indirawati Zahid* of Academy of Malay Studies, for providing fruitful suggestions and advise during my first steps into Malay emotion studies.

I am deeply grateful to *Dr. Normaziah Nordin*, Project Manager of Pervasive Computing in MIMOS, and *Mr. Kow Weng Onn*, researcher of Language Engineering Group. He always kept an eye on the progress of my work and was always available when I needed his advises.

My sincere thanks to all those researchers I have had valuable discussions with and whom promptly answered my questions and shared their publications,: *Nur Hana*

Samsudin and *Sabrina Tiun* of Computer Aided Translation Unit, University Science Malaysia, *Encik Arry Arman*, developer of Indo-TTS, *Paul Boersma* and *David Weenink*, developer of Praat and all the others I may have missed. Not forgetting King's Scholar Mohd Norhakim for reading the manuscript and his funny jokes that makes me laugh some of the time.

This episode of acknowledgement would not be complete without mentioning my parents, who taught me the value of hard work by their own example. They rendered me enormous support during the whole tenure of my research, especially for being there for my daughter when she needed my company most.

Lastly, my deepest gratitude goes to to my husband *Imran Amin August de Roode,* not only for the inspiration and moral support he provided throughout my research work, but also for all the time he has spent providing me professional guidance and assistance in technical effort. I am also truly thankful to him for his patience over the long period of my absence. Without his loving support and understanding I would never have completed my present work. Particularly, I also owed to my three year old little girl, *Waheedah Amani* for all her cutest sayings and doings that have boost up my enthusiasm.

Finally, I would like to thank all whose direct and indirect support helped me complete my dissertation in time.

Syaheerah Lebai Lutfi

Master of Software Engineering
Faculty of Computer Science and Information Technology
University Malaya
Kuala Lumpur, Malaysia
21st March 2006

ABSTRACT

This study describes the addition of an affective component to the Malays TTS system in order to produce a system that is more expressive in nature. It introduces a new method for generating expressive speech by embedding an 'emotion layer' called **eX**pressive **T**ext **R**eader **A**utomation Layer, abbreviated as eXTRA. The emotion generation method is template-driven. The templates are diphone-based and each template carries unique affective data. The two types of emotions created for the system are anger and sadness. To ensure naturalness, the input sentence from user is matched with the template that consist of a sentence with the same syllable structure of the input sentence, allowing the emotion parameters from the template to be applied to the input at the level of phonemes. This syllable-sensitive matching process requires analysis of each syllable's consonant or vowel pattern.

The module is an independent component that can serve as an extension to any Malay TTS system that uses Multiband Resynthesis Overlap Add (MBROLA) engine for diphone concatenation. In a pilot project, the prototype is used with *Fasih*, the first Malay Text-to-Speech system developed by MIMOS Berhad, which can read unrestricted Malay text.

eXTRA is evaluated through perception tests. The results show more than sixty percent of recognition rate, which confirm the satisfactory performance of this approach. .The solution should provide improvement to output of Malay TTS system.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENT iii

ABSTRACT v

TABLE OF CONTENTS vi

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF ABREVIATIONS xii

1 INTRODUCTION 1

1.1 RESEARCH MOTIVATIONS AND OVERVIEW 1
1.2 ORGANISATION OF THESIS 2
1.3 BACKGROUND TO STUDY 4
1.4 RESEARCH OBJECTIVES 4
1.5 RESEARCH METHODOLOGY 5
1.5.1 INPUT GATHERING AND ANALYSIS 5
1.5.2 DESIGN AND DEVELOPMENT 6
1.5.3 EVALUATION 7
1.6 EXPECTED RESEARCH OUTCOMES 7
1.7 SCOPE OF STUDY 8
1.8 SIGNIFICANCE OF STUDY 8
1.9 APPLICATIONS OF EMOTIONAL SYNTHESIZED SPEECH 9
1.9.1 APPLICATION FOR THE LESS-ABLED 10
1.9.1.1 Application for the Visually Impaired and the Blind 10
1.9.1.2 Application for People with Dyslexia and Learning Disabilities (LD) 10
1.9.1.3 Application for Amyotrophic Lateral Sclerosis (ALS) Patients 11
1.9.2 APPLICATION FOR E-COMMERCE AND INSTANT MESSAGING 12
1.9.3 APPLICATION FOR PROOF READING 13
1.9.4 APPLICATION FOR WARNING AND ALARM SYSTEMS 13
1.9.5 APPLICATION AS SUBSTITUTE OF VOICE ACTORS 13
1.10 PUBLICATIONS 14
1.11 SUMMARY 16

2 EMOTIONS AND SPEECH SYNTHESIS 17

2.1 INTRODUCTION 17
2.2 UNDERSTANDING EMOTIONS 18
2.2.1 COMPONENTIAL PERSPECTIVES OF EMOTION 19
2.2.2 THEORIES OF EMOTION 20
2.2.3 LINKING EMOTIONAL SYNTHESIZED SPEECH TO EMOTION THEORY 22
2.3 VOCAL AFFECT 23
2.3.1 VOCAL AFFECT IN MALAY CULTURE IN RELATION TO ANGER AND SADNESS 25
2.4 ISSUES AND CHALLENGES OF EMOTIONALIZING SYNTHESIZED SPEECH 27
2.4.1 VARIABILITY IN EMOTION 27
2.4.1.1 Circumflex Model of Affect 28
2.4.2 VARIABILITY IN SPEECH 30
2.5 OVERVIEW OF MALAY TTS 31

2.6 SUMMARY 33

3 LITERATURE REVIEW 34

3.1 INTRODUCTION 34
3.2 EMOTIONAL SYNTHESIS TECHNOLOGIES 34
3.2.1 RULE-BASED SYNTHESIS 35
3.2.2 UNIT SELECTION SYNTHESIS 36
3.2.3 DIPHONE SYNTHESIS 37
3.2.3.1 The Multi-band Resynthesis Over-Lap Add (MBROLA) Synthesizer 38
3.2.3.2 How MBROLA Works 40
3.2.4 TEMPLATE-DRIVEN CONCATENATIVE SYNTHESIS 41
3.3 THE STRATEGIES OF INFUSING EMOTION IN MALAY SYNTHESIZED SPEECH 42
3.3.1 MALAY LANGUAGE SYLLABLE STRUCTURE 42
3.3.2 THE PROPOSED METHOD 43
3.4 SUMMARY 45

4 BUILDING EMOTION TEMPLATES TECHNIQUES 46

4.1 INTRODUCTION 46
4.2 TEMPLATE-DRIVEN EMOTIONS GENERATION 46
4.3 BUILDING THE SPEECH CORPUS 48
4.3.1 DESIGNING THE SENTENCE 49
4.4 VOICE RECORDING USING ACTED SPEECH 50
4.4.1 SPEAKER SELECTION 51
4.4.2 RECORDING SESSION 51
4.5 EVALUATION OF SPEECH SAMPLES 53
4.5.1 PERCEPTUAL TESTS 53
4.5.2 ANALYSIS OF PERCEPTION TESTS 55
4.6 PERCEPTION TEST RESULTS FOR HUMAN SPEECH SAMPLES 56
4.6.1 INITIAL STAGE (OPEN TEST) 56
4.6.2 FINAL STAGE (FIXED OPTIONS) 57
4.6.3 SUMMARY OF RESULTS AND DISCUSSIONS ON FINDINGS 58
4.7 ANALYSIS OF SPEECH SAMPLES 59
4.7.1 EXTRACTION OF RELEVANT PROSODIC FEATURES 59
4.8 SUMMARY 65

5 SYSTEM REQUIREMENTS, DESIGN AND PROTOTYPE 66

5.1 INTRODUCTION 66
5.2 SYSTEM REQUIREMENTS 66
5.2.1 USE CASE DIAGRAM 68
5.2.2 USE CASE DETAILS 68
5.3 SYSTEM DESIGN AND DEVELOPMENT 70
5.3.1 OVERVIEW 70
5.3.2 SYSTEM CHARACTERISTICS 71
5.3.3 SYSTEM ARCHITECTURE 72
5.3.4 SYSTEM DETAILED DESIGN 75
5.3.4.1 Block Diagram 75
5.3.4.2 Functional Decomposition Chart 76
5.3.4.3 Dataflow Diagram 76
5.3.4.4 Class Diagram 77
5.3.4.5 Outstanding Issues 80
5.3.4.6 Sequence Diagram 80
5.3.4.7 Object Model 82
5.3.4.8 The Process of Matching Sentences 83
5.3.4.9 Object Constraints 89

5.4 eXTRA PROTOTYPE **90**
5.4.1 GUI OF FASIH 90
5.4.2 GUI OF FASIH EXTENDED WITH eXTRA 91
5.4.2.1 eXTRA module's API 92
5.4.3 SCREENSHOT OF DEBUGGER 94
5.5 SUMMARY **94**

6 EVALUTION, RESULTS AND DISCUSSIONS **96**

6.1 INTRODUCTION **96**
6.2 EVALUATION OF eXTRA **96**
6.2.1 EVALUATION METHOD 96
6.3 PERCEPTION TEST RESULTS FOR SYNTHESIZED EMOTIONAL SPEECH GENERATED BY FASIH WITH eXTRA LAYER **97**
6.3.1 SUMMARY OF RESULTS AND DISCUSSIONS 97
6.4 SUMMARY **98**

7 CONCLUSIONS AND FUTURE WORK **99**

7.1 DISCUSSIONS OF FINDINGS AND APPLICABILITY **99**
7.2 CONTRIBUTIONS OF STUDY **100**
7.3 FUTURE WORK **101**
7.3.1 EXPANDING THE CORPUS / SPEECH DATABASE 101
7.3.2 EXTENDING THE EMOTION TYPES AND GENDER 102
7.3.4 IMPROVING THE CONTEXT-SENSITIVE INTONATIONS 103
7.4 REMARK **103**

8 REFERENCES **105**

9 APPENDIX A: EMOTIONAL EFFECTS CORRELATES TO SPEECH **111**

10 APPENDIX B: SENTENCES USED AS STIMULI FOR ANGER AND SADNESS **112**

11 APPENDIX C1: FORM USED FOR FIRST PERCEPTION TEST TO EVALUATE ACTOR'S SPEECH SAMPLES **114**

12 APPENDIX C2: FORM USED FOR FINAL PERCEPTION TEST TO EVALUATE ACTOR'S SPEECH SAMPLES **116**

13 APPENDIX D: NEUTRAL SENTENCES **118**

14 APPENDIX E: A LIST OF PHONEMES AND THEIR SAMPA CORELATES **119**

15 APPENDIX F: EMOTION TEMPLATE DATABASE **120**

16 APPENDIX G: PAPER SUBMITTED TO JOURNAL **121**

LIST OF TABLES

Page

Table 1.1: The Details of Publications 15

Table 2.1: Several Terms Relating to Anger (Wazir-Jahan, 1990) 28

Table 4.1: Sample Sentences That Satisfies Different Syllable Combinations for Anger Template. 50

Table 4.2: Effort Scale 54

Table 5.1: User and Functional Requirements 67

Table 5.2: Use Case Descriptions 69

Table 5.3: A Description of Basic Components 73

Table 5.4: A Description of Detailed Components 74

Table 5.5: A Description of the Blocks from the Block Diagram 75

Table 5.6: The Organization of Matching Between the Template and the Input Contents 88

LIST OF FIGURES

Page

Figure 1.1: Context Diagram Visualizing Possible Applications of eXTRA 9

Figure 2.1: Brunswickian Lens Model Adapted by Shrerer (1978) 24

Figure 2.2: Affective Space Called the Circumflex Model of Affect by Russell (1980) .. 29

Figure 2.3: Top-down Architecture of *Fasih* 33

Figure 3.1: MBROLA Input for Isolated Word 'Saya' 40

Figure 3.2: Simple Proposed Architecture of Affective *Fasih* 44

Figure 4.1: Simplified Framework of Emotion Layer Module 47

Figure 4.2: Results from the First Perception Test by Native Listeners 56

Figure 4.3: Results from the First Perception Test by Non-native Listeners 56

Figure 4.4: Results from the Final Perception Test by Native Listeners 57

Figure 4.5: Results from the Final Perception Test by Non-native Listeners 57

Figure 4.6: The Process of Extraction of Relevant Prosodic Features 60

Figure 4.7: Three-layer Annotation of Speech Signal in Praat Textgrid Format 61

Figure 4.8: Extraction of Significant Pitch Points and Their Durations for Phoneme /a/ in the Word 'kamu' (you) 62

Figure 4.9: Samples of Anger Templates 63

Figure 4.10: Samples of Sadness Template 64

Figure 5.1: Use Case Diagram for eXTRA 68

Figure 5.2: Major Components That Produce Emotional Speech 72

Figure 5.3: High-level Architecture of eXTRA 72

Figure 5.4: Low-level Architecture of eXTRA 74

Figure 5.5: System Block Diagram Showing the Data Processing 75

Figure 5.6: The Functional Decomposition of eXTRA 76

Figure5.7: Dataflow Diagram showing eXTRA's Internal Data Flow 77

Figure 5.8: Class Diagram of eXTRA 79

Figure 5.9: Sequence Diagram for "Get Emotionized Speech" 81

Figure 5.10: The `Sentence` Object Model 82

Figure 5.11: A Visualization of One of the Matching Processes 83

Figure 5.12: A Visualization of the Phonemes Matching Process 85

Figure 5.13: Emotional Prosodic Parameters Transfer from Template Sentence to Input Sentence 87

Figure 5.14: A Visualization of Related Objects and Their Constraints 89

Figure 5.15: ***Fasih***'s GUI 90

Figure 5.16: GUI of ***Fasih*** with eXTRA 91

Figure 5.17: Debugging Information Produced for Sentence "anda belajar sama dia" ... 94

Figure 6.5: Results for Both Synthesized Emotional Speech Using Neutral and Emotionally-inherent Content. 97

LIST OF ABBREVIATIONS

Abbreviations	Full Term
ADR	: Automated Dialogue Replacement
ALS	: Amyotrophic Lateral Sclerosis APIIT
ASCII	: American Standard Code for International Interchange
APIIT	: Asia Pacific Institute of Information Technology
BDD	: Bridging the Digital Divide
BM	: Bahasa Melayu
COCOSDA	: International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques
DECtalk	: Text-to-speech system by **D**igital **E**quipment **C**orporation
CRC	: Class-Responsibility-Collaboration
CV	: Consonant-Vowel
CVC	: Consonant-Vowel-Consonant
eXTRA	: eXpressive Text Reader Automation Layer
FCSIT	: Faculty of Computer Science and Information Technology
F0	: Fundamental frequency to create pitch
GUI	: Graphical User Interface
HAMLET	: Helpful Automatic Machine for Language and Emotional Talk Instant Messaging
IPA	: International Phonetic Alphabet
IRC	: Internet Relay Chatting
LD	: Learning Disabilities
MBROLA	: Multiband Resynthesis Overlap Add
MDG	: MBROLA-based speech Data Generator

MIMOS	: The Malaysian Institute of Microelectronic Systems
MP	: MBROLA-based speech Processor
ODBC	: Open Database Connectivity
OCR	: Optical Character Recognition
PHO	: Gerber Photoplot File (file extension)
POS	: Part of Speech
RP	: Received Pronounciation
SAMPA	: Speech Assessment Methods Phonetic Alphabet
SFS	: Speech Filing System
TACS	: Traffic Avoidance Collision System
TTS	: Text-to-Speech
UML	: Unified Modelling Language
UTMK	: Unit Transilasi Menggunakan Komputer or Computer Aided Translation Unit
UM	: University of Malaya
USM	: University Science of Malaysia
VC	: Vowel-Consonant

CHAPTER 1

Introduction

1.1 Research Motivations and Overview

When emotions are communicated, enhanced meaning and patterning are perceived and an atmosphere harmonious to natural learning is created. In particular, vocal emotions supply significant cues to delivering accurate, effective messages. However, the ability to express emotions distinguishes human speech from synthetic speech. The robotic and rather unnatural output quality of current Text-to-Speech (TTS) systems would be a factor that restrict the application of this technology. Therefore, a solution is needed in order to improve the output of TTS system to optimize its use

Human beings have emotions while machines do not. A Text-to-Speech system without an affective component may constraint communications in certain context of usage. This study is concerned with an application component for an effective communication system that can further extend concatenative Malay TTS systems. The aim is to propose a way of incorporating emotions to the Malay TTS. The method assumes that the TTS system is being built on the MBROLA (Multi-Band Resynthesis OverLap Add) engine for diphone concatenation. The application component, which is added as an affective layer to the Malay TTS, uses diphone-based emotion templates that contain various prosodic cues correlated with different emotional states. This study examines two

emotions, namely anger and sadness. The prototype, called eXpressive Text Reader Automation (eXTRA) has been designed and tested for use with ***Fasih,*** the first Malay concatenative speech synthesis system at MIMOS. An overview of ***Fasih*** is given at the end of Chapter 2.

Adding emotions to an existing TTS system may look straightforward at first, as it was assumed that this process is simple, but a closer look reveals many difficulties. In human speech, tones come out naturally, however this is difficult to incorporate tones in a non-human entity. The first problem is defining emotions themselves. The issue concerning the way emotion should be described in order to reveal relationships between utterances and specific emotions also poses a problem. In contrast with George Hofer's assumption that the expression of emotion is universal (Hofer, 2004, p. 16), Chapters 2 and 3 in this dissertation reveal that there is no single definition of emotion. It is not at all clear what emotional speech of people from one social environment or culture sounds like to another. This leads to the understanding that people globally can recognize emotional speech but have a hard time describing it. Despite all this, various attempts were made to infuse emotions in synthesized speech (refer section 3.4 in Chapter 3). To improve the quality of human-like sounding emotions, findings used in this dissertation derive from cross-disciplinary work that takes advantage of theoretic frameworks developed by psychologists to describe emotions.

1.2 Organisation of Thesis

Chapter 1 contains the introduction to the issues with which the research is concerned, the aims and objectives of the study and an outline of the research approach.

Chapter 2 introduces literature work concerned with emotions including vocal emotions from different perspectives, emotions in the Malay culture and issues pertaining to emotionalizing synthesized speech. In addition, this chapter presents the frameworks of emotions by various speech researchers used in emotive speech synthesis.

Chapter 3 presents the surveys of previous literature and studies relevant to emotive speech synthesis, as well as comparisons of different methods in emotive speech synthesis.

Chapter 4 describes the methodology employed in the study. The sub-topics include the research design and procedures adopted to build emotion templates. The methods used to collect the target speech data and evaluate them are also explained.

Chapter 5 explains the system requirements and shows detailed functional and technical designs of the eXTRA module that were derived from those requirements. The implementation of GUI and how eXTRA is hooked in the existing TTS, ***Fasih***, is explained at the end of this chapter.

Chapter 6 reports on the assessment of system evaluation. The results and findings generated from tests conducted for both human speech and synthesized speech are shown with the aid of charts. The interpretations of the results and the comparisons with those of previous studies are also presented here.

Chapter 7 discusses the findings and their implications. At the end of this chapter, suggestions for future researchers are presented.

1.3 Background to Study

One of the major complaints on existing TTS systems is the robotic quality of speech output. The quality of synthetic speech is not as good as what people expect it to be and can be tiring to listen to and therefore limits its application. This problem can partly be solved through the application of prosody, which is "the rhythmic and intonational aspect of language" (Merriam-Webster Online, 2005) into the speech output. Vocal emotions are a combination of different prosodic cues. Prosody is a combination of pitch, duration and paralinguistic features. In synthetic speech, the prosodic pattern is created by varying the combination mentioned. The variation of pitch (or fundamental frequency, F0) and duration is one of the main elements of a prosodic pattern that could avoid synthetic speech from sounding monotonous and hence, become more natural (Chen et. al, 2002; Cahn, 1990). Perceptual studies also signify that these two are the core conveyers of affect.

Similar to a synthesized dialogue system, Tatham, Morton and Lewis (1999) pointed out that users often become annoyed, bored and uninterested with unnatural and monotonous speech output. In contrast, a user's level of attention and understanding increases when machines he or she is interacting with elicit human-like sounds.

This project strives to eliminate the limitations mentioned above in synthesized Malay speech by enhancing the quality of the synthesized speech through the infusion of human emotions.

1.4 Research Objectives

The main aim of this project is to add an emotional layer to the Malay text to speech conversion system that uses diphone-based templates of anger and sadness to convey

these two emotional states. This layer supports a diphone-concatenation method. The more specific objectives are as follows:

- To study the acoustic correlates of the two most fundamental emotional states, i.e. anger and sadness, in terms of pitch and duration from real human voice samples. This can be done by investigating the way these two emotions affect Malay speech.
- To develop a *syllable-sensitive* algorithm that match user input speech data with template speech data in order to produce emotion-blended speech output.

1.5 Research Methodology

This section summarizes the applied research methodology. Detailed methods used to build emotion templates, findings that constructed the foundation of the design and development, and evaluation methods used to determine recognition of emotions in both human and synthesized speech are presented in Chapters 4 and 5.

1.5.1 Input Gathering and Analysis

To define the research problem, the researcher had several meetings and discussions with the relevant researchers from MIMOS BERHAD, Universiti Sains Malaysia's speech researchers and local linguist experts. Conference calls were also conducted to communicate with researchers in speech science from local and foreign countries in addition to the ongoing participation in discussion forums that helped to gather information. In addition to the availability of and accessibility to data, prior studies, research papers, library and internet searching was conducted). This helped to narrow down the problem and help the researcher to have deeper understanding of the current issues in speech synthesis as well as to adopt the methodology used by previous

researchers in this area. Consequently, these studies formed ideas that led to the proposal of a novel method to generate emotions; template-driven emotions generation (presented in Chapter 4). Specifically, previous linguistic and speech studies provided clearer ideas on speech features that are beyond formal grammar, as well as introducing factors that_discriminate natural speech as machine response. Previous studies on emotion theories were examined to draw up assumptions about emotions. These studies also include the findings from observations by anthropologists and researchers on cultural-specific emotions, particularly, the Malay culture. These findings were used to elicit the requirements for the emotion templates that depict the affective tones in eXTRA.

To develop better understanding of human vocalization (particularly in terms of changes in prosodic parameters), the researcher also listens to, analyzes and compares both natural and acted emotional voice samples from different genders, cultures and languages. Interesting findings include the fact that female voice tends to have higher pitch than male in every emotion, and expressions of emotions are dependent on languages and cultures. Multiple voice recordings by professional actors were conducted and samples were analyzed to extract emotional prosodic information.

1.5.2 Design and Development

First, the requirements were identified and requirement-statements were drafted. An informal technique, Class-Responsibility-Collaboration (CRC) cards was used to define the classes involved. The basic architecture was drafted based on the insight gained from the involved classes. Subsequently class relations and the object model were determined to support the emotionizer algorithm. Open Database Connectivity (ODBC) and a simple database storage mechanism was chosen to ensure that in a later stage this

can be replaced by a more advanced database if necessary. The design for the prototype module takes advantage of Object-Oriented design techniques for the purpose of flexible fine-tuning all parameters to best match the requirements. To maintain semantic consistency, while allowing the design to maintain its flexibility, the concepts of polymorphism and encapsulation were employed. Unified Modeling Language (UML) oriented tools were used for the design. The eXTRA prototype was written in Java ™ version 1.5., giving the advantage of easy hookup with targeted TTS, that are mostly written in the same language.

1.5.3 Evaluation

Perceptual tests (listening tests) were conducted twice. The first test was conducted at the beginning of the research to evaluate recorded human voice samples. The purpose is to extract emotional prosodic parameters from human voice samples that obtained high recognitions and use them to build emotion templates. Later, another perceptual test was conducted at the end of the research to evaluate the emotional synthesized speech output generated by the TTS extended with eXTRA. This determines the success of eXTRA. Apart from perceptual tests, this module was evaluated through demos in different stages with the researchers from MIMOS Berhad.

1.6 Expected Research Outcomes

The main contribution of this project is to provide an automated expressive text reader module that shall be attached as a layer to any MBROLA based concatenative synthesized speech system. The purpose is to increase the quality in synthesized speech systems by producing a natural sounding output that has varying degrees of emotions. In addition, it is hoped that different expressed emotions can be distinguished by users,

thus, improving believability in the system. The prototype initially provides two types of emotions; anger and sadness to be applied to the neutral output of the base system.

1.7 Scope of Study

As a pilot study, the prototype features anger and sadness. Users are able to apply these two emotions to speech output. The stimuli used for emotional diphone templates are a series of four-word Malay declarative utterances that are limited to two and three syllables, as Malay language mostly contains two and three-syllable words.

1.8 Significance of Study

Adding emotions to synthesized speech means that the latter can verbalize written language with the kind of emotion appropriate for a particular occasion (e.g. announcing bad news in a sad voice). Speech articulated with the appropriate prosodic cues can sound more convincing and may catch the listener's attention, and in extreme cases, can even avoid tragedies. A warning announcement such as "This tape will self-destruct in five seconds" will be more effective if delivered in an anxious tone. A lesson learned from a mid-air crash tragedy exposed by the National Geographic Air Crash Investigation series (Barro & Vaillot, 2004), is that the pilots in this particular tragedy were more inclined to listen to the wrong information given by the human flight controller instead of the Traffic Avoidance Collision System (TACS). The controller and TACS both gave instructions in order to avoid mid-air collision with another flight. However, the instructions were contradictory; the controller radioed them to fly upwards, while the TACS instructed them to fly lower, in its robotic voice. Though the TACS was accurate, the pilots were influenced by the intensity and emotions expressed in the voice of the human controller, and decided to move upwards which caused

collision that claimed hundreds of passenger's lives. It is said that if the system's voice output was less robotic and had a more appropriate emotion correlating with its instructions, the tragedy could have been avoided.

1.9 Applications of Emotional Synthesized Speech

In principle, expressive or emotional synthesized speech can be used in any human-machine interaction devices. Improved synthesized speech can benefit other speech-based systems that carry out various tasks such as weather information over the telephone, auditory presentation of instructions for complex hand free tasks (Tams & Thatam, 2000) and as proofing tool, where text from columns and newsletters are read aloud to help users find errors that their eyes might not spot even with repeated readings. A TTS system that is equipped with eXTRA may be used in a range of applications. To show the system in its context of possible usage, a context diagram (Figure 1.1) is presented next, followed with elaborated descriptions of the current and future use of emotional synthesized speech in different domains.

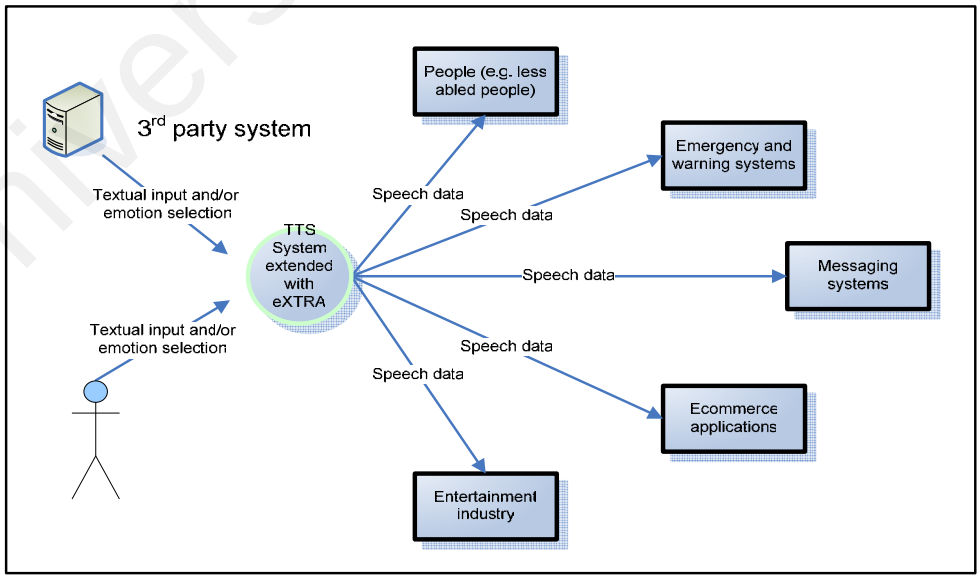


Figure 1.1: Context Diagram Visualizing Possible Applications of eXTRA

1.9.1 Application for the Less-abled

1.9.1.1 Application for the Visually Impaired and the Blind

Emotive synthesized speech can be used to replace specific audio books as reading and communication aids for the visually impaired and the blind. This can be done by combining a speech synthesis system with Optical Character Recognition (OCR) to allow for the conversion of printed text into concatenated intelligible speech. By adding emotions to synthetic speech, the result becomes more acceptable; monotone effects can be avoided.

Additionally, it can also help to convey messages more accurately. For example, some important information of bold and underlined text may be communicated using a slight change in pitch or loudness. Another important feature is that the speech system can communicate advance information such as the estimated duration of reading to the listener, in an attempt to overcome a blind person's disadvantage of being unable to see the length of an input text.

1.9.1.2 Application for People with Dyslexia and Learning Disabilities (LD)

For people who are dyslexic, it is almost impossible to learn reading and writing without spoken help. With proper computer software, unsupervised training for these problems is easy and inexpensive to arrange. Most traditional computer speech software for dyslexic patients use pre-recorded or digitized speech. Though pre-recorded speech sounds the most natural, it takes up a lot of disk space and can only read out what has been recorded. Therefore synthesized speech systems provide a great leap forward in making text accessible for dyslexic people. Dyslexic children have difficulty segmenting words into individual syllables or phonemes and have trouble blending speech sounds into words. Emotive TTS can especially help these children to have a

phonemic awareness by improving their understanding of and access to the sound structure of language. A TTS that mimics human speech tones may also act as a dictation system for dyslexic children and help them with reading, spelling and word finding problems. It is ideal for teaching pronunciation as the words are pronounced correctly with the correct pitch and loudness, and words with similar spellings but with different pronunciation (‘read’ /rid/ and its past form ‘read’ /red/) are pronounced accordingly. Apart from this, an emotive TTS can also function as phonetic spell checkers for poor spellers or dyslexic people. Dyslexic people spell phonetically, that is they attempt to spell the word as it sounds. Though Malay language is phonetically regular, some loan words may be wrongly spelt. An example is "fisik' for 'fizik’. An emotive TTS can help to avoid this kind of errors if it is extended with a phonetic spell checker feature.

1.9.1.3 Application for Amyotrophic Lateral Sclerosis (ALS) Patients

ALS is a disorder of motor nerves, resulting in muscle weakness. This occur in the regions of the in arms and legs and in swallowing, speech, etc. Many patients with ALS experience difficulty with speaking because of their inability to project the voice or the inability to form the words. Research shows that many ALS patients feel overwhelmed when members of their family feel tempted to speak to others on their behalf if they know what the patients are attempting to say (Forshew, 1999). There are many alternate communication systems available for ALS patients. However, an emotive TTS can make a significant difference. It can provide a more effective and intimate communication by enabling patients to not only pass their messages but also to indicate their emotional state by conveying the messages using the appropriate intonation

1.9.2 Application for E-commerce and Instant Messaging

Computerized Instant Messaging (IM) and Internet Relay Chatting (IRC) use real time communication that is usually limited to sending text as a means of communication. Affective instant messaging applications allow users to express their emotional state by detecting tagged emotional inherent words or symbols and automatically displaying the related emoticons with certain facial expression. These IM applications can be improved by replacing text message output by emotional synthetic speech output. This way, messages will be perceived more accurately in their original context.

The E-commerce world may be expanded and improved by expressive speech applications. Online shops may use different voices and tones (e.g. introvert or extrovert voices) instead of just text or flat-voice reviews to describe merchandise. Not only will merchandisers be able to improve their existing markets, but markets may become more accessible to the visually impaired and the blind, as well as to non-literate people. Experiments by Nass and Lee (2002) revealed that participants in a mock Web auction and mock Web book store are more likely to buy products being sold when information is communicated by an attractive voice that expresses personality, despite many reminders from the researchers and the synthesized voice itself, explaining that the voice was not a real human's voice. Additionally, participants appreciated the text or content being read much more, and even found the writer to be more credible and likeable. Thus, confirming the general observation that computers and computer-synthesized voices are to a certain extent, "social actors". In other words, people respond to a computerized voice that sounds like them just as they would to a real person who sounds like them. Just as with real people, they prefer consistency in behavior (behave similarly) because it's easier to understand and predict.

1.9.3 Application for Proof Reading

An expressive speech synthesizer may be used with a word processor as an aid to proof reading. Many users find it easier to detect grammatical and stylistic problems when listening to natural speech instead of reading text. This also makes common spelling errors easier to detect.

1.9.4 Application for Warning and Alarm Systems

In warning and alarm systems, expressive or emotional synthesized speech may be used to give more accurate information of the current situation. Using emotional speech (e.g. fast or loud speech) instead of warning lights or buzzers gives an opportunity to reach the warning signal for example from a distance. A spoken alarm that instructs a person to observe distance from a protected object is more likely to succeed in preventing the person from inflicting damage to the object compared to sirens. Findings from a synthetic speech warning system research (Lee *et al.*, 2004) indicates that vocal cues offer a promising alternative to other types of auditory warnings that could increase response time and reduce annoyance.

1.9.5 Application as Substitute of Voice Actors

This is most useful in the entertainment industry. In making a film or motion picture a worldwide attraction, a voice-over dubbing or translation is often applied to replace the original dialogue. In Malaysia, foreign-language television series, soap operas or anime, especially from Japan, Korea, South America and Spain are often dubbed into Malay to make them more accessible to the local audience. Voice artists are used to speak the new voice track and they can be very expensive. Sometimes, expensive technique such as Automated Dialogue Replacement (ADR) is applied in dubbing to avoid

uncontrollable issues such as traffic or animal noise, during principle photography,. This technique involves the re-recording of dialogue by original actor *after* photography. One way to cut cost is to replace voice actors and related techniques with synthesized speech. Communicating emotions is a core requirement for synthesized speech to compete with human voices.

1.10 Publications

Along the research path, several papers were presented and published in international conference proceedings and also proposed for submission to an international journal.. The two first papers primarily introduced the template-driven method and reported on preliminary experiments and the results. The initial findings obtained from the experiments are used to draw up requirements and rationale the design decisions of eXTRA. Consequently, the last paper reports the complete work, and will be presented in Penang, Malaysia on December 2006. The complete details of the publications are listed in Table 1.1 in the next page:

Table 1.1: The Details of Publications

Paper Presentation I	
Title	*Adding Emotions to Malay Synthesized Speech Using Diphone-Based Templates*
Author	Syaheerah L Lutfi, Raja Noor Ainon, Salimah Mokhtar, Zuraidah M. Don
Conference Proceeding	iiWAS '05: 7th International Conference on Information and Web-based Applications & Services, Kuala Lumpur Malaysia, (pp 269-276), September 19-21.
Paper Presentation II	
Title	*Template-Driven Emotions Generation in Malay Text-to-Speech: A Preliminary Experiment*
Author	Syaheerah L Lutfi, Raja Noor Ainon, Salimah Mokhtar, Zuraidah M. Don
Conference Proceeding	CITA '05: 4th International Conference of Information Technology in Asia, Kuching, Sarawak, Malaysia (pp.144-149), December 12-15.
Paper Presentation III	
Title	*eXpressive Text Reader Automation Layer (eXTRA): Template-driven Generation of Emotions in Malay Synthesized Speech* (Paper accepted and in press)
Author	Syaheerah L Lutfi, Raja Noor Ainon, Zuraidah M. Don
Conference Proceeding	Oriental COCOSDA '06: 9th International Conference of Speech Databases and Assesment, Penang, Malaysia (in press), December 9-11.
Paper Submitted to Journal	
Title	*Adding Emotions to Malay Synthesized Speech Using Diphone-Based Templates*
Author	Syaheerah L Lutfi, Raja Noor Ainon, Salimah Mokhtar, Zuraidah M. Don
Journal	iiWAS Special Issue Journal 2006

1.11 Summary

This chapter has presented the motivation for the research objectives, significance, methodology and also the target use of the prototype In summary, developing the automated expressive text reader module will reveal the software engineering approach in optimizing the use of Malay TTS system with the addition of affective component. The aim is to produce an output which is as natural and meaningful as possible. In the next chapter, background research that was conducted regarding emotions will be presented. An overview of the first Malay TTS will be given in the last section of Chapter 2.

CHAPTER 2

Emotions and Speech Synthesis

2.1 Introduction

It is reported that the addition of prosodic parameters is insufficient to make the output of synthesized speech indistinguishable from human speech. Referring to paper works and research conducted on tonal languages, such as Mandarin and Korean synthesized speeches, (Chen *et al.*, 1998, 2002; Lee and Oh, 1999; Hu and Chen, 1999; Cai *et al.,* 1998) it is discovered that although many prosodic system models were introduced in the hope of providing a high degree of naturalness, there is a dearth of prerequisite studies on various human emotions.

Instead, more efforts were put into studying linguistic features and translating these into prosodic patterns, particularly F0 contour (computational frequency to create pitch). The prosodic patterns were then repeatedly applied to speech. While this may be good for a small amount of sentences, repetitive tones become boring and tedious for reading whole paragraphs of text. Applying fixed qualitative rules to prosodic variation patterns comes with great limitations. One way of solving this invariability problem is through the infusion of emotion into the speech. TTS systems generally do well in reading out factual sentences but not for conversations in a novel, for example. Adding emotion can

significantly improve the quality of applications like talking agents and virtual characters.

This chapter reviews the theories of emotions, focusing in particular on anger and sadness because these two emotions are the most fundamental and generally, the most recognized emotions. This will stop us from the pitfall of repeating the efforts of the extensive emotion study resources and will help us to understand what is important in describing a particular emotion. Besides that, the review of relevant work will provide information on how emotions are perceived and how variations of emotions are being classified and organized. Further, the chapter describes how emotion relates to speech and move on to the major problems that are faced by researches in order to produce synthesized emotional speech.

2.2 Understanding Emotions

Emotions influence human's life so greatly to the extent that either too much or too little of it impairs thinking and decision making. Picard (1997) illustrates this by quoting Aristotle's *Rhetoric*:

"Indeed they are always in sympathy with an emotional speaker even when there is nothing in what he says; and that is why many an orator tries to stun the audience with sound and fury"

Emotions play a significant role as an affective influence, especially in capturing attention (Brave and Nass, 2003). In fact, the significance is so great that it leads brain-based researchers such as Marilee Sprenger (2002) to assert that emotions mediate just

about anything, as she puts it, “emotions drive attention, and attention drives learning, memory and just about everything else"., Anything that does not have emotions is often perceived as boring or unconvincing. When confronted with an interface, users actually constantly (non-consciously) monitor cues to the affective state of their interaction partner. (Reeves and Nass, 1996 as cited in Brave and Nass, 2003). This obviously shows that a natural and efficient interface requires not only recognizing emotion in users, but also to *express* emotion.

2.2.1 Componential Perspectives of Emotion

Researchers agree that emotions are not just feelings, subjective experiences or something that is thought of. For example, even *tiredness* is classified as an emotion by some researchers (Murray and Arnott, 1993). To get a clearer picture of what emotions are, theories of emotion are studied to find out the different aspects or components that formulate different assumptions on the factors that describe emotions:

- Physiological or biological change patterns, e.g. changes in the facial muscles indicating smiley face, pupil dilation, etc.
- Cognitive or Appraisal where an individual evaluates what is right and wrong according to the situation or event,
- Social or cultural influences.
- Subjective feeling (Russell, 1980)
- Vocal expression (Banse and Scherer, 1996)

How these components are achieved will be explained in the next section.

2.2.2 Theories of Emotion

Theories of emotion expand over time and are always referred to, as they are essential to no matter what aspect of emotion being studied, including speech synthesis. However, despite the different theories and terms listed for emotions, there is no wide acceptation on a universal definition of emotion (Picard, 1997; Mozziconacci, 2002). Cornelius (1996) described the four traditional perspectives of emotion theories starting with Darwinian, Jamesian, Cognitive and the most recent, Social Constructivist. Each of these traditions has its own strength, and there is a fair degree of overlap among the perspectives. He also pointed out that those researchers who study speech and emotion automatically becomes a subscriber of one or more of the theoretical perspectives and research traditions mentioned above.

The Darwinian Theory is the foundation of other theories. It conveys that there are certain universal basic facial expressions that distinguish various emotions. Following this assumption, some researchers identified six basic emotions according to six universal facial expressions that are recognized in many cultures; the rest categorized seven and some other, eight. In general, most emotional-based investigations consider six basic emotions, termed as the Big Six by Cornelius (1996). They are ***happiness, sadness, fear, disgust, anger*** and ***surprise***. Majority speech technologists usually add ***neutral*** as fundamental emotion because neutral often becomes a base acoustic reference used for comparisons with other emotions' acoustical signals. These emotions are considered to be fundamental because each fulfills a specific role in helping living things deal with key survival issues posed by the environment (Plutchik, 1980), while other emotions can be derived from them (Cornelius, 2002). An example provided by Bulut, Narayanan and Syrdal (2002) is that both hot-anger (rage) and cold-anger (hostility) are regarded in the same category, although they show different acoustical

and psychological characteristics. Therefore, a perceived emotion can be interpreted as one of the seven basic emotions.

The Jamesian approach assumes that facial expression may also be accompanied by physiological response. James is a major proponent of emotion as an experience of bodily changes, such as heart rate increasing or perspiring hands (Picard, 2000). While emotions seen within the most dominant perspective; the Cognitive approach, are more dependent, whereby emotions are said to be inseparable with thoughts. This is also known as *appraisal* theory, in which its major assumption is that emotions are elicited and differentiated on the basis of the person's subjective interpretation or evaluation of a situation, object or event (Schrerer, 2000). Perceived emotions are intellectualized in this approach.

The most recent theory that also emphasizes mental components, the Social Constructivist, points out that physiological factors alone (such as bodily movements and facial expressions) are insufficient to explain emotions. Instead, the social environment influences the construction of emotions; hence 'socially constructed'; taking into account gender, personality and cultural environment (Cornelius, 2002). This theory takes into account non-standard emotions that may be culturally-specific, for example; envy, hope, guilt, even culturally-specific ***anger*** etc. A study about German intonation by Gibbon (1998) established the fact that the interpretation of spoken utterances is also influenced by the socio-cultural expectations of the spectators/listeners. The study demonstrated many differences, even across languages which are closely related such as Standard Northern German and Received Pronounciation (RP) English. According to Gibbon, superimposing the typically higher

pitch rage of English females onto a German female voice makes the voice sound "aggressive and over-excited".

In line with the Social Constructivist perspective, anthropological studies on emotions in Malay culture are discussed in section 2.4.1 The findings of the studies are used to 'customize' the emotion templates so that they describe tones that the local listeners are more familiar with.

2.2.3 Linking Emotional Synthesized Speech to Emotion Theory

It is understood that although emotion is ill-defined and imprecise, the intuitive concept of emotion in most of us is well-established. This section explains how a synthesizer system relates to the theoretical background laid out above.

The trend of the traditions above show that emotions are observed from three major perspectives; ***physical, cognitive*** and more recently, ***social***. The correlation of these aspects to speech is obvious. When we get emotional due to some cognitive interpretation, it reflects on physiological changes that further correspond on speech that show up primarily in F0 and timing (duration). However, the system is not embodied and therefore the Jamesian tradition is not relevant to this research. The cognitive perspective is related to the system because the intended perceptions of certain emotion concepts are evoked in the listener's mind. The social constructivist perspective is very important because the aim of this project is to portray emotions in relation to anger and sadness close to the Malay culture. This theory also fulfills the result achieved from the listening test that shows differences in perceptions between native and non-native speakers of the Malay language.

2.3 Vocal Affect

Emotions can be conveyed verbally or non-verbally in communications. One of the most affective signals in emotion is expressed through speech, the principle mode of communication. For example, slower, lower-pitched speech, with little high frequency energy, generally conveys sadness, while louder and faster speech, with strong high-frequency energy and more explicit enunciation, typically accompanies joy (Picard, 1997). See Appendix A for more examples of Emotional Effects Correlates to Speech adapted from Murray and Arnott (1993). In fact, many studies demonstrated that it is possible to identify the various aspects of speaker's physical and emotional state, including age, sex, appearance, intelligence and personality by voice alone.

The interest in the mechanism of human speech has led to speech studies which are later channeled to adaptation of speech by electronic devices. Studies on vocal expressions of emotions are developed from two approaches, perception-oriented and acoustic-oriented (Razak, Abidin and Komiya 2003; Cahn 1990; Scherer (1978) as quoted by Hofer 2004). The perception-oriented approach is listener centered and is concerned with how the listeners ***perceive*** the emotions, whereas the acoustic-oriented approach is speaker centered and is concerned with the analysis of vocal parameters of ***expressed*** speech that links to emotion. Acoustic features in speech are further divided into ***prosodic*** and ***phonetic*** classes. Phonetic features deal with basic sounds such as sounds produced by vowels or consonants in speech and their pronunciations. Prosodic features are composed of pitch, temporal (duration) and amplitude structure that correlate to the intonation or rhythmic aspects of speech such as stress word in utterance, the raising and falling of pitch and accents (Razak, Abidin and Komiya, 2003; Murray and Arnott, 1993). Prosodic features of speech contribute more to emotional expressions in speech than phonetic features.

The Bruswickian lens shown next that is adapted by Shrerer (1978) describes the expression and perception of emotion.

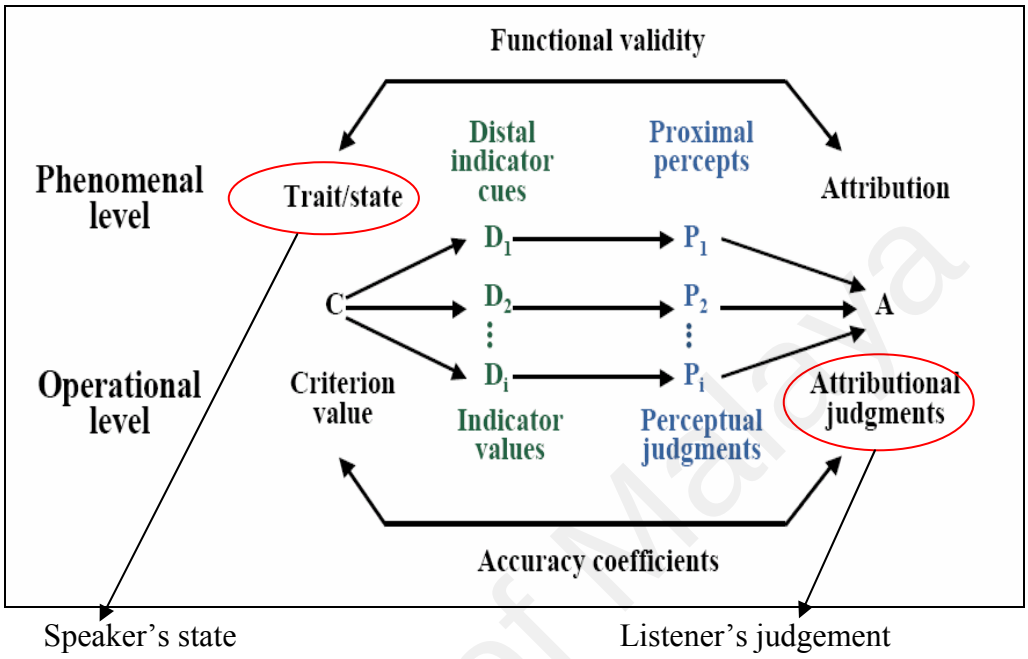


Figure 2.1: Brunswickian Lens Model Adapted by Shrerer (1978)

Figure 2.1 gives visual support to the acoustic (speaker) and perception (listener) oriented researches. It reads like this: Speaker's emotional state C is expressed through "Distal indicator cues" D which corresponds to the various acoustic parameters belonging to a certain emotion. The cues perceived by the listener as "proximal percepts" P are the pitch and other voice parameters. The attribution A is the final perceived speaker's emotion.

In addition, Murray and Arnott (1993, 1996) also stressed on voice quality as one of the factors that correlates voice to emotion. Naturalness in emotional synthesized voice depends on the underlying "personality" of the human voice recorded which systems are based on. In this project, the system's original recorded voice can be changed by parameters manipulation to mimic the actor's voice. This could also improve the TTS

system's voice, which is usually unclear because of default voice parameters. Vocal effects associated with several basic emotions are accounted in detail in Appendix A, adapted from Murray and Arnott (1993)

The next section discusses voice affect that is culturally focused.

2.3.1 Vocal Affect in Malay Culture In Relation to Anger and Sadness

The expression and perception of emotions may vary from one culture to another (Silzer, 2001). In an attempt to develop emotions in a synthesized speech which are indistinguishable from human emotions, as outlined in the project goal, characteristics of anger and sadness in Malay culture are studied. However, there is not much literature on vocal emotions in Malay, and so far, no study on anger and sadness from a Malay perspective is found. In an attempt to understand emotions from the Malay perspective especially with regard to anger and sadness, the work by Wazir-Jahan (1990) is referred to quite substantially. According to her, the description of the emotions in Malay was not based on empirical research but based on passing observations and intuitive reasoning. She concedes that many studies have been carried out on latah (for women) and amuk (for men, English amok), since these two expressions of emotion are closely related to the understanding of the 'Malay mind' then brought about by rebellious reactions against colonization. Wazir-Jahan examined the observations of the Malay mind by several western anthropologists who believe that the Malay people look 'externally impassive' but are actually very sensitive even to something as normal as 'the accidents of every day life'. Evidence gathered from past observations seem to show that the Malays are inclined to keep their emotions in check until the time when they cannot contain them anymore and that is when they explode.

These observations seem to be in line with what is expressed by the former Prime Minister, Tun Dr. Mahathir in his book The Malay Dilemma, "the transition from the self-effacing courteous Malay to the amok is always a slow process. It is so slow that it may never come about at all. He may go to his grave before the turmoil in him explodes" (Mohamad, 1981). In this thesis, the interest does not lie in the phenomenon of amok in itself but in its expression since it bears elements of a culturally specific form of anger.

A study carried out by Silzer (2001) illustrates that the expression of human emotions are cultural specific, e.g. how anger is expressed in English is different from how 'marah' is expressed in Malay. He explains that the causal component of marah is more specific such that 'marah' "is the result of intentional personal offence, where the offender knowingly committed the "bad" act, while realizing it would cause the other person unpleasant feeling". This causes the offended party to inform the offender in a certain tone of voice that he or she has done something wrong, and should rectify the problem. It is also observed that when expressing anger, Malays are inclined to shout. This way of expressing anger could probably caused by the accumulation of negative feelings which when released manifest in the form of shouting or yelling.

From the preliminary studies, it is also found that when uttered in anger, Malay utterances tend to have a slightly higher overall pitch while sadness is accompanied by lower overall pitch when compared to English utterances (Razak, Abidin and Komiya, 2003).

2.4 Issues and Challenges of Emotionalizing Synthesized Speech

Intelligibility, naturalness and variability are three aspects that differentiate human speech and synthetic speech (Murray, Arnott and Rohwer, 1996). Intelligibility is concerned with how far the words spoken by machine are understood and the different speaking styles. Modern TTS systems have no problem with the former, but the latter is a challenge. One aspect of naturalness is the variability introduced by the emotional state of the speaker and related pragmatic effects. However, the fact that the complex nature of human speech depends on the variability of speaking style and emotion of the speaker, makes it complicated for naturalness to be imitated by machine (Bulut, Narayanan and Syrdal, 2002; Murray and Arnott, 1996). This section reports the two major obstacles faced by researchers as well as in this very project, in the attempt of emotionalizing speech. These obstacles are intertwined with each other, which are variability in emotion and variability in speech

2.4.1 Variability in Emotion

Although section 2.4.1 explained that ***marah*** is more specific than ***anger***, it does not alter the fact that there are still different descriptions and interpretations for every subset of emotion that causes precise conceptualization of the underlying emotional states almost impossible. This makes modeling emotional speech very complex. Although labels are assigned to them like emotion-denoting words, as shown in Table 2.1, some terms may fit better to a certain emotion than others. Concentrating on anger alone, there are several emotive states within, which may be combined or appear exclusively. It largely depends on which context it is looked upon, which in turn, depends on the extent to which a person demonstrates his behaviour that is indicative to any of the emotive states (Wazir-Jahan, 1990). In her book, Wazir-Jahan has listed several positive

and negative emotional states. The negative emotional states relate to anger, which are familiar to the Malays in different terms, as extracted in Table 2.1:

Table 2.1: Several Terms Relating to Anger (Wazir-Jahan, 1990)

Negative Emotive States	
Malay	English
Iri hati	Envious
Sakit hati	Angry, with a tendency for revenge; vindictive
Busuk hati	Unpleasant trait with a combination of feelings of jealousy and malice
Main hati	Casual flirtation; not to be taken seriously

2.4.1.1 Circumflex Model of Affect

Because there are so many variations of emotion, researches design several semantic models of emotions with emotions clustered in a certain space. One of the popular ones is the Circumflex Model of Affect by Russell (1980) as shown in Figure 2.2 in the next page. Instead of independent emotion categories, this model uses circular ordering of emotion dimensions that straightforwardly classify an emotion as close or distant from another one. By having subjects rate the similarity of different emotion words and converting the ratings into angles, a circular ordering emerged (Plutchik, 1994 as cited in Hofer, 2004). Thus, the circular ordering functions as an affective space.

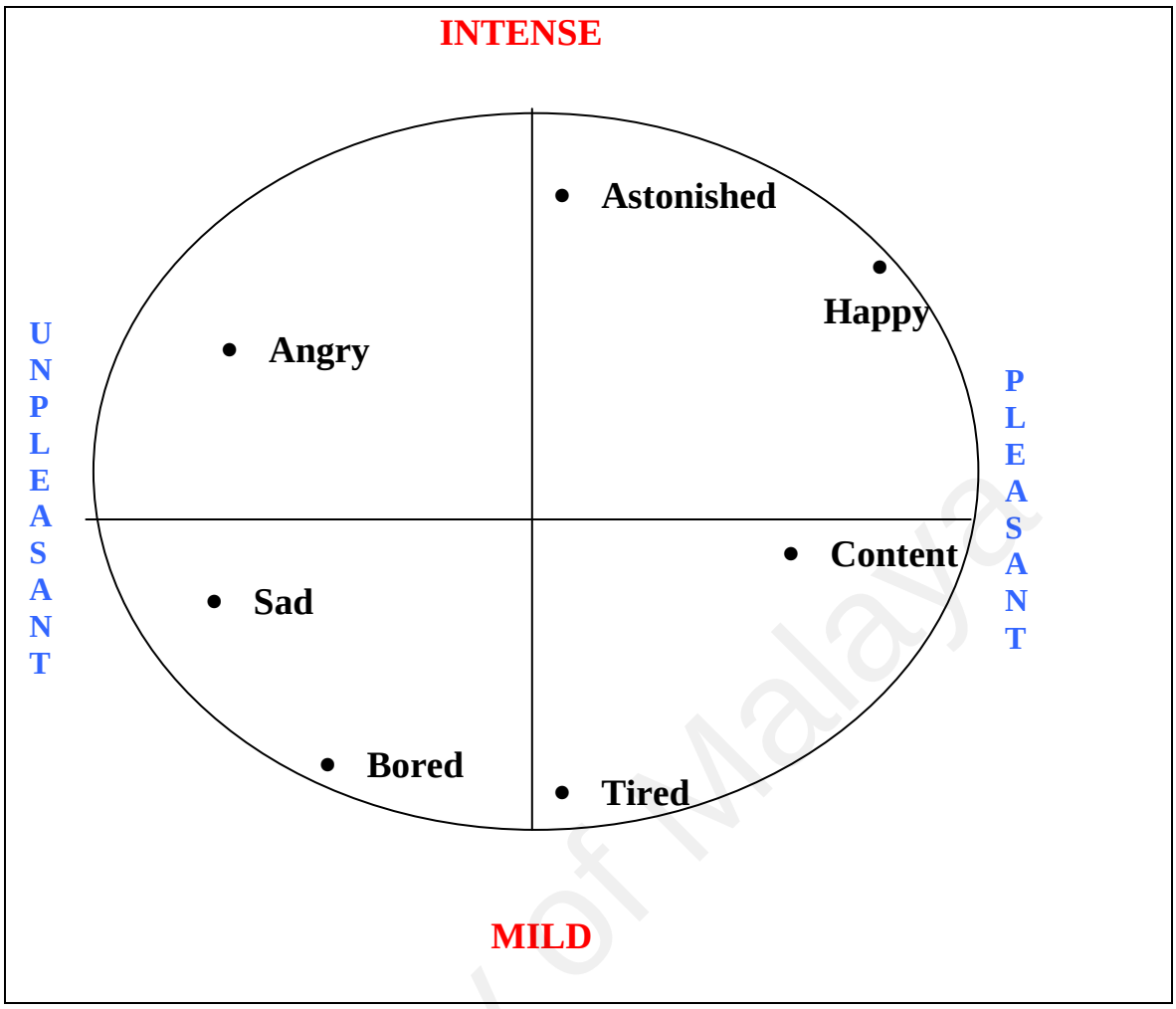


Figure 2.2: Affective Space Called the Circumflex Model of Affect by Russell (1980)

Technically, this model is irrelevant to the production of emotion in *Fasih*, at least at this early stage since what is more important is to create acoustical templates that contains the right emotional voice parameters. Semantic models of emotion independent of acoustical correlates are significant only when automatic control of affect is a key (Cahn, 1990). In emotional synthesized speech models where speech is quantified, the subjective semantic affect of emotion is often ignored. However, it is used to assist the data analysis from perception tests as explained in Chapter 4 (section 4.5.1), to determine in which dimension a listeners' perceptions falls into.

2.4.2 Variability in Speech

A speaker may not repeat what he says in the same way; he may not use the same words to say the same thing twice knowingly or not (even in read speech) (Murray and Arnott, 1996). One can also say the same word in many different ways depending on the context. Therefore, the instances of the same word will not be acoustically identical. This is quite difficult to map in a TTS system, especially when using qualitative rules, which causes the repetition of the same set of prosody when reading long sentences.

Murray and Arnott (1996) listed three factors that lead to variability in speech, namely:

i. **Speaking style**: Way of speaking is altered in response to conditions related to the speakers' environment and their status relative to those whom they are speaking to. (E.g. speaking differently to a child than to a peer).

ii. **Emotion and mood**: Speech production mechanism of a speaker is affected by different emotional states, or longer-term states; mood.

iii. **Physiological stress**: Wide-ranging factors relating to physiological arousal such as fatigue, illness, effects of drugs and workload, which are very difficult to include in synthesized speech.

Other than the above, the below factors discussed in Shih and Kochansky (2002) also add variability in speech:

iv. **Word-prominence in an utterance**: Depending on the context, a sentence can be said in different ways. How it is being said, (which word is stressed) could already provide enough information for the listener to interpret the meanings. Consider the example by Shih and Kochansky (2002) below:

One can say;

a. **I** did not eat the melon. (She did, instead.)
b. I did **not** eat the melon. (Strong denial, after an accusation.)
c. I did not **eat** the melon. (I took it, and fed it to the dog, instead.)
d. I did not eat the **melon.** (I munched out on the cookies in the cabinet.)

v. **Stress language:** This is different from the above. English is an example of stress-language, while Malay is a stress-free language (Zuraidah, Knowles and Yong, 2005). This feature plays a strong role on accents. Stress location is part of the lexical entry of each English word. Example given by Shih and Kochansky (2002) is that, "apple" and "orange" both have stress on the first syllable, while "banana" has stress on the second syllable. When an English word is spoken in isolation in declarative intonation, it is said that f0 typically peaks on the stressed syllable, and this element is associated with accent. Accent is stress associated with a pitch movement in connected speech (Zuraidah, 1996). This is however not a concern in Malay, but suppose an Englishman speaks Malay associating the language to the stress words of his own, then listeners may notice accent.

The next section will present the overview of Malay TTS system, eventually leading to a brief introduction to *Fasih*, the first Malay TTS system in Malaysia. *Fasih* is used to host eXTRA to test the module's capabilities.

2.5 Overview of Malay TTS

Based on the goal to develop a Malay TTS system capable of reading unrestricted Malay sentences correctly and naturally, the Malay Speech Interface Group of MIMOS developed an evolutionary prototype called *Fasih* (a Malay word for "fluent"). *Fasih* is

based on IBM's ViaVoice English TTS that started to be developed two years earlier. Project *Fasih* was initiated in 2002 as part of MIMOS' Bridging the Digital Divide (BDD) initiative to help the blind access information using speech. Over time, project *Fasih* expanded into other domains.

MIMOS worked on a few other TTS system projects. One of these was Speak, initiated in 2002, which uses a syllable-based concatenative synthesizer. Another one, called BM Baca, was developed in 2003 in collaboration with the Asia Pacific Institute of Information Technology (APIIT). The latter uses Microsoft Speech engine, a ruled-based synthesizer, also known as formant synthesizer. Following suggestions by APIIT, researchers in MIMOS used Multi-band Resynthesis Overlap Add (MBROLA) synthesizer for the BM Baca project. Soon after this BM Baca became an official Bahasa Melayu (BM) Text-to-Speech project. However, this Malay TTS system first used an English female voice to speak out Malay words and it was found that the pronunciation was Anglicized and mismatched. This problem led to the recording of new sound using a Malay female voice to be used with the TTS system. Later, in 2004, the project was renamed to *Fasih*. *Fasih* has four core modules as shown in Figure 2.3, namely the Normalizer, Phonetizer, Prosody and MBROLA.

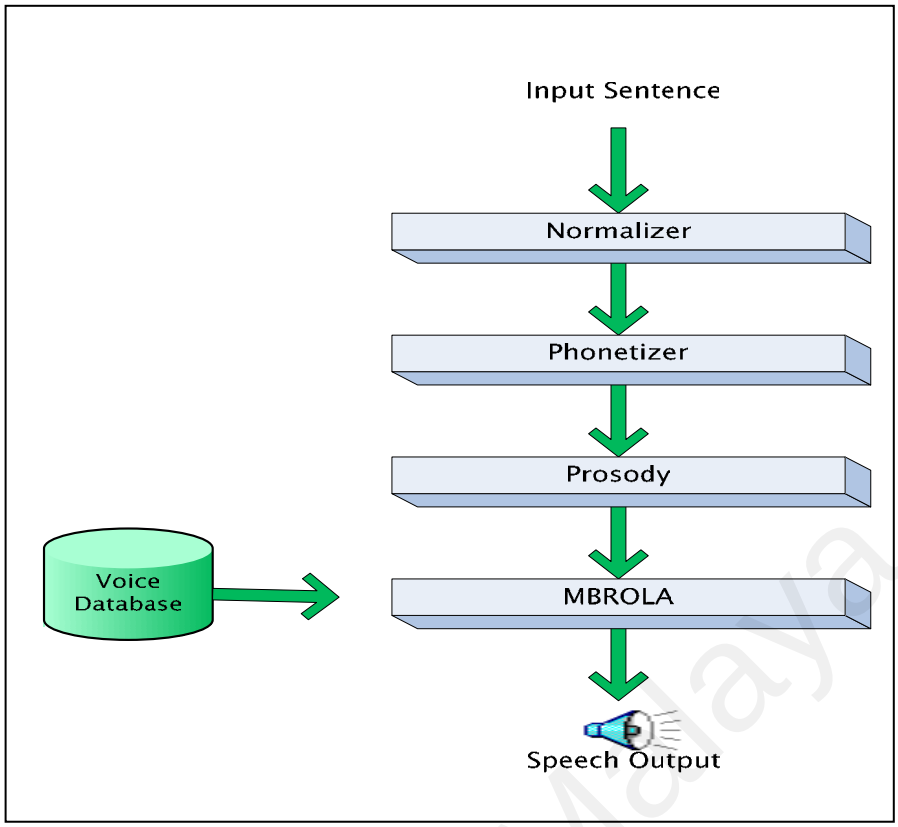


Figure 2.3: Top-down Architecture of *Fasih*

2.6 Summary

Introduction to the studies of emotions, particularly vocal emotions in this chapter will promote the understanding of the requirements of implementing an affective component to current Malay TTS systems. The outcome of culturally specific emotions study reveals that it is crucial to build a more familiarized set of emotions to the local users of a TTS system. A TTS system that produces affective output that is better 'recognized' by people would have a reduced artificiality and increased spontaneousness, hence, offering users more comfort when interacting with the TTS system. The discussion on issues and challenges perhaps will increase the awareness to the stumbling blocks that exist in emotive speech synthesis and help grasp the important substances for implementing an affective component to a TTS system. The next chapter discusses the techniques used in speech synthesis work, related researchers conducted and different emotive speech synthesis products.

CHAPTER 3

Literature Review

3.1 Introduction

This chapter briefly discusses the popular techniques of emotional speech synthesis.. Several related work and products are also being compared to discover the advantages or drawbacks in the synthesized speech output. The subsequent sectionsection gives a detailed explanation of the MBROLA, the speech synthesizer of which the first Malay TTS system, ***Fasih,*** is based on. The last section unravels on how the analysis of these different techniques emerges a new method to generate a set of synthetic emotions that is more natural….

3.2 Emotional Synthesis Technologies

Different synthesizers use different techniques to allow control over voice parameters. In most previous systems, three to nine full-blown emotions were modeled. This section discusses the three most widely used techniques for speech synthesis and focuses in more detail on the diphone synthesis technique, since the system described in this project uses this technique.

3.2.1 Rule-based Synthesis

As the term suggests, this technique produces speech by applying rules to acoustic correlates of speech sounds. Also known as formant synthesis, speech sounds using this technique may be experienced as unnatural and mechanized. However, an advantage of such system is that many voice parameters can be varied easily, making modeling emotions very interesting.

DECtalk (DEC stands for Digital Equipment Corporation) is an example of a rule-based system that serves as the underlying synthesis system for Janet Cahn's Affect Editor (Cahn, 1989). Another example is the Helpful Automatic Machine for Language and Emotional Talk (HAMLET) by researchers Abadjieva, Murray and Arnott (Abadjieva ***et al***., 1993). These systems are based around a series of rules that systematically alter the voice of the synthesizer in ways appropriate to the emotion being simulated. The emotions were selected explicitly by the user and the emotion effects were applied to the input of unrestricted text. However, in the Affect Editor, the rules can process manually tagged input text.

In both cases, the approach taken to generate emotions was to derive the acoustic parameter setting correlating to each emotion from literature and study of actor-generated emotions. Some researchers however, use different approaches to generate voice parameters. For example, by using perception tests to find the best parameter settings for different emotions. Although both ways are able to produce some colouring of emotion in speech, the output quality can still be unnatural. This may be so because the same sets of rules are produced for assigning limited intonation contours to the speech but the contents vary. Varying contents means that the placement of word in an utterance varies and therefore the intonation contour applied to the utterance may be

incorrect. Moreover, the rules may fit only limited length of utterances and not unrestricted text input. This could be because the rules are based on the parameter settings derived from original recordings that uses utterances of limited lengths as stimuli. Incorrect intonation contour suggests unnatural voice (Murray and Arnott, 1996). These authors also indicated that the "limitations in the word parsing and intonation rules mean that no system can correctly assign the correct contour for *every* possible utterance". Furthermore, there are many various linguistic features that interactively effect phonological characteristics (Wu and Chen, 1999), which restricts the collection of appropriate and complete rules to describe the prosody diversity.

3.2.2 Unit Selection Synthesis

Unit selection synthesis is said to be the most natural sounding speech synthesis available today. This technique uses a large speech database, usually more than one hour of recorded speech. "Unit" in this technique is referring to segments that are used as a basic synthesis reference in the database, such as individual phones, diphones, syllables, morphemes, words, phrases, and sentences. Each unit has varied acoustic parameters and the selected segments or units are concatenated to form speech. In other words, at runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). The process of selecting the best units usually uses a specially-weighted decision tree. The larger the unit selection speech database is, the more natural the output can be. The primary motivation for the use of a large database is that with a large number of units available with varied prosodic and spectral characteristics, it should be possible to synthesize a more natural sounding speech than can be produced with a small set of controlled units (such as diphones).

However, one of the biggest drawbacks of unit selection is that most such systems do not have an implementation that allows for explicit modification of acoustic parameters, therefore making it inappropriate to use in emotional speech synthesis. This is because no signal processing is involved in unit selection technique. Even conceptually, controlling prosody in such systems is only possible within the limits of the units recorded in the database, unless signal manipulation techniques are used, which invariably lead to a reduction in the quality of the synthesized speech. Thus, though it can produce natural-sounding speech such as how a TV newscaster sounds like, modeling distinctive emotions based on such systems is unsuitable. This is probably the reason that led some expressive speech synthesis work (Black, 2003; Eide and Aaron, 2004) using unit selection technique to train prosody engines that speak in different speaking styles (e.g. delivering good news in sounds that are more upbeat and delivering bad news in sounds that are more subdued) rather than portraying emotion per se.

3.2.3 Diphone Synthesis

Diphone synthesis uses recorded speech in a given language that is concatenated. This recorded speech is stored in a database and is usually minimal. In phonetics, a diphone refers to a recording of the transition between two phones. In other words, a diphone stretches from the middle of one phone to the middle of the next. The diphone recording is usually made in a monotonous pitch and an F0 contour is generated through signal processing at synthesis time.

For the generation of highly natural synthetic speech, the control of prosodic parameters is of primary importance. This technique allows fine-grained control over the prosodic parameters of the synthesized voice, unlike the other techniques above. Therefore attempts to express emotions that use a model-based mapping of human voice to acoustic correlates of synthesized speech is best when using this technique. This approach is applied in this project. However, control over voice quality seems to be poor.

Bulut, Narayanan and Srydal, (2002) used Copy synthesis to model emotion using diphone synthesis. The F0 and duration for each phone is measured based on the actor's portrayal of a given emotion in an utterance and correlated to diphones for synthesis.

3.2.3.1 The Multi-band Resynthesis Over-Lap Add (MBROLA) Synthesizer

MBROLA is a speech synthesizer based on the concatenation of diphones (MBROLA, 2005), available for free provided for non-commercial use. It takes a list of phonemes as input, together with their prosodic parameters - *duration of phonemes* and a *piecewise linear description of pitch*, and produces speech samples on 16 bits (linear), at the sampling frequency of the diphone database used. A language-specific diphone database tailored to be used with MBROLA is needed for the diphone concatenation to produce accurate speech sounds in the particular language. For ***Fasih***, a Malay diphone database is used with MBROLA. The database is transcribed with Speech Assessment Methods Phonetic Alphabet (SAMPA) symbols. SAMPA is a computer-readable phonetic script based on the International Phonetic Alphabet (IPA). It provides an official way to transcribe phonetics for many languages into American Standard Code for International Interchange (ASCII) codes.

The flexibility offered by MBROLA in modifying acoustic parameters produced some notable work such as Marc Schroder and Martine Grice's (Schroder and Grice, 2003) emotional speech synthesis where they model the vocal correlates of emotions using emotion dimensions. This approach uses two emotion dimensions; valence (or pleasure) and arousal, to represent emotional *tendency*, rather than of specific emotions. The reason given is that in concatenative synthesis, the voice quality inherent in the diphones is inappropriate for certain emotions. Emotions from some point of the two emotion-dimension space are mapped to the corresponding acoustic correlates and realized in synthetic speech. Their evaluation showed that the vocal effort is perceived as intended.

Emofilt, by Burkhardt (2005) is another example of software that stimulates emotional arousal with speech synthesis based on MBROLA engine, which supports 35 languages to date. Similar to the role of affective layer in this project, Emofilt acts as a "filter" in MBROLA framework that adds certain emotion to speech that is being generated. This software acts as a transformer between the phonemisation and the speech-generation component. It also provides graphical user interface (GUI) that allows users to manipulate a set of speech parameters before generation of speech.

Sadly, Emofilt's output samples provided in its website (Burkhardt, 2005b) sound very artificial and unnatural. Bukhardt claimed that the limitation in naturalness is caused by constraints in diphone-concatenation approach. Nonetheless, it is hoped that this project will challenge the above conventional method and introduce improvements in emotional speech output using the same approach.

3.2.3.2 How MBROLA Works

Saya.pho

```
s       122  50 320 100 324
V       100 23 111 78 222
J       70 50 223 100 146
@       91 0 456 100 378
__1000
```

Figure 3.1: MBROLA Input for Isolated Word 'Saya'

Figure 3.11 above shows the format of the input data required by MBROLA. MBROLA input is in PHO format. Each line contains a SAMPA phoneme symbol, duration (in ms), and a series (possibly none) of pitch points. A pitch point comes in a pair: the position of where the pitch rely within the phoneme (in percentage of its total duration), and the pitch value (in Hz) at this position. Multiple pitch points make up a pitch pattern. For example, the first line of *saya.pho*:

s 122 50 320 100 324

tells the synthesizer to produce the sound for phone *s* of 122 ms, and to put a pitch point of 320 Hz at 50% of 122 ms (mid-utterance) and a pitch pattern point of 324 Hz at the end of the sound. Pitch pattern points define a piecewise linear pitch curve. Notice that the pitch pattern defined is continuous, since the program automatically drops pitch information when synthesizing unvoiced phones. For a pause, an underscore is inserted followed by the duration of pause in ms, for example:

_1000

is a *pause* for 1000 ms.

The synthesizer outputs chunks of synthetic speech determined as sections of the piecewise linear pitch curve. Phones inside a section of this curve are synthesized all at once. The last one of each chunk, however, cannot be properly synthesized while the next phone is not known (since the program uses diphones as base speech units)

3.2.4 Template-driven Concatenative Synthesis

To invoke a better imitation of human speech tones, Wu and Chen (1999) proposed a template-driven generation of prosodic information for their Chinese TTS conversion. They took heed of the characteristics of Mandarin Chinese which, it being a tonal language (five basic tones) and that the words are monosyllabic to arrive at the conception of using templates to generate more natural speech. The conversion system went a step ahead by also accounting the various linguistic features in Mandarin. The information of Chinese linguistic features that are relevant to the information of word prosody such as tone combination, word length, part of speech (POS) and word position in a sentence, which the other techniques mentioned above have difficulties dealing with, can instead be realized in synthetic speech using templates. This was done by establishing templates that contains prosodic information on word level. A sentence intonation module and a template selection module were also proposed to select the target prosodic templates to be applied to input speech. The result of their evaluation confirmed that the output speech was highly intelligible and natural.

3.3 The Strategies of Infusing Emotion in Malay Synthesized Speech

The objective of infusing emotion in synthesized speech is to produce a better quality speech output. The limitations and advantages of the different techniques discussed in the previous section has brought into awareness that there are two major factors that need to be considered in order to build an affective layer prototype to be used with a Malay synthesized speech; the Malay language phonological feature and a technique (or combination of techniques) that shall optimize its function. The next section explains the syllabic structure in Malay language, followed by the proposed method.

3.3.1 Malay Language Syllable Structure

It is observed that in Malay language, the structure of syllables is straightforward. In addition, the intonational or prosodic relationship between syllables within a word is more obvious than between two words. The simple syllable structure that the Malay language is based on allows for the use of an algorithm that focuses on the number of syllables rather than other linguistic features. In Malay, the syllable structure units are as follows:

- **CVC** (Consonant-Vowel-Consonant)
- **CV** (Consonant-Vowel)
- **VC** (Vowel-Consonant)

A consonant is identified by either one or two letters that produces a ***single speech sound***. (Merriam-Webster, 2005). For example, /r/ or two-letter consonant /ng/ in the word “Ngeri” (eerie) both produces one sound. A vowel produces the most prominent sound in a syllable.

Therefore, when synthesizing some Malay loan words that are of English origin, such as "drama" or "trauma" [dra-ma (**CCV**-CV), trau-ma (**CCV**-CV)] the sound produced may sound strange. This is because these words do not satisfy any of the Malay language's syllable unit as shown above. The first syllable in both words when combined, actually produce two sounds – 'de | ra' and 'te | ra'.

3.3.2 The Proposed Method

A hybrid technique is proposed for optimizing the functions of the affective layer. The rationales are discussed below:

i. Since the hosting system uses diphone concatenative synthesis, the employment of this technique is compulsory.

ii. The facts about Malay language syllable structure discussed in 3.36.1, added with the restrictions of phonological rule-based systems mentioned in section 3.2.1(), shaped the idea to create *syllable-sensitive* rule-based system.

iii. The effectiveness of the template-driven method proposed by Wu and Chen (1999) has brought the idea to adapt this method and combine it with the techniques in (i) and (ii).

This combination of the techniques in (i), (ii) and (iii) above derives the eXpressive Text Reader Automation Layer, or eXTRA. Figure 3.2 below shows the proposed architecture of *Fasih* with eXTRA being the affective component added.

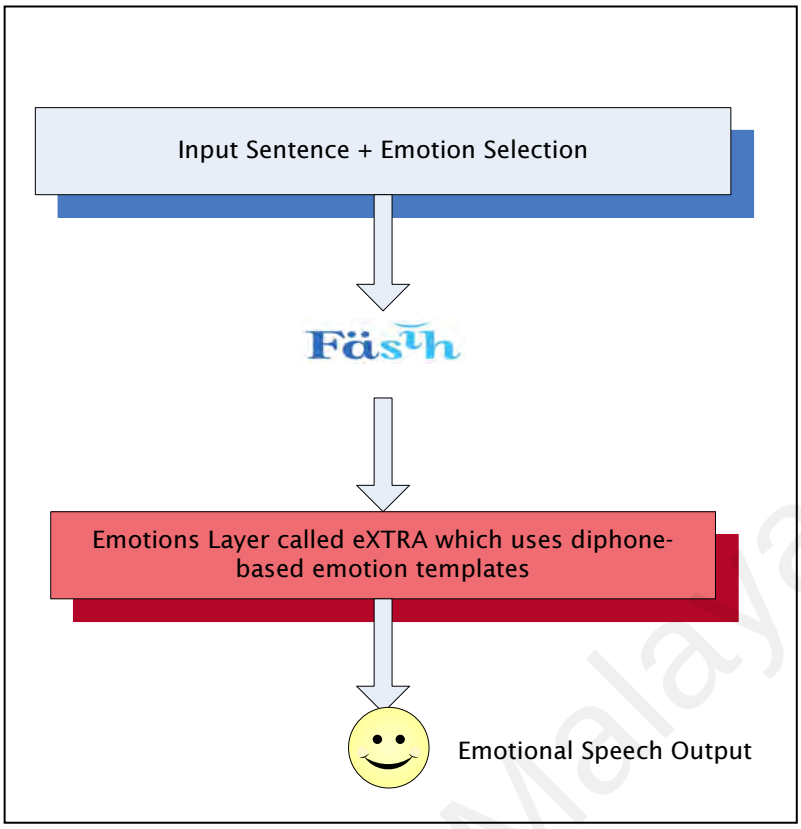


Figure 3.2: Simple Proposed Architecture of Affective *Fasih*

As the figure above illustrates, eXTRA shall manipulate *Fasih*'s synthetic output into emotional speech by applying a set of *emotion templates*. Since *Fasih* uses MBROLA engine, which uses diphones as based unit, the emotion templates in eXTRA should also be based on diphones.

These emotion templates shall describe varying intonationss according to targeted emotion that can be applied to input sentences on phoneme level. The selection of target template is carried out by a Template Selector module, based on syllable sequence of the input .In other words, a targeted templatetemplate should have the same number of words and syllables as a particular input text. It is assumed that when each set of phonemes in a syllable is fed with its own 'emotional' parameter values, the intonation contours assigned are more accurate, thus, producing more natural speech. The role of

emotion templates in the template-driven method employed by eXTRA is explained in the beginning of next chapter.

3.4 Summary

This chapter has revealed the popular techniques of emotional speech synthesis and their minus and plus points. In addition of the emotion studies in Chapter two, the analysis of these synthesis techniques and the Malay linguistic structure cultivated the requirements of an affective component's framework, which in turn determined the methods for implementing the eXTRA prototype. The use of a hybrid technique, with the main method for creating emotional speech being template-driven, led the generation of emotion templates, which is presented in the next chapter.

CHAPTER 4

Building Emotion Templates Techniques

4.1 Introduction

This chapter explains the role of emotion templates in eXTRA and presents the work that has been accomplished in order to build an emotion template database. Apart from that, it reports the observations, difficulties and results obtained which rationales the design decisions of eXTRA.

4.2 Template-driven Emotions Generation

As mentioned in the previous chapter (section 3.2.1), diphone synthesis allows maximum control of prosodic parameters. Therefore, attempts to model the emotions in eXTRA could use a model-base mapping. The model here is a real human. Extracting the exact affective information from a human's voice and transferring it into acoustic data in templates ensures more natural emotional-blended speech when the target template is applied to the speech.

The templates describe various tones that reflect a particular emotional state. An algorithm that allows an input sentence from user to be read using the appropriate tones

that convey the emotional state the user chooses is written (Chapter 5, section 5.4.2.1). This is done in such way that the input sentence is matched with the correct emotion template. To accomplish this, a syllable-sensitive rule-based algorithm is applied.

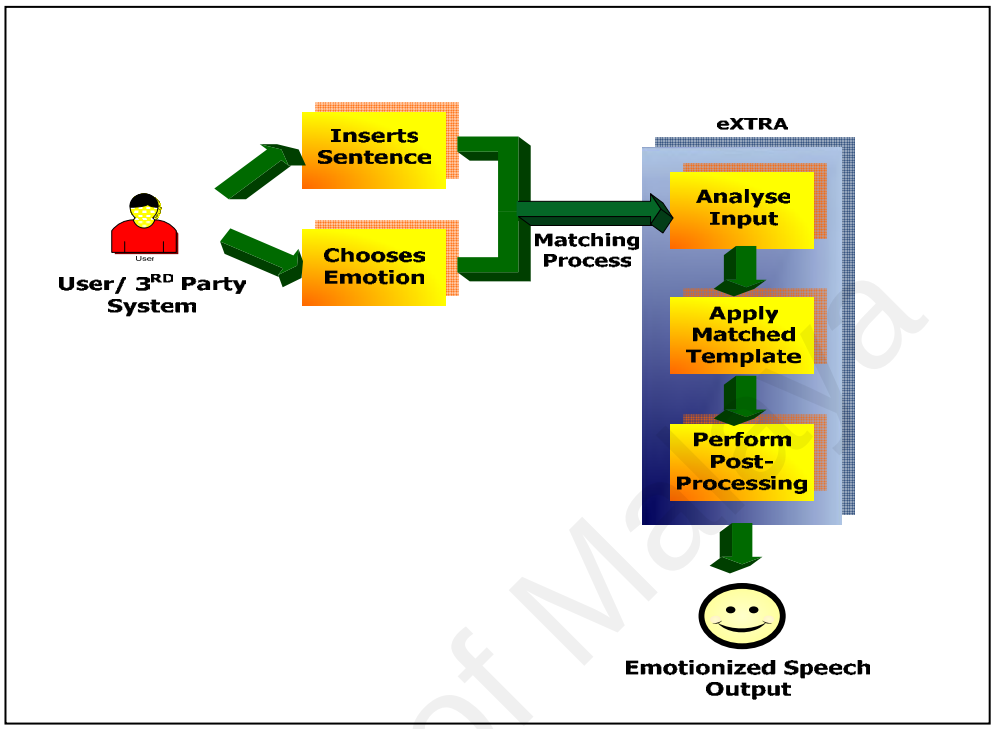


Figure 4.1: Simplified Framework of Emotion Layer Module

Figure 4.1 provides the visual illustrations of eXTRA's framework. Using the syllable-sensitive algorithm, each word from user input is analyzed and chunked into syllables in reverse order (stack) to determine syllable count; the input sentence is processed from the last word to the first. The result is then matched against the emotion template that contains the sentence with the same syllable-count and sequence. In other words, the template selection is done by identifying the integers that represent the syllable sequence of the template-sentence – "2222", "2332" etc. This is done by using a template selector module.

Consider the input sentence " Awak tidak tahu malu" (you have no shame) is to be spoken in anger. This sentence has a syllable sequence set of "2222". Therefore, the anger template that will be selected from the database (Appendix F) also comprises the

syllable sequence set “2222 ”. The sentence in this template is “Kamu sungguh kurang ajar” (You are so rude). Consequently, the anger template is applied to the input sentence to produce an emotionized output. This is done by matching the emotional prosodic parameters from the template-sentence to the input-sentence at the level of phonemes. The matching process is explained in Chapter 5, section 5.3.4.8. To ensure a more natural tune, the post-processing is done. It involves assigning silence and default parameters to additional phonemes correlating to each word wherever necessary. An explanation that is more technical and detailed is presented in Chapter 5.

The next sections explain the procedures and work undertaken to build the emotion templates.

4.3 Building the Speech Corpus

For this project, a small database containing short sentences was built. These sentences are a set of Malay utterances used as emotional stimuli that were read and acted by a speaker in the state of anger or sadness. The readings were recorded and accepted samples were used for building synthesized emotional templates. Initially, the utterances chosen were semantically neutral. However, after conducting a few preliminary tests (results displayed in Appendix G, section 3.3.3), it was realized that this was not necessary, as the emotion templates that are based on these stimuli should show some distinctive colouring of anger and sadness. Moreover, it was quite difficult for the speaker to portray the intended emotion while reading semantically neutral sentences. Semantically neutral utterances are often used by emotional synthesis speech researchers to facilitate segmenting voice for synthesizing process in different emotions. This is more convenient than repeating the process with different sets of sentences that

consist of emotionally-inherent contents. However, in this research, templates are used to be applied with speech that is already synthesized. How "emotional" the templates sound would depend highly on how good the speakers portray the intended emotions in the recorded utterances. Therefore, new utterances with emotionally inherent contents were chosen to stimulate speaker's portrayal of intended emotions.

4.3.1 Designing the Sentence

In this prototype, the sentences chosen are ***declarative*** and only contain ***four*** words. The sentences are limited to ***two and three syllables***, as the Malay language mostly contains two and three-syllable words. The approach used to select these words was by comparing the word structure to syllable sequences. The possibilities of two and three-syllable words that satisfy all different positions in a four-word sentence is derived by using the below formula:

$$a^n,$$

where,

a is the total number of syllable-set and,

n is the number of words.

Therefore, there are

$$2^4 = 16 \text{ sentences}$$

that correspond to one emotional state where two and three-syllable words cover all different positions in four-word sentences. Table 4.1 shows some samples of anger sentences. The complete sets of sentences for both emotions are presented in Appendix B.

Table 4.1: Sample Sentences That Satisfies Different Syllable Combinations for Anger Template.

No	Sequence of Syllable		Word Position			Scenario
		1	**2**	**3**	**4**	
1	**2 2 2 2**	Kamu	sungguh	kurang	ajar	*Reaksi terhadap anak murid yang menendang kerusi guru dgn sengaja*
2	**3 2 2 2**	Sekarang	baru	engkau	tahu	*Diajukan kepada si pengacau selepas menampar mukanya*
3	**2 3 2 2**	Baru	sekarang	aku	faham	*Akhirnya rancangan jahat kawan sendiri terbongkar*
Translation: *1. You are so rude* *2. Now you get to know (who I am)* *3. Only now I understand (his plans)*						

All sentences were proof read and approved by an expert linguist to determine their accuracy.

4.4 Voice Recording Using Acted Speech

The use of acted speech, though widely criticized (Campbell, 2000), is chosen over natural speech as the author's goal is to have the emotions elicited in the read speech to be highly recognized and clearly distinguished from one another. Two most important factors in building an emotion template database is to ensure that the emotions are properly and accurately extracted from the human voice samples and that the speaker should be able to portray convincible emotions. Therefore a speaker with excellent voice quality is needed for the purpose of obtaining recorded samples using the 32

sentences presented in Appendix B. To determine the quality of a speaker, auditions were held.

4.4.1 Speaker Selection

It is learned that from previous tests (Appendix G), using multiple speakers cause the synthesized speeches have varying frequency and "voices", mimicking the original speaker voices. At this point it is realized that to maintain voice consistency it is enough to have one only speaker and this will also ensure the consistency of the sounds stored in the emotions templates. Therefore, only one female actor was needed to ensure that the emotional parameters recorded based on her voice are consistent, especially in terms of pitch and intensity.

Three professional female actors were tested during an audition. Two of them are Bachelor of Arts graduates who majored in Drama acting and the other is a drama teacher, all from University Science of Malaysia (USM). During the audition, the speakers were asked to read 10 short sentences and 5 long sentences. They were asked to read the sentences three times, each in neutral, anger and sadness state while their voices were recorded. The speech samples were later examined and the speaker with the most suitable and 'clean' voice was chosen.

4.4.2 Recording Session

The chosen speaker has the clearest un-breathy voice and could deliver best emotions when acting her speech. This is probably because she has vast experience in theater and drama acting on varsity and national level. She

was provided with the two sets of prepared sentences a week earlier to allow her to practice reading the sentences with suitable emotional tones. In order to motivate and focus the speaker, each of the sentences was accompanied by a scenario. These brief scenarios were prepared for eliciting anger or sadness. For example: “Kamu sungguh kurang ajar” (You are so rude) and “Reaksi terhadap anak murid yang menendang kerusi guru dengan sengaja” (Reaction towards a student of yours who kicked your chair on purpose) were sentence and scenario, respectively. Having such elicitation scenario helps to reduce the interpretation variations. She was also briefed about the two emotional states in relation with the Malay culture as revealed in findings in Chapter 2 (section 2.4.1) so that she could specifically deliver tones that are more in tune with Malay.

Recording was done in Computer Aided Translation Unit (UTMK) in USM. A Plentronics DSP- 100 PC headset with attached microphone was used. This headset is specifically enhanced for speech recognition purpose and is equipped with noise canceling feature. Her voice was recorded in a 16-bit mono data setting. The speaker was free to move her body as gestures of acting but was asked to maintain the position of the microphone. She was also allowed to repeat any sentence as many times as she would like if she was not satisfied with the emotional tone she delivered for the sentence. The recording process went on smoothly as the speaker had no problem in acting out the suitable emotions. The content of the utterances also help the speaker to express the correct emotions during the recording. It took approximately one and a

half hour to complete. Each sentence was read in different tones, which means that there were sixteen tones recorded for each emotional state.

4.5 Evaluation of Speech Samples

4.5.1 Perceptual Tests

To ensure that the speech samples are reliable; the intended emotions elicited in the speech samples are recognized by listeners, a series of perception tests was conducted. The procedure of these tests was adopted from Zahid (1999) and modified so that it is more suitable for recognition of intended emotions in the speech samples in this experiment. A threshold of sixty percent (60%) of recognition rate was set and the speech samples that did not meet this rate were discarded. The basis of the threshold value derived from the fact that people only recognize sixty percent of emotion expressed in human voice (Shrerer, 1981 in Nass *et al.*, 2000).

The series of tests involved local and non-native listeners. Both groups of listeners are undergraduates and postgraduates of FCSIT, UM. Foreign listeners were asked to participate in this test to confirm the literature about cultural-specific emotions. Though it is understood that non-native listeners do not understand the content of the utterances, the recognition of specific tones in the speech samples by these listeners has validated whether the tones in the samples correlated with the right emotions in their

perspective. The tests were conducted in the Language Engineering lab of FCSIT, University Malaya in two stages as explained below:

i. Initial Stage

Twenty listeners who were not aware of the test stimuli; ten native speakers and ten foreigners were asked to listen to a series of stimuli using the headphone to evaluate the emotional states. Their answers were written on a form that has no answer options (refer Appendix C1). It is believed that forced choices of answer only restrict the decisions of the participants, thus have a high possibility of biased judgments. For more variation, fear and happiness speech samples were added to the anger and sadness samples and were randomly ordered. The closest answers that correctly describe the intended emotions were chosen and the rest were discarded. The semantic model mentioned in Chapter 2, the Circumflex Model of Affect (Russell, 1980) was used as a guideline to cluster the answers according to a range of emotions.

Other than evaluating emotional states, the subjects were also asked to evaluate the recognition effort by choosing only one option within the scale given, as shown in Table 4.2. This scale is adopted from Huang, Acero and Hon (2001) and modified for use with emotion recognition, to capture the effort needed for recognizing a particular emotion elicited in the utterances. In other words, it is used to determine how easy or difficult it is for the subjects to recognize the emotions elicited in the samples.

Table 4.2: Effort Scale

Effort Required to Recognize the Emotion/Expression Elicited in Utterances	
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No emotion recognized with any feasible effort	1

ii. Final Stage

The listeners whose answers were chosen were asked to evaluate the speech samples again after two weeks. The original order of the samples was also randomly reshuffled. The reason for the two-week gap and re-evaluating samples that are reshuffled was to ensure that the listeners do not recall the answers given during their participation in the first stage. This time, fixed options were provided on a form (refer Appendix C2) for them to choose their answers from. Again, the closest acceptable answers that describe the intended emotion were chosen. These final selections were assumed as the most accurate and reliable answers.

4.5.2 Analysis of Perception Tests

The perception test results were analyzed to determine the recognition rate. The reports on the results obtained and discussions on the findings are presented in the following section.

4.6 Perception Test Results for Human Speech Samples

4.6.1 Initial Stage (Open Test)

i. Native Listeners

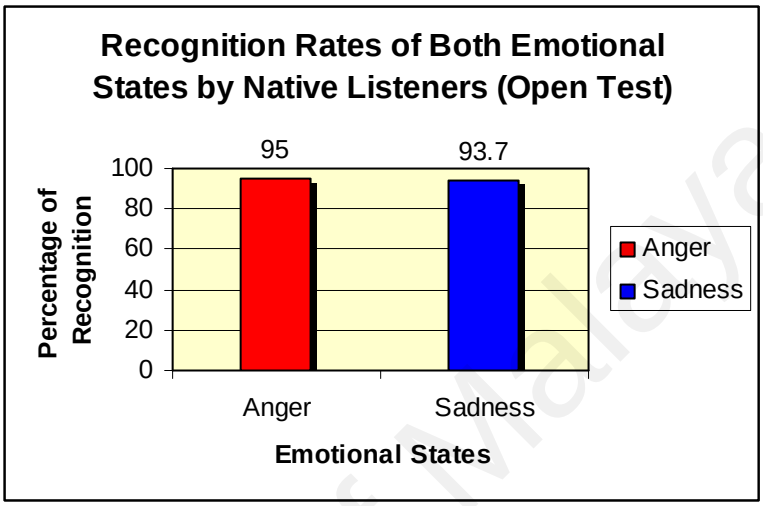


Figure 4.2: Results from the First Perception Test by Native Listeners

ii. Foreign Listeners

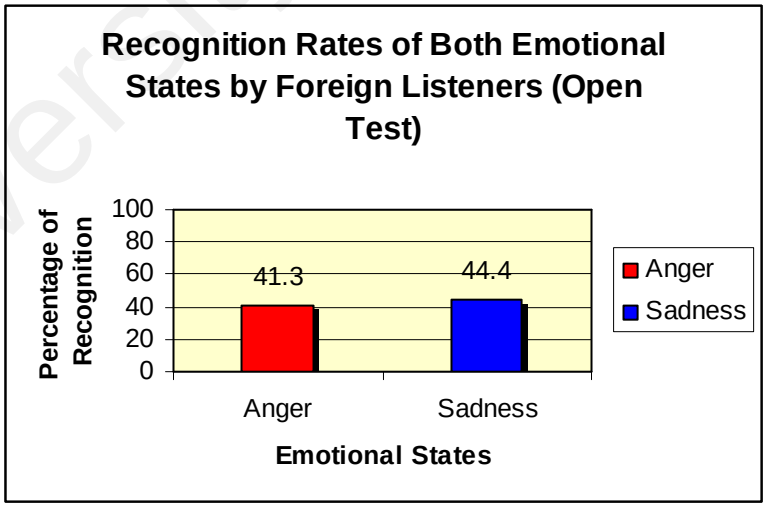


Figure 4.3: Results from the First Perception Test by Non-native Listeners

4.6.2 Final Stage (Fixed Options)

i. Native Listeners

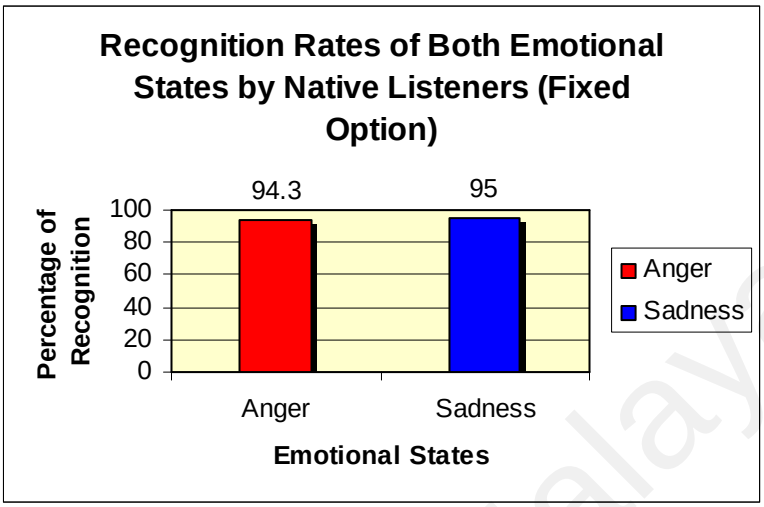


Figure 4.4: Results from the Final Perception Test by Native Listeners

ii. Foreign Listeners

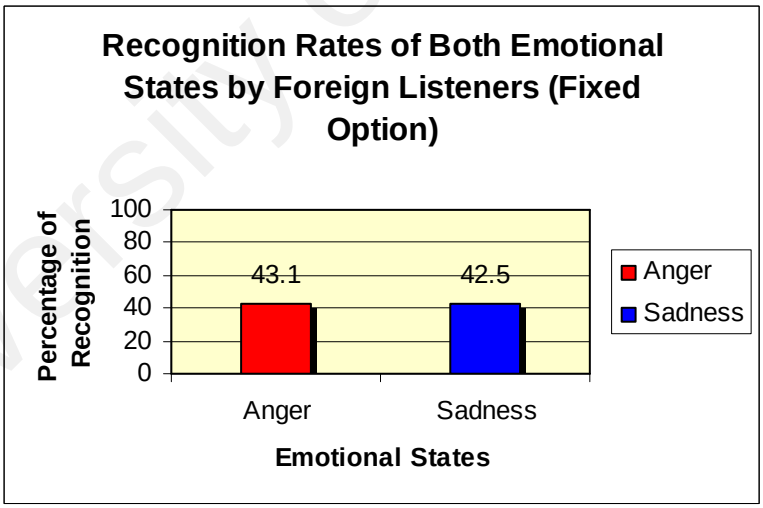


Figure 4.5: Results from the Final Perception Test by Non-native Listeners

4.6.3 Summary of Results and Discussions on Findings

The results indicated by figures 4.2 and 4.4 shows that the recognition rates for both the emotional states in human speech samples by native participants reached the threshold of sixty percent (60%). For the open test, Russell's affective space called the Circumflex Model of Affect (Russell, 1980) was used to determine which dimension the listeners' perceptions falls within. For example, answers that fall within the intense and unpleasant space, such as "aggravated" or "irritated" are grouped as Anger.

The detailed result obtained from analysis (as displayed in Appendix G, section 3.3.3) shows that *every* utterance obtained over sixty percent (60%) of recognition. Majority native listeners felt that no effort is required for describing the emotions elicited in almost all the speech samples. The result confirmed that all thirty two samples can be used as valid samples for emotional features extraction, for the purpose of building emotion templates.

On the other hand, a noticeable difference can be seen in the bar charts of Figures 4.3 and 4.5, representing results obtained from both tests participated by foreign listeners. The low recognition rates clearly show that the non-native participants may have different perceptions from the native participants. The findings based on this data also confirmed Silzer's statement that "expression and perception of emotions vary from one culture to another" (Silzer, 2001). However, results from foreign participants do not affect the selection of the speech samples for the purpose of extracting the emotional prosodic features. The data showed that most participants from the Middle East tend to perceive anger as "neutral", while participants from the Western countries presume sadness as "anger". An interesting discovery is that, in some cases; even samples that produced clear angry expressions were deemed as neutral by at least four out of ten foreign listeners.

The next stage involves analyzing the accepted speech samples, which is explained in the next section.

4.7 Analysis of Speech Samples

4.7.1 Extraction of Relevant Prosodic Features

The analysis stage concerns the process of extracting the prosodic features of each selected emotion. This part of the research took the longest time and required a lot of patience.

For this task, phonemes were used as a reference acoustic unit. The basis for this is because since diphone will be used to produce templates for both the emotional states, the parameters' requirements of diphone are the main consideration. As explained earlier, diphone uses pitch and duration *for each individual phoneme* as its main speech parameters.

The emotional prosodic features extraction process as shown in Figure 4.6 below is broken into three main stages as follows:

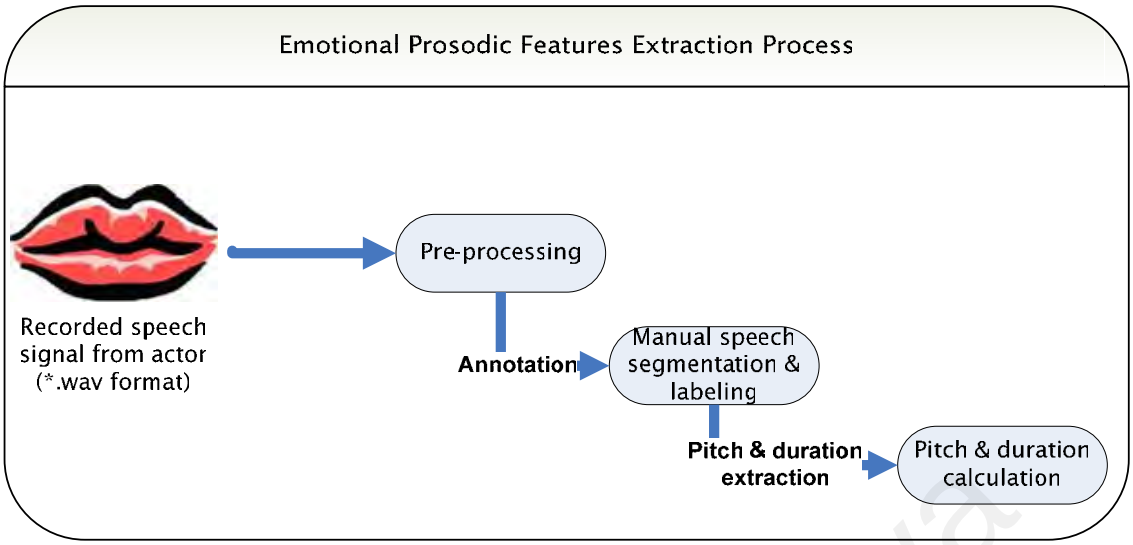


Figure 4.6: The Process of Extraction of Relevant Prosodic Features

4.7.1.1 Pre-processing

Pre-processing here involved the removal of noises from the speech signal using the Speech Filing System version 4.6 (SFS 4.6) (SFS, 2005).

4.7.1.2 Annotation Process

Annotation process involved the *segmenting and labeling* of speech data. The purpose is to determine the boundary of individual phoneme in the speech samples. The annotation of the samples was done manually since there was no Standard Malay language aligner trained to automate labeling of speech. These required very long hours of listening with high concentration to each chunk of phonemes from every single speech samples. The tool used to segment and label the data was speech analysis software called Praat version 4.3.17 (Boersma and Weenink, 2005). First, the voiced and unvoiced segments of the pre-processed sound files were determined. The unvoiced segments were removed and the first segment that is marked as voice is the beginning of the speech

period. Then each speech file was annotated into three layers in Praat's textgrid format according to *word*, *syllables* and *phonemes*, as shown in Figure 4.7 for the anger utterance *"Kamu sungguh kurang ajar" (You are very rude).* The phonemes were converted to the SAMPA phonetic symbols.

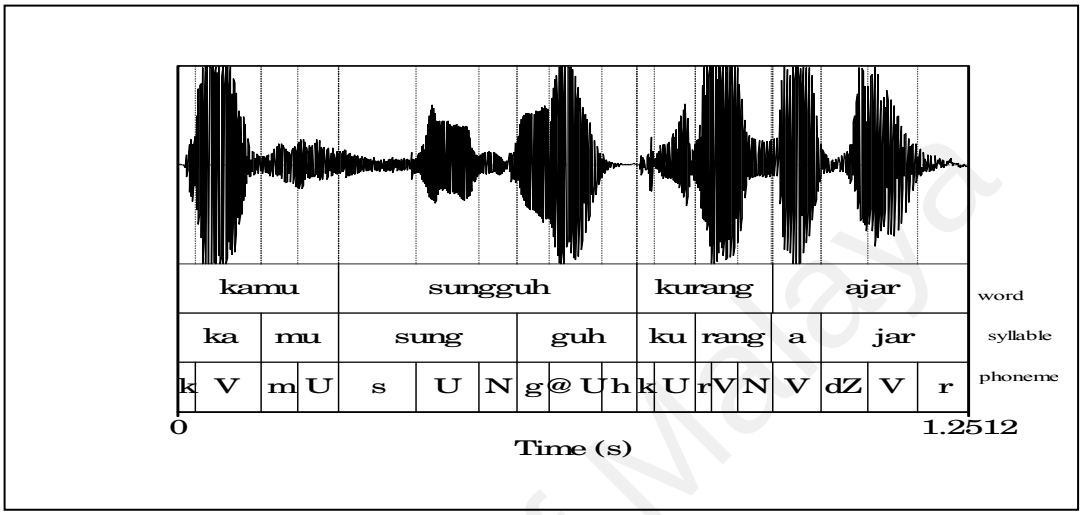


Figure 4.7: Three-layer Annotation of Speech Signal in Praat Textgrid Format

4.7.1.3 Pitch and Duration Extraction

Later, each segmented and labeled chunks of the phoneme layer were analyzed to extract its *duration* and significant *pitch points*, as shown in Figure 4.8. The positions of the pitch were also calculated in proportion of the duration of each phoneme and the value is inserted in percentage. This process of extraction was repeated for each phoneme until the whole sentence was completed, and was applied to every speech file of both anger and sadness emotional states.

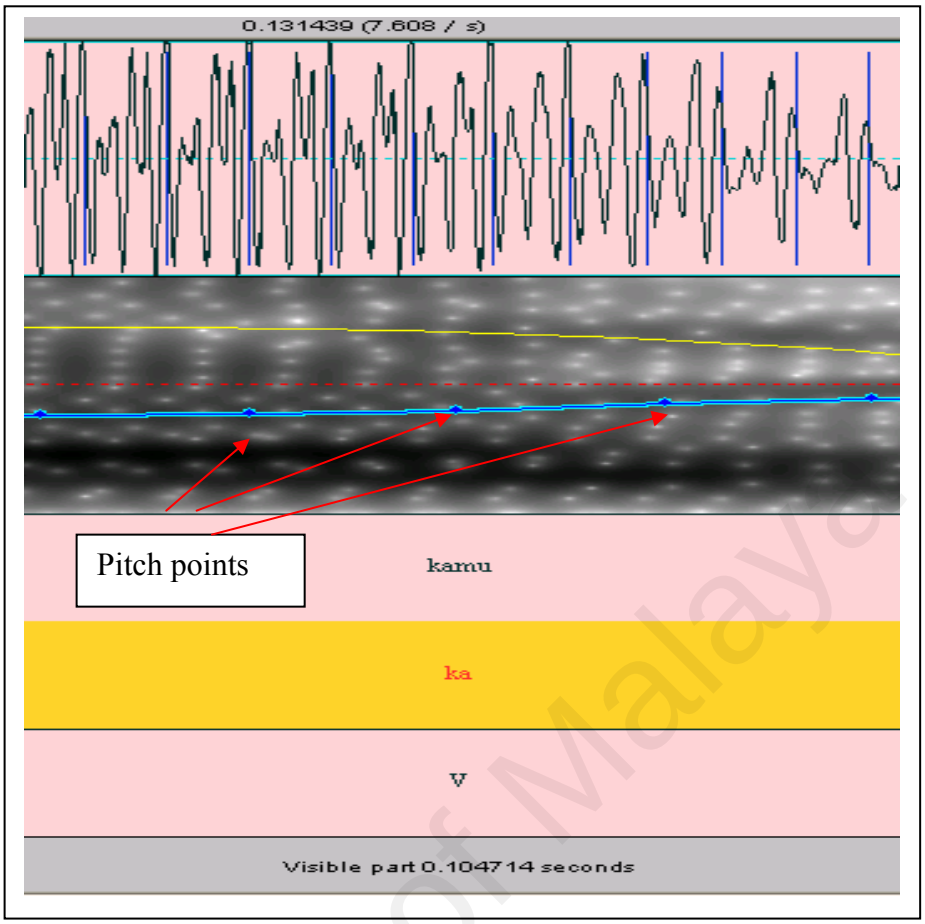


Figure 4.8: Extraction of Significant Pitch Points and Their Durations for Phoneme /a/ in the Word 'kamu' (you)

It is noted that vowels has longer duration compared to consonants, resulting in more pitch points. This suggests that vowels carry heavier cue for emotions in speech. Findings derived from experiments conducted by Zuraidah (1996) and Zuraidah, Knowles and Yong, (2005) also strongly support this suggestion.

The extracted emotional prosodic information is directly stored in an MBROLA player. Thus, each file of the player describes different emotional tones and functions as an emotion template. There are sixteen emotion templates describing sixteen emotional tones for each emotional state The template database which is a text file comprising information such as emotional state, syllable sequence, template name and template

sentence is attached in Appendix F. The figures below show a few samples of the emotion templates. Two samples depict anger (figure 4.9) and the other two (figure 4.10) depicts sadness

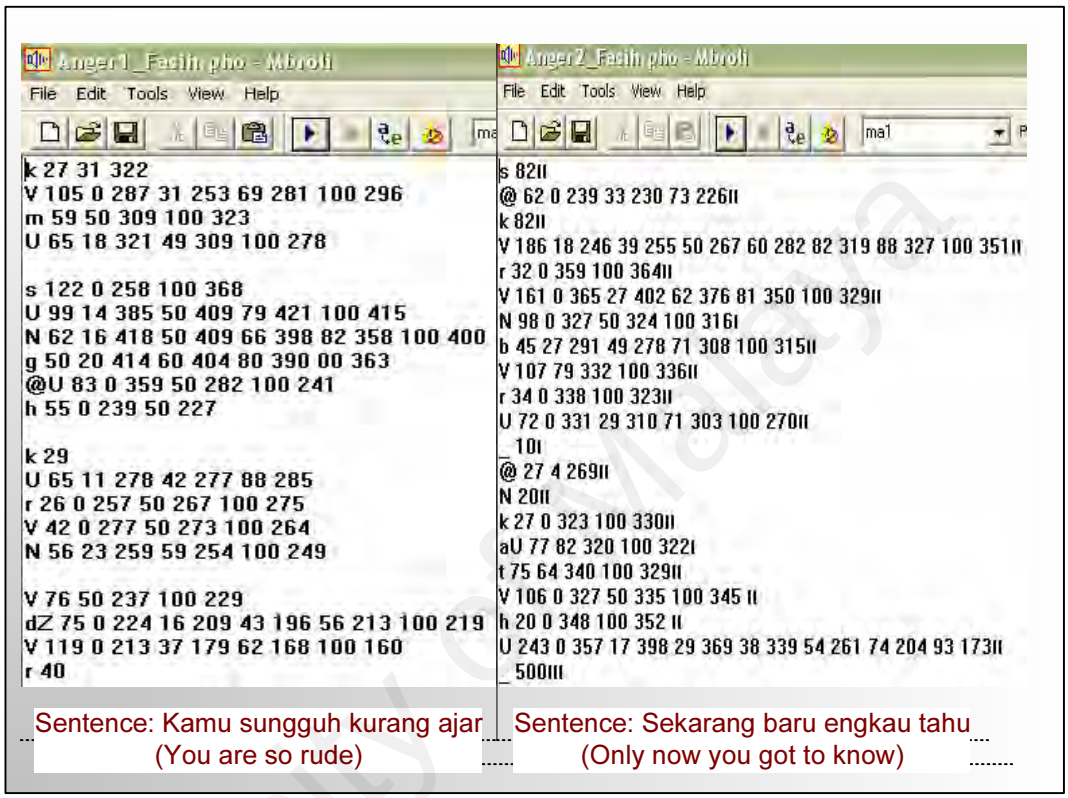


Figure 4.9: Samples of Anger Templates

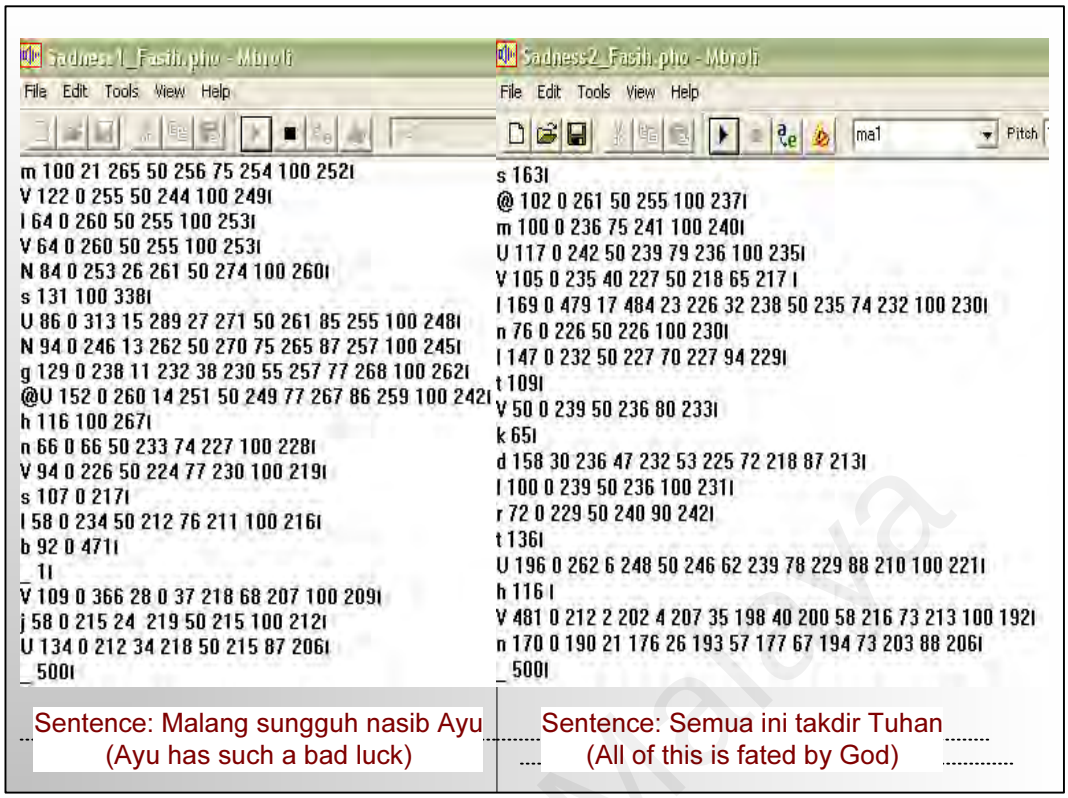


Figure 4.10: Samples of Sadness Template

Note that the phonemes (listed vertically) in the templates' sentences are converted to SAMPA format. Each phoneme consists of its own emotional prosodic parameter values (listed horizontally). The meaning of the values is previously explained in Chapter 3, section 3.2.3.2. Each template is identified by its syllable sequence. For example, the templates on the left side in the figure has a sequence of "2222" while the one on the right has a sequence of "3222". When the templates are played, it was discovered that the synthesized speeches sound almost similar to the original speech samples. However, the real challenge is to match these prosodic parameters with input text by users. The matching process is presented in next chapter.

4.8 Summary

This chapter has described the methods used to build an emotion template database. Much effort has been taken to create templates consisting of emotion information which resembles that of a real human. The flexibilities offered by MBROLA, such that the ability to directly manipulate the prosodic parameter values, has the advantage of capturing vocal emotions of real human to the exact acoustic correlates into templates. Apart from this, any modifications can also be done straightforwardly. It is believed that this would minimize the major problem discussed in Chapter 2, which is the lack of natural emotion in synthetic speech . The next chapter is about the software requirements, design, development and the prototype of eXTRA.

CHAPTER 5

System Requirements, Design and Prototype

5.1 Introduction

This chapter is divided into three major parts. The first part explains the user and functional requirements derived from the findings in previous chapters. The second part provides the reader with an understanding of the technical design of **eX**pressive **T**ext **R**eader **A**utomation Layer (eXTRA), the module that acts as a layer to infuse emotions into synthesized speech output. The design is shown from high level to low level with the aid of diagrams. A targeted TTS system called *Fasih* is used to illustrate its functionality. The final part of this chapter introduces eXTRA prototype and the corresponding implementation codes by showing screen layouts..

5.2 System Requirements

To determine the software requirements, the fundamental requirements for establishing objectives must be defined. User requirements and functional requirements are elicited from findings and conclusions in previous chapters and analysed further for the purpose of implementing them in a consistent software product that interacts with *Fasih*. Continuing on these efforts, this section serves as the initial step in constructing a model

that allows for the automation of eXTRA. The table below shows the list of user and functional requirements.

Table 5.1: User and Functional Requirements

User Requirements	Functional Requirements
1. Shall get emotionized speech by submitting a text and selecting various speech parameters (emotional state, speaker sex, speech rate, pitch level and baku activation)	1. Shall add affective features to speech data 2. Shall be able to process and produce MBROLA-compliant speech data for the Malay language 3. Shall process a text sentence of exactly 4 words, each having maximum 3 syllables 4. Shall implement an algorithm that abides by the *rules* as discussed in Chapter 4, section 4.2. 5. Shall maintain speech intelligibility by maintaining minimum phoneme durations 6. Shall be written in source-code native to MBROLA to allow for optimal integration 7. Shall allow for integration with a receiving TTS through a single interface 8. Shall provide an object model that allows for easy manipulation of phonemes

5.2.1 Use Case Diagram

Next is to aid the understanding in requirements of eXTRA with the conceptual model by applying a use case diagram (Figure 5.1).

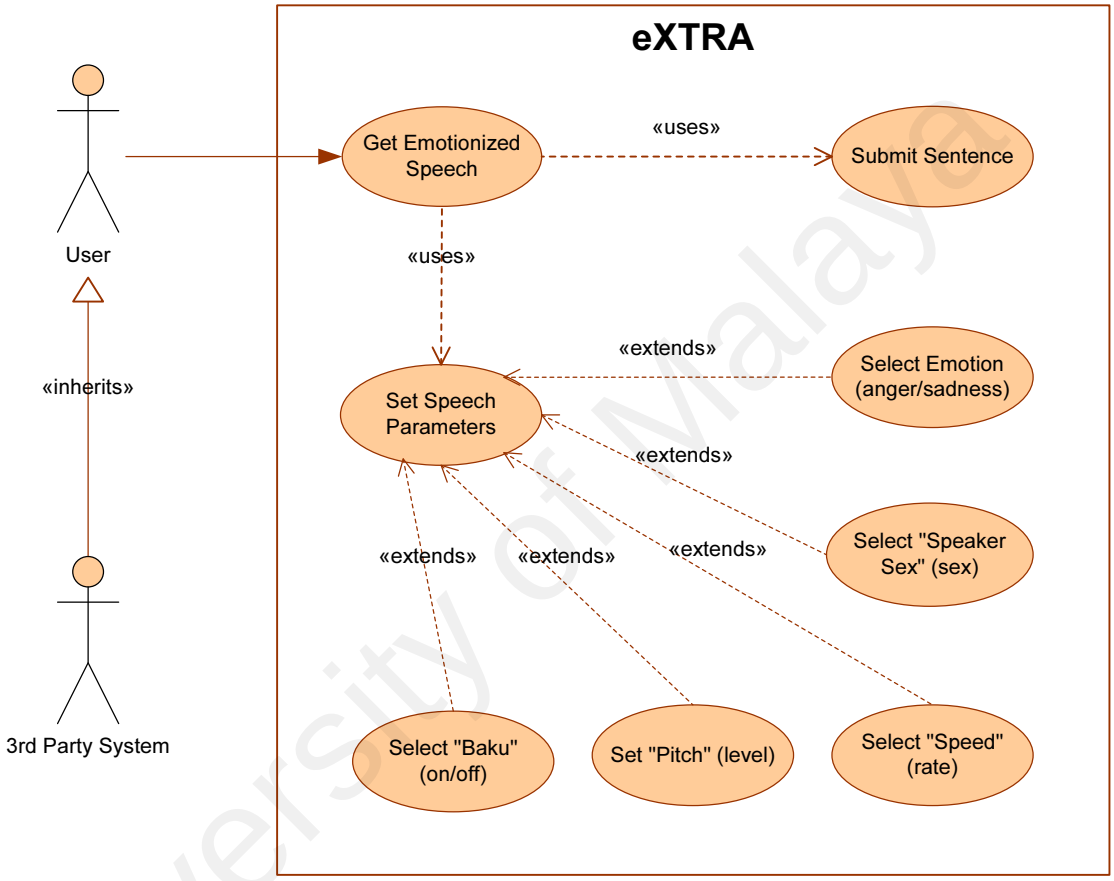


Figure 5.1: Use Case Diagram for eXTRA

5.2.2 Use Case Details

The module requires two types of input; *textual input* representing human readable texts and the *desired selection of speech parameters* such as emotional state selection (e.g. Anger or Sadness) and speaker sex selection (that determines different voices). The users are possibly a human user, or an external system. This allows for distributed input, where either the user or the external system may submit the required data in full or they

each separately submit part of it (e.g. human user submits emotion and external system submits textual input) to complete the module's input requirements. This implicates a variety of possibilities for input to the module. Table 5.2 below provides a detailed description of the use case "Get Emotionized Speech" presented in Figure 5.1.

Table 5.2: Use Case Descriptions

User	System
	1. System has been started 2. The following parameters are displayed in default values at starting point: ▪ Emotion state (Neutral) ▪ Speaker (Female 1) ▪ Speed (Medium) ▪ Pitch (Medium) ▪ Baku (checked)
3. Submits sentence of four words, each word consisting of two or three syllables. 5. Sets following parameters: ▪ Selects desired emotion or types in the emotion-symbol after input sentence. ▪ Selects desired speaker voice (male/female) ▪ Sets desired speech pitch level ▪ Sets desired speech speed 6. Clicks or enters "Read".	4. Displays "Read Sentence" on the screen 7. Pronounces the typed sentence in Bahasa Melayu Baku, and in desired emotion and voice.
Assumption: User leaves on the "Baku" option checked.	

5.3 System Design and Development

5.3.1 Overview

In the previous section, software requirements were transformed into a use case model. This section will transform software requirements into components grouped by its generalization of tasks and operations. Specifically, this section enlightens (replace 'enlightens' with a more suitable word) the system design by providing a technical explanation on how the eXTRA module manipulates SAMPA-formatted speech data to enrich it with affective components.

The first part of this chapter explains the scope of the system. Then, it defines the fundamental architectural overview. This is obtained by partitioning each part of the module identified during requirements analysis into functional subjects and assigning a set of functions to each subject. These high level functions are then refined into a more detailed architecture and the high-level dataflow is set out in a block diagram. Communication boundaries between eXTRA and the hosting TTS system are addressed by interfaces and an explanation on how the system may interact with them is given. In addition, template storage and retrieval is explained. In the second part of this design section, focus is increasingly directed towards diagrams to provide insight into the template-processing algorithm. During development object-oriented techniques were applied and have been described in this chapter by means of Unified Modeling Language (UML). Eventually, outstanding issues are covered and resolutions are considered.

For the purpose of implementing the prototype, the Java ™ platform was chosen based on three major reasons;

i. Both MBROLA and *Fasih* are Java-based applications, making the Java ™ language the preferred choice to develop an extension to *Fasih*.

ii. The Java ™ language offers a wide coverage of Object-Oriented design features, which provides flexibility to ease programming.

iii. The Java ™ platform's independent characteristics.

5.3.2 System Characteristics

The eXTRA module is probably best understood when it is viewed as an *extension* or *layer* to a TTS system. This module is designed to extend MBROLA-based TTS systems with the ability to add affective components to speech. This is done through post-processing MBROLA-generated speech data. The MBROLA engine stores speech data in SAMPA format, a computer readable phonetic script. SAMPA uses a combination of ASCII characters to identify individual phonemes, for example '**V'** for the letter '**A**'. A complete SAMPA list for Standard Malay is presented in Appendix E. The eXTRA module then enhances this speech data to become emotional speech data by applying an emotion template. Thus, the generating of emotional speech output requires three essential components: *input data* and *template data* (both representing speech data) and a *rule-based algorithm* that renders the data into output.

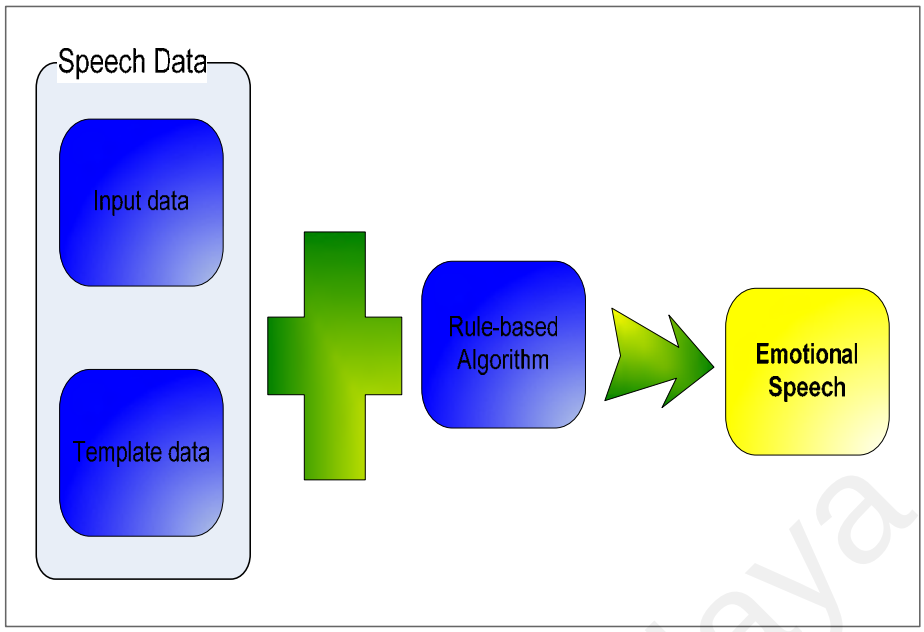


Figure 5.2: Major Components That Produce Emotional Speech

5.3.3 System Architecture

The TTS system generates and stores on disk the initial speech data (pre-processing) and then outputs this speech data into audible voice (post-processing). The eXTRA module manipulates the pre-processed speech data immediately after the TTS system completes its pre-processing.

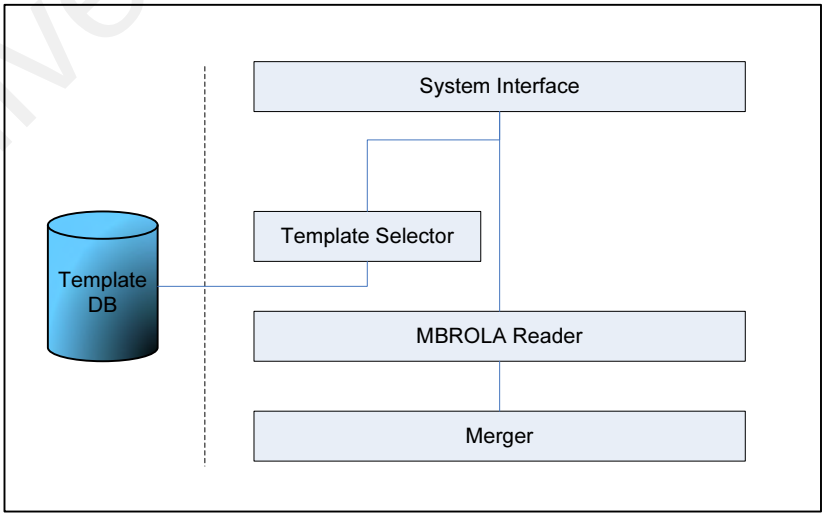


Figure 5.3: High-level Architecture of eXTRA

Figure 5.3 above shows a high-level architectural overview of the module, exposing its most basic component structure and how these components are connected to each other. The TTS System Interface is the initial starting point and Merger is the destination. Table 5.3 below details the responsibilities of each component.

Table 5.3: A Description of Basic Components

Component	Responsibility
System Interface	Speech parameters are passed through here.
Template Selector	This component is responsible for choosing the correct emotion template from the database, based on textual input characteristics (number of words, syllable sequence etc.) and the user-requested emotional state.
MBROLA Reader	This component reads the textual user input from hosting TTS and its initial (non-emotional) MBROLA speech data output from disk to memory.
Merger	This component eventually models phonemes characteristics into an object structure that allows for flexible manipulation. It models template speech data into a similar object structure. Then, it processes a rule-based algorithm to merge template speech data and user speech data (input) into emotional speech output.

Figure 5.4 below is an expanded version of Figure 5.3, exposing eXTRA module's detailed internal architecture, while Table 5.4 explains the responsibilities of each of the child component.

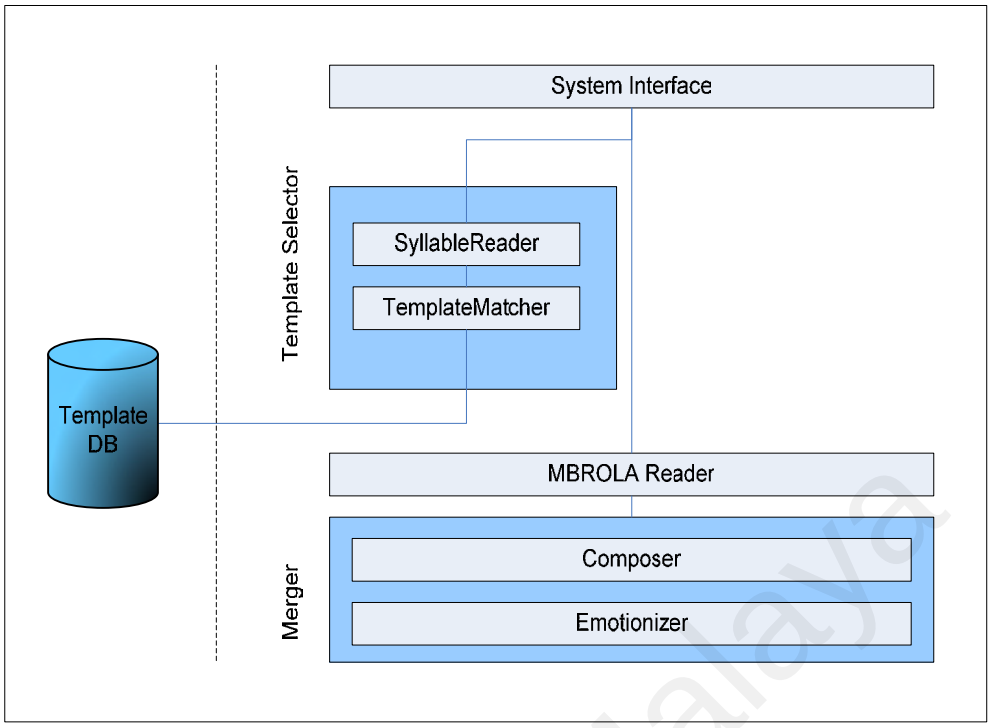


Figure 5.4: Low-level Architecture of eXTRA

Table 5.4: A Description of Detailed Components

Child Components	Responsibility
of Template Selector:	
SyllableReader	This component reads and analyses syllables
TemplateMatcher	This component matches the results of the analysis, with data from the database in order to select the correct emotion template.
of Merger:	
Composer	This component provides for navigable structures of phonemes (input and template are separately handled).
Emotionizer	This component applies a rule-based algorithm to Composer data in order to merge input-derived data with template data.

5.3.4 System Detailed Design

5.3.4.1 Block Diagram

The system block diagram in Figure 5.5 envisions the hosting of TTS's interaction with the eXTRA module, followed by a description of the blocks in Table 5.5.

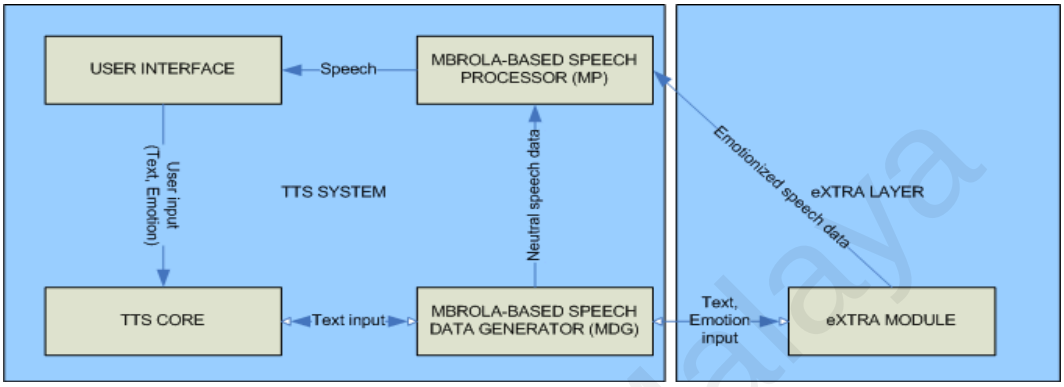


Figure 5.5: System Block Diagram Showing the Data Processing

Table 5.5: A Description of the Blocks from the Block Diagram

Block	Description
TTS SYSTEM	MBROLA-based TTS system for Standard Malay
eXTRA LAYER	Layer that extends the TTS system to produce emotional speech
USER INTERFACE	User submits textual input and selects speech parameters which include emotional state.
TTS CORE	Interacts with MDG to process Malay text into speech data
MDG	Produces MBROLA-readable speech data
MP	Processes speech data into speech output (sound)
eXTRA MODULE	Manipulates speech data into emotional speech data, based on user input and MDG output.

As illustrated in the system block diagram, neutral speech data is processed immediately after it is generated by MDG. In order to produce emotional speech however, speech

data and user input data must first be passed to eXTRA for manipulation. After eXTRA has completed its processing, the resulting speech data is passed to MP in the same way as it would be for neutral speech.

5.3.4.2 Functional Decomposition Chart

The functional decomposition breaks down the module into distinguishable functional module-parts. This decomposition serves as an important source of analysis for modeling classes and objects in the development phase. The eXTRA module allows for a functional decomposition that is straight-forward and can therefore be represented in a single flow chart as exposed in Figure 5.6.

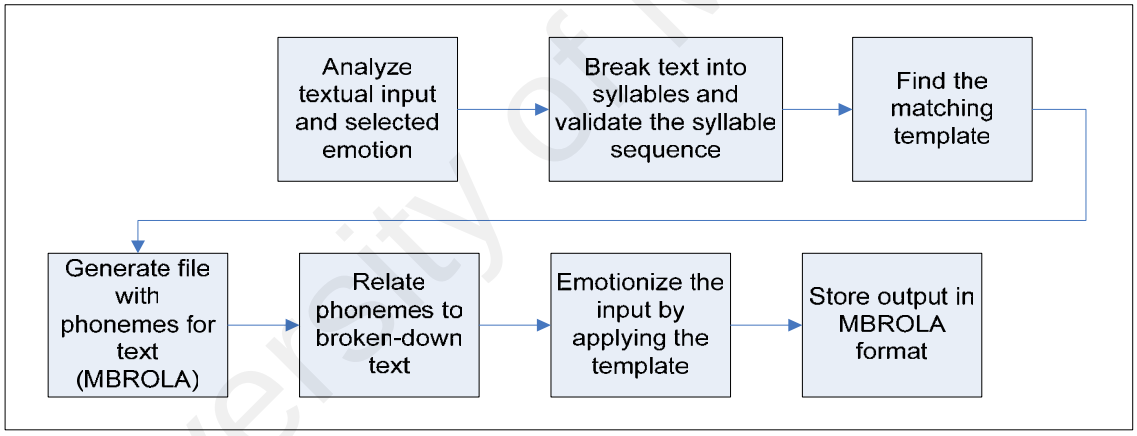


Figure 5.6: The Functional Decomposition of eXTRA

5.3.4.3 Dataflow Diagram

The dataflow diagram below (Figure 5.7) shows how data is processed by the module, starting from the user's input. Two streams of data can be distinguished:

- the stream of data that results in *determining* the emotion template and,
- the stream of data that is used to manipulate MBROLA speech data (the hosting TTS' output) by *applying* the emotion template.

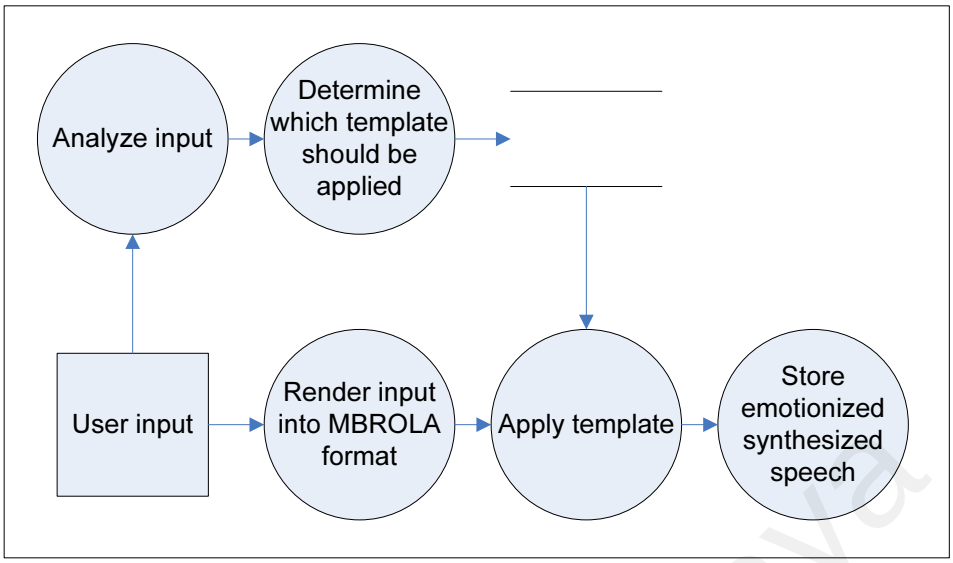


Figure5.7: Dataflow Diagram showing eXTRA's Internal Data Flow

The data flow diagram in Figure 5.7 is described as follows:

1. User input is submitted.
2. Textual input is analyzed by cutting it into syllable chunks and determining the resulting syllable sequence.
3. Based on the syllable sequence and emotion state, it is determined which template to apply
4. User input is rendered into default (neutral) user speech and stored in MBROLA format by the TTS system.
5. The selected template is applied to the user speech data
6. Modified (emotionized) user speech data is stored on disk

5.3.4.4 Class Diagram

The next page displays the collection of classes of eXTRA module (Figure 5.8). In addition to the basic classes, several classes were added to cater for flexibility and consistency. As shown in the diagram, the eXTRA class serves as the main class from

where most objects are instantiated. Since the class representation of a neutral sentence is very similar to that of an emotionized sentence, a model was adapted that provides for a universal `sentence` object. This `sentence` object is used for any type of sentence and provides access to, as suggested by the classes' names; its `words`, `syllables`, `phonemes` and `pitchpair` (pitch levels). A `Syllablizer` class provides functions that are used to cut words into syllables in order to determine a sentence's syllable sequence. The `SyllableSequenceMatcher` class in conjunction with `SyllableSequenceReader` class is responsible for identifying a matching template, by matching the user input sentence with syllable patterns from emotion templates.

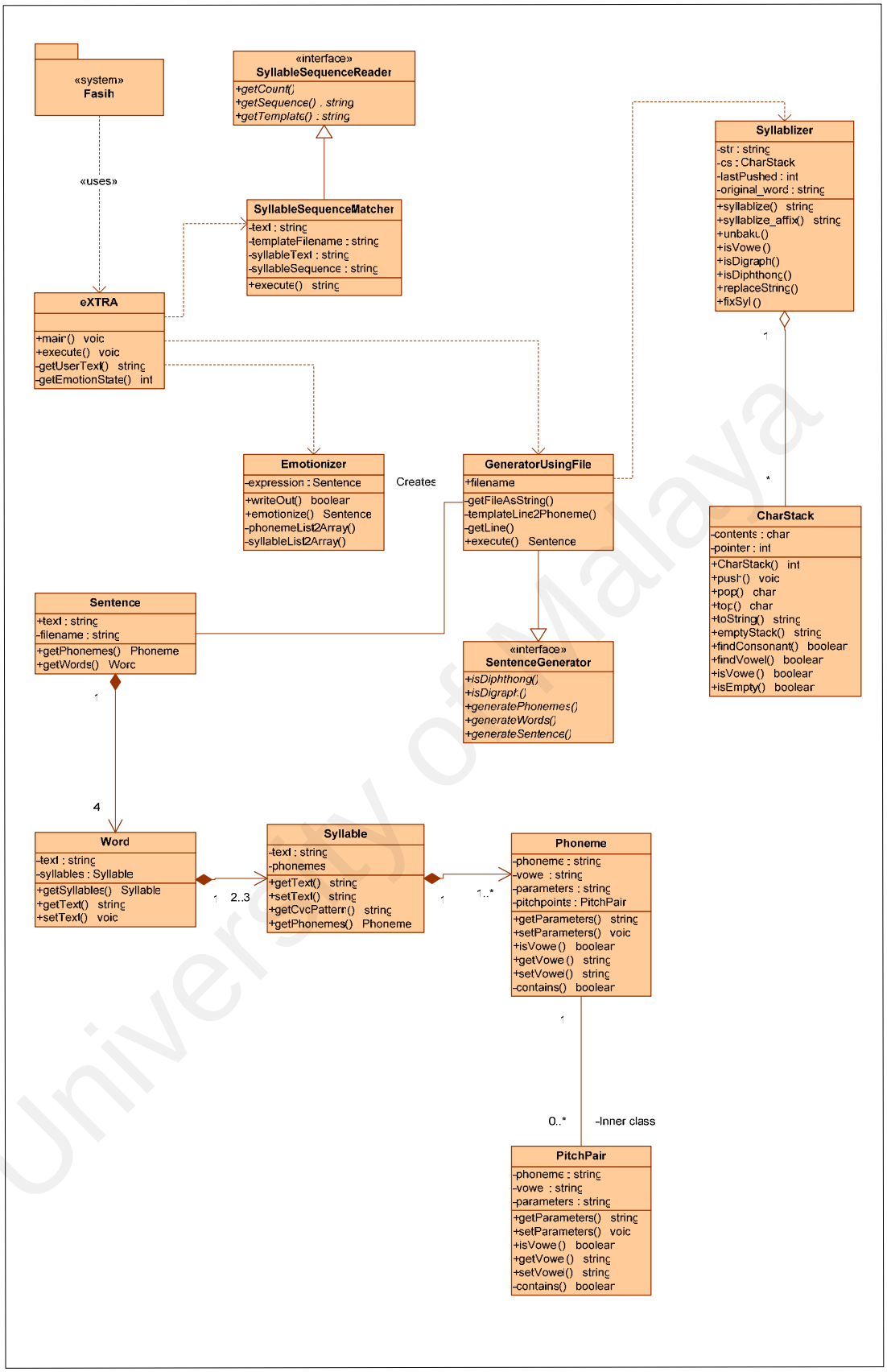


Figure 5.8: Class Diagram of eXTRA

5.3.4.5 Outstanding Issues

The prototype was realized in a pragmatic way, focusing on an object model that allows for flexibility of phonetics manipulation. Therefore the current design allows for some optimization, in particular when it comes to the following:

- method `GeneratorUsingFile.execute()`
- constructor `Emotionizer`(Sentence template, Sentence userInput)
- use of inner classes (`Phoneme`)

The `GeneratorUsingFile.execute()` method's current file reading process assumes a strict linefeed and carriage return sequence for each end of line. It should allow for more flexibility in reading template files, since the usage of linefeed and carriage return may differ between text editors available in the market. In addition, this method as well as the mentioned constructor may be 'normalized', by distributing several tasks into a set of utility methods to allow for implementing future enhancements more easily, should it be required.

The use of nested classes (in Java referred to as *inner* classes) may be avoided by applying inheritance principles. This concerns `PitchPair` inner class currently related to the class `Phoneme`.

5.3.4.6 Sequence Diagram

The use case "Get Emotionized Speech" from the use case model (Figure 5.1) is transformed to the sequence diagram below (Figure 5.9) to demonstrate the system's sequence.

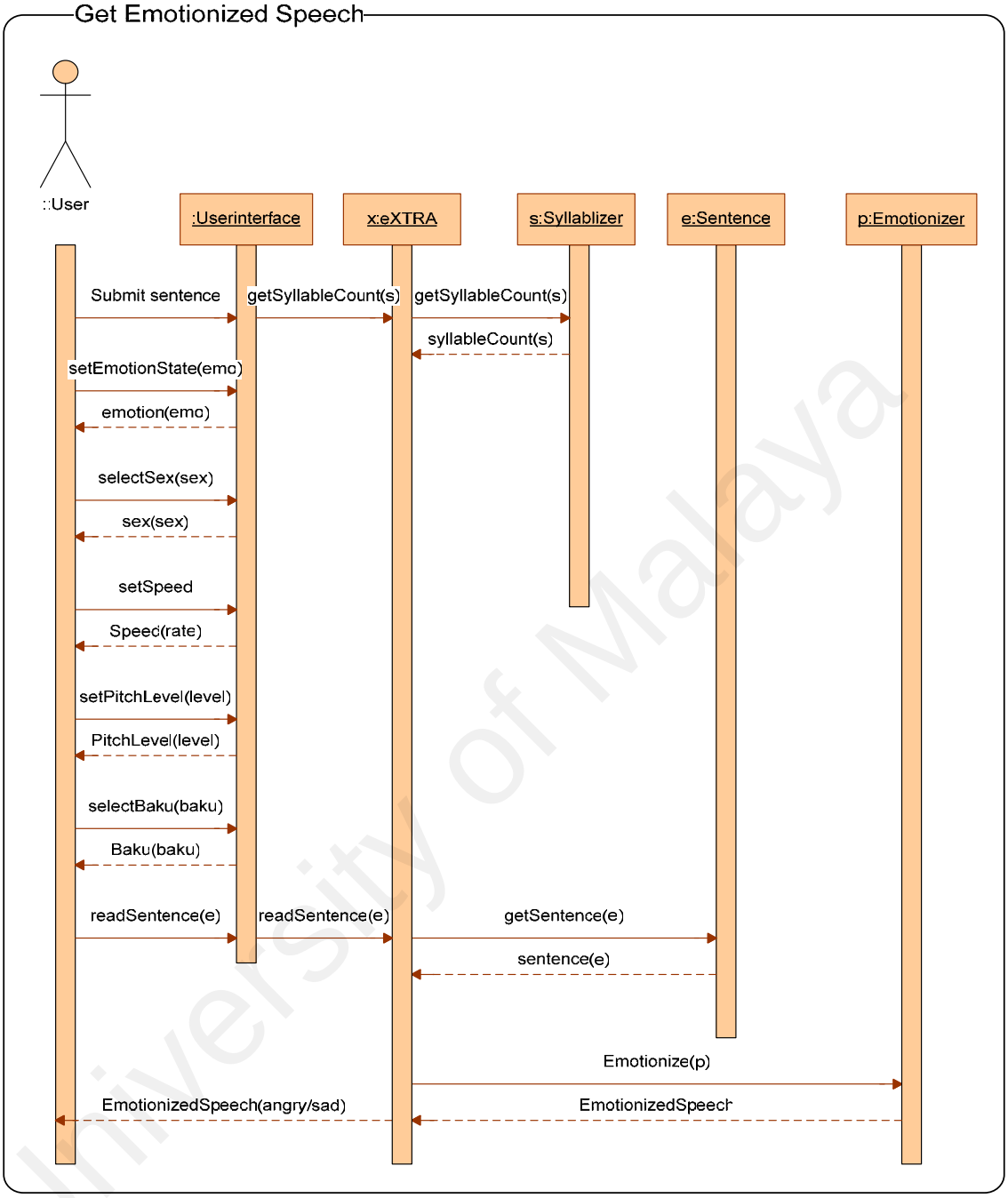


Figure 5.9: Sequence Diagram for "Get Emotionized Speech"

The sequence diagram shows that the user mainly interacts with the hosting TTS' graphical user interface by providing desired speech parameter settings. The `readSentence` action triggers a sequence of background processing that includes interaction with the hosting TTS system, resulting in audible emotional speech output.

5.3.4.7 Object Model

This is an object model of object `Sentence` that navigates through a tree of objects in order to manipulate their values. Figure 5.10 below displays the object model's hierarchy.

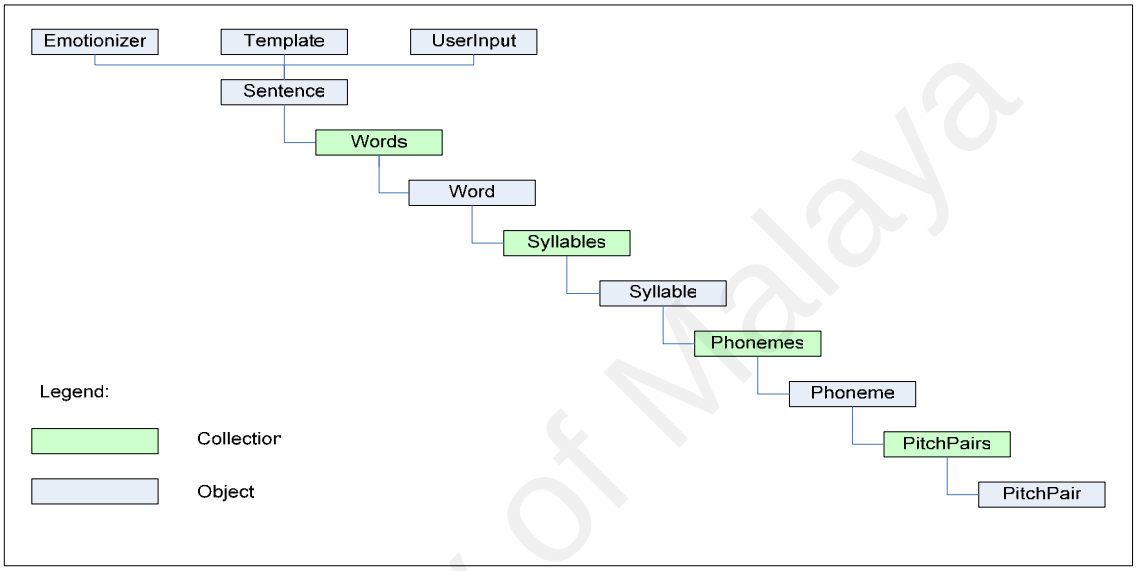


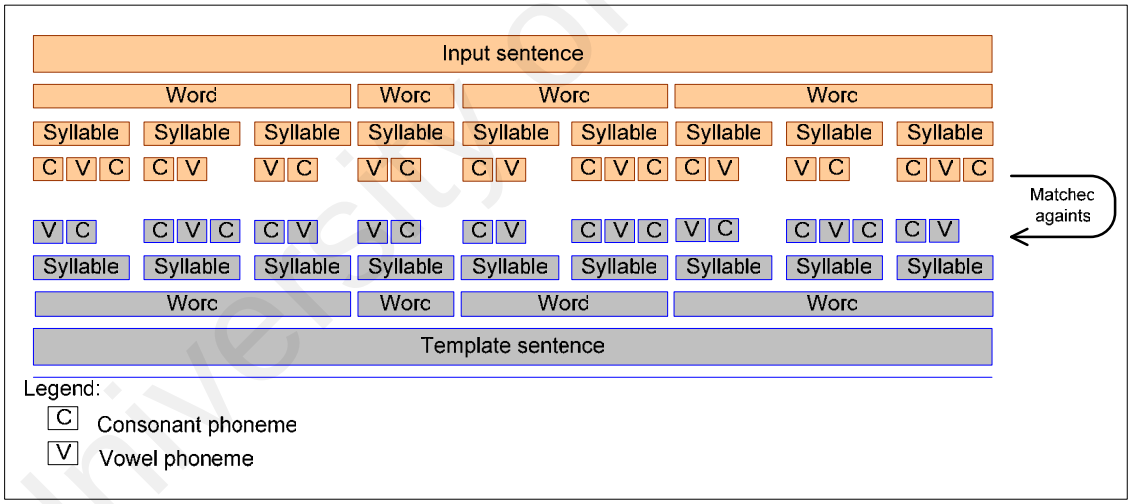
Figure 5.10: The `Sentence` Object Model

As the module design is part of a prototype that was used to observe phoneme characteristics, in particular those that are associated with emotions, the object model's rationale focused on flexibility in manipulating phonemes. A sentence is divided into words. In turn words are divided into syllables and syllables into phonemes, each containing their own pitchpair parameters. Each of these divisions is represented in objects in the object model. Much attention has been paid to the navigability of objects to achieve a high level of flexibility. As explained earlier in Chapter 4, eventually sentences need to be compared and their data to be matched on a syllable basis. To facilitate this matching process, a `sentence` object may represent either:

- A template (emotion speech data retrieved from a database),
- A user input (text converted into neutral speech data), or
- An emotionized sentence (emotionized speech data derived from a template and userinput)

5.3.4.8 The Process of Matching Sentences

The fact that a `sentence` object may play different roles (user input or template), allows for a systematic matching or comparing of their characteristics during the manipulation process . Figure 5.11 below visualizes a generic matching process at the level of phonemes.



5.11: A Visualization of One of the Matching Processes

The figure above displays a systematically organized structure of two `Sentence` objects and their derived objects being matched against each other by an `Emotionizer` object. The first `Sentence` object represents user input (Input sentence) while the second `Sentence` object represents a template from the database (Template sentence). In both sentence objects, the words are cut into syllables, which in turn are chunked into

phonemes before they are matched against each other. When matched, the emotional prosodic parameters of the template phonemes are transferred into the input phonemes Figure 5.12 further details the matching process by using an example;

Consider the input sentence is “Awak tidak tahu malu” and the selected emotional state is anger. This sentence has a syllable sequence set of 2222, therefore the matched anger template would be the template that has the same syllable sequence as well. From the template database (Appendix F), the matched template is Anger1_Fasih.pho, consisting of the sentence “Kamu sungguh kurang ajar”. Figure 5.12 in the next page shows how the input phonemes are matched against the template phonemes.

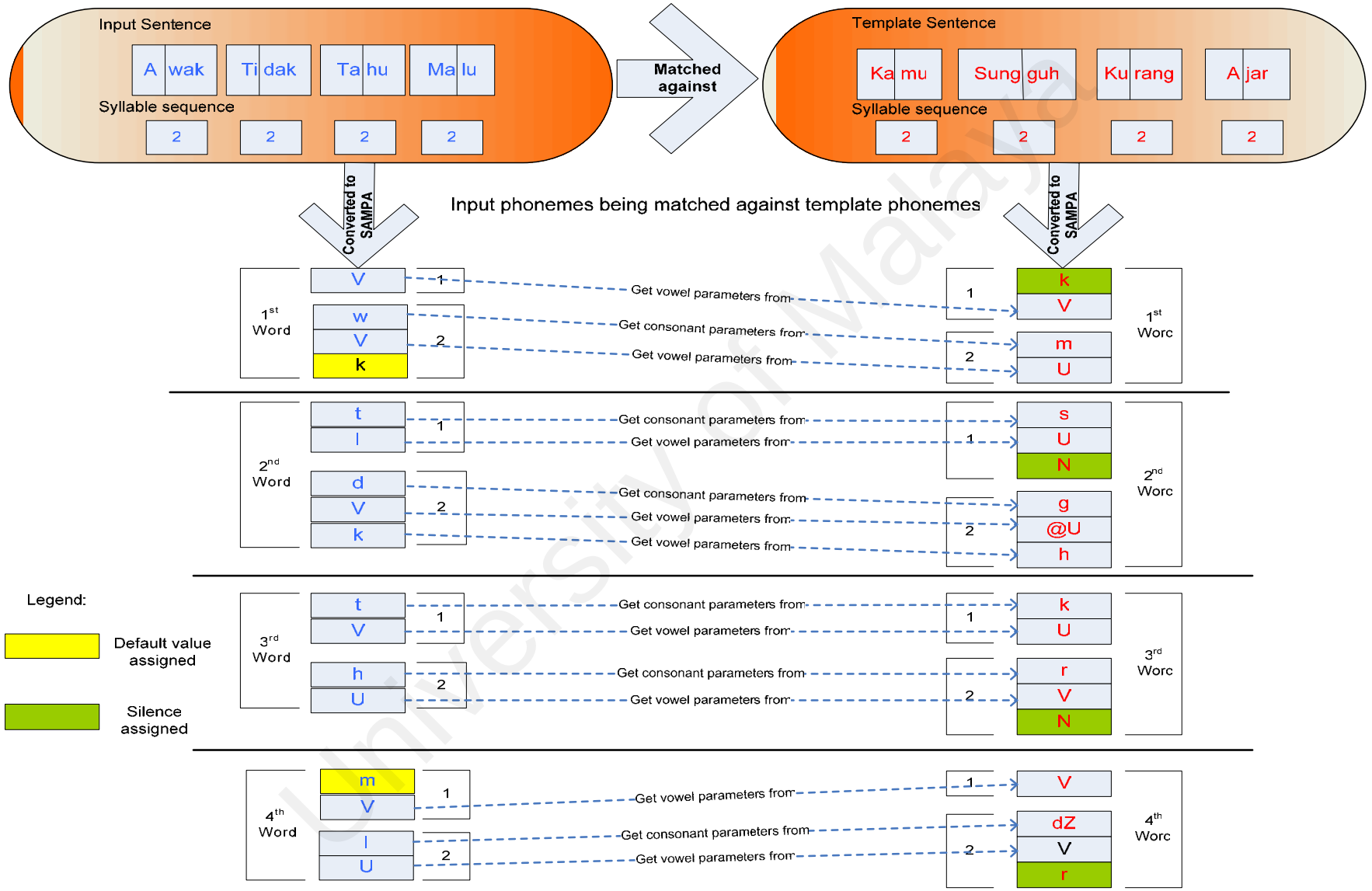


Figure 5.12: A Visualization of the Phonemes Matching Process

Vowels usually have longer duration than consonants, thus, contributing to more pitch points. However, vowel pitch points are not suitable to be transfered to consonants, since this may produce longer sound than expected. To solve this issue, syllable and categorical matching are applied. Syllabic matching refers to the matching of phonemes between the input and template according to syllables. In other words, a pattern of syllables from the sentence is first identified in order to establish a match against another sentence's syllable pattern. Categorical matching refers to the matching of phonemes of the same type; vowels are matched against vowels while consonants are matched against consonants. This is illustrated in Figure 5.12, where the vowels from the input sentence are matched against the vowels from the template sentence according to syllables. This also applies for consonants.

In the case where a phoneme is left without a match, a default duration value or silencing is assigned. A default duration value is assigned to the unmatched phonemes in the input sentence while the unmatched phonemes in the template are put to silence.

The next figure illustrates the transferring of prosodical information from the template to the input.

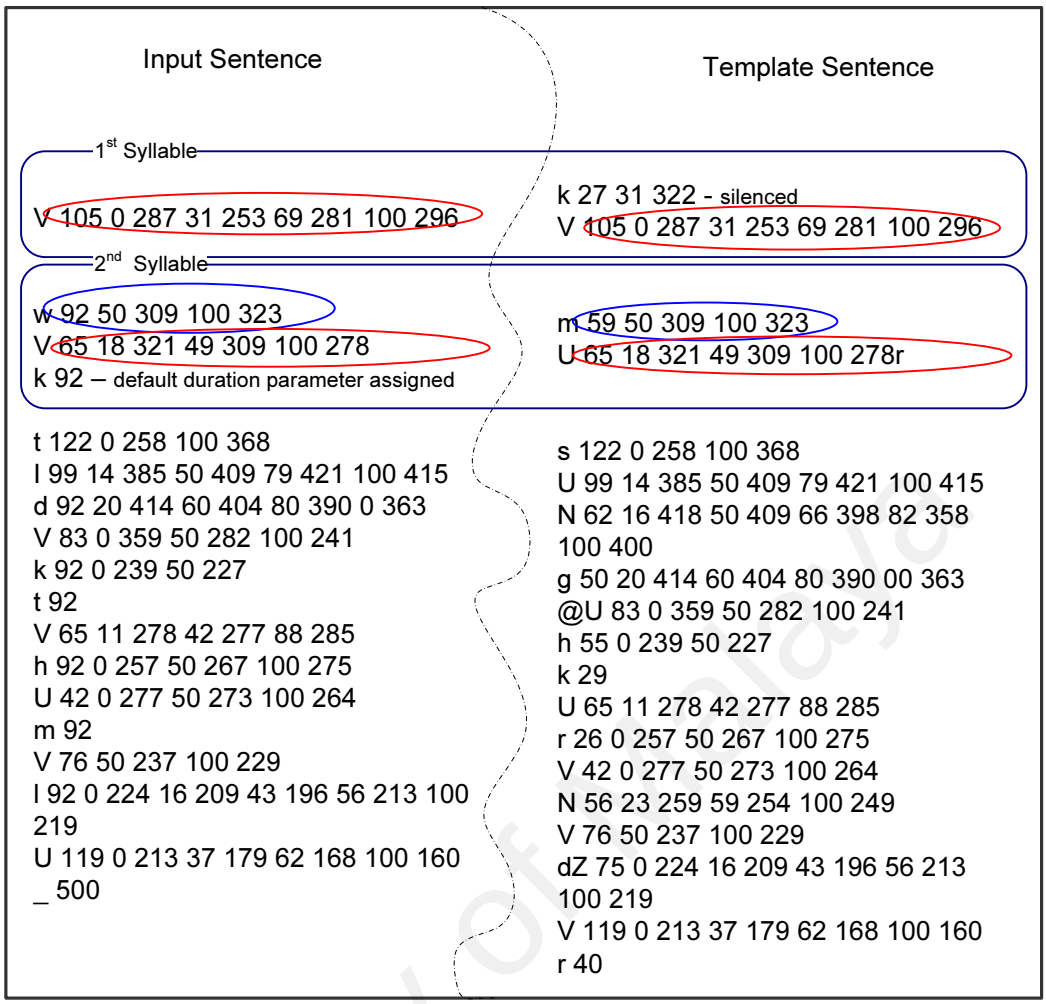


Figure 5.13: Emotional Prosodic Parameters Transfer from Template Sentence to Input Sentence

Figure 5.13 depicts on how the prosodic parameters from the template are transferred to the input, particularly at the first word. In a particular syllable, template-vowel parameters (in red circle) are transferred to input-vowels and template-consonant parameters (in blue circle) are transferred to input-consonants. However, in the case where a consonant from the template has duration of less than 92, the value in the matched input-consonant will be adjusted to default value of 92. The rationale for this adjustment is that, when tested with several input sentences, it is found that certain consonants produce sounds that are too short when concatenated. This causes the synthesized speech to be unclear.

Table 5.6 shows the corresponding organization of the matching process in Figure 5.13.

Table 5.6: The Organization of Matching Between the Template and the Input Contents

Line No.	Contents of Template Sentence			Contents of Input Sentence		
	SAMPA symbol	Parameters		SAMPA symbol	Parameters	
		Duration(ms)	Pitch Points pairs		Duration(ms)	Pitch Points pairs
1	k	silenced				
2	V	105	0 287 31 253 69 281 100 296	V	105	0 287 31 253 69 281 100 296
3	m	59 (<92)	50 309 100 323	w	92	50 309 100 323
4	U	65	18 321 49 309 100 278	V	65	18 321 49 309 100 278
				k	92 (default)	

Legend:

Unmatched phoneme

Table 5.6 shows that the relevant prosodic parameters from the phonemes in the template are transferred to the matched phonemes in the input. A post-processing is also done for the purpose of assigning silence and default values to the 'left-over', unmatched phonemes. The example shows that consonant /k/ in the template is put to silence while consonant /k/ in input is given a default value of 92 for the opposite reason. Such value is given so that the consonant produces a basic sound when concatenated. This value is copied from Fasih, which assigns *only* duration parameter to its consonants.

It is observed that the output speech sounds quite emotional when synthesized. This may be due to the organization of the high-level matching process. The performance of post-processing, which assigns silence and default values to unmatched phonemes also

contributes to such result. To confirm the recognition of the emotions elicited in the synthesized output, a listening (perceptual) test was conducted. The results are presented in Chapter 6 (section 6.3).

5.3.4.9 Object Constraints

For this prototype, several constraints were applied to (a collections of) objects. For example, to 'emotionize' a sentence, the module allows a maximum of only four words in a sentence and each word shall contain only two or three syllables. Figure 5.14 aids in showing the constraints that apply to the relationships between objects.

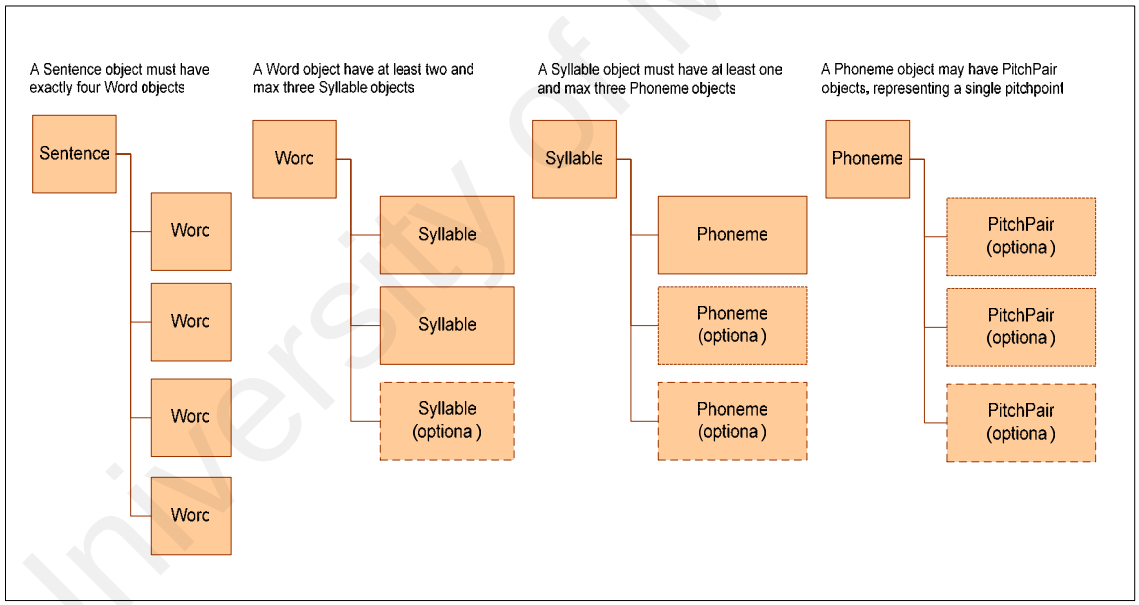


Figure 5.14: A Visualization of Related Objects and Their Constraints

5.4 eXTRA Prototype

This section will describe the conformity of the developed prototype to the proposed software requirements (Section 5.2). The design model described in the previous sections is transformed to screen layouts that show the functionality of the prototype. The prototype is based on the hosting TTS called *Fasih*, developed by MIMOS Berhad.

5.4.1 GUI of Fasih

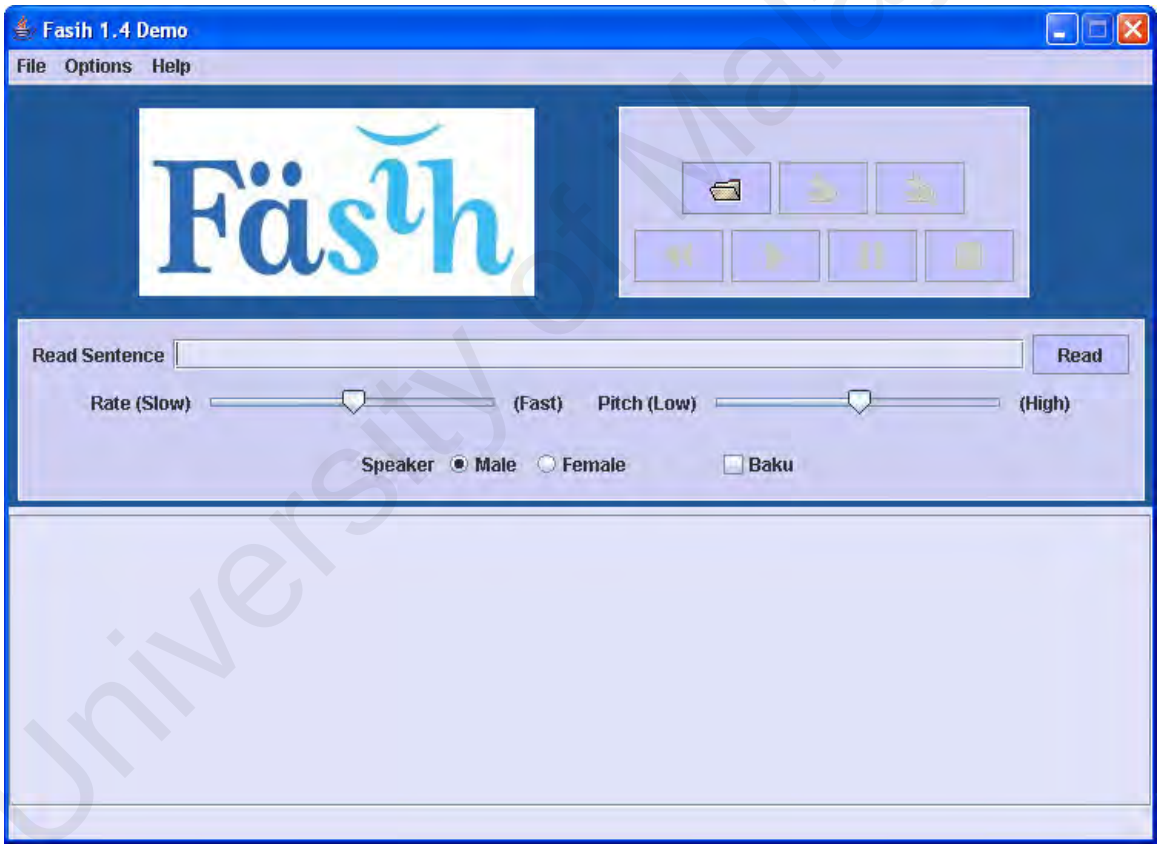


Figure 5.15: *Fasih*'s GUI

This is a screenshot of the original *Fasih*'s GUI. There is a textbox that allows users to enter a sentence and a 'Read' button to submit the input. In addition, it allows for the selections and settings of several speech parameters to manipulate voice output, such as 'rate', 'pitch', 'speaker' (sex) and 'baku'.

5.4.2 *GUI of Fasih Extended With eXTRA*

To enable users to use the eXTRA functionality, radio buttons that indicate different emotional states were added to the prototype (as shown by the arrow in Figure 5.16). Following is a screenshot of the prototype's GUI:

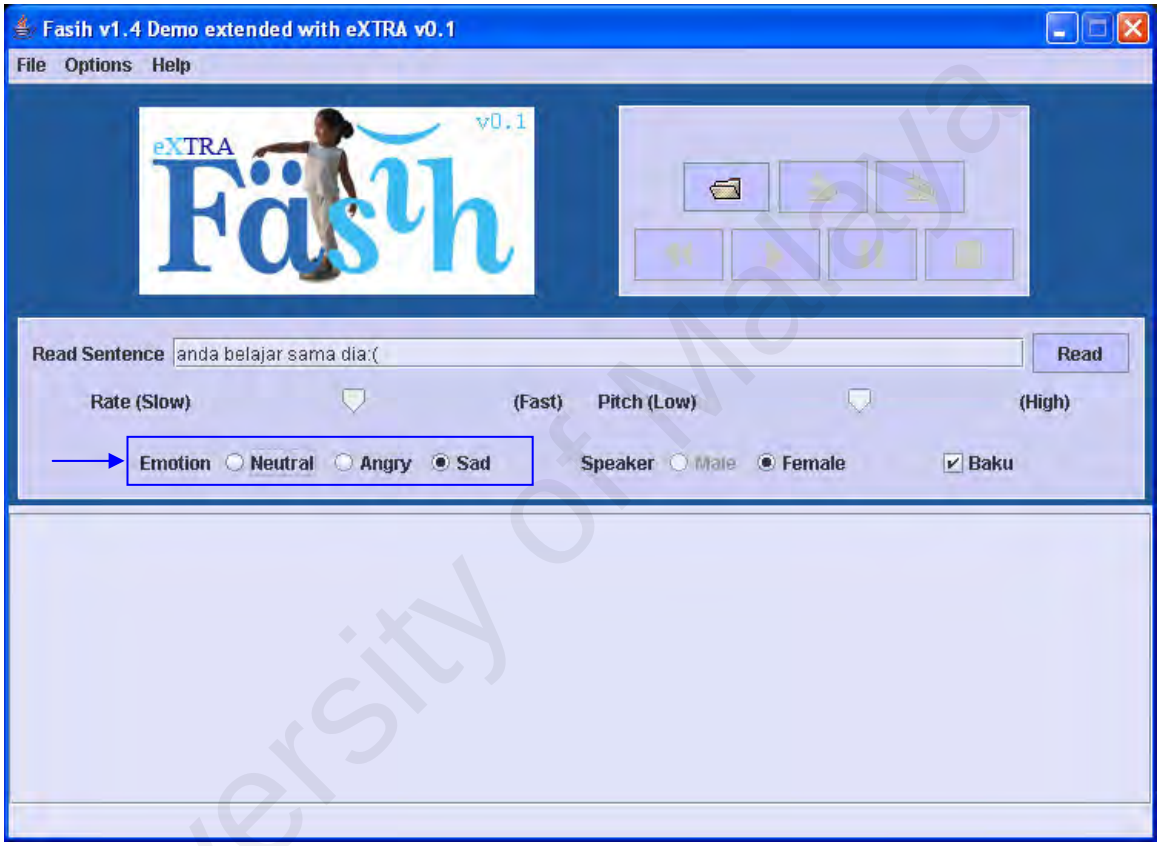


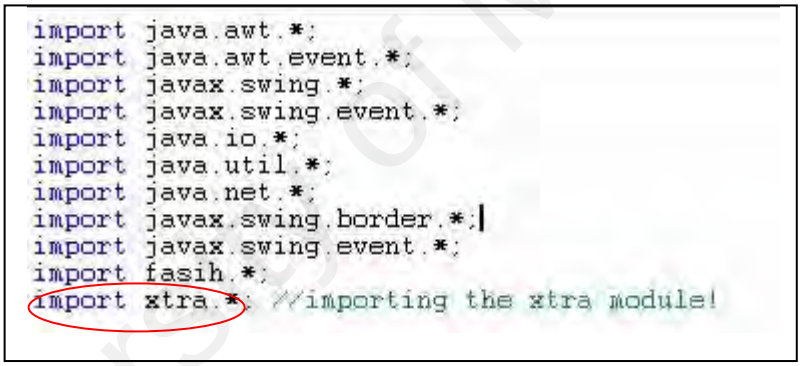
Figure 5.16: GUI of *Fasih* with eXTRA

The radio buttons displaying emotional states Neutral, Angry and Sad enable users to set the chosen 'emotion' parameter. Alternatively, users can type in special symbols corresponding to each emotional state, as shown in Figure 5.16.

Users can also choose different speakers according to gender to read the input. However, since the modeling of the synthesized voice were based on a female voice characteristics; pitch frequencies, intensity etc. Therefore, choosing the female speaker

would give a better sound. Because of this reason also, the male speaker seem to produce a 'softer' voice, similar to a woman's. To some users, he sounds quite funny. This is because the pitch extracted from real human samples to be used with the synthesized voice, and the frequency set in Praat is according to the female voice.

The following screenshot shows how the `xtra` package, which contains the Java classes for eXTRA, is imported to *Fasih*. The code is implemented in a *Fasih* GUI class called ***Fasih***`_TextReader,` which is responsible for a major part of the TTS system's interaction capabilities.



5.4.2.1 eXTRA module's API

To allow *Fasih* to conduct operations that were implemented by eXTRA, *Fasih* calls methods from eXTRA's Application Program Interface (API). To establish this, a slight modification to the TTS system's source code is needed. Specifically, the following method was added to *Fasih*'s source code:

```
xtra.Verbalizer.execute(userText, emotionState);
```

The above code consists of two parameters that are passed to eXTRA module's `Verbalizer` class:

`userText:` a string containing user-submitted text , e.g. "anda belajar sama dia" as shown in Figure 5.14

`emotionState:` an integer representing the emotion state submitted by the user;; 1(Angry) or 2 (Sad)

Here is the corresponding Java code (circled) that shows the lines where eXTRA method is called from *Fasih* GUI. A `Verbalizer` object called `emox` is instantiated. The `emox` object is used to call the execute method, by passing the two parameters mentioned above.

```
public void keyPressed(KeyEvent evt) {
    if (evt.getKeyCode() == KeyEvent.VK_ENTER) {
        if (evt.getSource() == tfCurrentSentence & tfCurrentSentence.getText() != "") {
            String inputstr = tfCurrentSentence.getText();
            if(inputstr.substring(inputstr.length()-2,inputstr.length()).equals(":)") || inputstr.substring(inputstr.length()-2,inputstr.length()).equals(":(")
                inputstr=inputstr.substring(0,inputstr.length()-2); //emotion-related
            currentVoice = getVoice();
            fasih.readSentence(currentVoice,inputstr, "", cbBaku.isSelected(), sliderRate.getValue(), sliderPitch.getValue(), "userdict.txt", false);
            // Insert emotion handling code here
            if(!jbNeutral.isSelected()){
                Verbalizer emox=new Verbalizer();
                emox.execute(inputstr,((jbSad.isSelected())?1:0)+1);
            }
            play("out_0.pho",""); // change out_0.pho to the new pho file
        }
    }
}
```

5.4.3 *Screenshot of Debugger*

By default, the prototype outputs information for debugging purposes to the Java console. This debugging information may be used to detect mismatching of existing and new templates that may be added. Figure 5.17 below displays a screenshot from a console window with debug information.

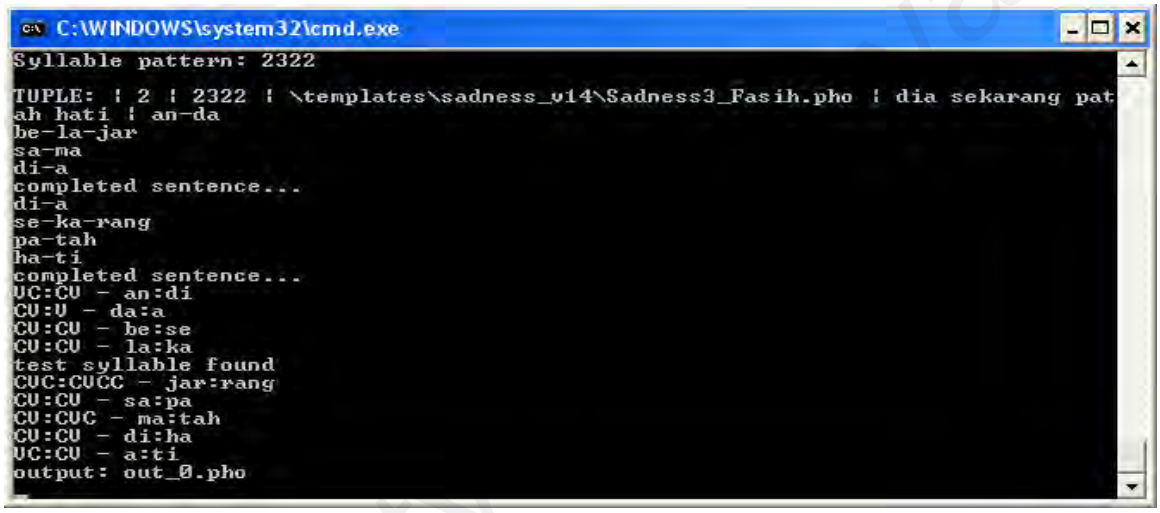


Figure 5.17: Debugging Information Produced for Sentence "anda belajar sama dia"

5.5 Summary

The interrelation between the four elements; software requirements, architecture, components and interface in eXTRA has been discussed accordingly. The affective layer framework derived in the previous chapter led to the construction of software requirements for eXTRA. These requirements, modeled with the aid of UML, in turn moulded the design decisions of eXTRA. The main goal of this project, which is to render the hosting TTS system output to emotion-blended speech, has engineered the architecture of eXTRA. The behavior of components and interface were presented in the sequence diagram while the detailed functions of different components were presented

in various diagrams to aid understanding. The last section revealed the result of software prototyping. The main purpose is to validate the requirements and design model provided in the earlier sections. The prototype is a preliminary type, that serves as a model for later stages for the final, complete version of the system. The prototype which is regarded as the incomplete version of the system may have limitations in its implementation as an affective component. The prototype's evaluation is discussed in the next chapter.

CHAPTER 6

Evaluation, Results and Discussions

6.1 Introduction

The first part of this chapter explains the evaluation method of eXTRA. The subsequent section reports on the summary of results derived from perception test for synthesized speech generated by *Fasih* extended with eXTRA layer.

6.2 Evaluation of eXTRA

6.2.1 Evaluation Method

The method used was similar to the perceptual test conducted for human speech sample. However, the test was only conducted once. Ten native listeners who were university undergraduates and postgraduates participated in this test. They were not aware of the test stimuli. First, they were asked to listen to sixteen neutral utterances, as presented in Appendix D, that were played twice, once in anger tone and once in sadness tone in random order. Participants were reminded to concentrate on the intonation instead of the content to identify the emotional states the utterances were corresponding to. Next, they were asked to listen to a series of emotionally-inherent utterances, also in random order.

These utterances were also played in anger and sadness states. The participants were provided with a fixed-option form, as in Appendix C2 to choose their answers from. Options also included Fear and Happiness emotional states for the purpose of adding variety and avoiding bias judgments. Results based on neutral and emotionally-inherent contents were presented in the following section.

6.3 Perception Test Results for Synthesized Emotional Speech Generated by Fasih with eXTRA layer

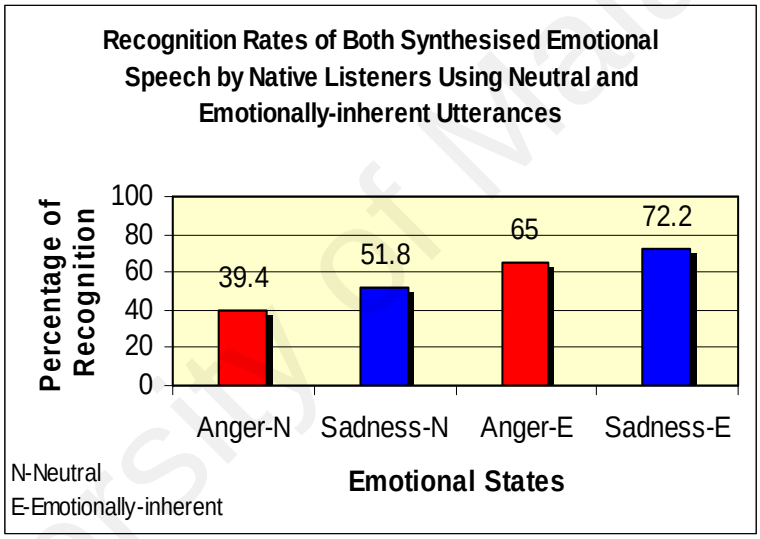


Figure 6.5: Results for Both Synthesized Emotional Speech Using Neutral and Emotionally-inherent Content. ('N' denotes Neutral utterances and 'E' denotes Emotionally-inherent utterances.)

6.3.1 Summary of Results and Discussions

Recognition rates for synthesized utterances that use neutral content were quite low, while the recognition effort scaled within one (1) to three (3). For emotionally-inherent content, the recognition accuracy was significantly higher. It is observed that

participants tend to focus on the contents rather than the tones elicited despite repeated reminders. This is because they were seen discussing with each other in between the listening activity, on the context of the sentence. Nevertheless, this kind of response is expected, because in real life situations, meaning and context are a bigger clue to the emotional state of the speaker. There were also significant differences in recognition accuracy among the two emotions using emotionally inherent contents: recognition rates observed for Anger set were significantly lower than Sadness sets; either the speaker was relatively less successful in expressing anger in some utterances, or anger is more difficult to recognize in certain utterances. Recognition effort result shows that most participants felt that there was no effort required for recognizing the emotions in most utterances, with very few requiring little attention.

Overall, the recognition rates show higher figures compared to previous research work (Bulut, Narayanan and Syrdal, 2002; Nass ***et al***., 2000; Murray and Arnott, 1993, 1996). Basically, these results indicated over sixty percent recognition rates for both intended emotions expressed in the synthesized utterances, which are impressive, considering that people recognize only sixty percent emotion in ***human*** voice (Shrerer, 1981 in Nass ***et al.***, 2000).

6.4 Summary

In this chapter, functional evaluation of eXTRA has been discussed. Next is the discussion of its efficiency and limitations which will be described in the final chapter.

CHAPTER 7

Conclusions and Future Work

7.1 Discussions of Findings and Applicability

This section discusses the research findings and puts forward the conclusion derived from the discussion.

The fact that Malay has no stress or word tone or any other intermediate level of prosodic organization (Zuraidah and Knowles, 2006) makes it a language suitable for computational linguistic research. The simple syllable structure that the Malay language is based on allows for a method that focuses on the number of syllables rather than other linguistic features. The high percentage of recognition by listeners shows that that the emotional synthesized Malay speech can be produced in such a way that it is indistinguishable by users. Thus, the objective of this research has been achieved. Besides that, the emotional tones derived from the actor's voice which provided the base for the tones in the emotion templates proved that the use of acted speech reached a consensus on what general cultural-specific vocal emotions sound like. Thus, the use of emotion templates with acted speech is preferred to using natural 'emotional' speech as it provides clearer affective tones and convenience.

The significant differences shown in the results from the experiments between neutral and emotionally-inherent contents proved that utterances that have no conflicting content is more suitable for use in building templates. It is also observed that emotions are vocally conveyed chiefly through vowels rather than consonants. This is shown by the number of pitch points extracted from each vowel-phoneme.

As highlighted in the beginning of Chapter 2, most present emotive synthesis systems were modeled based on prosodic patterns that lack prerequisite studies on human emotions, thus, they do not exhibit the variability as explained in section 2.4. Variability is present in all natural human speech, and thus, is important to be incorporated into synthesized speech in some way to simulate its natural sounding. eXTRA attempts to exhibit this variability by introducing multiple tones in a single emotional state. Furthermore, the limitations derived from the use of computed prosodic parameters can be overcome by using prosodic parameters from actual human voice samples. This can be effectively done by the use of emotion templates. However, in order to make further progress, a robust descriptive framework for speech, which includes variability factors, is required. Production of this framework will require a greater co-operation among disciplines related to speech science.

7.2 Contributions of Study

The aim of this study is to improve spoken synthesized Malay speech through the incorporation of emotions. The specific research objectives of this dissertation have been addressed. It is concluded that the aim of this study has been met as well. This is supported by encouraging results from end users, as reported in Chapter 6. Essentially, this study introduced a novel approach; template-driven emotion generation method to

produce emotive synthesized speech from text. Compared to previous studies, this approach is more effective in producing the correct intonations, due to high-level matching on the syntactic form of sentence (declarative, imperative, interrogative).In addition, it is learned that the basis of modeling emotive synthesized speech using templates is the design of the corpus to be used as stimuli. Organized data that form the corpus addresses the special requirements of building culturally specific emotional tones, other than allowing data processing to be done smoothly.

Most importantly, this approach allows for the details of affections elicited in a real human's voice to be 'injected' into the synthetic speech at the level of syllables, the basic rhythmical pronunciation unit in Malay. By doing it this way, the variability in human speech is also infused, attributing to a human-like emotional speech produced by the TTS.

7.3 Future Work

The limitations in the prototype suggest a natural direction for future work**.** This section outlines the future research directions.

7.3.1 Expanding the Corpus / Speech Database

So far, focus was set on adding emotions to declarative sentences that consist of four words, each consisting of two or three syllables. While these restrictions may allow for producing satisfactory results in the preliminary phase, it does not address a complete coverage of spoken words in Malay language. It is suggested that capabilities be extended towards encompassing the complete range of syllable structures for words,

different syntactic forms of sentences (interrogative, imperative) and word categories (noun, verb etc.) to allow unrestricted text to have the added emotions.. The parameters from an emotion template are more likely to be compatible with an input sentence of the same syntactic form. This is analogous to an input sentence that contains similar positions of word categories. This would result in speech that contains more appropriate intonation patterns. A complete coverage of syllable structures can be achieved by expanding the corpus used as stimuli for templates, which requires interdisciplinary efforts including close cooperation with linguists and committees such as The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA). It should be noted that expanded corpus requires increased work in data analysis and speech annotation, since at the time of this writing there is no tool available to automate the analysis of Malay speech data.

7.3.2 Extending the Emotion Types and Gender

The emotion templates can also be extended to a greater number of distinguishable emotions, even for less explicitly expressed emotions such as tiredness and guilt. In addition, an expression of a particular emotion can be blended with another emotion by aggregating their emotional parameters. This process can be automated. For example, anger may be blended with sadness, producing a new expression that derives from both of these emotion templates' parameters.

The stimuli used in this study were based on only female voices. In the future, male voice should also be included. The differences in frequency, pitch and loudness will also be studied.

7.3.4 Improving the Context-Sensitive Intonations

Another suggestion to improve context-sensitive intonations is the tagging of words at word class level in order to communicate emotions by correlating them to content. Consider :

Sentence 1: Saya tak ***makan*** buah itu.

Sentence 2: Saya tak makan ***buah*** itu

Sentence 3: ***Saya*** tak makan buah itu

The translated sentence is "I did not eat the fruit". Assume that in all the three sentences, the italic word is given prominence or stressed. In 1, the speaker stresses on "makan" which probably suggests that he or she ***did*** something else with the fruit. In 2, the speaker probably suggests that it is not the fruit that is being eaten, but ***something else***. Lastly, in 3, the speaker is denying that ***he or she*** ate the fruit; somebody else did.

Tagged words would produce the acquired prominence in specified words which would in turn give a more meaningful utterance, as well as aid users when using anaphoric or reduced descriptions when communicating.

7.4 Remark

This chapter concludes the deliverable of an affective layer prototype (eXTRA) for MBROLA-based Malay TTS systems. Historically, eXTRA was invented based on the idea to develop an affective component to the first Malay TTS by MIMOS Berhad, ***Fasih***. The intention was to enhance the toneless synthesized speech of ***Fasih*** by infusing naturalness via the use of emotions. The component was only to be used with ***Fasih***. At present, the idea of emotional speech synthesis is to be shared with computational speech practitioners, who possess common interest in this field, particularly in Malay speech synthesis. Therefore, the features of the affective layer

were expanded in order to cater for any TTS systems that operate on MBROLA. Finally, it crafts eXTRA as an acronym for **eX**pressive **T**ext **R**eader **A**utomation layer that shall improve the output of MBROLA-based Malay Text-to-Speech systems.

8 REFERENCES

Abadjieva, E., Murray, I. R., & Arnott, J. L. (1993). *Applying Analysis of Human Emotional Speech to Enhance Synthetic Speech.* **In** Proc. of Eurospeech '93, Berlin, Germany, (pp.909-912).

Banse, R., & Scherer, K. R. (1996). Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology,* 70(3), pp.614-636. Retrieved 20th September 2005, from APA Online

Barro, A., & Vaillot, B. (2004). Flying Blind [Television Programme, Astro Channel 50]. On *Air Crash Investigations Series, National Geographic Channel*.

Black, A. W. (2003). *Unit Selection and Emotional Speech.* **In** Proc. of Eurospeech 2003, Geneva, Switzerland.

Boersma, P., & Weenink, D. (2005). Praat (Version 4.3.17). Amsterdam, NL.

Brave, S., & Nass, C. (2003). Emotion in Human-Computer Interaction. In J. A. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 81-93). Mahwah, NJ: Laurence Elbaum Associates (LEA).

Bulut, M., Narayanan, S., & Syrdal, A. K. (2002). *Expressive Speech Synthesis Using a Concatenative Synthesizer.* **In** Proc. of ICSLP, Denver, CO.

Burkhardt, F. (2005). *Emofilt: The Simulation of Emotional Speech by Prosody-Transformation.* **In** Proc. of 9th European Conference on Speech Communication and Technology (Interspeech 2005), Lisbon, Portugal.

Burkhardt, F. (2005b). Emofilt. Retrieved 12th July 2005, from http://emofilt.sourceforge.net/.

Cahn, J. E. (1989). *Generating Expression in Synthesized Speech.* Unpublished Masters thesis, Massasuchets Institute of Technology (MIT), Boston.

Cahn, J. E. (1990). The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society, 8*, 1-19.

Cai, L., Zhang, W., & Hu, Q. (1998). Prosody Learning and Simulation for Chinese Text to Speech System. *Journal of Tsinghua University,* 38(S1), pp.92-95. Retrieved 13th March 2006, from http://hcsi.cs.tsinghua.edu.cn/english/Paper1998.htm

Campbell, N. (2000). *Databases of Emotional Speech.* **In** Proc. of ISCA Workshop on Speech and Emotion, Belfast.

Chen, S. H., Huang, S. H., & Wang, Y. R. (1998). An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech. *IEEE Transaction on Speech and Audio Processing,* 6(3), pp.226-239. Retrieved 23rd March 2005, from IEEE Explore Online

Chen, Y., *et al.* (2000). Learning Prosodic Patterns for Mandarin Speech Synthesis. *Journal of Intelligent Information Systems,* 19(2), pp.95-109. Retrieved 18th Jan 2005, from Kluwer Academic Publishers Online, NL

Cornelius, R. R. (1996). *The Science of Emotion: Research and Tradition in the Psychology of Emotion*. NJ: Prentice-Hall.

Cornelius, R. R. (2000). *Theoretical Approaches to Emotion.* **In** Proc. of ISCA Workshop on Speech and Emotion, Belfast.

Eide, E. (2002). *Preservation, Identification and Use of Emotion in a Text-to-Speech System.* **In** Proc. of IEEE 2002 Workshop on Speech Synthesis, Santa Monica, CA, USA.

Eide, E., & Aaron, A. (2004). *A Corpus-Based Approach to Expressive Speeh Synthesis.* **In** Proc. of 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA.

El-Imam, Y. A., & Don, Z. M. (2000). Text-to-Speech Conversion of Standard Malay. *International Journal of Speech Technology, 3*(2), 129-146.

Forshew, D. (1999). Basic Home Care for ALS Patients: The ALS Association Guide for Patients and Families. Retrieved 10th February 2006, from http://www.ALSa.org/files/cms/Resources/basic_home_care.pdf.

Gibbon, D. (1998). German Intonation. In D. Hirst & A. de. Cristo (Eds.), *Intonation Systems*. Cambridge, MA: Cambridge University Press.

Hofer, G. O. (2004). *Emotional Speech Synthesis.* Unpublished Masters Degree Thesis, University of Edinburgh.

Huang X., Acero,A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. NJ: Prentice Hall

Iida, A. (2002). *Corpus-Based Speech Synthesis with Emotion.* Unpublished PhD thesis, University of Keio, Japan.

Lee, J. D., Hoffman, J., & Hayes, E. (2004). *Collision Warning Design to Mitigate Driver Destructionn.* **In** Proc. of CHI 2004, Vienna, Austria.

Li, Y., Lee, T., & Qian, Y. (2004). Analysis and Modelling of F0 Contours for Cantonese Text-to-Speech, *ACM Transactions on Asian Language Information Processing, 3*(3), pp.169-180. Retrieved 13th March 2005, from ACM Online

Liow, S. J. R., & Lee, L. C. (2004). Metalinguistic Awareness and Semi-Syllabic Scripts: Children's Spelling Errors in Malay. *Reading and Writing,* 17(1-2), pp.17-26. Retrieved 13th March 2006, from SpringerLink Online

MBROLA. (2005). The MBROLA Project. Retrieved 20th May 2005, from http://tcts.fpms.ac.be/synthesis/mbrola.html.

Merriam-Webster. (2005). Merriam-Webster Collegiate Dictionary and Thesaurus Online. Retrieved 5th March 2005, from www.m-w.com.

Mohamad, M. (1981). *The Malay Dilemma*. Kuala Lumpur: Federal Publications.

Morton, J. B., Trehub, S. E., & Zelazo, P. D. (2003). Sources of Inflexibility in 6-Year-Olds: Understanding of Emotion in Speech. *Child Development,* 74(6), pp.1857-1868. Retrieved 21st July 2005, from http://pdfserve.galegroup.com/pdfserve/get_item/1/Se97087w1_4/SB695_04.pdf

Mozziconacci, S. (2002). Prosody and Emotions, *Speech Prosody 2002*. Aix-en-Provence, France: ISCA.

Mullenix, J. W., *et al.* (2002). Effects of Variation in Emotional Tone of Voice on Speech Perception. *Language and Speech, 45*(3), 255-283.

Murray, I. R., & Arnott, J. L. (1993). Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal Acoustical Society of America, 93*(2), 1097-1108.

Murray, I. R., & Arnott, J. L. (1996, October 3-6th). *Synthesizing Emotions in Speech: Is It Time to Get Excited?* **In** Proc. of 4th International Conference on Spoken Language Processing 1996, (pp.1816-1819).

Murray, I. R., Arnott, J. L., & Rohwer, E. A. (1996). Emotional Stress in Synthethic Speech: Progress and Future Directions. Speech Communication, 20(1-2), pp.85-91. Retrieved 17th March 2005, from Elsevier.com

Nafisah, A. (1999). Romanization of Multiscript/Multilingual Materials: Experiences of Malaysia. *65th IFLA Council and General Conference.* Retrieved 14th June 2006, from http://www.ifla.org/IV/ifla65/papers/150-155e.htm.

Nass, C.*, et al.* (2000). *The Effects of Emotion of Voice in Synthesized and Recorded Speech*.Unpublished manuscript, Standford, CA.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: The MIT Press.

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. NY: Harper & Row.

Razak, A. A., Abidin, M. I. Z., & Komiya, R. (2003a, 21-24 Sept). *Emotion Pitch Variation Analysis in Malay and English Voice Samples.* **In** Proc. of The 9th Asia-Pacific Conference on Communications 2003, (pp.108-112).

Razak, A. A., Yusof, M. H. M., & Komiya, R. (2003b). *Towards Automatic Recognition of Emotion in Speech.* **In** Proc. of the 3rd IEEE International Symposium on Signal Processing and Information Technology, (pp.548-551).

Russell, J. A. (1980). A Circumflex Model of Affect. *Journal of Personality and Social Psychology, 39*, 1161-1178.

Russell, J. A. (1991). In Defense of a Prototype Approach to Emotion Concepts. *Journal of Personality and Social Psychology,* 60(1), pp.37-47. Retrieved 4th July 2005, from http://www.ocf.berkeley.edu/~june/Research%20-%20Russell%20(1991).pdf

Scherer, K. R. (1978). Personality Inference from Voice Quality: The Loud Voice of Extroversion. *European Journal of Social Psychology, 8*, 467-487.

Scherer, K. R. (2000). *Emotion Effects on Voice and Speech: Paradigms and Approaches to Evaluation [PowerPoint Presentation].* **In** Proc. of ISCA Workshop on Speech and Emotion, Belfast.

SFS. (2005). Speech Filing System (Version 4.6). London, UK: Department of Phonetics and Linguistics, University College London (UCL).

Shaver, P.*, et al.* (1987). Emotion Knowledge: Further Exploration of a Prototype Approach, *Journal of Personality and Social Psychology*, pp.1061-1086. Retrieved 3rd July 2005, from APA Online

Shih, C., & Kochansky, G. (2002). Prosody and Prosodic Models, *ICSLP Conference, Denver, CO*.

Shroder, M., & Grice, M. (2003). *Expressing Vocal Effort in Concatenative Synthesis.* **In** Proc. of 15th ICPhS, Barcelona.

Silzer, P. J. (2001, Oct 31st-Nov 3rd). *Miffed, Upset, Angry or Furious? Translating Emotion Words.* **In** Proc. of ATA 42nd Annual Conference, Lost Angeles, CA, (pp.1-6).

Springer ,M. (2002). Becoming a "Wiz" at Brain-based Teaching: How to Make Every Year Your Best Year (2nd ed.). California: Corwin Press.

Sundaram, S., & Narayanan, S. (2003). *An Empirical Text Transformation Method for Spontaneous Speech Synthesizers.* **In** Proc. of Eurospeech, Geneva.

Syaheerah, L. Lutfi.*, et al.* (2005a, 12-15th December). *Template-Driven Emotions Generation in Malay Text-to-Speech: A Preliminary Experiment.* **In** Proc. of 4th International Conference of Information Technology in Asia (CITA 05), Kuching, Sarawak, (pp.144-149).

Syaheerah, L. Lutfi.*, et al.* (2005, 19-21st September). *Adding Emotions to Malay Synthesized Speech Using Diphone-Based Templates.* **In** Proc. of 7th International Conference on Information and Web-based Applications & Services (iiWAS 05), Kuala Lumpur, Malaysia, (pp.269-276).

Tams, A., & Tatham, M. A. A. (2000). Intonation for Synthesis of Speaking Styles, *IEE Seminar "State-Of-The-Art In Speech Synthesis"*. London: IEE.

Tatham, M. A. A., Morton, K., & Lewis, E. (1999). *Assignment of Intonation in a High-Level Speech Synthesizer.* **In** Proc. of the Institute of Acoustics, (pp.255-262).

Tiun, S., & Kong, T. E. (2005). Building a Speech Corpus for Malay TTS System, *National Computer Science Postgraduate Colloquium 2005 (NaCPS'05)*.

Wazir-Jahan, K. (Ed.). (1990). *Emotions of Culture: A Malay Perspective*. NY: Oxford University Press.

Wu, C. H., & Chen, J. H. (1999). *Template-Driven Generation of Prosodic Information for Chinese Concatenate Synthesis*. **In** Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, Phoenix, Arizona, (pp.65-68).

Zahid, I. (2003). *Kajian Intonasi Fonetik Eksperimental:Realisasi Makna Emosi Filem Sembilu 1 Dan 2.* Unpublished PhD. Thesis, University Malaya, Kuala Lumpur.

Zovato, E.*, et al.* (2004). *Towards Emotional Speech Synthesis: A Rule Based Approach.* **In** Proc. of 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA, (pp.219-220).

Zuraidah, M. D. (1996). *Prosody in Malay: An Analysis of Broadcast Interviews*. Unpublished PhD Thesis, University Malaya, Kuala Lumpur

Zuraidah, M. D., Knowles, G., & Yong, J. (2005) How Words Can Be Misleading: A study of Syllable Timing and 'Stress' in Malay: *Submitted to the Journal of Phonetics*

Zuraidah, M. D., & Knowles, G. (2006). Prosody and Turn-taking in Malay Broadcast Interviews. *Journal of Pragmatics*. 38: 490-512.

.