

Abstract

This dissertation aims to add an emotional filter to existing Malay Text-To-Speech (TTS) system, FASIH to produce emotional output thus contributing to the enhancement of naturalness to the existing system. This research focuses on four types of emotions; happiness, anger, sadness and fear. This research uses a new method to develop an emotion filter by adopting rule base approach to manipulate the pitch and duration of standard TTS output to produce the emotional output. To understand human emotions, this research obtained quality human emotions source, verified those sources, analyzed the variation of speech pattern and used these information to develop the emotions filter. The emotional speech output from the modified TTS synthesizer is tested for verification by conducting listening tests and acceptance tests. The results showed that the synthesized speech was emotionally identifiable with a favorable impression. The emotions filter developed worked successfully with FASIH to produce an emotional output. It is concluded that this emotions filter can work with any TTS system that uses prosody (pitch and duration) to synthesize the output for any type of emotions based on the rules provided for each emotions.

Acknowledgments

There are many people that I want to thank in relation to this project. I will not have completed this project without their help. Firstly, I like to express my sincere appreciation to my supervisor, Associate Professor Raja Noor Ainon from University of Malaya. Professor Raja Noor Ainon has given the guidance needed in completing this dissertation. Her invaluable direction, patience and encouragement help me to complete this dissertation. I like to thank my co-supervisor Prof. Dr. Zuraidah M. Don who has given me invaluable advice and expertise that enable me to complete this project successfully. I also like to thank Mr. Kow Weng Onn from MIMOS whom have given me support and valuable advice in completing this project. I am also thankful to Malini whom has expressed her emotions to the success of this project. I am very grateful to students of University Malaya for participating in perceptual experiments. Finally I like to thank my husband, Mansoor Ali , my mother Madam Noorjahan and my brother Shahjahan for their love, support and encouragement.

November , 2006

Mumtaz Begum

Table of Contents

	Page
Chapter 1: Introduction	1
1.1. Research inspiration	1
1.2. Challenges	2
1.3. Research objectives	3
1.4. Research methodology	3
1.5. Overview of the research	4
1.6. The organization of the dissertation	5
1.7. Terminology	6
Chapter 2: Research Background	7
2.0 Overview	7
2.1 Understanding emotions	7
2.1.1 Theories of emotions	8
2.1.2 Classifying emotions	9
2.1.3 Speech and emotions	11
2.1.4 Emotions in writing	12
2.2 Text-to-speech system	14
2.2.1 Text analysis	15
2.2.2 Audio generation (Digital Signal Processing)	16
2.2.2.1 Diphone concatenation synthesizer	17
2.2.2.2 Unit selection synthesizer	18
2.2.2.3 Formant synthesizer	19
2.2.2.4 Other form of synthesizer	19
2.2.3 TTS systems in Malaysia	20
2.2.4 FASIH TTS system	21
Chapter 3: Related Research	23
3.0 Overview	23
3.1 Related research and solutions in adding emotions to TTS	23
3.1.1 Diphone concatenation synthesizer	23
3.1.2 Unit selection synthesis	26
3.1.3 Formant synthesis	27
3.1.4 Summary of existing emotional synthesizers	27
3.2 Studies on human speech corpus	28
3.2.1 Sentences for recording	29
3.2.2 Recording	29
3.2.3 Perception test	30
3.3 Conclusion	31

Chapter 4: Research Methodology	32
4.0 Overview	32
4.1 Research approach	32
4.2 Information gathering	32
4.3 Creating human speech corpus	33
4.3.1 Creating sentences	33
4.3.2 Recording session	34
4.3.3 Perception test of recorded speech corpus	35
4.4 Analyzing recorded human speech corpus	38
4.5 Designing the emotions filter	39
4.5.1 Identification of different emotions	39
4.5.2 Modification of pitch	39
4.5.3 Modification of duration	40
4.6 Testing and review	40
4.6.1 Perception test of the synthesized output	40
4.6.2 Acceptance test	41
Chapter 5: Analysis and Findings	42
5.0 Overview	42
5.1 Analysis of perception test of recorded human speech	42
5.2 PRAAT analysis of recorded sentences	44
5.2.1 Happiness	46
5.2.2 Anger	49
5.2.3 Sadness	50
5.2.4 Fear	50
Chapter 6: Designing Emotions Filter for FASIH TTS System	52
6.0 Overview	52
6.1 Justification of proposed method	52
6.2 Overall view of the proposed emotions filter	52
6.3 Recognizing emotions	53
6.4 The execution process of the emotions filter	54
6.4.1 Identification of happy and angry emotions	55
6.4.2 Modification of duration	55
6.4.3 Modification of pitch	56
Chapter 7: Implementation of Emotions Filter for FASIH TTS System	58
7.0 Overview	58
7.1 The overall implementation of emotions filter	58
7.1.1 Use case diagram	58
7.1.2 Sequence diagram	59
7.1.3 Class diagram	60
7.1.4 System architecture: Pipes and filters architecture	61
7.2 Identifying individual words and exclamation marks	62
7.2.1 The identification of individual words	63
7.2.2 The identification of exclamation mark	63

7.2.3	Adding the emotions buttons to FASIH interface	64
7.3	Commencement of the emotions filter	65
7.3.1	General module	65
7.3.2	PHO file module	66
7.3.3	Setting file module	66
7.3.4	Individual emotions module	67
7.3.5	Syllable module	67
7.4	Generating happy and angry emotions	68
7.5	Generating sad and fear emotions	70
7.6	Connection to FASIH	73
Chapter 8:	Overview of Present Research	74
8.0	Summary of research	74
8.1	The perception test	74
8.2	The acceptance test	76
8.3	Conclusion	77
8.4	Future recommendations	79
Appendix A: Speech Corpus		
Appendix B: Perception and acceptance test forms		
Appendix C: Analysis of speech corpus		
Appendix D: Graphical user interface design		
Appendix E: Predetermined rate for modification stored in txt files.		
Appendix F: List of achievements of this research		

List of Figures

	Page	
2.1	The basic function of TTS system	15
2.2	The overall structure of TTS system	17
2.3	The functional diagram of diphone concatenation synthesizer	18
2.4	The functional diagram of unit selection synthesizer	18
2.5	The functional diagram of formant synthesizer	19
2.6	Top level architecture of FASIH	22
2.7	The functional diagram of FASIH TTS System	22
3.1	The diagrammatic function of Emofilt	25
4.1	The recording interface of PRAAT	35
4.2	Media player used during the perception test	37
5.1	The average recognition rate for both local and foreign listeners	43
5.2	The average effort rate for both local and foreign listeners	44
5.3	The segmentation process for an angry sentence	44
5.4	The segmentation process for a sad sentence	45
6.1	Adding emotions filter to FASIH	53
6.2	Overall execution process of emotions filter	54
6.3	The modification of duration for the sentence 'saya telah menang!'	55
6.4	The detail modification of duration for the word 'saya'	56
6.5	The modification of pitch for the sentence 'saya telah menang!'	57
6.6	The detail modification of pitch for the word 'saya'	57
7.1	Use case diagram for the emotions filter	59
7.2	Sequential diagram for end users	60
7.3	Class diagram for the emotions filter	61
7.4	The system architecture of the emotions filter	62
7.5	The instruction on FASIH to segregate individual words	63
7.6	The instruction on FASIH to identify exclamation mark	63
7.7	The FASIH interface before modification	64
7.8	The FASIH interface after modification	64
7.9	Commencement of emotions filter	65
7.10	Tagging of word, syllable, consonant and vowel	66
7.11	Reading FASIH standard output	66
7.12	Read modified PHO file	66
7.13	Reading parameter from txt files	67
7.14	Identification of syllable	68
7.15	The process of modification	68
7.16	Determining the type of happy emotions	69
7.17	The tagging of first syllable for happy module	69
7.18	Obtaining the rate factor for happy module	70
7.19	The duration modification for happy	70
7.20	The pitch modification for happy	70
7.21	Identification of vowel for sad module	71
7.22	The logic flow of emotions module	72
7.23	The lines of command connecting FASIH and the emotions filter	73

8.1	The average recognition rate for the synthesized output	76
8.2	The acceptance test results	77

Appendix D: Graphical user interface design

Figure D1: The FASIH interface before modification

Figure D2: The FASIH interface after modification

Appendix E: Predetermined Rate for Modification stored in txt files.

Figure E1: Txt file for happy sentence with 1 exclamation mark

Figure E2: Txt file for happy sentence with 2 exclamation marks

Figure E3: Txt file for angry sentence with 1 exclamation mark

Figure E4: Txt file for angry sentence with 2 exclamation marks

Figure E5: Txt file for sad sentence

Figure E6: Txt file for fear sentence

List of Tables

	Page	
1.1	Potential users that benefit from the emotions TTS system	2
1.2	Terminology of this research	6
2.1	The effect of emotions on human voice (Stallo, 2000)	12
2.2	Few locally produced TTS system	20
3.1	The overall function of Emofilt	25
3.2	Summary of potential solution from previous research	27
4.1	Sample sentences used for recording purposes	34
4.2	Listeners of perception test one	36
4.3	Listeners of perception test two	41
4.4	Breakdown of participant for acceptance test	41
5.1	Recognition of emotion from perception test one	43
5.2	PRAAT analysis for the recorded sentences	46
5.3	Comparing standard FASIH output with recorded speech	46
5.4	Average pitch and duration differences for vowel	47
5.5	Duration analysis for happy emotion	48
5.6	Pitch analysis for happy emotion	48
5.7	Duration analysis for angry emotion	49
5.8	Pitch analysis for angry emotion	49
5.9	Duration analysis for sad emotion	50
5.10	Pitch analysis for sad emotion	50
5.11	Duration analysis for fear emotion	51
5.12	Pitch analysis for fear emotion	51
7.1	System configuration	58
8.1	Recognition of emotions from perception test 2	75
8.2	Ethnic breakdown of participant for acceptance test	76
8.3	Profession breakdown of participant for acceptance test	76

Appendix A: Speech Corpus

Table A 1: Sentences for speech recording session (Happy)

Table A 2: Sentences for recording session (Angry)

Table A 3: Sentences for recording session (Sad)

Table A 4: Sentences for recording session (Fear)

Table A 5: Analysis of Malaysian Novels and Story books

Appendix C: Analysis of speech corpus

Table C 1: Analysis of perception test for happy sentences

Table C 2: Analysis of perception test for fear sentences

Table C 3: Analysis of perception test for angry sentences

Table C 4: Analysis of perception test for sad sentences

List of Abbreviations

1. **DSP: Digital Signal Processing**
2. **NLP: Natural Language Processing**
3. **TTS: Text-To-Speech System**
4. **NLP: Natural Language Processing**
5. **DSP: Digital speech processing**
6. **PAT: Parametric Artificial Talker**
7. **LPC: Linear Prediction Coding**
8. **HMM: Hidden Markov Models**
9. **HAMLET: Helpful Automated Machine for Language and Emotional Talk**
10. **CVC: Consonant Vowel Consonant**
11. **CV: Consonant Vowel**