

IMPLEMENTATION OF SIMULATED ANNEALING IN UNIT SELECTION  
FOR MALAY TEXT-TO-SPEECH SYSTEM

LIM YEE CHEA

A dissertation submitted in fulfillment of the  
requirements for the award of the degree of  
Master of Science (Mathematics)

Faculty of Science  
Universiti Teknologi Malaysia

NOVEMBER 2009

Dedicated to Jesus Christ,  
my personal Lord and Savior,  
my pastor, Church members,  
my beloved mum, dad, brother and sister.

## ACKNOWLEDGEMENTS

“Let us then with confidence draw near to the throne of grace, that we may receive mercy and find grace to help in time of need.” First and foremost, I want to thank Jesus for His grace and mercy throughout this project. It is by His hand and wisdom in guiding me to finish my work.

I would like to extend my appreciation to my honorable supervisor, Dr. Zaitul Marlizawati Zainuddin and my co-supervisor, Dr. Tan Tian Swee, for their academic guidance, suggestions, support and encouragement shown during the course of my study. The patience, tolerance, diligence and dedication shown to me have given me great encouragement and a good example to follow after.

Finally, I would love to convey my gratitude to my beloved family members and church members for their love and care shown to me along the process of the study. They have given me so much assistance, comfort and prayer support, either financially or spiritually, of which words could not express and will forever be remembered in my heart. Here I want to especially appreciate Mohd Redzuan bin Jamaludin, his willingness and guidance in doing Matlab.

## ABSTRACT

Unit selection method has become the predominant approach in speech synthesis. The quality of unit selection based concatenative speech synthesis primarily governed by how well two successive units can be joined together. Therefore, the main purpose of unit selection is to minimize the audible discontinuities. The process of unit selection is based on phonetic context and Simulated Annealing that selects units from large database with the minimization of a criterion, which is often called cost. This dissertation presents a variable-length unit selection Malay text to speech system that is capable of providing more natural and accurate unit selection for synthesized speech. To provide the capability of selecting a speech unit not only limited to phoneme, diphone or triphone but also a string of phonemes that can be matched directly to the database, unit selection methods have been implemented. The Mel Frequency Cepstral Coefficients (MFCC) as spectral parameters have been introduced in the unit selection based speech synthesis. Distance measurement is needed to measure the difference between two vectors of this speech feature. The spectral distance used is Euclidean Distance.

## ABSTRAK

Kaedah pilihan unit telah menjadi cara utama dalam sintesis pertuturan. Kualiti untuk pilihan unit dalam penyambungan perkataan adalah berpandukan kepada betapa baiknya kedua-dua unit menyambung bersama. Oleh itu, matlamat utama dalam pilihan unit adalah untuk mengurangkan komposisi jarak. Process untuk pilihan unit adalah bergantung pada konteks fonetik dan Simulated Annealing yang memilih unit dari database dengan meminimumkan satu criteria, yang selalunya dipanggil kos. Disertasi ini melaksanakan satu pemilihan unit berlainan panjang yang mampu memberikan pemilihan unit yang lebih tepat dan semulajadi untuk pertuturan sintesis. Untuk mengadakan pemilihan pertuturan unit yang berupaya bukan hanya terhad kepada foneme, dua fonem atau tiga fonem tetapi juga satu rangkaian fonem yang boleh terusdipadankan kepada pangkalan data, kaedah pemilihan unit telah dilaksanakan. Mel Frequency Cepstral Coefficients (MFCC) sebagai spektra parameter telah diperkenalkan dalam pemilihan unit pertuturan sintesis. Pengiraan jarak adalah diperlukan untuk mengira jarak antara dua vector ini. Spectra jarak yang digunakan adalah Jarak Euclidean.

## TABLE OF CONTENT

| <b>CHAPTER</b> | <b>TITLE</b>                  | <b>PAGE</b> |
|----------------|-------------------------------|-------------|
|                | <b>TITLE PAGE</b>             | i           |
|                | <b>DECLARATION PAGE</b>       | ii          |
|                | <b>DEDICATION</b>             | iii         |
|                | <b>ACKNOWLEDGEMENT</b>        | iv          |
|                | <b>ABSTRACT</b>               | v           |
|                | <b>ABSTRAK</b>                | vi          |
|                | <b>TABLE OF CONTENTS</b>      | vii         |
|                | <b>LIST OF TABLES</b>         | xi          |
|                | <b>LIST OF FIGURES</b>        | xiii        |
|                | <b>LIST OF SYMBOLS</b>        | xvi         |
|                | <b>LIST OF APPENDICES</b>     | xviii       |
| <br>           |                               |             |
| <b>1</b>       | <b>INTRODUCTION</b>           | <b>1</b>    |
|                | 1.0 Introduction              | 1           |
|                | 1.1 Background of the Problem | 2           |
|                | 1.2 Problem Statement         | 3           |
|                | 1.3 Objective of the Study    | 3           |
|                | 1.4 Scopes of the Study       | 3           |
|                | 1.5 Significance of the Study | 4           |
|                | 1.6 Research Methodology      | 4           |
|                | 1.7 Dissertation Layout       | 5           |

|          |   |          |
|----------|---|----------|
| <b>2</b> | <b>LITERATURE REVIEW</b>                        | <b>5</b> |
| 2.1      | Speech synthesis                                | 6        |
| 2.1.1    | Concatenative Speech Synthesis                  | 7        |
| 2.2      | Unit Selection                                  | 9        |
| 2.2.1    | Non-Uniformed or Variable Length Unit Selection | 11       |
| 2.2.2    | Corpus-based Unit Selection                     | 12       |
| 2.3      | Cost function for unit selection                | 14       |
| 2.3.1    | The Acoustic Parameters                         | 16       |
| 2.3.2    | Linguistic Features                             | 16       |
| 2.3.3    | Local cost                                      | 17       |
| 2.3.3.1  | Sub-cost on prosody                             | 19       |
| 2.3.3.2  | Sub-cost on discontinuity                       | 20       |
| 2.3.3.3  | Sub-cost on phonetic environment                | 20       |
| 2.3.3.4  | Sub-cost on spectral discontinuity              | 21       |
| 2.3.3.5  | Sub-cost on phonetic appropriateness            | 22       |
| 2.3.3.6  | Other sub-costs                                 | 23       |
| 2.3.3.7  | Integrated cost                                 | 23       |
| 2.4      | Cost weighting                                  | 24       |
| 2.5      | Target cost                                     | 25       |
| 2.6      | Concatenation cost                              | 26       |
| 2.7      | Spectral Distances                              | 29       |
| 2.8      | Feature Extraction                              | 30       |
| 2.8.1    | MFCC  | 30       |
| 2.9      | Distance Measures                               | 32       |
| 2.9.1    | Simple Distance Measures                        | 33       |
| 2.9.1.1  | Absolute Distance                               | 33       |
| 2.9.1.2  | Euclidean Distance                              | 34       |
| 2.9.2    | Statistically Motivated Distance Measures       | 34       |
| 2.9.2.1  | Mahalanobis Distance                            | 34       |
| 2.9.2.2  | Kullback–Leibler (KL) distance                  | 35       |

|          |  |           |
|----------|--|-----------|
| 2.10     | Heuristic Method   | 36        |
|          | 2.10.1 Simulated Annealing                               | 37        |
|          | 2.10.2 Approaches to improve SA algorithm                | 39        |
|          | 2.10.3 Polynomial approximation                          | 40        |
|          | 2.10.4 Annealing Schedule                                | 41        |
|          | 2.10.4.1 Theoretically optimum cooling schedule          | 41        |
|          | 2.10.4.2 Geometric cooling schedule                      | 42        |
|          | 2.10.4.3 Cooling schedule of Van Laarhoven <i>et al.</i> | 42        |
|          | 2.10.4.4 Cooling schedule of Otten <i>et al.</i>         | 43        |
|          | 2.10.4.5 Cooling schedule of Huang <i>et al.</i>         | 43        |
|          | 2.10.4.6 Adaptive cooling schedules                      | 44        |
|          | 2.10.4.7 A new adaptive cooling schedule                 | 44        |
| 2.11     | Parallel SA  | 46        |
| 2.12     | Segmented Simulated Annealing                            | 47        |
| <b>3</b> | <b>PROPOSED SYSTEM AND IMPLEMENTATION</b>                | <b>49</b> |
| 3.0      | Introduction   | 49        |
| 3.1      | System Design Flow                                       | 50        |
| 3.2      | Malay Phonetics and Phone Sets                           | 51        |
| 3.3      | Malay Phoneme  | 51        |
|          | 3.3.1 Malay Vowels                                       | 51        |
|          | 3.3.2 Malay Consonant                                    | 51        |
| 3.4      | Phoneme Units Database                                   | 52        |
| 3.5      | Feature Extraction                                       | 55        |
| 3.6      | Phonetic context   | 58        |
| 3.7      | Unit Selection   | 59        |
| 3.8      | Concatenation  | 60        |
| <b>4</b> | <b>SIMULATED ANNEALING</b>                               | <b>63</b> |
| 4.0      | Introduction   | 63        |
| 4.1      | Procedure of Simulated Annealing                         | 65        |
| 4.2      | Initial Solution   | 67        |
| 4.3      | The cooling schedule                                     | 67        |
|          | 4.3.1 Markov chain                                       | 70        |

|          |  |                 |
|----------|--|-----------------|
| 4.4      | Neighbourhood Generation Mechanism         | 70              |
| 4.5      | Metropolis's criterion                     | 80              |
| 4.6      | Stopping criteria                          | 82              |
| 4.7      | Unit Selection                             | 82              |
| 4.7.1    | Phonetic context                           | 82              |
| 4.7.2    | Concatenation Cost                         | 88              |
| 4.7.2.1  | Concatenation cost for Move 1              | 89              |
| 4.7.2.2  | Concatenation cost for Move 2              | 90              |
| 4.7.2.3  | Concatenation cost for Move 3              | 90              |
| 4.7.2.4  | Concatenation cost for Move 4              | 91              |
| 4.8      | Concatenation                              | 100             |
| <b>5</b> | <b>TESTING, ANALYSIS AND RESULT</b>        | <b>107</b>      |
| 5.1      | Experiment                                 | 107             |
| 5.2      | Test Materials                             | 107             |
| 5.3      | Test Conditions                            | 107             |
| 5.4      | Test Procedure                             | 108             |
| 5.5      | Profiles of Listeners                      | 109             |
| 5.5.1    | Percentages of Listeners by Gender         | 110             |
| 5.5.2    | Percentage of Listeners by Race            | 111             |
| 5.5.3    | Percentage of Listeners by State of Origin | 112             |
| 5.6      | Result and Analysis                        | 113             |
| 5.6.1    | Word Level Testing                         | 113             |
| 5.6.2    | Mean Opinion Score                         | 114             |
| <b>6</b> | <b>CONCLUSION AND RECOMMENDATION</b>       | <b>117</b>      |
| 6.1      | Conclusion                                 | 117             |
| 6.2      | Suggestion for Future Work                 | 120             |
|          | <b>REFERENCES</b>                          | <b>121</b>      |
|          | <b>APPENDICES A-E</b>                      | <b>131 -163</b> |

## LIST OF TABLES

| <b>TABLE NO.</b> | <b>TITLE</b>  | <b>PAGE</b> |
|------------------|---|-------------|
| <b>2.1</b>       | Sub-cost functions  | 17          |
| <b>3.1</b>       | Total units after extracting the phoneme units from the carrier sentences   | 54          |
| <b>4.1</b>       | Maximum number of iterations for Markov Chain length 1<br>and 2 to reach final temperature greater than 0.1.            | 70          |
| <b>4.2</b>       | The information of the 10 words before filter using phonetic context.   | 86          |
| <b>4.3</b>       | The information of the 10 words after filter using partially matched<br>phonetic context (left phonetic context).       | 87          |
| <b>4.4</b>       | The information of the 10 words after filter using fully matched phonetic<br>context (left and right phonetic context). | 88          |
| <b>4.5</b>       | Information of concatenation cost (Move 1) with temperature<br>reduction rate, $\alpha = 0.90$                          | 89          |
| <b>4.6</b>       | Information of concatenation cost (Move 2) with temperature<br>reduction rate, $\alpha = 0.90$                          | 90          |
| <b>4.7</b>       | Information of concatenation cost (Move 3) with temperature<br>reduction rate, $\alpha = 0.90$                          | 90          |
| <b>4.8</b>       | Information of concatenation cost (Move 4) with temperature<br>reduction rate, $\alpha = 0.90$                          | 91          |
| <b>4.9</b>       | Information of concatenation cost with temperature<br>reduction rate, $\alpha = 0.95$                                   | 92          |
| <b>4.10</b>      | Information of concatenation cost with temperature<br>reduction rate, $\alpha = 0.85$                                   | 93          |
| <b>4.11</b>      | Information of concatenation cost with temperature<br>reduction rate, $\alpha = 0.80$                                   | 94          |

|             |  |     |
|-------------|--|-----|
| <b>4.12</b> | Information of concatenation cost with temperature reduction rate, $\alpha = 0.95$ | 95  |
| <b>4.13</b> | Information of concatenation cost with temperature reduction rate, $\alpha = 0.90$ | 96  |
| <b>4.14</b> | Information of concatenation cost with temperature reduction rate, $\alpha = 0.85$ | 97  |
| <b>4.15</b> | Information of concatenation cost with temperature reduction rate, $\alpha = 0.80$ | 98  |
| <b>4.16</b> | The sequences of the 10 selected words.  | 100 |
| <b>5.1</b>  | Profiles of Listeners  | 109 |
| <b>5.2</b>  | Words selected for listening test.   | 113 |
| <b>5.3</b>  | The score line of synthesis words with considers the concatenation cost.           | 115 |
| <b>5.4</b>  | The score line of the 10 synthesis words with considers the concatenation cost.    | 116 |

## LIST OF FIGURE

| FIGURE NO. | TITLE  | PAGE |
|------------|--|------|
| 2.1        | Classes of waveform synthesis methods for speech synthesis.  | 7    |
| 2.2        | Viterbi search.  | 8    |
| 2.3        | Architecture of corpus-based unit selection concatenative<br>speech synthesizer.   | 13   |
| 2.4        | Schematic diagram of cost function   | 15   |
| 2.5        | Example of unit search algorithm. The shortest path is marked in blue.   | 28   |
| 2.6        | Example of unit search algorithm. The difference in cost between the<br>optimal sequences of two graphs is evaluated for $d_3$ in pre-selection. | 29   |
| 2.7        | Objective Spectral distances   | 30   |
| 2.8        | Block diagram of the conventional MFCC extraction algorithm  | 31   |
| 2.9        | Parallel Simulated Annealing Taxonomy  | 46   |
| 2.10       | Segmented simulated annealing  | 48   |
| 3.1        | Block Diagram of System Design Flow.   | 50   |
| 3.2        | A set of coefficient transform from MFCC algorithm.  | 53   |
| 3.3        | Speech unit database.  | 54   |
| 3.4        | The GUI to extract MFCCs coefficients.   | 55   |
| 3.5        | The GUI to extract MFCCs coefficients.   | 56   |
| 3.6        | The 12 coefficients extracted for phoneme “_m”.  | 56   |
| 3.7        | The 12 coefficients extracted for phoneme “a”.   | 57   |
| 3.8        | Distance measure and speech feature.   | 57   |
| 3.9        | The candidate unit for phoneme “_n” that matched right phonetic context.   | 58   |
| 3.10       | The candidate unit for phoneme “a” that matched left and right phonetic<br>context.  | 59   |
| 3.11       | Unit selection   | 60   |

|             |  |     |
|-------------|--|-----|
| <b>3.12</b> | Waveform for phoneme “_n”.   | 61  |
| <b>3.13</b> | Waveform for phoneme “a”.  | 61  |
| <b>3.14</b> | Waveform for phoneme “s”.  | 61  |
| <b>3.15</b> | Waveform for phoneme “i”.  | 62  |
| <b>3.16</b> | Concatenation of the best matching units for the word “nasi”.  | 62  |
| <b>4.1</b>  | SA flow diagram to find best speech unit sequence.   | 66  |
| <b>4.2</b>  | Temperature reduction pattern for various reduction rates with Markov Chain length 1.  | 69  |
| <b>4.3</b>  | Temperature reduction pattern for various reduction rate with Markov Chain length 2.   | 69  |
| <b>4.4</b>  | Metropolis criterion   | 81  |
| <b>4.5</b>  | The feasible search region to form a Malay word “kampung” before filter using phonetic context.  | 84  |
| <b>4.6</b>  | The feasible search region to form a Malay word “kampung” after filter using partially matched phonetic context (left phonetic context).       | 85  |
| <b>4.7</b>  | The feasible search region to form a Malay word “kampung” after filter using fully matched phonetic context (left and right phonetic context). | 85  |
| <b>4.8</b>  | SA best solutions, mean and worst solutions for ten problems from Table 4.12.  | 99  |
| <b>4.9</b>  | Waveform “_s1”.  | 101 |
| <b>4.10</b> | Waveform “e537”  | 101 |
| <b>4.11</b> | Waveform “l362”  | 101 |
| <b>4.12</b> | Waveform “a2710”   | 101 |
| <b>4.13</b> | Waveform “n1031”   | 102 |
| <b>4.14</b> | Waveform “j7”  | 102 |
| <b>4.15</b> | Waveform “u206”  | 102 |
| <b>4.16</b> | Waveform “t142”  | 102 |
| <b>4.17</b> | Waveform “ny1”   | 103 |
| <b>4.18</b> | Waveform “a2060”   | 103 |
| <b>4.19</b> | Concatenation waveform for the word “selanjutnya”.   | 103 |
| <b>4.20</b> | Spectrogram for the word “nasi”.   | 104 |
| <b>4.21</b> | Spectrogram for the word “berpengetahuan”.   | 104 |
| <b>4.22</b> | Spectrogram for the word “demikian”.   | 105 |

|             |  |     |
|-------------|--|-----|
| <b>4.23</b> | Spectrogram for the word “demikian” that do not consider concatenation cost. | 105 |
| <b>4.24</b> | Spectrogram zoom in for the word “demikian” from Figure 4.22.                | 106 |
| <b>4.25</b> | Spectrogram zoom in for the word “demikian” from Figure 4.23.                | 106 |
| <b>5.1</b>  | Percentage of listeners by gender.   | 110 |
| <b>5.2</b>  | Percentage of listeners by race.   | 111 |
| <b>5.3</b>  | Percentage of listeners by state of origin.                                  | 112 |
| <b>5.4</b>  | Level of intelligibility of the 10 selected words.                           | 114 |
| <b>5.5</b>  | Results of the mean opinion score.   | 115 |

## LIST OF SYMBOLS/ ABBREVIATIONS

|                        |  |
|------------------------|--|
| $AC$                   | Average cost   |
| $k_b$                  | Boltzmann constant                                   |
| $S$                    | Configuration set                                    |
| $C$                    | Cost function  |
| $E$                    | Energy   |
| $C_{\max}$             | Estimation of the maximum value of the cost function |
| $\langle f(T) \rangle$ | Expected cost in equilibrium                         |
| FFT                    | Fast Fourier Transform                               |
| $F_0$                  | Fundamental Frequency                                |
| GUI                    | Graphical User Interface                             |
| KL                     | Kullback-Leibler                                     |
| LSF                    | Line spectral frequencies                            |
| LP                     | Linear prediction                                    |
| LPC                    | Linear Predictive Coefficients                       |
| $LC$                   | Local cost   |
| $MC$                   | Maximum cost   |
| MOS                    | Mean Opinion Score                                   |
| $MCD$                  | Mel-cepstral distortion                              |
| MFCCs                  | Mel-Frequency Cepstral Coefficients                  |
| $Mel(f)$               | Mel scale  |
| MCA                    | Multiple centroid analysis                           |
| $NC_p$                 | Norm cost  |
| $N$                    | Neighbourhood structure                              |
| PLP                    | Perceptual linear prediction                         |

|               |                                      |
|---------------|--------------------------------------|
| $P(E)$        | Probabilities of acceptance          |
| $\delta$      | Real number                          |
| $C_{pro}$     | Sub-cost on prosody                  |
| $C_{F_0}$     | Sub-cost on $F_0$ discontinuity      |
| $C_{env}$     | Sub-cost on phonetic environment     |
| $C_{spec}$    | Sub-cost on spectral discontinuity   |
| $C_{app}$     | Sub-cost on phonetic appropriateness |
| $T$           | Temperature                          |
| $\alpha$      | Temperature reduction rate           |
| TTS           | Text-to-speech                       |
| TSP           | Travelling salesman problem          |
| $U$           | Upper bound                          |
| $\sigma^2(T)$ | Variance in the cost at equilibrium  |

**LIST OF APPENDICES**

| <b>APPENDIX</b> | <b>TITLE</b>                                | <b>PAGE</b> |
|-----------------|---|-------------|
| A               | Source Code of MFCC                         | 131         |
| B               | Source Code of Simulated Annealing (Move 1) | 137         |
| C               | Source Code of Simulated Annealing (Move 2) | 145         |
| D               | Source Code of Simulated Annealing (Move 3) | 153         |
| E               | Evaluation Questionnaire                    | 161         |

## CHAPTER 1

### INTRODUCTION

#### 1.0 Introduction

Corpus-based concatenative synthesis has become the major trend recently because the resulted speech sounds more natural than that produced by parameter-driven production models (Chou, 1999). Unit selection synthesizers in the current state produce highly intelligible, near natural synthetic speech (Tsiakoulis *et al.*, 2008). This method creates speech by re-sequencing pre-recorded speech units selected from a very large speech database (Cepko *et al.*, 2008). Speech is produced by searching through large speech database (corpus) and concatenating selected units, thus forming the output signal. This approach shows its superiority over formant and articulatory synthesis, because it tends to concatenate natural acoustic units with no modification. Thus, offering better speech quality (Janicki *et al.*, 2008). Text to speech synthetic is produced by concatenating speech unit from a very large speech corpus containing enough prosodic and spectral varieties for all synthetic units (Vepa *et al.*, 2002; Vepa and King, 2004). Hence, it is possible to synthesize highly natural-sounding speech by selecting an appropriate sequence of units (Vepa *et al.*, 2002). The selection of the best unit sequence from the database can be treated as a search problem which has the lowest overall distance. Since the quality of the resulting synthetic speech will depend to a large extent on the variability and availability of representative units, therefore, it is crucial to design a corpus that covers all speech units and most of their variations in a feasible size (Min *et al.*, 2001). The unit selection process is based on the cost function that consists of target cost and join cost. The join cost is measurement of the acoustic smoothness between

the concatenated units (Dong and Li, 2008). This dissertation will focus on concatenation costs which generally use a distance measure on a parameterization of the speech signal. MFCCs are chosen as spectral parameters as they are most commonly used in state-of-the-art recognizers (Rabiner and Juang, 1993). Distance measurement is needed to measure the difference between two vectors of this speech feature. The spectral distance used is Euclidean Distance. Mel Frequency Cepstral Coefficients were derived using standard methods commonly used in speech recognition. MFCCs are representative of the real cepstrum for a windowed short time signal derived from the Fast Fourier Transform (FFT) of the speech signal (Wei *et al.*, 2006).

## 1.1 Background of the Problem

The main problem with the existing Malay text-to-speech (TTS) synthesis system is the poor quality of the generated speech sound. This poor quality is caused by the inability of traditional TTS system to provide multiple choices of unit for generating an accurate synthesized speech (Tan and Sheikh, 2008b). Most of the current Malay TTS systems are utilizing diphone concatenation that only supports a single unit for each existing diphone, the selection of speech unit for concatenation may not be accurate enough (Tan and Sheikh, 2008b). The current trend in high quality text-to-speech systems (TTS) is to concatenate acoustic units selected from large-scaled corpus of continuous read speech. Thus, a robust unit selection is needed to handle the huge volume of data in the database (Blouin *et al.*, 2002). There exist artifacts such as phase mismatches and discontinuities in spectral shape since units are extracted from disjoint phonetic contexts which can have a deleterious effect on perception (Hunt and Black, 1996). It is nominally cast as a multivariate optimization task, where the available unit inventory is searched for the “best” sequence of units which makes up the target utterance. This optimization relies on suitable cost criteria to characterize relevant aspects of acoustic and prosodic context (Bellegarda, 2008).

## **1.2 Problem Statement**

The task of the research is to use Simulated Annealing to find the minimum path for the speech units.

## **1.3 Objective of the Study**

The dissertation aims to achieve the three objectives outlined in this section

- i) To implement Mel Frequency Cepstral Coefficients (MFCCs) in unit selection.
- ii) To implement heuristic optimization method in unit selection.
- iii) To evaluate the performance of the heuristic optimization method in unit selection.

## **1.4 Scopes of the Study**

This dissertation presents a variable-length unit selection scheme to select text-to-speech (TTS) synthesis units from phoneme based corpus which supporting phoneme pattern in Malay Text to Speech. Speech feature selected are MFCCs. Spectral distance used is Euclidean distance. Heuristic methods namely Simulated Annealing is implemented in unit selection to select the best sequence of unit.

## 1.5 Significance of the Study

For Malay TTS system, this is the first version of implementation of unit selection using heuristic method which is Simulated Annealing. The performance of this kind of algorithm and methods will be evaluated based on values of cost functions obtained and listening test. By doing so, the advantages and disadvantages of this method will be known if compared to other existing unit selection methods.

## 1.6 Research Methodology

The variable length unit selection is capable of providing more natural and accurate unit selection for synthesized speech and has been implemented in Malay text to speech system in this project (Tan and Sheikh, 2008b). During synthesis, proper units are selected by searching the closest database units to the symbolic target sequence using the Simulated Annealing. The number of possible units at a given time can number in the tens of thousands if a database is built from a 100-hour corpus (Nishizawa and Kawai, 2006). Therefore, heuristic optimization method is needed to select the appropriate units without having to go through all possible combination of units sequences. The C++ programming codes for Simulated Annealing was developed. To make the acoustic distortion measures correspond to human perception more consistently, the Mel Frequency Cepstral Coefficients (MFCC) as spectral parameters have been introduced in the unit selection based speech synthesis (John *et al.*, 1993). Distance measurement is needed to measure the difference between two vectors of this speech feature. The spectral distance used is Euclidean Distance. The smaller the magnitude in Euclidean Distance means closer the concatenation point and thus generated better speech sound. The performance of the heuristic method and other unit selection method were evaluated based on values of cost functions obtained and listening test.

## **1.7 Dissertation Layout**

This dissertation is divided into six major parts. Chapter 1 includes introduction, background, objective and scope of the thesis. The purpose is to show how this research is different from other conventional method.

Chapter 2 provides the comprehensive study in various unit selection methods. The focus will be on the cost function for unit selection, speech features and spectral distance. It will also include a discussion for Simulated Annealing (SA) with the purpose of laying a foundation for the possible approach to improve the performance of SA.

Chapter 3 describes on the proposed system and implementation. It will discuss the process involved in generating the waveform for synthesis word from contextual linguistic, selection of speech units, concatenation and output sound.

Chapter 4 describes the procedure for SA. It will also describe the procedure in unit selection from contextual linguistic, SA to concatenation. Various parameter setting and neighbourhood generation mechanism for SA will be used to investigate the performance of SA.

Chapter 5 is listening test for the synthesis words based on result in Chapter 4. The purpose is to justify the contribution of concatenation cost in improving the speech quality.

Chapter 6 provides the conclusion for the system. It will also give some recommendation for further improvement of the system.