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 raingkaian 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

LIST OF TABLES

LIST OF FIGURE

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
δ	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
α	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

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.