A PROTOTYPE OF THAI TEXT-TO-SPEECH SYNTHESIS BASED ON LINEAR PREDICTIVE CODING (LPC) METHOD

SUTTISUN SUTHAD NA AYUDTHYA

A THESIS SUBMITTED IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2001

ISBN 974-040-743-9

4037532 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc (COMPUTER)
KEYWORDS : TEXT-TO-SPEECH / SPEECH SYNTHESIS / LPC
SUTTISUN SUTHAD NA AYUDTHYA : A PROTYPE OF THAI TEXT-TO-SPEECH SYNTHESIS BASED ON LINEAR PREDICTIVE CODING METHOD. THESIS ADVISORS : ASSOC. PROF. SUPACHAI TANG WONGSAN, Ph.D., ASST. PROF. CHOMTIP PORN PANOMCHAI, Ph.D. 79 p. ISBN 974-040-743-9

This research project was aimed to develop a prototype of a Thai Text-to-Speech System, which could synthesize speech in Thai from a text by mean of computers. The voice of the synthesized speech should be correlated to Thai reading principles and be similar to the human voice.

The newly developed prototype consisted of two major systems: text processing and signal processing. The text processing system decomposed the input text into phonetic codes, which were further processed by the signal processing system. The signal processing system was composed of two processes: mid-tone syllable synthesis and tone transformation. Initially, mid-tone syllables were synthesized by concatenating of speech units recorded in the form of semi-syllables, which consist of initial part of 288 units and final part of 243 units. By using the concatenation method, 7,776 mid-tone syllables (32 first consonant sounds * 27 vowel sounds * 9 final consonant sounds) could be synthesized. Afterwards, the tone transformation process was done by modifying pitch values of the mid-tone syllable, which were computed by the Autocorrelation method. Then the modified pitch values were used to synthesize speech signal using the LPC (Linear Predictive Coding) method. By modifying the pitch values, the frequency of the speech signal was changed. As the result, a mid-tone syllable could be transformed to the low, falling, high and rising-tone syllable. Totally, 38,888 syllables (7,776 mid-tone syllables * 5 tones) could be synthesized.

In the experimental stage, the system produced speech signal of three types of text data: 1) the mid-tone syllables, 2) the tone transformed syllables, and 3) the sample meaningful sentences. The speech quality was evaluated using the Mean Opinion Score (MOS) method. The experimental result of the mid-tone syllables group is acceptable at 100%. Nevertheless, the result of the tone transformed syllables group is acceptable 61.25%, poor 9.75% and unacceptable 29.00%. The result of the sample meaningful sentences group included four evaluation aspects: pronunciation, distinctness, naturalness, and intelligibility, with scores of 2.58, 2.66 2.40 and 2.89 respectively. Therefore, it could be concluded that the present prototype had demonstrated the ability of speech synthesizing of Thai texts up to quite a satisfactory level.

4037532 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

ศุทธิสันต์ สุทัศน์ ณ อยุธยา : ต้นแบบการสังเคราะห์เสียงพูดภาษาไทย (A PROTOTYPE OF THAI TEXT-TO-SPEECH SYNTHESIS BASED ON LINEAR PREDICTIVE CODING METHOD). คณะกรรมการควบคุมวิทยานิพนธ์ ศุภชัยตั้งวงศ์สานต์, Ph.D., ชมทิพย์ ทรพนมชัย, Ph.D. 79 หน้า ISBN 974-040-743-9

ในงานวิจัยนี้ได้พัฒนาระบบต้นแบบ Thai Text-to-Speech คือระบบที่สามารถสังเคราะห์เสียงภาษาไทยจากข้อความที่เก็บในรูปแบบของ Text ในระบบคอมพิวเตอร์ โดยเสียงที่สังเคราะห์ได้นั้นควรที่จะถูกหลักการอ่านออกเสียงในภาษาไทย และมีลักษณะคล้ายเสียงพูดของมนุษย์ผู้เป็นเจ้าของภาษา

สำหรับระบบที่ได้พัฒนาขึ้นนี้ประกอบไปด้วยสองส่วนหลัก คือ Text processing และ Signal processing โดยในส่วนของ Text Processing นั้นจะทำการแปลงข้อความภาษาไทยแต่ละพยางค์ให้เป็นรหัสแทนองค์ประกอบเสียง จากนั้นรหัสเหล่านี้จะถูกประมวลผลในส่วนของ Signal processing สำหรับการทำงานในส่วนของ Signal processing นั้นจะมีสองขั้นตอนคือ การสังเคราะห์เสียงพยางค์วรรณยุกต์สามัญ และ การผันเสียงวรรณยุกต์ ในการสังเคราะห์เสียงพยางค์วรรณยุกต์สามัญได้ใช้วิธีการ Concatenation ของหน่วยเสียงที่เก็บไว้ในรูปแบบ Semi-syllable โดยมี Initial part จำนวน 288 หน่วยเสียง และ Final part จำนวน 243 หน่วยเสียง ทำให้สามารถสังเคราะห์พยางค์วรรณยุกต์สามัญได้จำนวน 7,776 พยางค์ (32 เสียงพยัญชนะต้น * 27 เสียงสระ * 9 เสียงพยัญชนะท้าย) สำหรับในส่วนของการผันเสียงวรรณยุกต์ได้ใช้วิธีการปรับค่า Pitch ของเสียงพยางค์วรรณยุกต์สามัญ โดยใช้วิธีการ Autocorrelation ในการหาค่า Pitch และหลังจากปรับค่า Pitch แล้วก็ใช้วิธีการ LPC(Linear Predictive Coding) ในการสังเคราะห์เสียง เมื่อค่า Pitch เปลี่ยนไปจะทำให้ความถี่เปลี่ยนแปลงไปด้วย จากการเปลี่ยนแปลงความถี่ดังกล่าว ทำให้สามารถผันเสียงวรรณยุกต์อื่นๆ ในภาษาไทย ได้แก่ เอก โท ตรี และ จัตวา จึงทำให้สามารถสังเคราะห์เสียงได้ทั้งหมด 38,800 พยางค์ (7,776 พยางค์วรรณยุกต์สามัญ * 5 ระดับวรรณยุกต์)

การทดลองสังเคราะห์เสียงพูดภาษาไทยนั้น ได้ทำการอ่านออกเสียงจากตัวอย่างข้อความภาษาไทย 3 ลักษณะ คือ อ่านจากข้อความที่มีระดับเสียงสามัญทีละพยางค์ อ่านจากข้อความที่มีการผันวรรณยุกต์ทีละพยางค์ และ อ่านจากข้อความที่เป็นประโยคซึ่งมีความหมาย โดยประเมินคุณภาพเสียงจากความคิดเห็นของผู้ทดสอบ (Mean Opinion Score) ผลการทดลองที่ได้ คุณภาพเสียงพยางค์ที่เป็นวรรณยุกต์สามัญนั้นมีคุณภาพยอมรับได้ 100% แต่สำหรับเสียงที่มีการผันวรรณยุกต์คุณภาพของเสียงยอมรับได้ 61.25% คุณภาพต่ำ 9.75% และ ยอมรับไม่ได้ 29.00% ในการทดลองคุณภาพเสียงที่ได้จากการอ่านจากประโยคตัวอย่างซึ่งเป็นประโยคที่มีความหมายนั้น ได้ใช้ปัจจัยในการวัดคุณภาพของเสียง 4 ปัจจัย คือ การออกเสียงและสำเนียง การแยกแยะคำที่แตกต่าง ความเป็นธรรมชาติ และ ความสามารถในการรับฟังเป็นคำพูด โดยผลคะแนนที่ได้คือ 2.58, 2.66, 2.40 และ 2.89 ตามลำดับ จากผลการทดลองสรุปได้ว่าระบบสามารถสังเคราะห์เสียงได้ในระดับที่ยอมรับได้