A PROTOTYPE OF TEXT-TO-SPEECH FOR THAI BASED ON TIME DOMAIN PITCH-SYNCHRONOUS OVERLAP AND ADD

NATTHAKIJ ANGSUBHAKORN

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 974-665-966-9

3936984 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc.(COMPUTER)
KEY WORDS : TEXT-TO-SPEECH / SPEECH SYNTHESIS / TD-PSOLA
NATTHAKIJ ANGSUBHAKORN : A PROTOTYPE OF TEXT-TO-SPEECH FOR THAI BASED ON TIME DOMAIN PITCH-SYNCHRONOUS OVERLAP AND ADD. THESIS ADVISORS : ASSOC. PROF. SUPACHAI TANGWONGSAN, Ph.D., ASST. PROF. DAMRAS WONGSAWANG, Ph.D. 90 p. ISBN 974-665-966-9.

This research work presents a prototype of Text-to-Speech for Thai. Text-to-Speech (TTS) or Speech Synthesis is an application that can automatically generate human speech from the input text. It normalizes the complex word, exception word, abbreviation, number, and symbol into a simple form. In addition, the speech synthesis is used when visibility is problematic such as for the blind or when visibility is focused on something else. Moreover, it can be used to perceive information remotely such as via telephone.

The prototype was developed and based on the Time Domain Pitch-Synchronous Overlap and Add method. The model consists of two major modules: text analysis and speech signal processing. The text analysis module decomposes the input text into phonetic unit description parameters, which is composed of phonetic units and prosody information. Next, these parameters are further processed by a speech signal processing module, which is based on TD-PSOLA algorithm. The phonetic units are decomposed into a sequence of overlapping signals extracted pitch-synchronously by multiplying the signals to a window function. The window function is usually Hanning or Hamming type, which is typically centered at the pitch-mark position. The pitch-mark position can be obtained using a general pitch determination algorithm such as Autocorrelation, AMDF, and Cepstrum Analysis. In order to increase the pitch of speech, the duration between pitch-marks should be decreased. On the other hand, to decrease pitch of speech, the duration between pitch-marks is inversely increased. In addition, the duration modification of the signals can be performed by repeating or omitting pitch-marked signals. In order to increase duration, the signals are repeated, while to decrease duration, the signals are omitted. As a result, pitch contour of signals can be modified similar to Thai standard tone contours, which consists of low, falling, high, and rising levels.

The experiment was performed by generating a set of sentences and evaluating the overall quality of output speech using the Mean Opinion Score (MOS) method. The evaluation aspects considered in the experiments were pronunciation, distinctness, naturalness, and intelligibility. The evaluation was performed by 15 Thai native speakers in a controlled environment. The experimental result shows that the speech output generated from the prototype is considered intelligible and can be recognized as natural Thai pronunciation, spoken by a native speaker. The prototype is able to produce most commonly used words in CVC (Consonant-Vowel- Consonant) pattern.

3936984 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

ณัฐกิจ อังศุภากร : ต้นแบบการสังเคราะห์เสียงพูดในภาษาไทย (A PROTOTYPE OF TEXT-TO-SPEECH FOR THAI BASED ON TIME DOMAIN PITCH-SYNCHRONOUS OVERLAP AND ADD). คณะกรรมการควบคุมวิทยานิพนธ์ ศุภชัย ตั้งวงศ์ศานต์,Ph.D., ดำรัส วงศ์สว่าง ,Ph.D. 90 หน้า. ISBN 974-665-966-9.

การวิจัยเรื่องการสังเคราะห์เสียงพูดในภาษาไทย คือการแปลงข้อความภาษาไทยเป็นเสียงพูดในภาษาไทยแบบอัตโนมัติ โดยในการแปลงคำนั้นยังสามารถแปลงคำต่างๆเช่น คำที่มีความซับซ้อน, คำยกเว้น, คำย่อ, ตัวเลขและสัญลักษณ์ต่างๆได้ นอกจากนั้นการสังเคราะห์เสียงพูดยังสามารถใช้ในกรณีที่การมองเห็นเป็นปัญหา ไม่ว่าจะเป็นการไม่สามารถมองเห็นหรือไม่สามารถใช้สายตาในการรับรู้ข้อมูลได้ อีกทั้งยังสามารถใช้ในกรณีที่ต้องการรับฟังข้อมูลจากระยะไกล เช่น ผ่านระบบโทรศัพท์ เป็นต้น

ต้นแบบของการสังเคราะห์เสียงพูดในภาษาไทย ได้พัฒนาโดยใช้วิธีการ Time Domain Pitch-Synchronous Overlap and Add (TD-PSOLA) ระบบนี้จะประกอบไปด้วย 2 ส่วนหลักๆได้แก่ ส่วน text analysis และส่วน speech signal processing โดยส่วน text analysis จะทำการแปลงข้อความภาษาไทยให้เป็นสัญลักษณ์แทนเสียงพูด (Phonetic Unit Description) ซึ่งประกอบไปด้วยหน่วยเสียง (Phonetic Unit) และรูปแบบการออกเสียง (Prosody Information) จากนั้นหน่วยเสียงเหล่านี้จะถูกนำมาประมวลผลโดยใช้หลักการของ TD-PSOLA สัญญาณเสียงจะถูกแยกเป็นหน่วยสัญญาณเสียงย่อยๆที่ซ้อนทับกัน โดยการนำสัญญาณมาคูณกับ window function ซึ่งโดยทั่วไปจะใช้แบบ Hanning หรือ Hamming ซึ่งโดยทั่วไปแล้ว window function เหล่านี้จะถูกกำหนดจุดศูนย์กลางอยู่ที่ pitch-mark ของแต่ละหน่วยสัญญาณเสียงที่ซ้อนทับกันอยู่ โดยการหา pitch-mark นี้สามารถทำได้โดยการใช้วิธีการ pitch determination ทั่วไปได้ เช่น Autocorrelation, AMDF และ Cepstrum Analysis ในการที่จะเพิ่มความถี่ของเสียงทำได้โดยการปรับระยะห่างระหว่าง pitch-mark ให้สั้นลง ในทางกลับกัน การลดความถี่เสียงก็สามารถทำได้โดยการเพิ่มระยะห่างระหว่าง pitch-mark นอกจากนี้การปรับความยาวของสัญญาณเสียงก็สามารถทำได้โดยการเพิ่มหรือลดหน่วยของสัญญาณเสียงที่ซ้อนทับกันอยู่ ซึ่งทำให้ความยาวของสัญญาณเสียงเปลี่ยนไป จากการเปลี่ยนแปลงความถี่ดังกล่าว ทำให้สามารถเปลี่ยนเสียงตามแนวแกนเวลาซึ่งมีผลให้สามารถผันเสียงเป็นวรรณยุกต์ต่างๆได้ โดยต้นแบบการสังเคราะห์เสียงพูดสามารถผันเสียงวรรณยุกต์ต่างๆในภาษาไทยครบทุกเสียง ได้แก่เสียงเอก,เสียงโท,เสียงตรีและเสียงจัตวา

การทดลองการสังเคราะห์เสียงพูดในภาษาไทย โดยการสังเคราะห์เสียงพูดจากประโยคตัวอย่าง ได้ทำการประเมินคุณภาพของเสียงที่ได้โดยวิธีการให้คะแนนตามความคิดเห็น (Mean Opinion Score) โดยมีปัจจัยในการประเมินผลได้แก่ การออกเสียงและสำเนียง (Pronunciation), การแยกแยะคำที่แตกต่าง (Distinctness), ความเป็นธรรมชาติ (Naturalness), รวมถึงความสามารถในการรับฟังเป็นคำพูดของภาษานั้นๆ (Intelligibility) โดยการประเมินจะใช้ผู้ฟังจำนวน 15 คน ซึ่งฟังในห้องฟังที่มีการควบคุมสภาวะแวดล้อม จากผลการทดลองแสดงให้เห็นว่าเสียงพูดที่ได้จากระบบสามารถรับฟังว่าเป็นการออกเสียงในภาษาไทย โดยระบบสามารถออกเสียงคำส่วนใหญ่ที่อยู่ในรูปแบบ "พยัญชนะ-สระ-ตัวสะกด" ได้