THAI SYLLABLE SPEECH RECOGNITION BY SEGMENTAL PROBABILITY MODEL

WIJIT THANASANURAK

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 974-04-0604-1

3937004 SCCS/M : MAJOR : COMPUTER SCIENCE; M.Sc. (COMPUTER SCIENCE)
KEY WORDS : SPEECH RECOGNITION / SEGMENTAL PROBABILITY MODEL / TONE RECOGNITION / LINEAR REGRESSION MODEL / VOWEL-FIRST-SEARCH TREE

WIJIT THANASANURAK : THAI SYLLABLE SPEECH RECOGNITION BY SEGMENTAL PROBABILITY MODEL. THESIS ADVISORS: SUPACHAI TANGWONGSAN, Ph.D., CHOMTIP PORNPANOMCHAI, Ph.D., 79 P. ISBN 974-04-0604-1

This research studies the solution to isolated syllable, speaker dependent speech recognition in Thai. Based on the assumption that Thai phonological structure is very similar to Chinese, the Segmental Probability Model (SPM) approach currently used in Chinese speech recognition has been adapted, employing the Expectation and Maximization (EM) algorithm to cluster a speech utterance. Vowel elimination is suggested to improve accuracy and reduce error rate in isolation of syllables with the same vowel. We also propose the Vowel-First-Search Tree (VFST) technique for faster matching of the syllable models, by looking for the vowel phoneme first, then the initial and final consonant phoneme, respectively. For tone recognition, we employ the linear regression model to determine the pitch direction at the end of the pitch contours of each syllable.

The experiment is carried out using 14 orders of Linear Predictive Coding Cepstrum (LPCC), with a sampling rate of 11,025 Hz, and a total of 553 syllables in the model base. The results yield accuracy at 91.68% for testing on one speaker, and 95.84% for two speakers. The average recognition time is 1.13 second. The results show the SPM to be a promising solution for Thai speech recognition, and also applicable to multi-speaker systems. The accuracy of tone recognition is 72.02%, 69.71%, 97.08%, 97.69% and 96.48% for five Thai tonemes called the mid, the low, the fall, the high, and the rise, respectively. The linear regression model can determine the fall, the high, and the rise tones correctly, but has some difficulties in determining the mid tone from the low tone because their pitch contours are very similar.

In conclusion, this model has proven to be effectively adaptable to use with Thai speech recognition systems. Some more improvements may be needed to work out the problems with the similarity in mid tone and low tone pitch contours.

3937004 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์; วท.ม. (วิทยาการคอมพิวเตอร์)
คำสำคัญ : การรู้จำเสียงพยัญชนะ / SEGMENTAL PROBABILITY MODEL / การรู้จำเสียงวรรณยุกต์ / วิธีการค้นหาพยางค์โดยเริ่มที่ส่วนของสระก่อน

วิจิตร ธเนศานุรักษ์ : การรู้จำเสียงพยางค์ภาษาไทยแบบขึ้นกับผู้พูดโดยใช้วิธี Segmental Probability Model (THAI SYLLABLE SPEECH RECOGNITION BY SEGMENTAL PROBABILITY MODEL). คณะกรรมการควบคุมวิทยานิพนธ์ : ศุภชัย ตั้งวงศ์ศานต์, Ph.D., ชมทิพ พรพนมชัย, Ph.D. 79 หน้า. ISBN 974-04-0604-1

ในงานวิจัยนี้ได้นำเสนอแนวทางการแก้ปัญหาของ Speech Recognition สำหรับภาษาไทย โดยกำหนดขอบเขตงานวิจัยเฉพาะเสียงของพยางค์เดี่ยว (Isolated Syllable) แบบขึ้นกับผู้พูด (Speaker Dependent) โดยใช้วิธี Segmental Probability Model (SPM) เป็นแนวทางการพัฒนา SPM ได้พัฒนาขึ้นเพื่อใช้กับภาษาจีนที่มีความคล้ายกับภาษาไทยอย่างมาก โดยงานวิจัยได้ดัดแปลงวิธีการบางส่วนของ SPM ดังนี้ การใช้ Expectation and Maximization (EM) Algorithm เพื่อแบ่งกลุ่มข้อมูลเสียง นอกจากนั้นได้ตัดส่วนของเสียงสระออกไปเพื่อลดความผิดพลาดในการแยกพยางค์ที่อยู่ในกลุ่มเสียงสระเดียวกัน (Vowel Elimination) ในงานวิจัยยังได้เสนอวิธี Vowel-First-Search Tree (VFST) ในการค้นหาคำตอบที่ถูกต้องอย่างรวดเร็ว VFST เป็นวิธีเปรียบเทียบเสียงพยางค์โดยเริ่มต้นจากส่วนของเสียงสระก่อนเนื่องจากให้ผลความถูกต้องสูงกว่าเสียงส่วนอื่น ต่อจากนั้นจึงค้นหาจากเสียงพยัญชนะส่วนต้นและพยัญชนะส่วนท้ายของพยางค์ตามลำดับ สุดท้ายงานวิจัยได้เสนอวิธีการแยกโทนเสียงทั้งห้าของภาษาไทยคือ เสียงสามัญ เอก โท ตรี และจัตวา ตามลำดับ โดยนำเอา Linear Regression Model มาใช้วิเคราะห์การเปลี่ยนแปลงของความถี่ Pitch ในพยางค์

การทดลองนี้ได้เลือกใช้ Linear Predictive Coding Cepstrum (LPCC) จำนวน 14 Order และใช้ Sampling Rate เท่ากับ 11,025 Hz ระบบมีจำนวนคำศัพท์ทั้งสิ้น 553 พยางค์ โดยมีผลการทดสอบความถูกต้อง 91.68% สำหรับเสียงของบุคคลเดี่ยวและ 95.84% สำหรับเสียงของบุคคลสองคน ในการทดสอบแต่ละครั้งใช้เวลาในการประมวลผลเฉลี่ยประมาณ 1.13 วินาที ส่วนการแยกแยะโทนเสียงได้ผลความถูกต้อง 72.02%, 69.71%, 97.08%, 97.69% และ 96.48% สำหรับโทนเสียงที่ 1, 2, 3, 4 และ 5 ตามลำดับ จากผลการทดลองแสดงให้เห็นว่า SPM สามารถนำมาประยุกต์ใช้กับภาษาไทยได้เป็นอย่างดีและสามารถใช้กับระบบ Multi-Speaker ได้อีกด้วย ส่วนการนำ Linear Regression Model มาใช้ในการวิเคราะห์โทนเสียงสามารถใช้ได้ดีกับโทนเสียงที่ 3, 4, 5 แต่ยังมีความผิดพลาดในการแยกโทนเสียงที่ 1 และ 2 เนื่องจากโทนเสียงทั้งสองมีโครงสร้าง Pitch ที่คล้ายกัน