ฐานข้อมูลเสียงขนาดใหญ่สำหรับระบบรู้จำเสียงพูดต่อเนื่องภาษาไทย

LOTUS: Large Vocabulary Thai Continuous Speech Recognition Corpus

ภัชริกา คชสำโรง, ตรีภพ สรรเพชรนิยม, ศวิต กาสุริยะ, ณัฐนันท์ ทัดพิทักษ์กุล, ชัย วุฒิวิวัฒน์ชัย

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

112 อุทยานวิทยาศาสตร์ประเทศไทย ถนนพหลโยธิน ตำบลคลองหนึ่ง อำเภอคลองหลวง

จังหวัดปทุมธานี 12120

ABSTRACT - Speech corpus is a major component in constructing an efficient and accurate speech recognition system. Even the best algorithm with carefully designed system cannot accomplish good performance if the system is trained by a poor corpus. Therefore, this paper aims to present our LOTUS (Large vOcabulary Thai continUous Speech recognition) corpus, which is a large-scale Thai speech corpus designed and collected as those performed in other standard corpora such as JNAS for Japanese and WSJCAM0 for American English. The LOTUS corpus consists of 2 sets. The first set is a phonetically-balanced and phonetically-distributed speech set. It is designed to cover all possible Thai phonemes with intensively tagged phoneme boundaries. The second set contains speech utterances that cover 5,000 most frequent words appearing in our largest Thai text corpus. The second set is clustered into 3 subsets, a training set, a development set, and an evaluation set. The LOTUS corpus is not only used for our main project of the first Thai dictation system, but also can be used for acoustic-phonetic research and for initializing other domain-dependent acoustic models. This paper explains details of LOTUS including the design, recording approach, problems during construction, tools implemented for corpus manipulation, and various ways to utilize the corpus.

KEY WORDS -- Thai speech corpus, LOTUS corpus, large vocabulary continuous speech recognition, dictation system

บทคัดย่อ -- ฐานข้อมูลเสียงพูดเป็นส่วนประกอบสำคัญที่จะทำให้ระบบรู้จำเสียงพูดมีประสิทธิภาพและให้ผลการรู้จำถูกต้องสูงสุด แม้ว่าในปัจจุบันได้มีการคิดค้นนำเสนอวิธีมากมายในการพัฒนาระบบรู้จำเสียงพูดให้ดีขึ้น แต่ระบบรู้จำที่ดียังคงต้องอาศัยฐานข้อมูลเสียงขนาดใหญ่ เพื่อใช้เป็นข้อมูลพื้นฐานในการวิเคราะห์และฝึกฝนระบบ บทความฉบับนี้นำเสนอฐานข้อมูลชื่อโลตัส (Large vOcabulary Thai continUous Speech recognition corpus: LOTUS) ซึ่งเป็นฐานข้อมูลเสียงพูดภาษาไทยขนาดใหญ่ ที่ผ่านการออกแบบและพัฒนาตามมาตรฐานในการสร้างฐานข้อมูลลักษณะเดียวกัน อาทิเช่น ฐานข้อมูลเสียงภาษาญี่ปุ่นที่ชื่อว่า JNAS และฐานข้อมูลเสียงภาษาอังกฤษที่ชื่อว่า WSJCAM0 ฐานข้อมูลโลตัสประกอบด้วย 2 ชุดข้อมูล ชุดแรกคือชุดข้อมูลซึ่งครอบคลุมหน่วยเสียงที่เกิดขึ้นทั้งหมดในภาษาไทย ชุดข้อมูลนี้จะมีการกระจายของหน่วยเสียงอย่างสมดุลและมีการกำกับขอบเขตของหน่วยเสียง ชุดที่สองคือชุดข้อมูลที่ครอบคลุมคำศัพท์ที่เกิดขึ้นบ่อยในภาษาไทยจำนวน 5,000 คำ โดยแบ่งออกเป็นชุดย่อยสำหรับฝึกฝน ชุดย่อยสำหรับทดสอบเพื่อพัฒนา และชุดย่อยสำหรับทดสอบเพื่อประเมิน ฐานข้อมูลนี้นอกจากจะใช้ในเป้าหมายหลักเพื่อวิจัยและพัฒนาระบบพิมพ์ตามเสียงพูด (Dictation system) ระบบแรกสำหรับภาษาไทยแล้ว ยังสามารถนำไปใช้ใน

งานวิจัยเกี่ยวกับสัทลักษณ์ และเป็นฐานข้อมูลหลักในการสร้างแบบจำลองหน่วยเสียงสำหรับระบบรู้จำเสียงในเนื้อหาเฉพาะอื่นๆ ได้ บทความฉบับนี้จะได้อธิบายถึงรายละเอียดต่างๆ ของฐานข้อมูลโลตัสเสียงพูดภาษาไทยที่สร้างขึ้นตั้งแต่การออกแบบ การบันทึกเสียง ปัญหาที่เกิดขึ้นในการสร้าง เครื่องมือที่ใช้ในการสร้าง และแนวทางในการประยุกต์ใช้ฐานข้อมูลในอนาคต

คำสำคัญ -- ฐานข้อมูลเสียงภาษาไทย, ฐานข้อมูลโลตัส, การรู้จำเสียงพูดต่อเนื่องที่ครอบคลุมคำศัพท์จำนวนมาก, ระบบพิมพ์ตามเสียงพูด

1. บทนำ

ก้าวแรกในการวิจัยและพัฒนาระบบรู้จำเสียงพูดต่อเนื่อง (Continuous speech recognition) คือการพัฒนาระบบพิมพ์ตามเสียงพูด (Dictation system) ในการพัฒนาระบบดังกล่าว นอกจากจะได้ริเริ่มออกแบบหน่วยเสียงสำหรับภาษาไทยแล้ว ยังได้เก็บฐานข้อมูลเสียงขนาดใหญ่ ซึ่งมีคุณค่าไม่เพียงแค่ใช้ในการสร้างระบบพิมพ์ตามเสียงพูดเท่านั้น แต่ยังสามารถใช้ในการสร้างแบบจำลองหน่วยเสียงสำหรับการรู้จำในขอบเขตเนื้อหาเฉพาะอื่นๆ ได้อีกด้วย บทความฉบับนี้ได้นำเสนอฐานข้อมูลโลตัส (Large vOcabulary Thai contInUous Speech recognition corpus: LOTUS) ซึ่งเป็นฐานข้อมูลเสียงพูดต่อเนื่องภาษาไทยขนาดใหญ่ที่ถูกออกแบบมาเพื่อใช้ในการวิจัยและพัฒนาระบบพิมพ์ตามเสียงพูดครอบคลุมคำศัพท์จำนวน 5,000 คำ โดยยึดหลักการออกแบบเช่นเดียวกับฐานข้อมูลเสียงในภาษาอื่นๆ เช่น ฐานข้อมูล JNAS สำหรับภาษาญี่ปุ่น [8] และฐานข้อมูล WSJCAM0 สำหรับภาษาอังกฤษ [9] ในอนาคตอันใกล้ ฐานข้อมูลนี้จะเผยแพร่ให้กับผู้ที่สนใจนำไปใช้ในการวิจัยและพัฒนาผ่านทางเว็บไซต์ http://www.nectec.or.th/rdi/lotus บทความฉบับนี้สรุปรายละเอียดในการออกแบบฐานข้อมูล การสร้างฐานข้อมูล ปัญหาที่พบในการสร้างฐานข้อมูล และเครื่องมือที่ใช้ในการพัฒนา พร้อมทั้งนำเสนอแนวทางการนำฐานข้อมูล

2. การออกแบบฐานข้อมูล.

การออกแบบฐานข้อมูลโลตัส มีวัตถุประสงค์สำคัญอยู่ 2 ประการ คือ 1) เพื่อเป็นข้อมูลชุดฝึกฝนและทดสอบในการสร้างแบบจำลองหน่วยเสียง (Acoustics model) ของภาษาไทย 2) เพื่อเป็นข้อมูลชุดฝึกฝนและทดสอบในการสร้างแบบจำลองทางภาษา (Language model) สำหรับระบบพิมพ์ตามเสียงพูดภาษาไทยที่ครอบคลุมคำศัพท์ 5,000 คำ โดยทำการคัดเลือกประโยคจากฐานข้อมูลบทความ ORCHID (Open Linguistic Resources Channelled toward InterDisciplinary research) [6] ซึ่งมีจำนวน 27,634 ประโยค และฐานข้อมูลบทความอื่นๆ จะได้คลังข้อความรวมทั้งสิ้น 180,504 ประโยค ซึ่งมีคำประมาณ 2,500,000 คำ (43,255 คำศัพท์) ฐานข้อมูลโลตัสประกอบด้วยชุดข้อมูล2 ชุดดังรายละเอียดที่จะกล่าวต่อไป

ตารางที่ 1 รายละเอียดของข้อมูลชุดหน่วยเสียงสมดุล

รายละเอียด	**ข้อมูลชุดหน่วยเสียงสมดุล**
จำนวนประโยค	802
จำนวนคำ	7,847
จำนวนพยางค์	12,702
จำนวนหน่วยเสียง	38,106
ครอบคลุมหน่วยเสียงคู่	1,628 (90.9%)
จำนวนผู้พูด (ล่าสุด)	48 คน
ขนาดของฐานข้อมูลเสียง	13 ชั่วโมง

ตารางที่ 2 รายละเอียดชุดครอบคลุมคำศัพท์ 5,000 คำ

รายละเอียด	***TR***	***DT***	***ET***
จำนวนประโยค	3,007	500	500
จำนวนคำศัพท์	5,000	1,622	1,630
จำนวนคำ	55,504	8,076	8,290
จำนวนผู้พูด (ล่าสุด)	24	12	12
ขนาดฐานข้อมูลเสียง	70 ชั่วโมง		

2.1 ชุดหน่วยเสียงสมดุล หรือ PD (Phonetically distribution set)

เป็นชุดประโยคที่ออกแบบการคัดเลือกประโยคมาเพื่อใช้สำหรับการฝึกฝนแบบจำลองหน่วยเสียงขั้นต้น ซึ่งจะครอบคลุมการเกิดของ “หน่วยเสียงคู่” (Biphone) ในภาษาไทยทั้งภายในพยางค์, ระหว่างพยางค์ และระหว่างคำ โดยไม่คำนึงถึงระดับเสียงวรรณยุกต์ (Tonal Level) การเกิดหน่วยเสียงคู่ในชุดนี้จะมีการกระจายสอดคล้องกับบทความที่ใช้ในการคัดเลือก ซึ่งก็คือคลังข้อความ ORCHID ผลการคัดประโยคจะได้ชุดประโยค PD ที่ครอบคลุมหน่วยเสียงคู่จำนวน 1,628 คู่ซึ่งคิดเป็น 90.9% ของหน่วยเสียงคู่ที่เกิดขึ้นได้ในภาษาไทย เสียงที่ได้จากชุดประโยคนี้จะถูกนำไปกำกับขอบเขตของหน่วยเสียงอย่างละเอียด รายละเอียดในการคัดเลือกประโยคชุด PD สามารถอ่านได้ใน [1] รายละเอียดโดยสรุปของชุด PD แสดงในตารางที่ 1

2.2 ชุดประโยคที่ครอบคลุมคำศัพท์ที่มีสถิติการใช้สูงสุด 5,000 คำ

เป็นชุดประโยคที่ถูกออกแบบมาเพื่อเป็นชุดฝึกฝนและทดสอบในการสร้างแบบจำลองทางภาษา (Languange model) สำหรับระบบพิมพ์ตามเสียงพูดภาษาไทย โดยทำการคัดเลือกประโยคที่ประกอบด้วยคำศัพท์ที่มีสถิติการใช้สูงสุด 5,000 ลำดับแรกจากคลังข้อความทั้งหมด รายละเอียดข้อมูลชุดคำศัพท์ได้แสดงไว้ในตารางที่ 2 โดยแบ่งข้อมูลออกเป็น 3 ชุด ดังต่อไปนี้

2.2.1 ชุดฝึกฝน (Training set, TR)

ใช้ในการฝึกฝนแบบจำลองหน่วยเสียงเพิ่มเติมและฝึกฝนแบบจำลองภาษาซึ่งครอบคลุมคำศัพท์จำนวน 5,000 คำ

2.2.2 ชุดทดสอบเพื่อพัฒนา (Development test set, DT)

ใช้ในการทดลองระบบพิมพ์ตามเสียงพูด ประกอบด้วยประโยคที่ได้รับการคัดเลือกมาอย่างเหมาะสมทั้งทางด้านความยาวของประโยค, ค่าเพอร์เพลกซิตี (Perplexity) และประกอบด้วยคำศัพท์ที่อยู่ในกลุ่มคำศัพท์ 5,000 คำที่มีในชุดฝึกฝน

2.2.3 ชุดทดสอบเพื่อประเมิน (Evaluation test set, ET)

ใช้ในการทดสอบขั้นสุดท้ายเพื่อประเมินความสามารถของระบบพิมพ์ตามเสียงพูด มีรายละเอียดวิธีการคัดประโยคเช่นเดียวกับชุดทดสอบสำหรับการพัฒนา (DT)

3. ความร่วมมือ

ในการสร้างฐานข้อมูลขนาดใหญ่ ยิ่งจำนวนข้อมูลเสียงมีจำนวนมาก ก็จะยิ่งเป็นประโยชน์ต่อการวิจัยระบบรู้จำมากยิ่งขึ้น ดังนั้นในการสร้างฐานข้อมูลให้มีจำนวนผู้พูดมากที่สุดเท่าที่จะมากได้ จึงต้องอาศัยความร่วมมือกัน ระหว่างศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค) กับมหาวิทยาลัย อีก 2 แห่ง คือ มหาวิทยาลัยสงขลานครินทร์ และ มหาวิทยาลัยเทคโนโลยีมหานคร รายละเอียดในการบันทึกเสียงในแต่ละแห่ง แสดงไว้ในตารางที่ 3

4. ระบบเสียงภาษาไทย (Thai Phonology)

ระบบเสียงภาษาไทย เป็นส่วนสำคัญที่ต้องคำนึงถึงในการสร้างฐานข้อมูล รูปแบบแทนพยางค์ของฐานข้อมูลโลตัสได้แก่ /C_i _V_T/ และ /C_i _V_C_f_T/ โดยที่ C_i แทนพยัญชนะต้น (พยัญชนะเดี่ยว หรือ พยัญชนะควบกล้ำ), V แทนสระ (เสียงสั้น เสียงยาว หรือสระผสม) , C_fแทนเสียงพยัญชนะสะกด (เฉพาะพยัญชนะต้นบางตัวเท่านั้น) , T แทนเสียงวรรณยุกต์ เช่น คำว่า “การ” มีรูปแบบของหน่วยเสียง คือ /k_aa_n^ _0/ โดยที่สัญลักษณ์ ^ ใช้เพื่อกำกับเสียงพยัญชนะท้าย ในระบบเสียงภาษาไทยนั้นมีพยัญชนะเดี่ยวอยู่ 21 หน่วยเสียง พยัญชนะควบกล้ำ 12 หน่วยเสียง สระ 24 หน่วยเสียง และหน่วยเสียงที่ใช้แทนเสียงคำในภาษาต่างประเทศอีก 5 หน่วยเสียง สัญลักษณ์แทนเสียงพยัญชนะต้นและพยัญชนะท้ายในภาษาไทยแสดงไว้ในตารางที่ 4 สัญลักษณ์แทนเสียงพยัญชนะควบกล้ำในภาษาไทยแสดงไว้ใน ตารางที่ 5 สัญลักษณ์แทนเสียงสระในภาษาไทยแสดงไว้ใน ตารางที่ 6 สัญลักษณ์แทนหน่วยเสียงที่ใช้ในงานวิจัยนี้อ้างอิงมาจากสัทลักษณ์มาตรฐานของ IPA (International Phonetic Association) [3] และปรับเปลี่ยน เพื่อให้เหมาะสมสำหรับการประมวลผลข้อมูลด้วยคอมพิวเตอร์ การถ่ายถอดเสียงคำอ่าน (Phonetic transcription) เป็นงานที่สำคัญอีกงานหนึ่ง เพราะในการประมวลผลข้อมูลเสียงจำเป็นต้องมีสัญลักษณ์แทนเสียง

เพื่อใช้ประกอบในการประมวลผล อาทิเช่น ในการคัดเลือกประโยคชุด PD จะต้องใช้เกณฑ์การคัดเลือกหน่วยเสียง ดังนั้นชุดประโยคต่างๆ จะต้องนำมาผ่านกระบวนการแปลงรูปเขียนเป็นคำอ่าน (Thai Grapheme To Phoneme, G2P) [2] แล้วจึงมาทำการคัดเลือกประโยคตามกระบวนการคัดเลือกประโยคต่อไป

5. การบันทึกเสียง

การบันทึกสียงของฐานข้อมูล ได้ถูกออกแบบให้ทำการบันทึกเสียงผ่าน DAT (Digital audio tape) ก่อน แล้วจึงทำการแปลงสัญญาณเสียง เป็นไฟล์อิเล็กทรอนิกส์มาตรฐานผ่านทางการ์ดเสียง (Sound card) ของเครื่องคอมพิวเตอร์ โดยสภาวะแวดล้อมการบันทึกเสียง จะแบ่งออกเป็น 2 แบบ (1) แบบห้องทำงาน (SNR 20 dB) และ (2) แบบห้องเงียบ (SNR 30dB) ผ่านไมโครโฟน 2 ตัวพร้อมกัน คือแบบ Dynamic Close-talk (TELEX H-41) ระดับคุณภาพสูง และแบบ Dynamic Unidirectional (SONY F-720) ระดับคุณภาพปานกลาง ผู้พูดทุกคนต้องทำการบันทึกเสียงทั้ง 2 สภาวะแวดล้อม ในแต่ละรอบของผู้พูดแต่ละคนจะได้รับการทดสอบและรับคำแนะนำเพื่อปรับวิธีการพูด ความดังในการพูดพร้อมทั้งแนะนำ ขั้นตอนในการพูดเพื่ออัดเสียงในการอัดเสียงสำหรับผู้พูดแต่ละคน จะทำการอัดเสียงของสภาพแวดล้อมก่อนเป็นเวลา 3 วินาที แล้วจึงเริ่มอัดเสียงพูดในลักษณะการอ่าน โดยรายละเอียดเพิ่มเติมจะได้นำเสนอไว้ในเอกสารประกอบฐานข้อมูลโลดัสต่อไป

6. ปัญหาที่พบ

6.1 ปัญหานิยามของคำและการตัดคำ

ในการคัดเลือกประโยคมีอุปสรรคสำคัญคือนิยามของคำในภาษาไทยและความผิดพลาดของการตัดคำ การตัดคำผิดมีผลต่อการคัดเลือกคำอย่างมาก ตัวอย่างปัญหาได้แก่ คำที่ควรเป็นคำประสมถูกตัดออกเป็นคำโดด มีผลทำให้ได้คำที่มีความหมายเปลี่ยนไป เช่น “เลือกตั้ง” ถูกตัดคำเป็น “เลือก” กับ “ตั้ง” หรือคำที่ควรเป็นคำโดดกลับถูกตัดคำเป็นคำประสม เช่น ควรตัดคำเป็น “ให้” กับ “การ” แต่กลับตัดคำเป็น “ให้การ” ซึ่งคำที่เขียนเหมือนกันเมื่ออยู่ในประโยคที่ต่างกัน อาจจะต้องตัดคำต่างกันตามความหมายในบริบทนั้นๆ ทำให้ค่าสถิติของการเกิดขึ้นของคำในฐานข้อมูลไม่ตรงกับข้อมูลจริง มีผลทำให้การคัดเลือกคำให้ผลที่ไม่เป็นจริงตามไปด้วย อาจทำให้คำที่ควรถูกเลือกเข้าไปในรายการคำศัพท์ 5,000 คำมีค่าสถิติน้อยจนไม่ถูกเลือกหรือไม่เกิดขึ้นเลยก็ได้ ดังนั้นเมื่อคัดเลือกรายการคำศัพท์ที่ค่าสถิติสูงสุด 5,000 คำ ได้แล้ว ต้องมีการตรวจสอบความถูกต้องของการตัดคำ เมื่อเทียบกับความหมายในการเกิดของคำในประโยคว่าถูกต้องหรือไม่ แล้วทำการแก้ไขการตัดคำใหม่ การตรวจสอบการตัดคำนี้ต้องอาศัยความรู้ความเข้าใจธรรมชาติของภาษา ซึ่งเป็นงานที่ต้องอาศัยเวลาและความละเอียดเป็นอย่างมาก การขยายขนาดของฐานข้อมูลให้ครอบคลุมมากกว่า 5,000 คำ จึงต้องอาศัยการวิจัยวิธีการที่เหมาะสมในการสร้างฐานข้อมูลที่ลดระยะเวลาและค่าใช้จ่ายในการตรวจสอบความถูกต้อง

6.2 ปัญหาการบันทึกเสียง

ปัญหาที่พบมากที่สุดอีกปัญหาหนึ่งคือการบันทึกเสียงจากผู้พูดที่อ่านรายการประโยคที่กำหนดให้ การบันทึกเสียงจำเป็นจะต้องกำหนดระดับความดังของการพูด กล่าวคือระดับความดังจะต้องไม่เกินขีดสูงสุดที่โปรแกรมสำหรับบันทึกเสียงจะรับได้ และจะต้องไม่เบาเกินไปจนกระทั่งมีค่ากำลังเสียงต่อกำลังสัญญาณรบกวน (Signal-to-noise ratio, SNR) ต่ำเกินไป โดยปกติในสภาพแวดล้อมที่ไร้เสียงรบกวน ควรจะต้องได้ SNR ไม่ต่ำกว่า 30 dB ในทางปฏิบัติความดังของการพูดจะเปลี่ยนแปลงตามระยะเวลาของการอ่านและคำที่อ่าน ผู้บันทึกจำเป็นจะต้องดูแลอย่างใกล้ชิดเพื่อให้สัญญาณเสียงที่ได้อยู่ในช่วงที่ยอมรับได้

6.3 ปัญหาการกำกับขอบเขตหน่วยเสียง

ในการพูดต่อเนื่อง สัญญาณเสียงของหน่วยเสียงสองหน่วยที่อยู่คู่กันมักจะต่อเนื่องกันจนกระทั่งในหลายๆ กรณี ไม่สามารถหาจุดแบ่งระหว่างทั้งสองหน่วยเสียงได้ ตัวอย่างที่เด่นชัดคือการต่อกันของหน่วยเสียงประเภท Nasal หรือ Semivowel ซึ่งรูปสัญญาณเสียงจะต่อเนื่องกันจนแยกไม่ออก อีกปัญหาหนึ่งคือการกำกับหน่วยเสียงเงียบสั้นๆ (Short pause) ซึ่งมักจะสับสนกับช่วงเงียบสั้นๆ ของหน่วย

เสียงประเภทเสียงกัก ปัญหาเหล่านี้จำเป็นต้องอาศัยการตีกรอบอย่างชัดเจนจากนักภาษาศาสตร์ ก่อนที่จะทำการกับเสียง รายละเอียดของข้อตกลงในการกำกับหน่วยเสียงคณะนักวิจัยจะได้เผยแพร่ต่อไปในอนาคต

7. เครื่องมือช่วยสำหรับการสร้างฐานข้อมูล

การจัดการกับข้อมูลเป็นจำนวนมาก เครื่องมือช่วยเป็นสิ่งที่จำเป็น สำหรับฐานข้อมูลโลตัส ทางทีมวิจัยได้พัฒนาเครื่องมือช่วยต่างๆ ขึ้นมาระหว่างที่ดำเนินการสร้างฐานข้อมูล เพื่อช่วยให้การจัดการกับข้อมูลต่างๆ ในแต่ละงานได้ง่ายขึ้น ได้แก่

7.1 เครื่องมือตัดคำภาษาไทย (Thai word segmentation)

SWATH (Smart word analysis for Thai) [7] เป็นโปรแกรมช่วยตัดคำสำหรับภาษาไทย โดยใช้วิธีการแบบจำลองเอ็นแกรม (Ngram model) และอาศัยพจนานุกรมคำศัพท์จำนวนมากช่วยในการตัดสินใจในการตัดคำ

7.2 เครื่องมือแปลงรูปเขียนเป็นคำอ่านสำหรับภาษาไทย (Thai grapheme to phoneme, G2P)

เป็นโปรแกรมช่วยในการแปลงตัวอักษรภาษาไทย ให้เป็นคำอ่าน โดยมีสัญลักษณ์แทนเสียงที่กำหนดไว้ดังที่กล่าวในหัวข้อที่ 4 โปรแกรม G2P อาศัย PGLR (Probabilistic generalized LR parser) ซึ่งได้รับการเรียนรู้จากคลังข้อความที่ผ่านการกำกับหน่วยเสียง [2]

7.3 เครื่องมือแก้ไขฐานข้อมูล (Corpus editor)

Corpus editor เป็นโปรแกรมช่วยในการจัดการกับคลังข้อความ ตรวจสอบการตัดคำ และตรวจสอบการถ่ายถอดเสียงของประโยค (Phonetic transcripton) ซึ่งต้องทำขึ้นควบคู่กับข้อมูลเสียง ทางทีมวิจัยเล็งเห็นถึงความสำคัญของการจัดทำเครื่องมือช่วยนี้ จึงได้พัฒนา Corpus editor ขึ้น โดยทำการตัดคำด้วยโปรแกรม SWATH และสร้างคำอ่านด้วยโปรแกรม G2P เพื่อให้ง่ายต่อการตรวจสอบความถูกต้องของคำอ่าน จึงได้เพิ่มความสามารถในการสังเคราะห์เสียงเข้าไว้ในโปรแกรมด้วย นอกจากนี้ยังได้ออกแบบช่องสำหรับประเภทของคำ (Part-of-speech, POS) ไว้ด้วย เพื่อรองรับการเพิ่มเติมที่อาจมีขึ้นในอนาคต รูปที่ 1 แสดงตัวอย่างหน้าจอของ Corpus Editor

7.4 เครื่องมือตัดหน่วยเสียง (Automatic phoneme segmentation)

เป็นโปรแกรมที่พัฒนาโดยใช้ Hidden Markov toolkit (HTK) [5] โปรแกรมนี้จะรับข้อมูลเสียงและคำอ่านที่ได้มาจากโปรแกรม G2P ผ่านทำการตรวจสอบความถูกต้องของคำอ่านโดยใช้ Corpus editor แล้วทำการกำกับขอบเขตของหน่วยเสียงแบบวนซ้ำ (Re-label) จนได้ข้อมูลเสียงพร้อมกับข้อมูลกำกับทางด้านเวลาของหน่วยเสียงที่ดีที่สุด ในระหว่างการกำกับขอบเขตหน่วยเสียงนั้น จะมีขั้นตอนย่อยที่สำคัญๆ คือการแก้ไขคำอ่าน (Pronunciation correction) การเติมเสียงเงียบสั้นๆ (Short pause insertion) และการเทียบเคียงหน่วยเสียงเพื่อหาขอบเขตของหน่วยเสียง (Phonetic alignment) โดยที่ทั้ง 3 ขั้นตอนนี้จะได้ทำอย่างอัตโนมัติโดยใช้เครื่องมือนี้ อ่านรายละเอียดเพิ่มเติมได้จาก [10]

8. การนำฐานข้อมูลไปใช้

ในอนาคตอันใกล้ ฐานข้อมูลโลตัสจะเผยแพร่ให้กับนักวิจัยและผู้ที่สนใจผ่านทางเว็บไซต์ http://www.nectec.or.th/ rdi/lotus นักวิจัยสามารถนำไปใช้เพื่อการวิจัยได้หลายทาง ได้แก่

1) ใช้ในการวิจัยและพัฒนา ระบบพิมพ์ตามเสียงพูด (Dictation system) ซึ่งเป็นเป้าหมายหลักของฐานข้อมูลนี้ นักวิจัยสามารถใช้ฐานข้อมูลในส่วนชุดข้อมูล PD ในการสร้างแบบจำลองหน่วยเสียงขั้นต้น ซึ่งจะครอบคลุมหน่วยเสียงในภาษาไทยได้ทั้งหมด และใช้ชุดข้อมูล TR ในการฝึกฝนแบบจำลองหน่วยเสียงเพิ่มเติม และใช้ชุดประโยค TR ในการสร้างแบบจำลองภาษาสำหรับระบบพิมพ์ตามเสียงพูดครอบคลุมคำศัพท์ 5,000 คำ ชุด DT จะใช้ในการทดสอบระบบเพื่อปรับปรุงขั้นตอนการสร้างระบบ และชุด ET จะใช้ในการประเมินความสามารถของระบบในขั้นสุดท้าย คณะนักวิจัยมีความหวังว่าฐานข้อมูลนี้จะเป็นมาตรฐานอันหนึ่งในการวิจัยและพัฒนาระบบดังกล่าวในอนาคตอันใกล้

2) ใช้ในการวิจัยเกี่ยวกับความสัมพันธ์ระหว่างเสียงกับหน่วยเสียง (Acoustic-phonetics) สำหรับภาษาไทย งานวิจัยเกี่ยวกับ Acoustic-phonetics ทำได้ยากเนื่องจากขาดแคลนฐานข้อมูลเสียงขนาดใหญ่ที่ใช้ในการวิเคราะห์ ชุดข้อมูล PD รวมถึงชุดข้อมูลอื่นๆ ในฐานข้อมูลนี้จะเป็นแหล่งข้อมูลสำคัญในการวิเคราะห์รูปแบบการออกเสียงของคนไทยจำนวนมาก

3) ใช้ในการวิจัยทางสัทลักษณ์ (Prosody) เช่นการวิเคราะห์เรื่องความยาวของการออกเสียงในแต่ละหน่วยเสียง (Duration) การหยุด (Pause) การเน้นเสียง (Stress) ระดับเสียงในระดับวลี (Intonation) และระดับเสียงในระดับพยางค์ (Tone) ซึ่งเป็นลักษณะเฉพาะของภาษาไทย ชุดข้อมูล PD ครอบคลุมคู่วรรณยุกต์ที่เกิดต่อกันในภาษาไทยครบทุกกรณี และครอบคลุมเสียงวรรณยุกต์ 3 ตัวต่อกันได้ถึง 91.2%

4) คลังข้อความที่ใช้ในการบันทึกเสียงในฐานข้อมูลโลตัสนับเป็นคลังข้อความขนาดใหญ่ของไทยที่ผ่านการตัดคำแปลงเป็นคำอ่าน และผ่านการตรวจสอบความถูกต้องอย่างละเอียดแล้ว สามารถนำไปใช้ในการวิจัยและพัฒนาเพื่อเพิ่มศักยภาพของเครื่องมือสำหรับตัดคำและเครื่องมือในการแปลงรูปเขียนเป็นคำอ่านได้ต่อไป

9. เอกสารอ้างอิง

[1] C. Wutiwatchai, P. Cotsomrong, S. Suebvisai, S. Kanokphara. 2002. *Phonetically Distributed Continuous Speech Corpus for Thai Language,* Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002), 869-872.

[2] P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt. 2001. *Thai Grapheme-to-Phoneme using Probabilistic GLR Parser,* Proceedings of Eurospeech 2001, 2: 1057-1060.

[3] S. Luksaneeyanawin, 1993. *Speech Computing and Speech Technology in Thailand*, Proceeding of the Symposim on Natural Language Proceeding in Thailand, 276-321.

[4] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul. 2003. *Thai Speech Corpus for Thai Speech Recognition*, Proceedings of the Oriental COCOSDA Workshop, 54-61

[5] S. Young D. Kershaw, J. Odell, D. Ollason, V. Valchev, P. Woodland. 2000. *The HTK book*, http://htk.eng.cam.ac.uk/docs/docs.shtml.

[6] V. Sornlertlamvanich, N. Takahashi, and H. Isahara. 1998. *Thai Part-Of-Speech tagged corpus: ORCHID*, Proceedings of the Oriental COCOSDA Workshop, 131-138.

[7] P. Charoenpornsawat. 1999. *Feature-based Thai Word Segmentation*, Master Thesis, Chulalongkorn University, Bangkok, Thailand. (in Thai).

[8] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi. 1999. *JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research*, In Journal of Acoustic Society of Japan, Vol. 20, No. 3.

[9] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young. 1994. *WSJCAM0 Corpus and Recording Description*, Cambridge University.

[10] P. Tarsaku, S. Kanokphara. 2002. *A study of HMM-based automatic segmentation for Thai continuous speech recognition system*, SNLP-O-COCOSDA.

ตารางที่ 3 ตารางแสดงการกระจายประโยคในแต่ละชุดสำหรับการบันทึกเสียงในแต่ละสถานที่

สถานที่บันทึกเสียง	***จำนวนผู้พูด***	***PD***	***TR***	***DT***	***ET***
เนคเทค	48 (ช.24 ญ24)	35	126	42	42
มหาวิทยาลัยสงขลานครินทร์	100 (ช.50 ญ.50)	20	101	50	50
มหาวิทยาลัยเทคโนโลยีมหานคร	100 (ช.50 ญ.50)	20	101	50	50

ตารางที่ 4 สัญลักษณ์แทนเสียงพยัญชนะต้นและพยัญชนะท้ายในภาษาไทย (26 หน่วยเสียง และ 12 หน่วยเสียง)

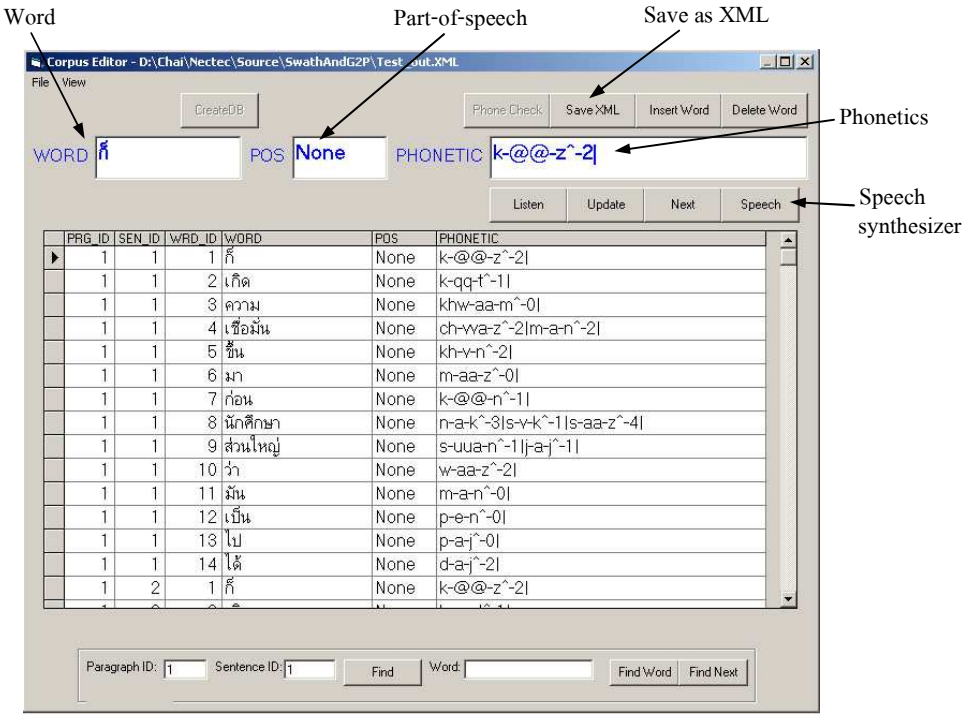
พยัญชนะ	***หน่วยเสียง***		***พยัญชนะ***	***หน่วยเสียง***	
	พยัญชนะต้น	*พยัญชนะท้าย*		*พยัญชนะต้น*	*พยัญชนะท้าย*
ก	k	k^	บ	b	p^
ข,ค,ฆ	kh	k^	ป	p	p^
ง	ng	ng^	ผ,พ,ภ	ph	p^
จ	c	t^	ฝ,ฟ	f	p^
ฉ,ฌ,ช	ch	t^	ม	m	m^
ซ,ศ,ษ,ส	s	t^	ร	r	n^
ญ,ย	j	j^	ล,ฬ	l	n^
ฎ,ด	d	t^	ว	w	w^
ฏ,ต	t	t^	ห,ฮ	h	-
ฐ,ฑ,ฒ,ถ,ท,ธ	th	t^	อ	z	-
ณ,น	n	n^	*หน่วยเสียง*	br,bl,fr,fl,dr	f^,s^,ch^,l^

ตารางที่ 5 สัญลักษณ์แทนเสียงพยัญชนะควบกล้ำในภาษาไทย (12 หน่วยเสียง)

พยัญชนะควบ	***หน่วยเสียง***	***พยัญชนะควบ***	***หน่วยเสียง***
ปร	pr	กร	kr
ปล	pl	กล	kl
พร	phr	กว	kw
พล	phl	คร	khr
ตร	tr	คล	khl
ทร	thr	ขว	khw

ตารางที่ 6 สัญลักษณ์แทนเสียงสระในภาษาไทย (24 หน่วยเสียง)

ตำแหน่งลิ้น / ***ความสูงของลิ้น***	***หน้า (สั้น/ยาว)***	***กลาง (สั้น/ยาว)***	***หลัง(สั้น/ยาว)***
สูง	i, ii (อิ, อี)	v, vv (อึ, อือ)	u, uu (อุ, อู)
กลาง	e, ee (เอะ, เอ)	q, qq (เออะ, เออ)	o, oo (โอะ, โอ)
ต่ำ	x, xx (แอะ, แอ)	a, aa (อะ, อา)	@, @@ (เอาะ, ออ)
สระผสม	ia, iia (เอียะ, เอีย)	va, vva (เอือะ, อัวะ)	ua, uua (อัวะ, อัว)



รูปที่ 1 ตัวอย่างหน้าจอของเครื่องมือแก้ไขฐานข้อมูล (Corpus editor)

ภัชริกา คชสำโรง
สำเร็จการศึกษาระดับปริญญาตรี สาขาภาษาศาสตร์ คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์ ในปี พ.ศ. 2540 และ ระดับปริญญาโท สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรีระดับปริญญาโท สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี ในปี พ.ศ. 2547 ปัจจุบันดำรงตำแหน่ง ผู้ช่วยนักวิจัย งานเทคโนโลยีเสียงพูด ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ มีความสนใจทางด้านภาษาศาสตร์ และ Speech corpus

ตรีภพ สรรเพชรนิยม
จบการศึกษาระดับปริญญาตรี สาขา คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ปัจจุบันดำรงตำแหน่ง ผู้ช่วยนักวิจัย งานเทคโนโลยีเสียงพูด ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ มีความสนใจ ทางด้าน Sound engineer และ Speech corpus

ศวิต กาสุริยะ
สำเร็จการศึกษาปริญญาโททางวิศวกรรมศาสตร์สาขาวิศวกรรมไฟฟ้าจากจุฬาลงกรณ์มหาวิทยาลัย เมื่อปี 2543 และได้ เข้ามาทำงานที่ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติในตำแหน่งผู้ช่วยนักวิจัย สังกัดงานวิจัยเทคโนโลยีเสียงพูด ฝ่ายวิจัยและพัฒนาเทคโนโลยีสารสนเทศ มีความสนใจในงานวิจัยทางด้านเทคโนโลยีเสียงพูด การรู้จำแบบ การประมวลผลสัญญาณ และปัญญาประดิษฐ์

ณัฐนันท์ ทัดพิทักษ์กุล
สำเร็จการศึกษาระดับปริญญาโทด้านวิศวกรรมไฟฟ้า จากมหาวิทยาลัยเทคโนโลยีสุรนารี ในปี 2545 ปัจจุบันดำรงตำแหน่งเป็นผู้ช่วยนักวิจัย ประจำศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ มีความสนใจทางด้าน Speech Recognition, Wavelet Transform และ Digital Signal Processing

ดร. ชัย วุฒิวิวัฒน์ชัย
สำเร็จการศึกษาระดับปริญญาตรี ด้านวิศวกรรมไฟฟ้า เกียรตินิยมอันดับหนึ่ง จากมหาวิทยาลัยธรรมศาสตร์ ในปีพ.ศ. 2537 และระดับปริญญาโทสาขาเดียวกันจากจุฬาลงกรณ์มหาวิทยาลัยในปี พ.ศ. 2541 ในปี พ.ศ. 2544 ได้รับทุนการศึกษาจากรัฐบาลญี่ปุ่นไปศึกษาต่อระดับปริญญาเอก ณ Tokyo Institute of Technology และได้รับปริญญาเอกในปี พ.ศ. 2547 ปัจจุบันดำรงตำแหน่งหัวหน้าส่วนงานเทคโนโลยีเสียงพูด ประจำศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ มีความสนใจทางด้าน Speaker recognition, Natural language processing และ Human-machine interaction.