เอกสารประกอบฐานข้อมูล

LOTUS Corpus

(**L**arge v**O**cabulary **T**hai contino**U**os **S**peech recognition **Corpus**)

สารบัญ

เอกสารประกอบฐานข้อมูล

1. บทนำ

ฐานข้อมูลเสียงขนาดใหญ่และมีคำศัพท์จำนวนมากมีความจำเป็นสำหรับการพัฒนาระบบรู้จำเสียงพูดต่อเนื่อง (Large Vocabulary Continuous Speech Recognition: LVCSR) ครอบคลุมถึงระบบพูดโต้ตอบอัตโนมัติ (Spoken Dialogue System), ระบบพูดแทนพิมพ์ (Speech Dictation) และระบบถอดความข่าว (Broadcast News Transcriber) บทความฉบับนี้เป็นเอกสารประกอบการสร้างฐานข้อมูลเสียงพูดภาษาไทยขนาดใหญ่เพื่อใช้ในงานวิจัยและพัฒนาระบบ LVCSR สำหรับภาษาไทย โดย มุ่งเน้นสำหรับพัฒนาระบบ Speech Dictation ซึ่งใช้ลักษณะการพูดแบบอ่าน (Reading stye)

ฐานข้อมูลประกอบด้วยชุดหน่วยเสียงสมดุล (Phonetically Balanced Set) ใช้สำหรับการฝึกฝน Acoustic Model, การกำกับหน่วยเสียงอัตโนมัติ (Automatic Phoneme Labeler) และเป็นชุดเสียงสำหรับการทดลองระบบที่มีปรับผู้พูด (Speaker Adaptation) ฐานข้อมูลยังประกอบด้วยชุดเสียงอีก 3 ชุดสำหรับฝึกฝน Acoustic Model และ Language Model ชุดสำหรับทดสอบเพื่อการพัฒนา และชุดสำหรับทดสอบเพื่อประเมินผล ฐานข้อมูลเสียงทั้ง 3 ชุดจะครอบคลุมคำศัพท์ภาษาไทย จำนวนไม่ต่ำกว่า 5,000 คำ[1] จากฐานข้อมูลบทความข่าวหรือบทความทั่วไป

ฐานข้อมูลยังประกอบด้วยเสียงพูดผ่านไมโครโฟน 2 ประเภท ประเภทแรกเป็นไมโครโฟน Close-talk คุณภาพสูง แบบ Unidirectional ระดับคุณภาพปานกลาง และทำการบันทึกเสียงใน 2 สภาพแวดล้อม คือ สภาพแวดล้อมแบบห้องเงียบ และ สภาพแวดล้อมแบบสำนักงาน โดยเก็บข้อมูลเสียงผ่าน Digital Audio Tape (DAT) ก่อนแปลงเป็นไฟล์อิเล็กทรอนิกส์ โดยมีผู้พูดทั้งเพศชายและหญิงในจำนวนเท่ากัน

ฐานข้อมูลนี้จะเป็นประโยชน์สำหรับนักวิจัย นักศึกษา และผู้ที่สนใจที่จะนำไปใช้ในการวิจัยและพัฒนาเทคโนโลยีทางด้านเสียงภาษาไทย และนำไปใช้ในการพัฒนาระบบต้นแบบ Speech recognition สำหรับภาษาไทย

2. รายละเอียดฐานข้อมูล

2.1 ชุดประโยค

ประโยค[2]ที่ใช้ในการอัดเสียงในฐานข้อมูลนี้จะถูกตัดเครื่องหมายหรือสัญลักษณ์พิเศษออกทั้งหมด (Non-Verbal) รายละเอียด จะกล่าวถึงในหัวข้อ *ข้อกำหนดบางประการของฐานข้อมูล* ซึ่งเป็นการจัดการข้อมูลเบื้องต้นก่อนที่จะนำประโยคมาบันทึกเสียง โดยมีการออกแบบชุดข้อมูลที่จะบันทึกเสียงแบ่งเป็นชุดประโยค 2 ชุดใหญ่ๆ ตามวัตถุประสงค์การนำไปใช้ ดังนี้

(1) ชุดหน่วยเสียงสมดุล (Phonetically Distributed Set) - PD

ใช้สำหรับการฝึกฝน Acoustic Model ขั้นต้น, ใช้ในการสร้างโปรแกรมกำกับขอบเขตหน่วยเสียง (Automatic Phoneme Alignment) หรือใช้ในการวิจัยเกี่ยวกับระบบแบบปรับผู้พูด (Speaker Adaptation) โดยประโยคในชุดหน่วยเสียงสมดุลจะ

[1] "คำ" ในที่นี้มีนิยามตามพจนานุกรมอิเล็กทรอนิกส์ที่ใช้ในขั้นตอนตัดคำ (Word Segmentation) ได้แก่ LEXITRON และ RI เป็นต้น

[2] "ประโยค" ในที่นี้อาจครอบคลุมถึงวลีหรือส่วนของประโยคได้ ทั้งนี้ขึ้นกับนิยามของการตัดประโยค

ครอบคลุมการเกิดของ “หน่วยเสียงคู่” (Biphone) ที่เกิดขึ้นในฐานข้อมูลข้อความภาษาไทยทั้งภายในพยางค์ ระหว่างพยางค์ และระหว่างคำ โดยไม่คำนึงถึงระดับเสียงวรรณยุกต์ (Tonal Level) การเกิดหน่วยเสียงคู่ในชุดนี้จะมีการกระจายสอดคล้องกับบทความที่ใช้ในการคัดเลือก โดยที่การคัดเลือกประโยคในชุด PD จะทำการคัดเลือกจากประโยคชุด PB (Phoneticlly Balance Set) ซึ่งประโยคชุด PB จะประกอบด้วยชุดประโยคที่มีหน่วยเสียงคู่ครบตามที่เกิดขึ้นในฐานข้อมูลบทความและมีการเกิดอย่างสมดุล โดยคัดเลือกขึ้นมาจากประโยคทั้งหมดทีละคู่ของหน่วยเสียงจนครบ จากนั้นจึงนำประโยคชุด PB มาทำการ คัดเลือกชุด PD โดยมีการคำนวณหาหน่วยเสียงคู่ที่เกิดขึ้นทั้งหมดในฐานข้อมูลข้อความ ORCHID ก่อนทำการคัดเลือก ประโยคเพิ่มเติมเข้าไปจนกว่าจะได้ประโยคที่มีการเกิดของหน่วยเสียงคู่ครบตามที่เกิดขึ้นจริง ในฐานข้อมูลข้อความ ORCHID และมีจำนวนประโยคน้อยที่สุด รายละเอียดในการคัดเลือกประโยคชุด PB และ PD สามารถอ่านเพิ่มเติมได้ใน C. Wutiwiwatchai, 2002.

คุณสมบัติของชุดหน่วยเสียงสมดุล (Phonetically Distributed Set) - PD

- ประกอบด้วยประโยคภาษาไทย ซึ่งคัดจากฐานข้อมูลข้อความ ORCHID
- ครอบคลุมการเกิด “หน่วยเสียงคู่” (Biphone)[3] ในภาษาไทยทั้งภายในพยางค์ ระหว่างพยางค์ และระหว่างคำ โดยไม่คำนึงถึงระดับเสียง (Tonal Level) (นิยามหน่วยเสียงเดี่ยวและหน่วยเสียงคู่สำหรับภาษาไทยแสดงไว้ในภาคผนวก)
- ครอบคลุมการเกิดหน่วยเสียงคู่โดยมีการกระจายสอดคล้องกับบทความที่ใช้ในการคัดเลือก
- จำนวนประโยค 398 ประโยค

(2) ชุดประโยคที่ครอบคลุมคำศัพท์ภาษาไทยจำนวน 5,000 คำ

ออกแบบมามาเพื่อฝึกฝนและพัฒนา Language Model สำหรับภาษาไทย ได้จากการเลือกประโยคที่ประกอบด้วยคำศัพท์ที่มีสถิติการใช้สูงสุด 5,000 ลำดับแรกจากคลังข้อความ รายละเอียดข้อมูลชุดคำศัพท์ได้แสดงไว้ในตารางที่ 1 โดยแบ่งข้อมูลออกเป็น 3 ชุด ดังต่อไปนี้

(2.1) ชุดฝึกฝน (Training Set) – TR

ชุดฝึกฝน ถูกคัดเลือกมาเพื่อใช้ในการฝึกฝน Acoustic Model เพิ่มเติมและฝึกฝน Language Model

คุณสมบัติของชุดฝึกฝน (Training Set) – TR

- ประกอบด้วยประโยคคัดจากฐานข้อมูลบทความขนาดใหญ่ ซึ่งอยู่ในขอบเขตของบทความทั่วไป หรือบทความข่าวขึ้นอยู่กับฐานข้อมูลบทความที่มี
- ครอบคลุมคำศัพท์ที่แสดงในพจนานุกรมที่ใช้ในการตัดคำจำนวนไม่ต่ำกว่า 5,000 คำ
- ข้อกำหนดของการคัดประโยคคือ มีความยาวของประโยคและค่า Perplexity[4] อยู่ในระดับปานกลาง โดยที่ค่า Perplexity ของประโยคแต่ละประโยคสามารถคำนวณได้จากแบบจำลองทางภาษา (Language Model) ชนิด Bigram[5] ของฐานข้อมูลบทความขนาดใหญ่ที่จัดไว้สำหรับการฝึกฝน Language Model
- ประกอบด้วยจำนวนประโยค 3,007 ประโยค

(2.2) ชุดทดสอบเพื่อพัฒนา (Development Test Set) - DT

---

[3] จำนวนหน่วยเสียงที่ต้องการให้ครอบคลุมมีตั้งแต่ หน่วยเสียงเดี่ยว (Monophone), หน่วยเสียงคู่ (Biphone) และหน่วยเสียงสาม (Triphone) เป็นต้น การเลือกจำนวนหน่วยเสียงขึ้นอยู่กับลักษณะความต่อเนื่องของการพูดในภาษานั้นๆ และขึ้นกับจำนวนข้อมูลบทความที่มี

[4] Perplexity เป็นค่าตัวเลขที่บ่งบอกถึงระดับความซับซ้อนของโครงสร้างภาษา

[5] จำนวน N ใน Language Model ชนิด N-gram เลือกโดยพิจารณาจากขนาดของฐานข้อมูลบทความที่มี

ใช้ในขั้นตอนการวิจัยระบบรู้จำเสียงพูด ชุดทดสอบนี้ ถูกคัดเลือกมาเพื่อใช้ในขั้นตอนการวิจัยระบบรู้จำ

คุณสมบัติ ของชุดทดสอบเพื่อพัฒนา (Development Test Set) - DT

- ประกอบด้วยประโยคที่มีคุณสมบัติเช่นเดียวกับประโยคในชุดฝึกฝนทั้งทางด้านความยาวของประโยค ค่า Perplexity และประกอบด้วยคำศัพท์ที่อยู่ในกลุ่มคำศัพท์ 5,000 คำที่มีในชุดฝึกฝน
- ประกอบด้วยประโยคจำนวน 500 ประโยค

(2.3) *ชุดทดสอบเพื่อประเมิน (Evaluation Test Set)* - ET

ชุดทดสอบนี้ใช้สำหรับการทดสอบขั้นสุดท้ายเพื่อประเมินความสามารถของระบบรู้จำ

คุณสมบัติ ของชุดทดสอบเพื่อประเมิน (Evaluation Test Set) – ET

- รายละเอียดเช่นเดียวกับชุดทดสอบ สำหรับการพัฒนา (DT) ทุกประการ
- ประกอบด้วยประโยคจำนวน 500 ประโยค

ตารางที่ 1 ตารางสรุปรายละเอียดของข้อมูลชุดต่างๆ ในฐานข้อมูล

รายละเอียดของฐานข้อมูล	*PD set*	*TR set*	*DT set*	*ET set*
จำนวนประโยค	801	3,007	500	500
จำนวนคำศัพท์	2,269	5,000	1,622	1,630
จำนวนคำ	7,847	55,504	8,076	8,290

2.2 ผู้พูด

ฐานข้อมูลเสียงพูดภาษาไทยขนาดใหญ่นี้ ได้พัฒนาขึ้นจากความร่วมมือของ 3 สถาบัน คือ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ มหาวิทยาลัยสงขลานครินทร์ และมหาวิทยาลัยเทคโนโลยีมหานคร โดยที่มหาวิทยาลัยทั้ง 2 แห่ง ดำเนินการบันทึกเสียงผู้พูด แห่งละ 100 คน โดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ บันทึกเสียงผู้พูด จำนวน 48 คน รวมผู้พูดทั้งสิ้น 248 คน (ผู้ชาย 124 คน, ผู้หญิง 124 คน) ผู้พูดแต่ละคนจะต้องบันทึกเสียงชุด PD และ TR หรือ DT หรือ ET ชุดใดชุดหนึ่งเท่านั้น

ตารางที่ 2 ตารางแสดงจำนวนผู้พูดในการบันทึกเสียงแต่ละสถานที่

สถานที่บันทึกเสียง	*ชาย*	*หญิง*	*รวม*
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (a)	24	24	48
มหาวิทยาลัยสงขลานครินทร์ (b)	50	50	100
มหาวิทยาลัยเทคโนโลยีมหานคร (c)	50	50	100

2.3 การจัดแบ่งชุดประโยคสำหรับการบันทึกเสียง

เนื่องจากจำนวนผู้พูดในแต่ละที่ที่ทำการบันทึกเสียงแตกต่างกัน จำนวนประโยคที่ผู้พูดต้องทำการบันทึกเสียงในแต่ละชุดนั้นก็จะแตกต่างกันไปด้วย เพื่อให้มีการกระจายการอ่านประโยคใดๆ ในแต่ละชุด ไม่ต่ำกว่า 1 ครั้ง ในแต่ละแห่งที่ทำการบันทึกเสียง โดยนำประโยคทั้งหมดมาจัดชุด โดยรายละเอียดการจัดชุดประโยคที่ทำการบันทึกเสียงในแต่ละแห่งมีรายละเอียดดังนี้

***ตารางที่ 3* ตารางแสดงการกระจายประโยคในแต่ละชุดสำหรับการบันทึกเสียงในแต่ละสถานที่**

สถานที่บันทึกเสียง	*PD*	*TR*	*DT*	*ET*
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (a)	35	126	42	42
มหาวิทยาลัยสงขลานครินทร์ (b)	20	101	50	50
มหาวิทยาลัยเทคโนโลยีมหานคร (c)	20	101	50	50

ผู้พูดในแต่ละที่จะถูกจัดแบ่งเป็น 3 กลุ่ม โดยมีการจัดแบ่งผู้พูดและชุดประโยคออกเป็นกลุ่มๆ แต่ละกลุ่มจะได้รับการอัดเสียงประโยคต่างกันดังนี้

กลุ่มที่ a1 : จำนวน 24 คน (ชาย 12 คน หญิง 12 คน)
- 35 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 126 ประโยคจากชุดประโยค TR คัดแบบสุ่ม

กลุ่มที่ a2 : จำนวน 12 คน (ชาย 6 คน หญิง 6 คน)
- 35 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 42 ประโยคจากชุดประโยค DT คัดแบบสุ่ม

กลุ่มที่ a3 : จำนวน 12 คน (ชาย 6 คน หญิง 6 คน)
- 35 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 42 ประโยคจากชุดประโยค ET คัดแบบสุ่ม

กลุ่มที่ b1: จำนวน 60 คน (ชาย 30 คน หญิง 30 คน)
- 20 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 101 ประโยคจากชุดประโยค TR คัดแบบสุ่ม

กลุ่มที่ b2: จำนวน 20 คน (ชาย 10 คน หญิง 10 คน)
- 20 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 50 ประโยคจากชุดประโยค DT คัดแบบสุ่ม

กลุ่มที่ b3: จำนวน 20 คน (ชาย 10 คน หญิง 10 คน)
- 20 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 50 ประโยคจากชุดประโยค ET คัดแบบสุ่ม

กลุ่มที่ c1: จำนวน 60 คน (ชาย 30 คน หญิง 30 คน)
- 20 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 101 ประโยคจากชุดประโยค TR คัดแบบสุ่ม

กลุ่มที่ c2: จำนวน 20 คน (ชาย 10 คน หญิง 10 คน)
- 20 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 50 ประโยคจากชุดประโยค DT คัดแบบสุ่ม

กลุ่มที่ c3: จำนวน 20 คน (ชาย 10 คน หญิง 10 คน)
- 20 ประโยคจากชุดประโยค PD คัดแบบสุ่ม
- 50 ประโยคจากชุดประโยค ET คัดแบบสุ่ม

รวมข้อมูลเสียงที่ได้จากการบันทึกเสียงในแต่ละแห่งทุกชุด *รวมทั้งสิ้น 51,664 ประโยค*

2.4 การบันทึกเสียง

(1) รายละเอียดการบันทึกเสียง (Recording Session)

ทำการอัดเสียง 2 รอบสำหรับผู้พูดแต่ละคน

- รอบแรกเป็นการอัดเสียงในสภาพแวดล้อมห้องเงียบ ผ่านไมโครโฟน 2 ตัวพร้อมกัน คือแบบ Dynamic Close-talk (TELEX H-41) ระดับคุณภาพสูง และแบบ Dynamic Unidirectional (SONY F-720) ระดับคุณภาพปานกลาง ข้อมูลเสียงที่ได้ในรอบนี้เป็น Clean Speech (CC สำหรับชุดที่ใช้ Close-talk และ CU สำหรับชุดที่ใช้ Unidirectional)
- รอบที่สองเป็นการอัดเสียงในสภาพแวดล้อมสำนักงานทั่วไป ผ่านไมโครโฟน 2 ตัวพร้อมกัน คือแบบ Dynamic Close-talk (TELEX H-41) และแบบ Dynamic Unidirectional (SONY F-720) ระดับคุณภาพปานกลาง ข้อมูลเสียงที่ได้ในรอบนี้เรียกว่าชุด Office Environment Speech (OC สำหรับชุดที่ใช้ Close-talk และ OU สำหรับชุดที่ใช้ Unidirectional)

ข้อมูลจะถูกเก็บในรูปแบบ DAT[6] ก่อนนำมาเชื่อมต่อเข้าคอมพิวเตอร์เพื่อแปลงสัญญาณเสียงเป็นไฟล์อิเล็กทรอนิกส์มาตรฐาน ซึ่งจะกล่าวในรายละเอียดต่อไปในหัวข้อ *รูปแบบของไฟล์*

ในแต่ละรอบของผู้พูดแต่ละคนจะได้รับการทดสอบและรับคำแนะนำเพื่อปรับวิธีการพูด ความดังในการพูดพร้อมทั้งแนะนำ ขั้นตอนในการพูดเพื่ออัดเสียง ในการอัดเสียงสำหรับผู้พูดแต่ละคน จะทำการอัดเสียงสภาพแวดล้อมก่อนเป็นเวลา 3 วินาที แล้วจึงเริ่มอัดเสียงพูดในลักษณะอ่าน (Reading) ประโยคดังที่กล่าวข้างต้น

การพูดจะพูดทีละประโยคโดยอ่านตามรูปอ่าน (Orthography) ของประโยคที่กำหนดให้และเว้นระยะระหว่างประโยค รายละเอียดของรูปอ่านของประโยคจะแสดงตัวอย่างให้เห็นในหัวข้อรูปแบบของไฟล์ ผู้พูดสามารถอัดเสียงประโยคเดิม ทับเสียงเก่าได้หากไม่พอใจในเสียงที่พูดหรือพูดผิดพลาด หลังจากนั้นจึงกรอกแบบสอบถามรายละเอียดของผู้พูด

(2) อุปกรณ์ในการอัดเสียง (Recording Equipment)

<u>ห้องอัดเสียง</u>

- สำหรับการอัดเสียงชุด Clean Speech (CL) ห้องอัดเสียงจะเป็นห้องเงียบ
- สำหรับการอัดเสียงชุด Office Environment (OF1 และ OF2) ห้องอัดเสียงจะเป็นสภาพแวดล้อมแบบสำนักงานทั่วไป เพื่อให้เกิดความหลากหลายของสภาพแวดล้อม จะกระจายสถานที่ในการอัดเสียงไปอย่างต่ำ 5 แห่ง

<u>ไมโครโฟน</u>

- สำหรับการอัดเสียงชุด CL จะใช้ไมโครโฟนชนิด Dynamic Close-Talk ระดับคุณภาพสูง
- สำหรับการอัดเสียงชุด OF1 จะกำหนดให้ใช้ไมโครโฟนประเภท Dynamic Close-talk คุณภาพปานกลาง
- สำหรับการอัดเสียงชุด OF2 จะกำหนดให้ใช้ไมโครโฟนประเภท Dynamic Unidirectional คุณภาพปานกลาง

อัดเสียงพูดผ่านไมโครโฟนซึ่งต่อเข้าหาเครื่องบันทึกเสียงระบบ DAT หลังจากนั้นจึงต่อเครื่องบันทึกเสียงระบบ DAT เข้าหาเครื่องคอมพิวเตอร์ผ่านการ์ดเสียงโดยใช้สายเชื่อมต่อชนิด Optic หรือ Coaxial หรือใช้ DAT-Link เพื่อแปลง สัญญาณเสียงจาก DAT มาเป็นไฟล์อิเล็กทรอนิกส์มาตรฐานแยกประโยคละหนึ่งไฟล์โดยใช้ซอฟต์แวร์หรือโปรแกรม ในการแปลงสัญญาณเสียง

2.5 โครงสร้างฐานข้อมูล

(1) *ไฟล์ที่ประกอบอยู่ในฐานข้อมูล*

[6] Digital Audio Tape (DAT) เป็นเทคโนโลยีในการเก็บข้อมูลเสียงดิจิตอล โดยไม่มีการบีบอัดใดๆ ทั้งสิ้น

ไฟล์ข้อมูล : รูปแบบของชื่อไฟล์คือ <M><E><G><CCC> _<S><U><DDD>_<XXX>.<YYY> มีรายละเอียดดังตารางที่3 ตัวอย่าง ข้อมูลไฟล์เสียงชุด PD สภาพแวดล้อมแบบห้องเงียบ ที่บันทึกโดยเนคเทค

CCM001_Pa001_001.wav

ตารางที่ 3 ตารางแสดงรายละเอียดของชื่อไฟล์ที่ประกอบอยู่ในฐานข้อมูล

ตำแหน่งชื่อ	**ความหมาย**	**รายละเอียด**
M	ชนิดของไมโครโฟน	C = Dynamic Close-talk U = Dynamic Unidirectional
E	สภาพแวดล้อมที่ทำการบันทึกเสียง	C = สภาพแวดล้อมแบบห้องเงียบ, O = สภาพแวดล้อมแบบสำนักงาน
G	เพศของผู้พูด	M = ผู้ชาย F = ผู้หญิง
CCC	ID ของผู้พูด	001-999
S	ชนิดของชุดประโยค	P คือประโยคจากชุดประโยค PB T คือประโยคจากชุดประโยค TR D คือประโยคจากชุดประโยค DT E คือประโยคจากชุดประโยค ET
U	Code ของแหล่งที่อัด	a = NECTEC b = PSU c = MUT
DDD	ID ของชุดประโยค	001 – 999
XXX	ID ของประโยค	001 – 999
YYY	ชนิดของไฟล์ข้อมูล	.wav คือไฟล์สัญญาณเสียง .lab คือไฟล์กำกับหน่วยเสียง

สำหรับไฟล์คำอ่านประกอบประโยค ของแต่ละไฟล์เสียงสามารถดูได้จากไฟล์ XXsen.txt โดยอ้างอิงจาก index.txt ดังนี้ เช่น ไฟล์เสียง CCMxxx_***Pa001_001***.wav สามารถไปดูคำอ่านประโยคได้จากไฟล์ Pdsen.txt เลขที่ประโยคที่ pd001

****

ชุดข้อมูล ลำดับที่ เลขที่ประโยค (ใช้อ้างอิงประกอบกับ XXsen.txt)

****

Pa001 **001** pd001

ไฟล์องค์ประกอบอื่นๆ :

นอกเหนือจากไฟล์ข้อมูลที่กล่าวมาแล้ว ฐานข้อมูลยังจะต้องประกอบด้วย

alltext5k.txt = คลังข้อความครอบคลุมคำศัพท์ 5,000 คำ ก่อนคัดเลือกมาเป็นชุด TR, DT และ ET ใช้ในการฝึกฝนแบบจำลองภาษา

dic5k.txt = ไฟล์พจนานุกรมการออกเสียงของคำศัพท์ที่ประกอบอยู่ในชุดประโยค TR, DT และ ET

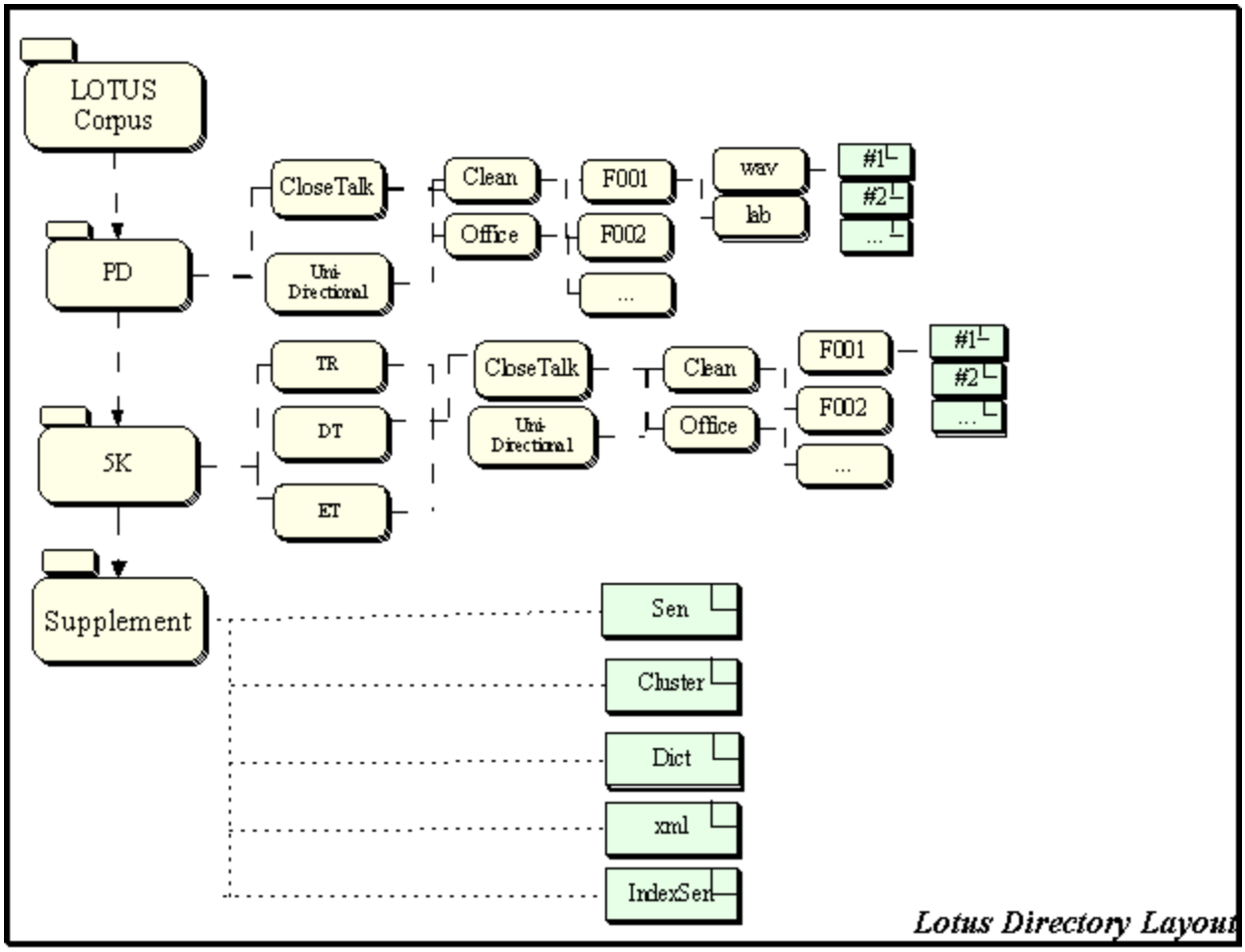
index.txt = ไฟล์แสดง index ของไฟล์เสียงและคำอ่านของประโยคที่ผู้พูดแต่ละคนได้พูด

spkinfo.txt = ไฟล์แสดงรายละเอียดของผู้พูดแต่ละคนพร้อม ID ของผู้พูด

XXsen.txt = ไฟล์แสดงรายละเอียดของประโยคในชุดประโยคต่างๆ พร้อม ID ของประโยค (XX แทนด้วยชุด PD, TR, DT, ET)

(2) *โครงสร้างไดเรคทอรี*

การจัดโครงสร้างไดเรคทอรีแสดงดังรูปที่ 1



รูปที่ 1 การจัดโครงสร้างไดเรคทอรี

2.6 รูปแบบของไฟล์

(1) ไฟล์สัญญาณเสียง (Speech Waveform)

เสียงพูดจะถูกบันทึกจากไมโครโฟน SONY F-720 Dynamic Microphone เข้าช่องสัญญาณขวา และจากไมโครโฟน TELEX H-41 เข้าช่องสัญญาณซ้าย เข้าสู่เครื่องบันทึกเสียง DAT SONY PCM-R300 ในรูปแบบ Sample Rate 48 kHz Resolution 16bit Channels Stereo (48 kHz 16bit Stereo) จากนั้นข้อมูลเสียงจะถูกถ่ายโอนส่งผ่านสาย Optical SPDIF ไปยัง Live! Drive ของ Sound Card รุ่น Sound Blaster Live! Platinum เพื่อบันทึกลงใน Computer โดยโปรแกรม Cool Edit 2000 ในรูปแบบ 48 kHz 16bit

Stereo Windows PCM จากนั้นใช้โปรแกรม Cool Edit 2000 Down Sample ลงเป็น 16 kHz 16bit Stereo Windows PCM โดยใช้ High Quality ที่ 999 และใช้ Pre/Post Filter

(2) ไฟล์แสดงรายละเอียดของประโยค (Sentences Transcription)
ไฟล์แสดงรายละเอียด ของประโยคเป็นไฟล์ข้อความ (Plane Text) ที่ประกอบด้วยรายการประโยคในแต่ละชุด พร้อมทั้ง Phonetics ของแต่ละประโยค โดยแยกเป็นชุดต่างๆ ดังนี้

<u>ชุดประโยค PD</u>

PDsen.txt

มีรูปแบบคือ <Index> <Sentence><Phonetics> <NEW-LINE>

ตัวอย่าง pd083 ศึกษา และ ปรึกษา ร่วมกัน s-v-k^-1 s-aa-z^-4|l-x-z^-3|pr-v-k^-1 s-aa-z^-4|r-uua-m^-2 k-a-n^-0|

โดยที่ <Index> จะนำไปไปใช้อ้างอิงในการกระจายชุดประโยคสำหรับการบันทึกเสียง และในส่วนของ <Sentence> จะมีการเว้นวรรคระหว่างคำ เพื่อแสดงถึงการตัดคำในประโยค ซึ่งจะสัมพันธ์กับส่วนของ <Phonetices> ซึ่งจะใช้สัญลักษณ์ | คั่นเพื่อแบ่งขอบเขตของคำในประโยค โดยในระดับพยางค์จะสังเกตุได้จากตัวเลขกำกับท้าย พยางค์และช่องว่างเมื่อมีการขึ้นพยางค์ใหม่ โดยตัวเลขกำกับท้ายพยางค์หมายถึงระดับวรรณยุกต์ของพยางค์นั้นๆ โดย
ตัวเลข 0 – 4 แสดงระดับวรรณยุกต์ ดังนี้

[-] = ไม่ใช่ตำแหน่งท้ายพยางค์
[0] = เสียงสามัญ (Middle Tone)
[1] = เสียงเอก (Low Tone)
[2] = เสียงโท (Falling Tone)
[3] = เสียงตรี (High Tone)
[4] = เสียงจัตวา (Rising Tone)

<u>ชุดประโยค TR, DT และ ET</u>

TRsen.txt

มีรูปแบบคือ <Index> <Sentence><Phonetics> <NEW-LINE>

ตัวอย่าง tr0009 ก็ คงจะ เห็น ได้ ไม่ ยาก k-@@-z^-2|kh-o-ng^-0 c-a-z^-1|h-e-n^-4|d-aa-j^-2|m-a-j^-2|j-aa-k^-2|

DTsen.txt

มีรูปแบบคือ <Index> <Sentence><Phonetics> <NEW-LINE>

ตัวอย่าง dt0024 เขา บอก ว่า ไม่ ต้อง หรอก kh-a-w^-4|b-@@-k^-1|w-aa-z^-2|m-a-j^-2|t-@-ng^-2|r-@@-k^-1|

ETsen.txt

มีรูปแบบคือ <Index> <Sentence><Phonetics> <NEW-LINE>

ตัวอย่าง et0019 แต่ ไม่ เก่ง ใน การ หา เงิน t-xx-z^-1|m-a-j^-2|k-e-ng^-1|n-a-j^-0|k-aa-n^-0|h-aa-z^-4|ng-q-n^-0|

รูปแบบของไฟล์กำกับจะคล้ายกับชุดประโยค PD แตกต่างกันที่ไม่มีการกำกับขอบเขตของคำหรือหน่วยเสียง และไม่มีการใส่ “sp” ระหว่างพยางค์ในประโยค

(3) ไฟล์พจนานุกรม (Dictionary or Lexicon)

ไฟล์พจนานุกรม dic5k.txt เป็นไฟล์ข้อความ (Plane Text) ประกอบด้วยคำศัพท์จำนวน 5,000 คำ ซึ่งครอบคลุมคำศัพท์ทั้งหมด ในชุดประโยค TR, DT และ ET พจนานุกรมนี้พัฒนามาจากพจนานุกรมอิเล็กทรอนิกส์ LEXITRON และ RI ของ NECTEC โดย คัดเฉพาะคำศัพท์ที่เกิดขึ้นในชุดประโยค และเพิ่มคำศัพท์ที่ไม่มีในพจนานุกรมแต่ปรากฏในชุดประโยค
รูปแบบของไฟล์พจนานุกรมคือ <Word> <Pronunciation with Tone> <NEW-LINE>
ส่วนหนึ่งของพจนานุกรมแสดงดังตัวอย่างต่อไปนี้

ก. k-@@-z^-0|
กฎ k-o-t^-1|
กฎหมาย k-o-t^-1 m-aa-j^-4|
กฎหมายอาญา k-o-t^-1 m-aa-j^-4 z-aa-z^-0 j-aa-z^-0|
กฎเกณฑ์ k-o-t^-1 k-ee-n^-0|
กด k-o-t^-1|
กดขี่ k-o-t^-1 kh-ii-z^-1|
กดดัน k-o-t^-1 d-a-n^-0|

หนึ่งบรรทัดจะแทนการออกเสียงหนึ่งเสียงโดยไม่สนใจรูปเขียน (Grapheme) คือ อาจมีคำที่มีรูปเขียนเหมือนกัน แต่รูปอ่าน (Phoneme) ไม่เหมือนกัน โดยจะมีการแบ่งพยางค์ ด้วยเครื่องหมาย “|” พร้อมทั้งระบุระดับเสียงวรรณยุกต์ที่ท้ายพยางค์ทุกพยางค์

(4) ไฟล์รายละเอียดผู้พูด (Speaker Information)

ไฟล์ spkinfo.txt เป็นรายละเอียดของผู้พูดแต่ละคน โดยจะให้รายละเอียดดังนี้
<แหล่งที่อัด> <Speaker-ID><เพศ><อายุ><ภูมิลำเนา>
ตัวอย่างเช่น

a	m001	ชาย	27	กรุงเทพฯ
a	m002	ชาย	30	นครราชสีมา
a	m003	ชาย	32	กรุงเทพฯ

(5) ไฟล์รายละเอียดประโยคทั้งหมดที่ครอบคลุมคำศัพท์ 5,000 คำ

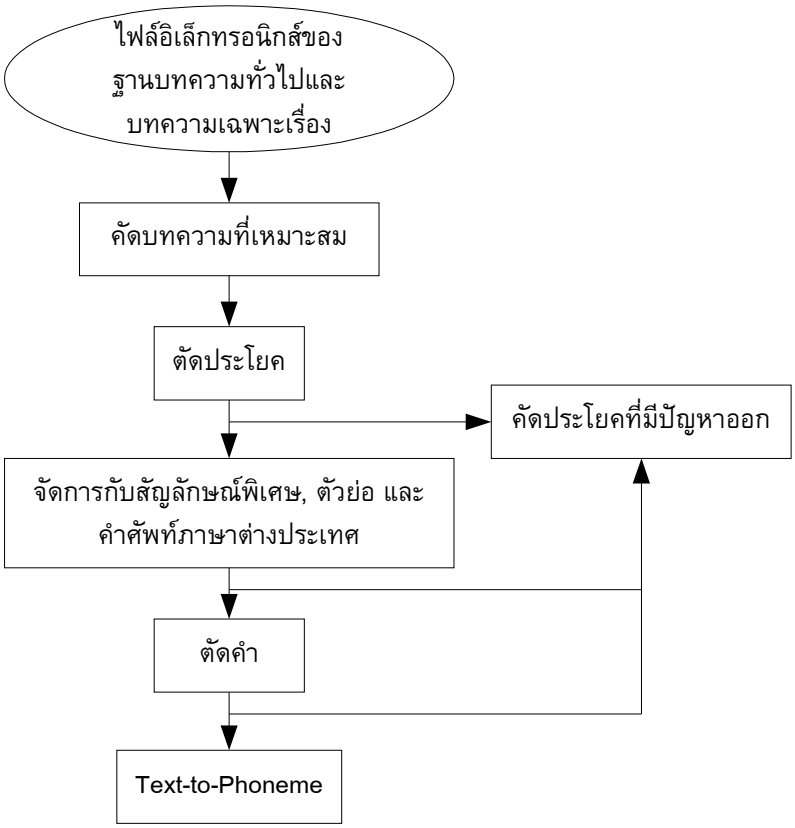
ไฟล์ alltext5k.txt ในขั้นตอนการคัดประโยคชุด TR, DT, ET ดังรูปที่ 4 จะมีการแบ่งชุดประโยคทั้งหมดออกเป็น 2 ส่วน 90% ใช้ในการคัด TR และ 10% ใช้ในการคัด DT และ ET ไฟล์นี้เป็นไฟล์ชุดที่ใช้คัด TR ซึ่งครอบคลุมคำศัพท์ 5,000 คำ ใน dic5k.txt จะใช้ไฟล์นี้ในการฝึกฝนแบบจำลองภาษาสำหรับการรู้จำได้

3. การสร้างฐานข้อมูล

ขั้นตอนการสร้างฐานข้อมูลแบ่งออกเป็น 3 ส่วนใหญ่ๆ คือ การจัดการบทความ การคัดประโยค, การจัดการเสียง และการสร้างองค์ประกอบอื่นๆ แต่ละส่วนมีรายละเอียดขั้นตอนดังต่อไปนี้

3.1 การจัดการบทความ

เป้าหมายของส่วนนี้คือ การแปลงบทความเป็นรูปหน่วยเสียง ขั้นตอนการคัดเลือกบทความแสดงดังรูปที่ 2

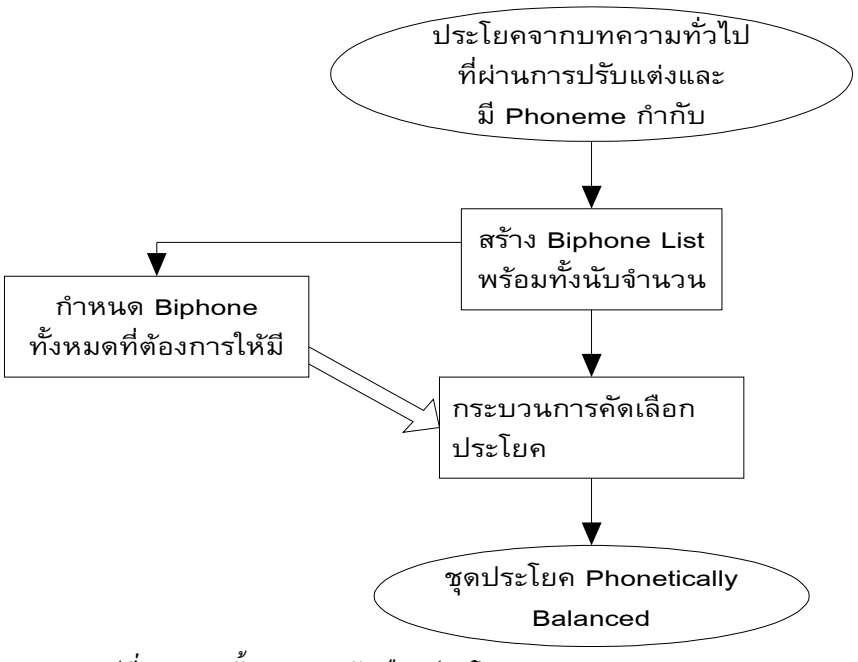


รูปที่ 2 แสดงขั้นตอนการคัดเลือกบทความก่อนนำมาแปลงเป็นหน่วยเสียงอ่าน

3.2 การคัดประโยค

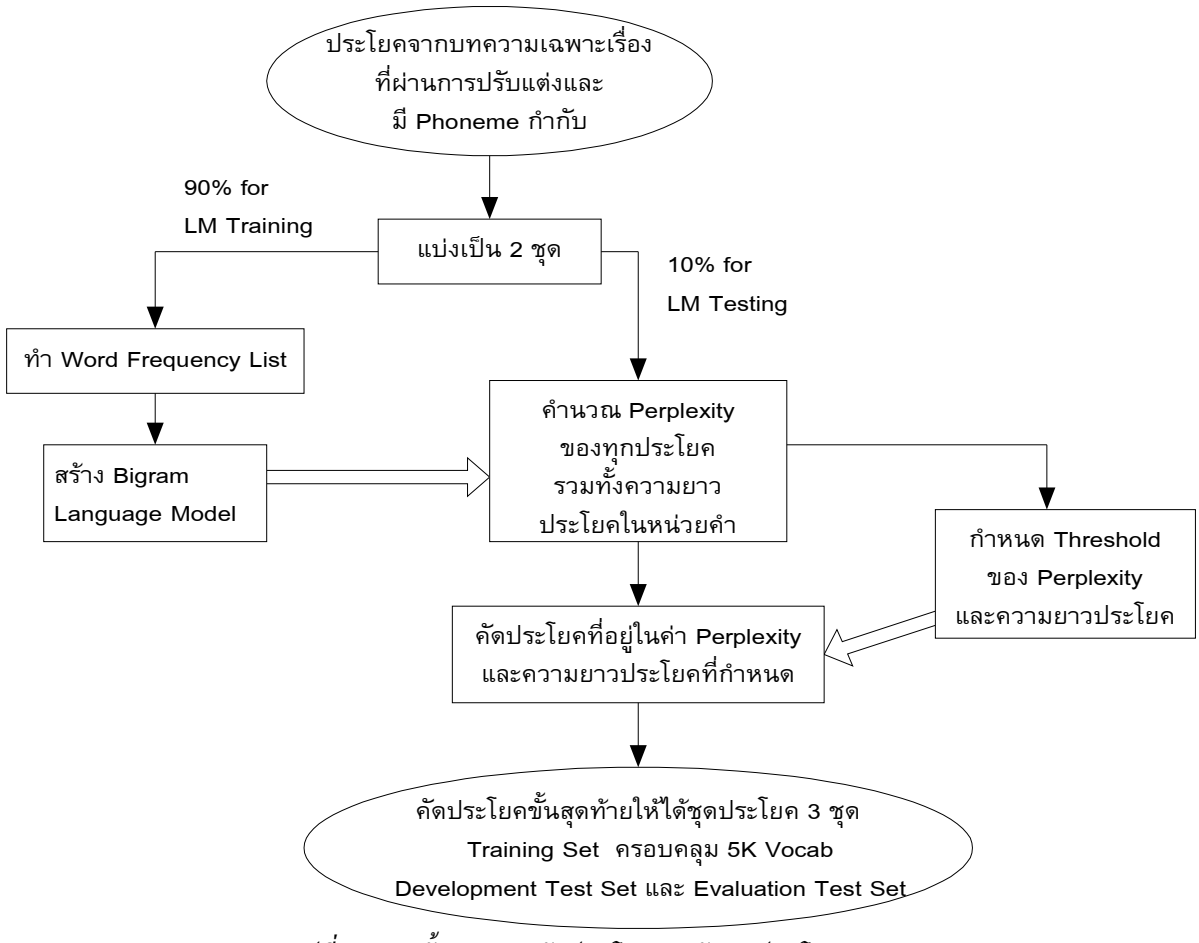
(1) การคัดประโยคสำหรับชุดประโยค PB

เมื่อได้ประโยคที่ทำการตัดประโยค ตัดคำ แล้วแปลงรูปเขียนเป็นคำอ่าน (Graphem to Phoneme) แล้วก็ทำการคัดเลือกประโยคชุด Phonetically Balanced ตามขั้นตอนดังรูปที่ 3 เมื่อได้ประโยคชุด PB แล้วจึงนำมาคัดเลือกประโยคเพิ่มเติมจนได้ชุดหน่วยเสียงสมดุล หรือ Phoneticcally Balanced Distribtuion set ต่อไป



รูปที่ 3 แสดงขั้นตอนการคัดเลือกประโยคชุด Phonetically Balanced

(2) การคัดประโยคสำหรับชุดประโยค TR, DT และ ET



รูปที่ 4 แสดงขั้นตอนการคัดประโยคสำหรับชุดประโยค TR, DT และ ET

3.3 การจัดการเสียง

(1) การบันทึกเสียง

ในส่วนนี้จะทำการแบ่งประโยคและจำนวนผู้พูดที่ต้องการให้พูดกระจายไปแต่ละสถานที่บันทึกเสียงตามรายละเอียดที่กล่าวไว้ในหัวข้อ *การบันทึกเสียง* ซึ่งเสียงที่ได้จากการบันทึกจะเก็บอยู่ใน DAT และนำมาแปลงเป็นไฟล์อิเล็กทรอนิกส์ต่อไป

(2) การกำกับขอบเขตเสียง

เสียงจาก DAT จะถูกแปลงเป็นไฟล์อิเล็กทรอนิกส์บนเครื่องคอมพิวเตอร์ตามรายละเอียดที่กล่าวในหัวข้อ *รูปแบบของไฟล์* และสร้างไฟล์หน่วยเสียงสำหรับแต่ละชุดประโยคโดยการกำกับจะทำโดยนักภาษาศาสตร์ ซึ่งจะทำการกำกับแบบกึ่งอัตโนมัติก่อนในขั้นต้น และตรวจสอบแก้ไขโดยนักภาษาศาสตร์อีกครั้ง

3.4 การสร้างองค์ประกอบอื่นๆ

ในส่วนนี้จะต้องสร้างองค์ประกอบอื่นๆ สำหรับฐานข้อมูล ได้แก่ ไฟล์พจนานุกรม ไฟล์รายละเอียดของผู้พูด รายละเอียดของประโยคและรายละเอียดประโยคที่ผู้พูดแต่ละคนได้พูด รวมทั้งจัดทำบทความเพื่ออธิบายรายละเอียดของฐานข้อมูลเพื่อเผยแพร่ต่อผู้สนใจต่อไป

3.5 ข้อกำหนดบางประการของชุดประโยคและการออกเสียง

(1) ข้อกำหนดสำหรับชุดประโยค

เนื่องจากประโยคเริ่มต้นได้มาจากฐานบทความ (Text Corpus) ทั่วไปหรือข่าว จำเป็นต้องมีการปรับแต่งและกำหนดการอ่าน ให้เป็นหนึ่งเดียวก่อนนำไปใช้สำหรับอัดเสียงจริง รายละเอียดบางประการของการปรับแต่งประโยคและข้อกำหนดของการอ่านได้แก่

- ปรับให้เป็นประโยคแบบ Non-Verbal กล่าวคือจะตัดสัญลักษณ์พิเศษออกหรือเปลี่ยนเป็นคำที่เหมาะสม สัญลักษณ์พิเศษได้แก่ Hyphen, ไปยาลน้อย, ไม้ยมก, ไปยาลใหญ่, Comma, Colon, Semicolon, Single quote, Double quote, Question mark, Exclamation mark, เครื่องหมายในการคำนวณต่างๆ เป็นต้น
- ประโยคในเครื่องหมาย “()” จะถูกตัดออกไปพร้อมกับเครื่องหมาย “()”
- แปลงรูปเขียนภาษาต่างประเทศเป็นคำภาษาไทยทับศัพท์
- คำย่อที่มิได้มีคำเต็มนำหน้าจะถูกเปลี่ยนเป็นคำเต็ม

(2) ข้อกำหนดสำหรับการอ่าน

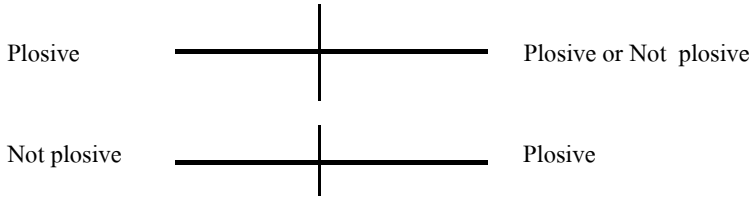
- การอ่านออกเสียงคำบางคำในภาษาไทยได้มีการเปลี่ยนแปลงไป หรือกล่าวได้ว่ารูปหน่วยเสียง (Phonetic) กับการออกหน่วยเสียง (Phonemic) แตกต่างกัน เช่นคำว่า “ท่าน” ซึ่งมีรูปหน่วยเสียงเป็น /th aa n^ 2/ แต่ ออกเสียงเป็น /th a n^ 2/ ฐานข้อมูลนี้จะกำหนดให้อ่านออกเสียงให้ตรงตามการออกหน่วยเสียง (Phonemic)
- คำพ้องรูป (Homograph) จะถูกกำหนดให้อ่านเพียงแบบเดียวก่อนการอัดเสียงจริง
- ปัจจุบันมักมีการออกเสียง “ร” เป็นเสียง “ล” หรือออกเสียงควบกล้ำไม่ได้ ซึ่งจะไม่มีการกำหนดให้ผู้พูดต้อง ออกเสียงให้ถูกต้อง
- เสียงทับศัพท์ภาษาต่างประเทศจะถูกกำหนดให้อ่านอย่างที่คนส่วนใหญ่อ่าน เช่นคำว่า “เอส” จะต้องอ่านออกเสียง /z ee s 3/ ตามแบบภาษาต่างประเทศ ไม่ใช่เสียง /z ee t^ 3/ ตามอย่างภาษาไทย[7]

[7] เสียงตัวสะกด /s^/, /ch^/, /l^/ และ /f^/ เกิดขึ้นในพยางค์ของภาษาต่างประเทศและไม่มีในภาษาไทย

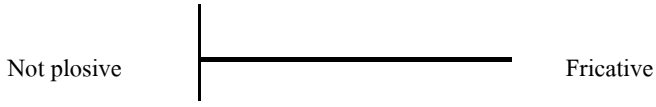
(3) ข้อกำหนดสำหรับการกำกับหน่วยเสียง

- เว้นระยะเสียงว่าง (silence) ตอนต้นและท้ายไฟล์ ประมาณ 300 ms
- เริ่มต้นและลงท้ายของทุกประโยคต้องมีการกำหนดช่วง silence /sil/ โดยที่ตอนต้นประโยคให้สิ้นสุดเสียง /sil/ ที่จุดเริ่มต้นของหน่วยเสียงแรกตอนท้ายประโยคตั้งแต่สิ้นสุดหน่วยเสียงสุดท้ายเป็นต้นไป
- ให้ตัดแบ่งแต่ละหน่วยเสียงโดยดูจากความเปลี่ยนแปลงของลักษณะคลื่นเสียง และสเปกโตรแกรมพิจารณาร่วมกันประกอบกับการฟัง โดยให้ตัดที่จุดเปลี่ยนแปลงของลักษณะเส้นเสียง โดยจุดตัดจะเลยจากจุดที่เปลี่ยนแปลงประมาณ 1-2 คลื่น ดังรูปที่ 1 และรูปที่ 2 โดยให้กำกับที่จุด zero crossing
- ถ้าระหว่าง phoneme มี **ช่องว่างระหว่างเสียง และช่องว่างนั้นไม่เกิน 20 มิลลิวินาที (ms)** ให้พิจารณาตามหลักเกณฑ์ต่อไปนี้
    - ถ้าเสียงข้างหน้าเป็นเสียงสระ ให้ตัดที่จุดสิ้นสุดของสระ แล้วทิ้งช่องว่างที่เหลือให้เป็นส่วนของเสียงพยัญชนะ

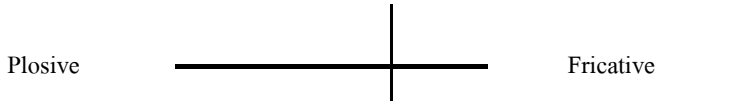
    - พยัญชนะกลุ่ม plosive ต่อกับกลุ่ม plosive ด้วยกัน หรือ พยัญชนะใดๆ ให้ตัดช่องว่างแบ่งครึ่งกัน



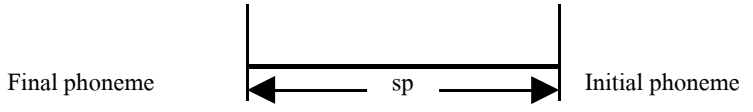
    - พยัญชนะใดๆ ยกเว้นกลุ่ม plosive ถ้ามีช่องว่างแล้วตามด้วยพยัญชนะกลุ่ม fricative ให้ตัดที่จุดสิ้นสุดของเสียงข้างหน้า แล้วทิ้งช่องว่างให้เป็นส่วนของเสียง Fricative



    - สำหรับพยัญชนะกลุ่ม plosive ที่ต่อกับเสียง fricative แล้วมีช่องว่างระหว่างเสียง ให้แบ่งตัดช่องว่างสัดส่วน 70:30



- ถ้าช่องว่างระหว่าง phoneme นั้น **มีความยาวเกิน 20 มิลลิวินาที (ms) และเป็นช่องว่างระหว่างพยางค์** ให้กำกับช่องว่างเป็น /sp/



- ในกรณีที่รูปของคลื่นและสเปคโตรแกรมไม่สามารถบ่งบอกจุดสิ้นสุดของหน่วยเสียงได้ ให้พิจารณาตามหลักเกณฑ์ต่อไปนี้

- เสียงสระอยู่ติดกับเสียง Approximant ใช้วิธีการฟังช่วยหาจุดสิ้นสุดของเสียง และดูความเปลี่ยนแปลงของ formant ประกอบกัน
- เสียงที่เกิดจากฐานเดียวกันอยู่ติดกัน เช่น Nasal ติดกับ Nasal แบ่งสัดส่วน 60:40 โดยให้เสียงท้ายยาวกว่าเสียงต้นพยางค์
- เสียง Nasal ติดกับเสียง Approximant แบ่งสัดส่วน 60:40 โดยให้เสียงท้ายยาวกว่าเสียงต้นพยางค์

4. บทความตีพิมพ์ภายใต้โครงการ

1. Rachod Thongprasirt, Thatsanee Charoenporn, Wasin Sinthupinyo and Virach Sornlertlamvanich.,"Development of Very Large Corpora in Thailand", Proceeding of Workshop, the Sixth Natural Language Processing Pacific Rim Symposium Post-Conference Workshop. Language Resource in Asia., November 30, 2001.

2. Rachod Thongprasirt, Virach Sornlertlamvanich, Patcharikra Cotsomrong, Sinaporn Subevisai and Supphanat Kanokphara, "Progress Report on Corpus Development and Speech Technology in Thailand," The Oriental COCOSDA 2002, May 9-11, 2002.

3. Chai Wutiwiwatchai, Patcharikra Cotsomrong,Sinaporn Subevisai and Supphanat Kanokphara, "Phonetically Distributed Continuous Speech Corpus for Thai Language," LREC 2002, Third International Conference on Language Resource and Evaluation., May 29-31, 2002., 869-872.

4. Sawit Kasuriya, Virach Sornlertlamvanich, Patcharika Cotsomrong, Supphanat Kanokphara, and Nattanun Thatphithakkul., "Thai Speech Corpus for Thai Speech Recognition," The Oriental COCOSDA 2003, October 1-3, 2003., 54-61

5. Thatsanee Charoenporn, Virach Sornlertlumvanich, Sawit Kasuriya, Chatchawan Hansakulbuntheung and Hitoshi Isahara ,"Open Collaborative Development of the Thai Language Resources for Natural Language Processing.", LREC2004, May 2004.

6. Patcharikra Cotsomrong, Treepop Sunpetchniyom, Sawit Kasuriya, Nattanun Thatphithakku, Chai Wutiwiwatchai, "LOTUS: Large vOcabulary Thai continUous Speech Recognition Corpus", NSTDA Annual Conference S&T in Thailand: Towards the Molecular Economy (NAC2005), March 2005.

เอกสารอ้างอิง

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

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

[3] “Handbook of the International Phonetic Association,” Cambridge University Press, 1999.

[4] J. Hamaker, R. J. Duncan, and J. Picone, “Japanese Electronic Industry Development Association’s Common Speech Data Corpus,” *prepared for Linguistic Data Consortium*, Institute of Signal and Information Processing, Mississippi State University, 1996.

[5] J. L. Shen, H. M. Wang, R. Y. Lyu, and L. S. Lee, “Automatic Selection of Phonetically Distributed Sentence sets for Speaker Adaptation with Application to Large Vocabulary Mandarin Speech Recognition,” In *Journal of Computer Speech and Language*, 1999.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," *NIST*, 1993.

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

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

[9] R. Rosenfield, "The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation," *Carnegie Mellon University*, 1994.

[10] 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

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

[12] 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.

[13] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Japanese Dictation Toolkit-1997 version-," In *Journal of Acoustic Society of Japan*, Vol. 20, No.3, May 1999.

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

[15] พิณทิพย์ ทวยเจริญ, "สัทศาสตร์และสัทศาสตร์ภาคปฏิบัติ," สำนักพิมพ์มหาวิทยาลัยธรรมศาสตร์, 2533.

ภาคผนวก ก.

หน่วยเสียงในภาษาไทย[8]

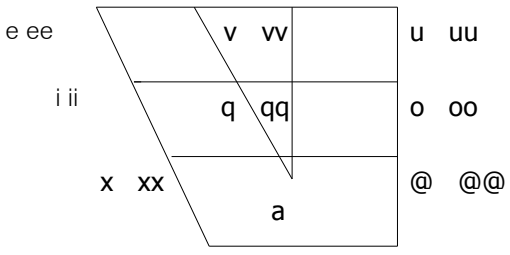
1. เสียงพยัญชนะต้นเดี่ยว (Initial consonant)

	Bilabial	Labio-dental	Alvelolar	Post-alveolar	Palatal	Velar	Glottal
Plosive	p ph b		t th d			k kh	z
Nasal	m		n			ng	
Fricative		f	s				h
Affricate				c ch			
Trill			r				
Approximant					j	w	
Lateral Approximant			l				

2. เสียงพยัญชนะท้ายเดี่ยว (Final consonant)

	Bilabial	Labio-dental	Alvelolar	Post-alveolar	Palatal	Velar	Glottal
Plosive	p^		t^			k^	
Nasal	m^		n^			ng^	
Fricative		f^	s^				
Affricate				ch^			
Trill							
Approximant					j^	w^	
Lateral Approximant			l^				

3. เสียงสระเดี่ยว (Vowel)



4. หน่วยเสียงผสม

- เสียงควบกล้ำ (Cluster consonant) ได้แก่ pr, phr, tr, kr, khr, pl, phl, kl, khl, kw, และ khw
- เสียงสระผสม (Dipthong) ได้แก่ ia, iia, va, vva, ua, และ uua

[8] เพื่อความสะดวกในการใช้งานในคอมพิวเตอร์ สัญลักษณ์ของหน่วยเสียงบางหน่วยเสียงจะถูกกำหนดเปลี่ยนแปลงไปจากระบบ IPA

นิยามหน่วยเสียงสำหรับใช้ในฐานข้อมูล LVCSR

1. หน่วยเสียงเดี่ยว (Monophone)

สำหรับฐานข้อมูลนี้ กำหนดให้เสียงพยางค์ภาษาไทยอยู่ในรูปแบบ / Ci V Cf /[9] มีจำนวนหน่วยเสียงเดี่ยว 68 หน่วยเสียง มีรายละเอียดดังนี้

พยัญชนะต้น (Ci)			
เดี่ยว	ตัวอย่าง	ผสม	ตัวอย่าง
p	ปาก	pr	ประสาน
t	เต้น, กุฏิ	phr	พราน
c	จะ	tr	เตรียม
k	ก่อน	kr	กราบ
z	อาน	khr	คร่า
ph	พบ, ภัย, ผ่าน	pl	ปลา
th	ทิ้ง, ธง, เฒ่า, ฐาน, มณโฑ	phl	พลาด
ch	ชอบ, เฌอ	thr	จันทรา
kh	คน, เชิน, ฆ่า	kl	เกลอ
b	บอก	khl	เคลื่อน
d	ด้าน, ชฎา	kw	กวาง
m	ไม่	khw	ขวา
n	นาน, เณร		
ng	เงิน	เสียงทับศัพท์	
l	เล่น, กีฬา	br	เบรน
r	รอ, ฤทัย	bl	บลู
f	ฝน, ฟัน	fr	ฟราย
s	สาย, ศิลา, รักษา, ซ่อน	fl	เฟลม
h	โหน, เฮฮา	dr	ดรากอน
w	ว่า		
j	ย้อน, หญิง	17 หน่วย	
21 หน่วย			

สระ (V)			
เดี่ยว	ตัวอย่าง	ผสม	ตัวอย่าง
a	อะ	ia	เอียะ
aa	อา	iia	เอีย
I	อิ	va	เอือะ
ii	อี	vva	เอือ
v	อึ	ua	อัวะ
vv	อื	uua	อัว
u	อุ	6 หน่วย	
uu	อู		
e	เอะ		
ee	เอ		
x	แอะ		
xx	แอ		
o	โอะ		
oo	โอ		
@	เอาะ		
@@	ออ		
q	เออะ		
qq	เออ		
18 หน่วย			

ตัวสะกด (Cf)	
เดี่ยว	ตัวอย่าง
p^	พบ
t^	เกร็ด
k^	ปาก
n^	หาร
m^	ลม
ng^	ฟาง
j^	ยาย
w^	กาว
เสียงทับศัพท์	
f^	กราฟ
l^	แอล
s^	เอส
ch^	คลัช
12 หน่วย	

[9] หน่วยเสียงที่แสดงจะไม่รวมสัญลักษณ์กำกับเสียงวรรณยุกต์ (Tone)

2. หน่วยเสียงคู่ (Biphone)

ในชุดประโยคหน่วยเสียงสมดุล (Phonetically Balanced Set) จะเป็นการคัดประโยคให้ครอบคลุมการเกิดหน่วยเสียงคู่ (Biphone) โดยที่หน่วยเสียงคู่ที่เกิดขึ้นได้ในชุดประโยคนี้มีจำนวนทั้งสิ้น 1,628 หน่วย ซึ่งคิดเป็น 90.9% ของจำนวนหน่วยเสียงคู่ที่เกิดขึ้นได้จริงในภาษาไทย การกระจายของหน่วยเสียงคู่ที่ปรากฏในชุดประโยคที่คัดได้เป็นดังตารางนี้

หน่วยเสียงคู่ (Biphone)	จำนวน
Ci V	583
V Cf	157
Cf Ci[10]	329
V Ci	559
รวม	1,628

[10] เกิดขึ้นระหว่างพยางค์หรือระหว่างคำ