รายงานวิจัยฉบับสมบูรณ์ประจำปีงบประมาณ 2546
โครงการวิจัยย่อยลำดับที่ 3 เรื่อง
ระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยแบบชุดคำศัพท์ขนาดใหญ่ ระยะที่ 2
(Thai Large-Vocabulary Continuous Speech Recognition Research Project: Phase II)

1. ผู้รับผิดชอบโครงการ

   รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล

2. วัตถุประสงค์ของโครงการ
   - 2.1 พัฒนาระบบสร้างแบบจำลองเสียงพูดให้สามารถใช้งานได้บนเสียงพูดแบบต่อเนื่อง
   - 2.2 พัฒนาระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยแบบชุดคำศัพท์ขนาดใหญ่
   - 2.3 พัฒนาระบบรู้จำเสียงวรรณยุกต์และทำนองเสียงพูดเพื่อใช้ในระบบรู้จำเสียงพูดต่อเนื่องภาษาไทย

3. ขอบเขตหรือเป้าหมายของโครงการ
   - 3.1. สร้างฐานข้อมูลเสียงพูดต่อเนื่องภาษาไทยระยะที่ 2
   - 3.2. สร้างระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยระยะที่ 2

4. ส่วนงานที่ได้ดำเนินการ

4.1 <u>การสร้างฐานข้อมูลเสียงพูดในระบบโทรศัพท์</u>

ในงานวิจัยนี้ได้มีการสร้างฐานข้อมูลเสียงพูดในระบบโทรศัพท์เคลื่อนที่แบบ CDMA โดยได้มีการบันทึกเสียงพูดจำนวน 300 คน จากผู้พูดเพศชาย 150 คน และเพศหญิงจำนวน 150 คนสามารถแบ่งตามสภาพแวดล้อมได้ดังนี้

	สัญญาณรบกวนน้อย (120)	สัญญาณรบกวนปานกลาง (120)	สัญญาณรบกวนมาก (60)
หยุดนิ่ง	ในสำนักงาน , ในบ้าน (90)	ริมถนน , ในห้าง (90)	ริมถนน , ย่านอึกทึก (30)
เคลื่อนที่	รถส่วนตัว (30)	รถปรับอากาศ , รถไฟฟ้า (30)	รถเมล์ (30)

ในระยะที่ 2 ของงานวิจัยนี้ได้ทำการบันทึกเสียงพูดดังกล่าวเสร็จเรียบร้อย

4.2 <u>การสร้างระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยระยะที่ 2</u>

การสร้างระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยในระยะที่ 2 ประกอบด้วย 3 ส่วนคือ

- การสร้างแบบจำลองเสียงพูดและการพัฒนาระบบรู้จำเสียงพูดต่อเนื่องส่วน decoder
- การรู้จำวรรณยุกต์เสียงพูดต่อเนื่องภาษาไทย

- การรู้จำทำนองเสียงพูดต่อเนื่องภาษาไทย

4.2.1 การสร้างแบบจำลองเสียงพูด

งานวิจัยในระยะที่ 2 นี้ได้พัฒนาโปรแกรมที่ใช้ในการสร้างแบบจำลองเสียงพูดขึ้นโดยใช้ฐานข้อมูลเสียงพูดต่อเนื่องแบบปราศจากสัญญาณรบกวนมาสร้างแบบจำลองเสียงพูด การสร้างแบบจำลองเสียงพูดจะใช้เสียงพูดต่อเนื่องและข้อความจากการกำกับเสียงพูด (Labeled Transcription) แบบจำลองหน่วยเสียงพูดแสดงในรูปที่ 1 ซึ่งได้จากการพัฒนาในโครงการระยะที่ 1

แบบจำลองเสียงพูดเบื้องต้นจะถูกฝึกฝนซ้ำกับข้อความเสียงพูดที่ไม่ต้องมีการกำกับ (Unlabeled Transcription) ในระบบที่พัฒนาในโครงการระยะที่สองนี้ ผลที่ได้จะเป็นแบบจำลองของหน่วยเสียงพูดซึ่งจะนำไปใช้ในการพัฒนาระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยต่อไป

ในงานวิจัยนี้ใช้โปรแกรม Hidden Markov Model Toolkit (HTK) จาก Cambridge University ประเทศอังกฤษ เป็นต้นแบบในการพัฒนา ซึ่งในระยะที่ 2 ได้พัฒนาส่วนการแบบจำลองหน่วยเสียงที่เหลือเกือบแล้วเสร็จ

ในงานวิจัยระยะที่ 2 นี้ได้ทำการทดลองเปรียบเทียบแบบจำลองหน่วยเสียงประเภทต่างๆดังแสดงในตารางที่ 1

ตารางที่ 1 ประเภทของหน่วยเสียงที่ใช้ในงานวิจัยนี้

Speech Unit	Possible Units
Monophone	58
intra-syllable triphone	7,769
inter-syllable triphone	64,475
CI Initial-Final	33I + 200F
CD Initial-Final	297I + 200F
CORM	297O + 200R
PORM	792O + 200R

จำนวนข้อมูลเสียงพูดที่ใช้สำหรับการสร้างแบบจำลองเสียงพูดและทดสอบในงานวิจัยนี้มี 29,669 ประโยค และ 3,000 ประโยคตามลำดับ โดยบันทึกเสียงจากผู้พูดเพศชายจำนวน 14 คน และเพศหญิงจำนวน 16 คน

ในงานวิจัยนี้ได้ทดสอบแบบจำลองที่สร้างขึ้นแบบที่ขึ้นกับเพศของผู้พูดและไม่ขึ้นกับเพศของผู้พูด ได้ผลการทดลองแสดงในตารางที่ 2

จากผลการทดลองสรุปได้ว่าแบบจำลองที่ขึ้นกับเพศของผู้พูดมีประสิทธิภาพที่ดีกว่าแบบจำลองที่ไม่ขึ้นกับเพศของผู้พูด ส่งผลให้ใช้แบบที่ขึ้นกับเพศของผู้พูดในการทดลองต่อๆไปภายหน้า

ตารางที่ 2 อัตราการรู้จำของแบบจำลองที่ขึ้นกับเพศของผู้พูดและไม่ขึ้นกับเพศของผู้พูด

Gender	Accuracy	
	Gender-dependent	Gender-independent
male	4.5 %	2.7 %
female	5.4 %	3.2 %

การทดลองเพื่อเปรียบเทียบประสิทธิภาพของหน่วยเสียงประเภทต่างๆ โดยใช้การจำลองเชิงกลสัทศาสตร์เพียง (Acoustic Modeling) อย่างเดียวมีผลการรู้จำแสดงในตารางที่ 3

ตารางที่ 3 อัตราการรู้จำของหน่วยเสียงประเภทต่างๆที่ใช้การจำลองเชิงกลสัทศาสตร์เพียงอย่างเดียว

Speech unit	Accuracy	
	Speaker-dependent	Speaker-independent
monophone	19.6 %	10.3 %
Intra-syllable triphone	34.8 %	19.6 %
Inter-syllable triphone	40.4 %	26.8 %
CI Initial-Final	38.4 %	26.1 %
CD Initial-Final	42.6 %	29.9 %
CORM	44.6 %	33.5 %
PORM	46.8 %	35.4 %

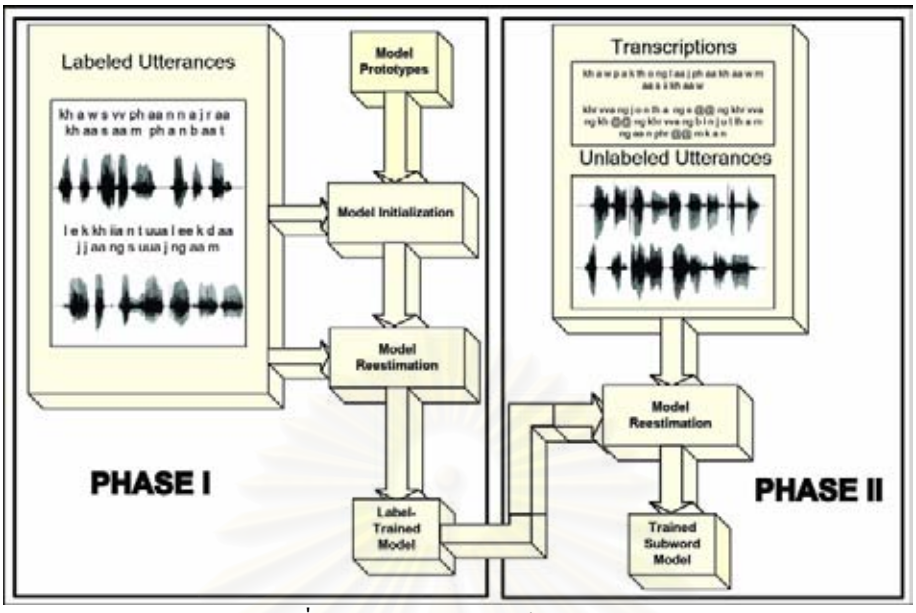
จากผลการทดลองแสดงให้เห็นว่าแบบจำลอง onset-rhyme (CORM และ PORM) ให้อัตราการรู้จำที่สูงกว่าแบบจำลองหน่วยเสียงอื่นๆ

การใช้การจำลองเชิงภาษา (Language Modeling) มาร่วมในระบบรู้จำเสียงพูดจะทำให้ประสิทธิภาพของระบบดีขึ้นดังแสดงในตารางที่ 4

ตารางที่ 4 อัตราการรู้จำของหน่วยเสียงประเภทต่างๆที่ใช้การจำลองเชิงกลสัทศาสตร์ และการใช้การจำลองเชิงภาษา

Speech unit	Accuracy	
	Speaker-dependent	Speaker-independent
monophone	42.8 %	37.0 %
Intra-syllable triphone	61.9 %	47.7 %
Inter-syllable triphone	69.5 %	54.3 %
CI Initial-Final	64.9 %	47.7 %
CD Initial-Final	71.2 %	57.1 %
CORM	73.6 %	61.8 %
PORM	75.4 %	63.8 %

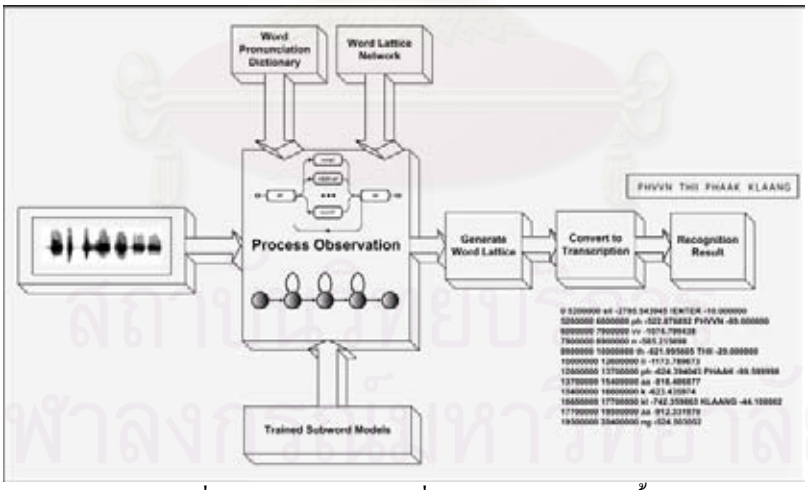
เมื่อใช้เกณฑ์ในการประเมินผลประสิทธิภาพของหน่วยเสียง 3 ประการคือ accuracy, generalization และ trainability พบว่าแบบจำลอง onset-rhyme (CORM และ PORM) มีประสิทธิภาพที่ดีกว่าแบบจำลองหน่วยเสียงประเภทอื่นๆ



รูปที่ 1 การพัฒนาระบบสร้างแบบจำลองเสียงพูด

4.2.2 <u>การพัฒนาระบบรู้จำเสียงพูดต่อเนื่องส่วน decoder</u>

นอกจากโปรแกรมที่ใช้ในการสร้างแบบจำลองหน่วยเสียงเสร็จแล้ว ยังได้พัฒนาระบบรู้จำเสียงพูดต่อเนื่องอีกด้วย เพื่อทดสอบแบบจำลองหน่วยเสียงที่สร้างขึ้นต้องใช้ระบบรู้จำเสียงพูดต่อเนื่อง ซึ่งในระยะที่ 2 นี้ได้ทำการพัฒนาระบบส่วนที่เหลือได้บางส่วน ซึ่งระบบรู้จำเสียงพูดต่อเนื่องมีรายละเอียดดังแสดงในรูปที่ 2



รูปที่ 2 ระบบรู้จำเสียงพูดที่จะพัฒนาในงานวิจัยนี้

งานวิจัยในระยะที่ 2 นี้ได้พัฒนาส่วนต่างดังนี้

- ส่วน word pronunciation dictionary
- ส่วน word lattice network
- ส่วน decoder โดยใช้กรรมวิธี Time Synchronous Viterbi Beam Search

ยังเหลือส่วนย่อยของระบบรู้จำเสียงพูดต่อเนื่องอีกหลายส่วนที่ต้องพัฒนาให้แล้วเสร็จ และยังจะต้องปรับปรุงพัฒนาส่วนย่อยที่ได้ดำเนิการไปแล้วให้มีประสิทธิภาพที่ดีขึ้น ซึ่งวาสงแผนจะดำเนินการต่อไปในโครงการปีที่ 3

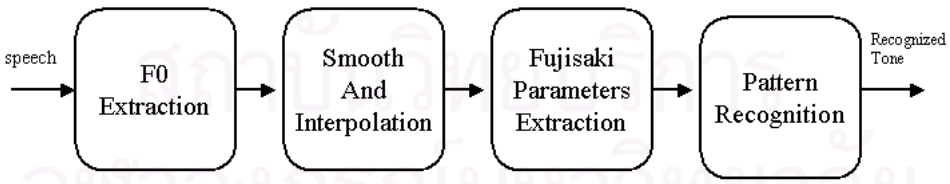
4.2.3 <u>การรู้จำวรรณยุกต์เสียงพูดต่อเนื่องภาษาไทยระยะที่ 2</u>

เนื่องจากภาษาไทยเป็นภาษาที่มีวรรณยุกต์จึงจำเป็นที่ต้องพัฒนาระบบรู้จำเสียงวรรณยุกต์ไทยนอกจากการสร้างแบบจำลองเสียงพูดดังกล่าวข้างต้น รายละเอียดของการพัฒนาระบบรู้จำเสียงวรรณยุกต์มีดังต่อไปนี้

งานวิจัยในส่วนนี้ เป็นการวิจัยเพื่อทำการรู้จำวรรณยุกต์เสียงพูดต่อเนื่องภาษาไทย โดยใช้แบบจำลองฟูจิซากิเป็นพื้นฐานในการหาค่า Feature และใช้ Neural Network ในการ Classified ออกเป็นวรรณยุกต์ต่าง ๆ

เหตุผลที่เลือกที่จะนำพารามิเตอร์ของแบบจำลอง Fujisaki มาใช้ในการวิจัยเพื่อทำการรู้จำวรรณยุกต์ของเสียงพูดต่อเนื่องภาษาไทยเนื่องจาก

- แบบจำลอง Fujisaki สร้างขึ้นจากการเลียนแบบการทำงานของอวัยวะในการสร้างเสียงพูด (ส่วนที่เกี่ยวข้องกับโทนเสียง) จริง จึงน่าที่จะใช้เป็นตัวแทนของโทนเสียงได้เป็นอย่างดี
- เหมาะสมกับเสียงพูดต่อเนื่องที่มีการพูดลักษณะเป็นธรรมชาติ
- มีจำนวนพารามิเตอร์น้อย ทำให้การคำนวณในส่วนของ Classification เป็นไปอย่างรวดเร็ว
- มีการพิจารณาถึง Declination Effect และ Tonal Assimilation Effect ไว้แล้วในแบบจำลอง
- ค่าของพารามิเตอร์หลังการ Normalize จะไม่ขึ้นกับความยาวของพยางค์ และไม่ขึ้นกับลักษณะเสียงสระสั้น-ยาว



รูปที่ 3 โครงสร้างระบบรู้จำวรรณยุกต์ของเสียงพูด

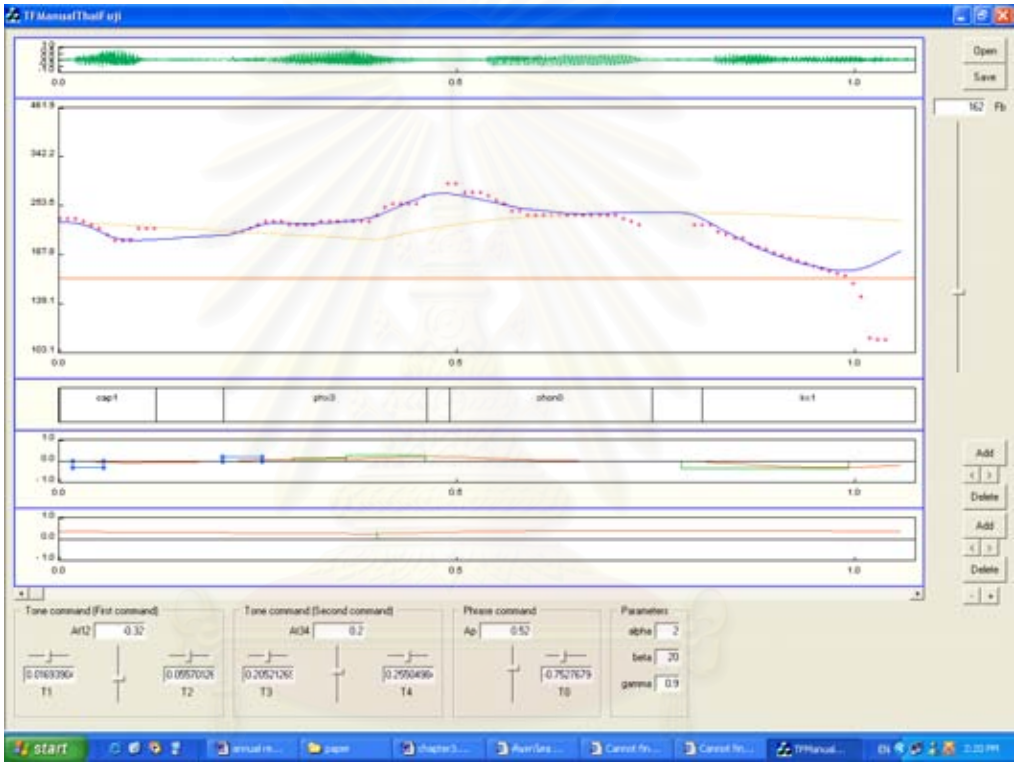
โครงสร้างของระบบรู้จำวรรณยุกต์ของเสียงพูดที่นำเสนอแสดงได้ดังรูปที่ 3 ข้างต้น โดยเสียงพูดจะถูกนำไปคำนวณหาเส้นโค้งความถี่มูลฐาน จากนั้นจึงผ่านการ smoothing และ Interpolation เพื่อลดสัญญาณที่อาจเกิดจากสัญญาณรบกวน ซึ่งถือเป็นกระบวนการประมวลผลก่อนหน้า จากนั้นทำการแยก Feature ด้วยการแยกพารามิเตอร์ของแบบจำลอง Fujisaki และขั้นตอนสุดท้ายเป็นการทำ Classification โดยใช้ Neural Network

กระบวนการแยกพารามิเตอร์ที่ทำการทดสอบประกอบด้วยการแยกพารามิเตอร์แบบ manual และการแยกพารามิเตอร์แบบอัตโนมัติอีก 3 วิธีคือการแยกพารามิเตอร์ตามวิธีของ Hansjorg Mixdorff, การแยกพารามิเตอร์โดยไม่ใช้ syllable boundary และการแยกพารามิเตอร์โดยใช้ syllable boundary

การแยกพารามิเตอร์แบบ manual นั้นเป็นการศึกษาความเป็นไปได้ในการนำพารามิเตอร์ของแบบจำลอง Fujisaki มาใช้เป็น Feature ในการรู้จำวรรณยุกต์ของเสียงพูด

<u>การแยกพารามิเตอร์แบบ Manual</u>

การแยกพารามิเตอร์แบบ Manual กระทำผ่านโปรแกรมที่พัฒนาขึ้นภายในห้องปฏิบัติการวิจัยกรรมวิธีสัญญาณดิจิทัลซึ่งแสดงหน้าจอการทำงานได้ดังรูป



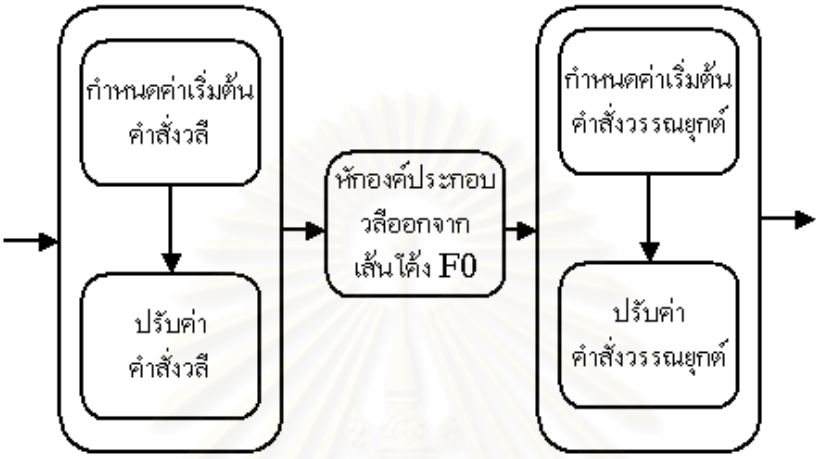
รูปที่ 4 หน้าจอการทำงานของ

การหาพารามิเตอร์แบบ Manual ประกอบด้วยขั้นตอนต่างๆ ดังนี้

- พิจารณาการเปลี่ยนแปลงโดยรวมของ F0 Contour ที่ได้จากเสียงพูด และกำหนดค่าเริ่มต้นสำหรับ Phrase command
- ปรับค่า Phrase command ทั้งในส่วนของเวลา และแมกนิจูดให้ Phrase component ที่สร้างขึ้นมีค่าใกล้เคียงกับการเปลี่ยนแปลงโดยรวมของ F0 Contour ที่ได้จากเสียงพูด
- นำ Phrase component ที่ได้จาก Phrase command ที่ปรับแล้วไปหักออกจาก F0 Contour
- กำหนดค่าเริ่มต้นสำหรับ Tone command โดยพิจารณาจากเส้นโค้งส่วนที่เหลือจากการหัก Phrase component ออกจาก F0 Contour

- ปรับค่า Tone command ทั้งเวลาเริ่มต้น เวลาสิ้นสุด และแมกนิจูดจนมี F0 Contour ที่สร้างขึ้นจากการรวมกันของ Phrase component และ Tone component ใกล้เคียงกับ F0 Contour ที่ได้จากเสียงพูดมากที่สุด

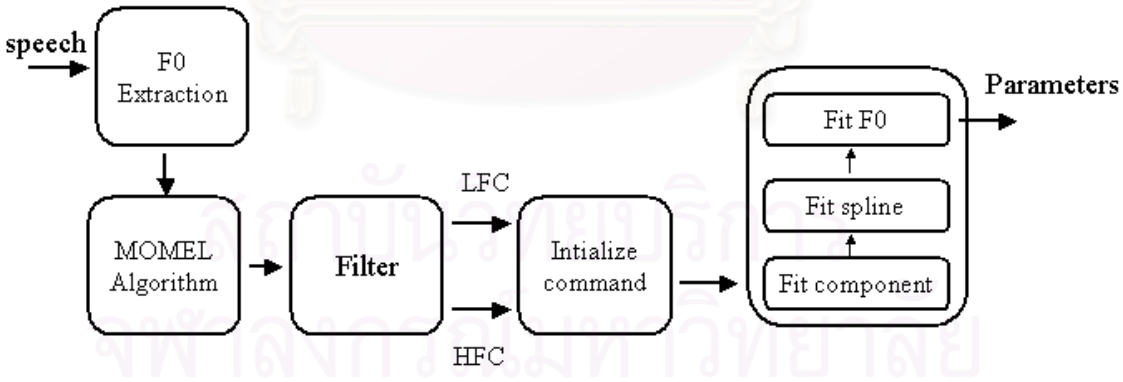
การหาพารามิเตอร์แบบ manual แสดงได้ดังรูปที่ 5



รูปที่ 5 ขั้นตอนการหาพารามิเตอร์แบบ manual

<u>การแยกพารามิเตอร์ตามวิธีของ Mixdorff</u>

การแยกพารามิเตอร์ตามวิธีของ Mixdorff เป็นวิธีการแยกพารามิเตอร์โดยอัตโนมัติ ซึ่งแสดงได้ดังรูป



รูปที่ 6 แผนภาพแสดงการแยกพารามิเตอร์ตามกรรมวิธีของ Mixdorff

ตามวิธีของ Mixdorff การแยกพารามิเตอร์สำหรับแบบจำลอง Fujisaki จาก F0 Contour ที่ได้ประกอบด้วย 4 ขั้นตอนหลักคือ

- Quadratic Spline Stylisation เป็นการทำ Smoothing และ Interpolation โดยใช้ MOMEL Algorithm

- Filtering and Component Separation เป็นการแยก F0 Contour ที่ได้หลังจากทำ MOMEL Algorithm ออกเป็น 2 ส่วนโดยใช้วงจรกรอง องค์ประกอบ 2 ส่วนที่แยกได้เรียกว่า LFC (Low Frequency Contour) และ HFC (High Frequency Contour)
- ให้ค่าเริ่มต้นของแบบจำลอง (Fujisaki Model Command Initialization) โดยค่าเริ่มต้นของ Phrase command พิจารณาจาก LFC และค่าเริ่มต้นของ Tone command พิจารณาจาก HFC
- ปรับค่าพารามิเตอร์ของแบบจำลอง โดยทำการปรับค่าทั้งสิ้น 3 ครั้งคือ
    - ปรับค่าเทียบกับ HFC และ LFC
    - ปรับค่าเทียบกับ F0 Contour ที่ได้หลังจาก MOMEL Algorithm
    - ปรับค่าเทียบกับ F0 Contour ตั้งต้น

การแยกพารามิเตอร์โดยไม่ใช้ Syllable boundary

รูปที่ 7 แผนภาพแสดงการแยกพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์

การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ เป็นการหาค่าพารามิเตอร์โดยอัตโนมัติ ซึ่งข้อมูลที่ใช้ในการหาพารามิเตอร์มีเพียง F0 Contour เพียงอย่างเดียว โดยเป็นการดัดแปลงวิธีจากการหาค่าพารามิเตอร์ตามวิธีของ Hansjorg Mixdorff ซึ่งข้อแตกต่างหลักคือการไม่ใช้อัลกอริทึม MOMEL เนื่องจากพบว่า การนำอัลกอริทึม MOMEL มาใช้กับเสียงในภาษาไทย จะทำให้การเปลี่ยนแปลงของ F0 Contour ถูกทำให้เรียบจนข้อมูลเกี่ยวกับวรรณยุกต์หายไปมาก

การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ประกอบด้วยขั้นตอนต่าง ๆ ดังนี้

- ทำ Neutralizetion และ Median Filtering เพื่อทำการลด Noise ต่าง ๆ ที่เกิดขึ้น
- ทำ Linear Interpolation ในส่วนที่ไม่ปรากฏค่า F0
- ทำการ Filter เพื่อแยก F0 Contour ที่ได้หลังจากการทำ Linear Interpolation ออกเป็น 2 ส่วนคือ HFC และ LFC
- กำหนดค่าตั้งต้นให้กับ Phrase command โดยพิจารณาจาก LFC
- กำหนดค่าตั้งต้นให้กับ Tone command โดยพิจารณาจาก HFC

- ปรับค่าพารามิเตอร์ของ Phrase command และ Tone command ให้ Phrase component ที่ได้จาก Phrase command มีค่าใกล้เคียงกับ LFC และ Tone component ที่ได้จาก Tone command มีค่าใกล้เคียงกับ HFC

การแยกพารามิเตอร์โดยใช้ Syllable boundary

การแยกพารามิเตอร์โดยใช้ Syllable boundary เป็นการแยกพารามิเตอร์โดยอัตโนมัติที่พัฒนาจากการแยกพารามิเตอร์โดยไม่ใช้ Syllable boundary โดยนำ Syllable boundary เข้าไปร่วมกำหนดค่าตั้งต้นของ Tone command ขั้นตอนต่าง ๆ จะเหมือนกับการแยกพารามิเตอร์โดยไม่ใช้ Syllable boundary ยกเว้นจะมีการกำหนดให้ใน 1 syllable มี 2 Tone command เสมอ และพิจารณาจาก HFC ในแต่ละช่วงพยางค์เพื่อกำหนดค่าเริ่มต้นให้เหมาะสม

ผลการทดสอบการรู้จำโดยใช้แบบจำลอง Fujisaki

ผลการทดสอบโดยใช้วิธี Five fold validation กับชุดข้อมูล Thai Proverb Corpus (TPC) เป็นดังนี้

วิธีที่ใช้	ผลที่ได้
Manual	96.35%
Hansjorg Mixdorff process	56.78%
Automatic Extract without Syllable Boundary Data	70.27%
Automatic Extract with Syllable Boundary Data	68.27%

สรุปผลการวิจัย

จากผลการทดสอบข้างต้นแสดงให้เห็นว่า พารามิเตอร์ของแบบจำลอง Fujisaki สามารถนำมาใช้เป็น Feature ในการรู้จำวรรณยุกต์เสียงต่อเนื่องภาษาไทยได้จริง แต่การนำพารามิเตอร์มาใช้นั้น ยังไม่มีกรรมวิธีในการหาพารามิเตอร์ที่แน่นอน และเป็นที่ยอมรับอย่างกว้างขวาง กรรมวิธีการหาพารามิเตอร์โดยอัตโนมัติที่ทำการวิจัยสามารถใช้ในการรู้จำวรรณยุกต์ได้ในระดับหนึ่ง แต่ยังได้ผลไม่น่าพอใจถึงขั้นที่สามารถนำไปใช้งานได้จริง อีกทั้งเวลาที่ใช้ในการคำนวณในขั้นตอนการรู้จำนั้น แม้จะใช้เวลาในการประมวลผลน้อย เนื่องจากจำนวนพารามิเตอร์ที่น้อย แต่การได้มาซึ่งพารามิเตอร์นั้นจะต้องอาศัยเวลาในการประมวลผลมาก

สิ่งที่น่าสนใจคือ การพัฒนากระบวนการแยกพารามิเตอร์ให้สามารถทำได้รวดเร็ว และใช้ในการรู้จำได้ดี และเป็นที่ยอมรับทั้งกับภาษาไทย และภาษาต่างประเทศส่วนงานที่จะดำเนินการต่อไป

4.2.4 การรู้จำทำนองเสียงพูดสำหรับเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม

งานวิจัยนี้มีวัตถุประสงค์เพื่อ ออกแบบคอนทัวร์ลักษณะ (feature contour) จากคอนทัวร์ความถี่มูลฐานของเสียงพูด (fundamental frequency: F_0) เพื่อนำมาใช้รู้จำทำนองเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม

แผนงานวิจัยด้านทำนองเสียงพูดภาษาไทย

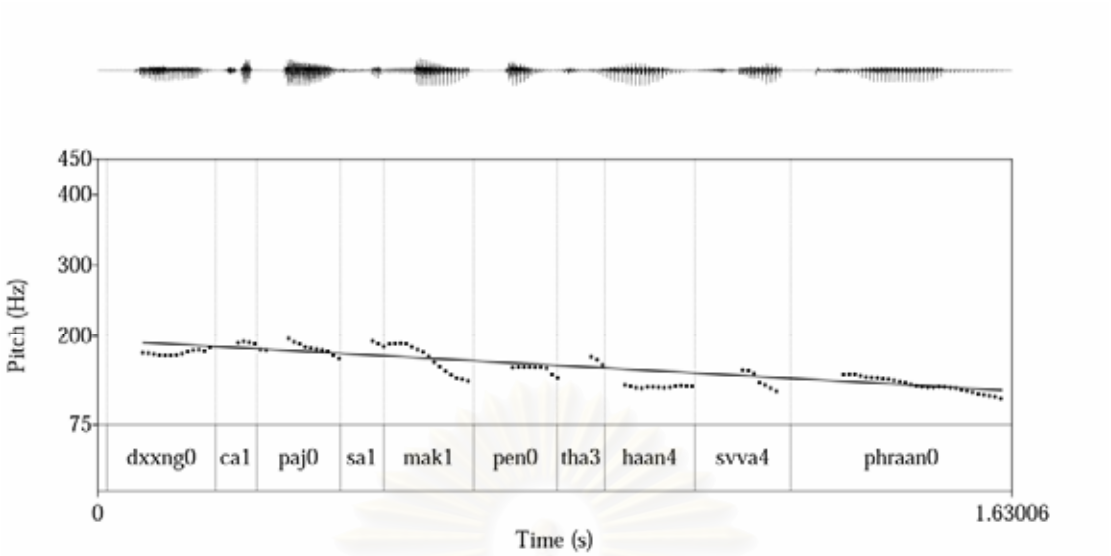
ทำนองเสียงพูดภาษาไทย แบ่งได้เป็น 3 ชนิด คือ ทำนองเสียงตก ทำนองเสียงขึ้น และทำนองเสียงผสม ลักษณะของประโยคเสียงพูดภาษาไทยแบบต่าง ๆ ซึ่งทำให้เกิดทำนองเสียงพูดทั้ง 3 ประเภท รวมทั้งผลของทำนองเสียงพูดต่อลักษณะของคอนทัวร์ F_0 มีดังนี้

- ทำนองเสียงตก (the Fall Class) เกิดเมื่อผู้พูดพูดประโยคบอกเล่า (statement) พูดชมเชย หรือพูดเป็นคำ ๆ (citation form) พูดแบบไม่แสดงความคิดเห็น (attitudinally unmarked) พูดแบบยอมจำนน (submissive) พูดแบบซ่อนความโกรธ (concealed anger) พูดแบบเบื่อ (bored) และพูดแบบแสดงอำนาจ (authoritative) เมื่อพิจารณารูปร่างของคอนทัวร์ F_0 แบบแคบ (ในระดับพยางค์) จะพบว่ารูปร่างของคอนทัวร์ F_0 ของแต่ละพยางค์ ขึ้นกับชนิดของเสียงวรรณยุกต์ของพยางค์นั้น แต่ระดับของ F_0 ของพยางค์ที่ถัดจากพยางค์แรกจะค่อย ๆ ลดระดับลงไปเรื่อย ๆ (downdrift) จนจบประโยค ดังนั้นเมื่อพิจารณาแบบกว้าง (ในระดับประโยค) จึงเห็นได้ว่าคอนทัวร์ F_0 ของทำนองเสียงตก จะค่อย ๆ ลดระดับลง (decline) เมื่อพูดไปเรื่อย ๆ จนจบประโยค

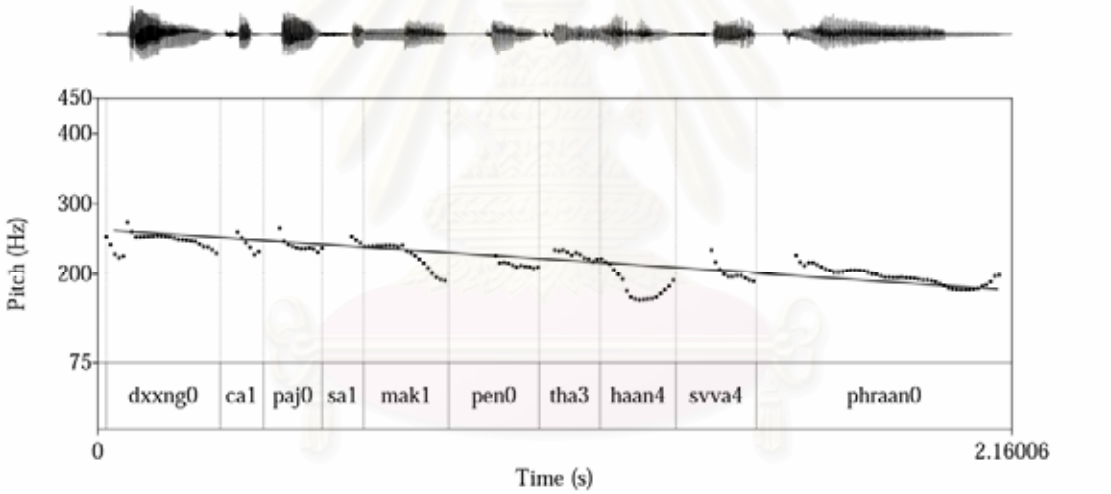
- ทำนองเสียงขึ้น (the Rise Class) เกิดเมื่อผู้พูดพูดประโยคที่แสดงให้เห็นถึงการถาม (question) แสดงความไม่เห็นด้วย (disagreeable) แสดงความไม่เชื่อ (disbelieving) แสดงความไม่คาดฝัน (surprised) และพูดแบบแสดงให้เห็นว่ายังพูดไม่จบความ (unfinished) เมื่อพิจารณารูปร่างของคอนทัวร์ F_0 แบบแคบ จะพบว่ารูปร่างของคอนทัวร์ F_0 ของแต่ละพยางค์ ขึ้นกับชนิดของเสียงวรรณยุกต์ เช่นเดียวกับทำนองเสียงตก แต่ช่วงค่า F_0 จะแคบกว่าเล็กน้อย และมีระดับของ F_0 ที่สูงกว่า จึงส่งผลให้รูปร่างของคอนทัวร์มีลักษณะสูงขึ้นกว่าทำนองเสียงตกเมื่อพิจารณาในระดับประโยค

- ทำนองเสียงผสม (the Convolution Class) เกิดเมื่อผู้พูดพูดเพื่อต้องการเน้นหนัก (emphatic) พูดแสดงความโกรธ (anger) แสดงความเห็นด้วยอย่างมาก (very agreeable) แสดงความสนใจมาก (very interested) และแสดงความเชื่อถือมาก (very believing) ในกรณีของทำนองเสียงผสม รูปร่างของคอนทัวร์ F_0 ในแต่ละพยางค์จะขึ้นกับประเภทของเสียงวรรณยุกต์ เช่นเดียวกับทั้ง 2 ทำนองเสียง ดังที่ได้กล่าวไปแล้ว แต่ช่วงค่า F_0 ของเสียงวรรณยุกต์ทุกเสียงจะกว้างกว่า ทำนองเสียงตก และทำนองเสียงขึ้น นอกจากนี้ในกรณีของเสียงวรรณยุกต์ สามัญ เอก โท และตรี จะพบว่า F_0 มีระดับที่สูงกว่า ทำนองเสียงตก แต่เสียงวรรณยุกต์จัตวาจะมีระดับ F_0 ที่ต่ำ ซึ่งส่งผลให้รูปร่างของคอนทัวร์ F_0 ในระดับประโยค มีลักษณะสูงกว่า ทำนองเสียงตก และช่วงการแกว่งกว้างกว่าทำนองเสียงตก และทำนองเสียงขึ้น

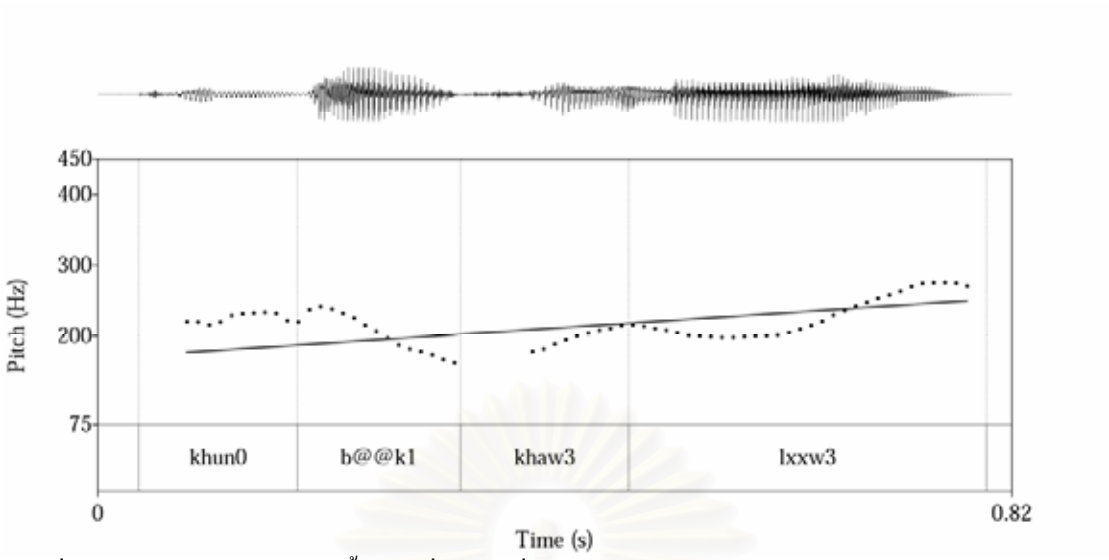
ตัวอย่างของคอนทัวร์ F_0 ของทำนองเสียงพูดทั้ง 3 แบบ ของเสียงผู้ชาย และเสียงผู้หญิง แสดงดังรูปที่ 8 ถึง 15



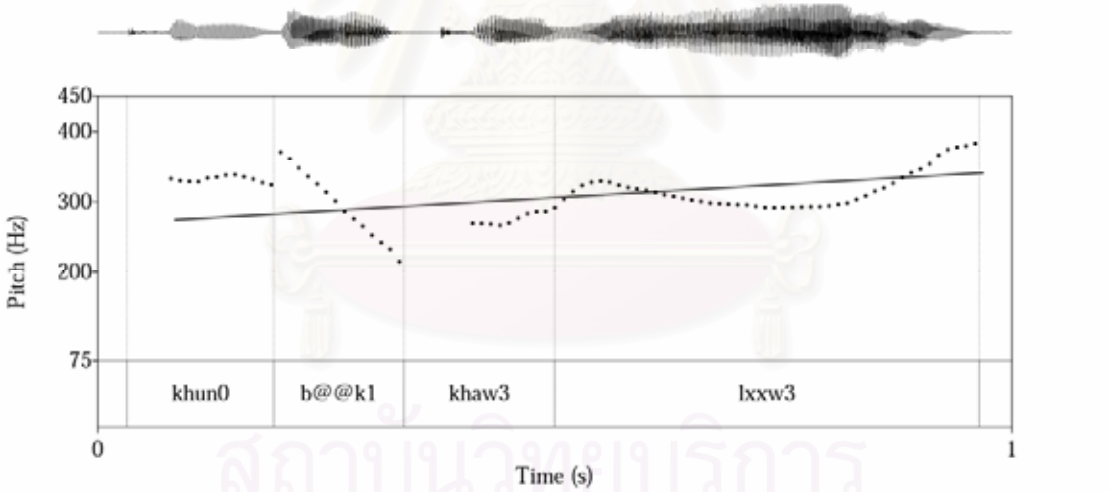
รูปที่ 8 ตัวอย่างของทำนองเสียงตก: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค "แดงจะไปสมัครเป็นทหารเสือพราน" เมื่อผู้พูดเป็นผู้ชาย



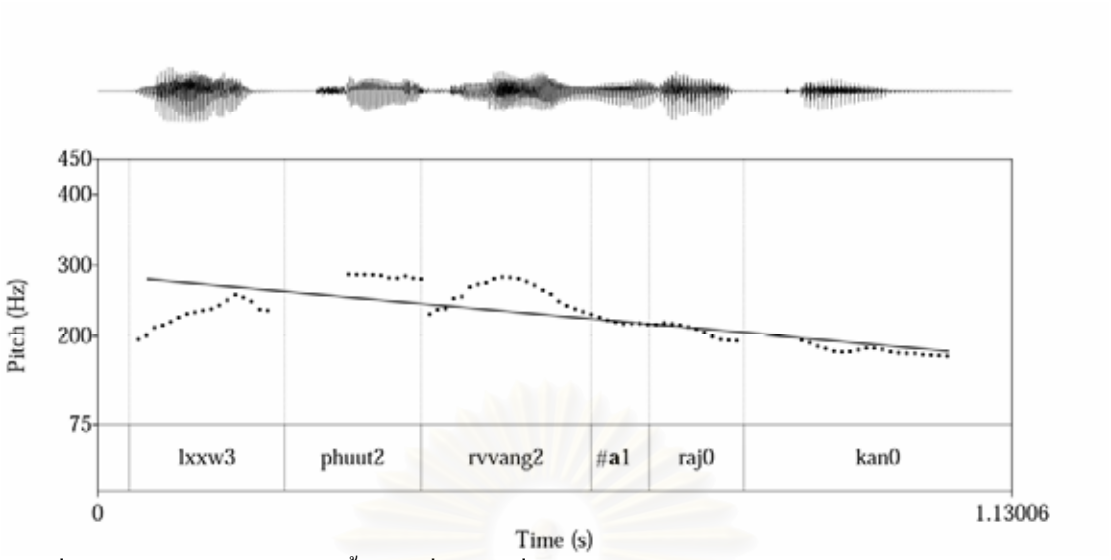
รูปที่ 9 ตัวอย่างของทำนองเสียงตก: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และ เส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค "แดงจะไปสมัครเป็นทหารเสือพราน" เมื่อผู้พูดเป็นผู้หญิง



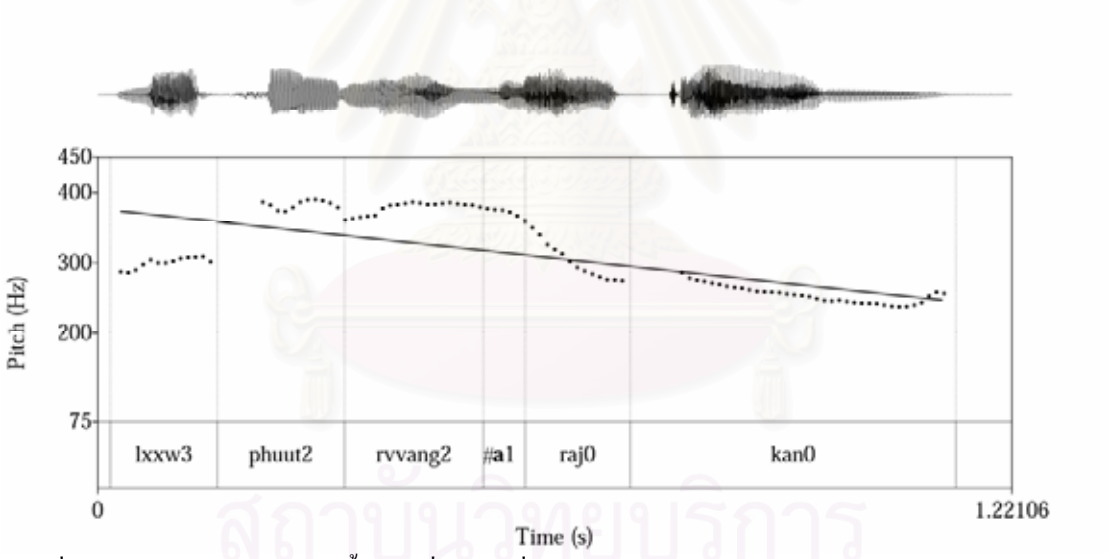
รูปที่ 10 ตัวอย่างของทำนองเสียงขึ้นแบบที่ 1: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค “คุณบอกเค้าแล้ว?” เมื่อผู้พูดเป็นผู้ชาย



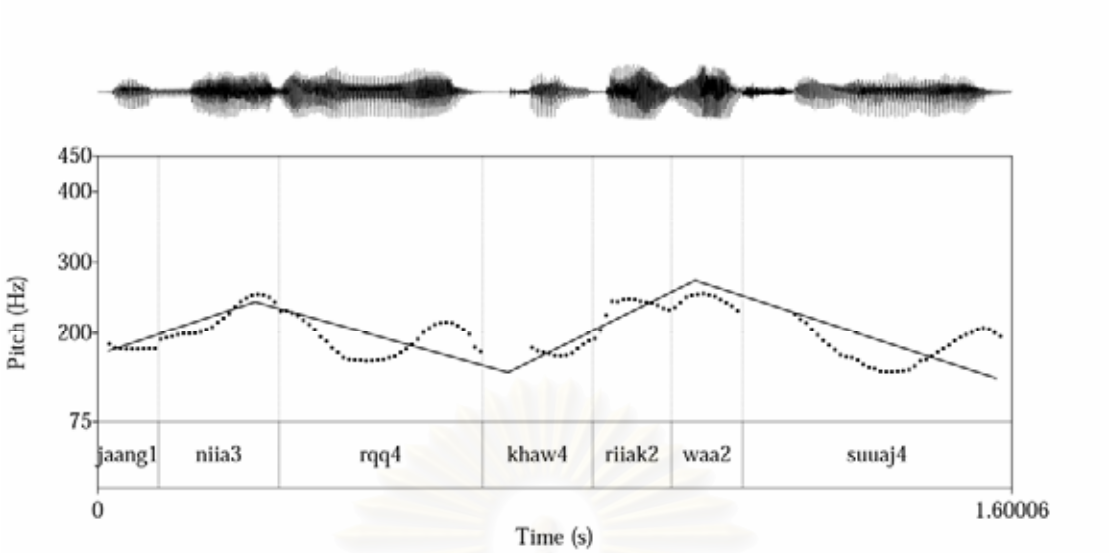
รูปที่ 11 ตัวอย่างของทำนองเสียงขึ้นแบบที่ 1: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค “คุณบอกเค้าแล้ว?” เมื่อผู้พูดเป็นผู้หญิง



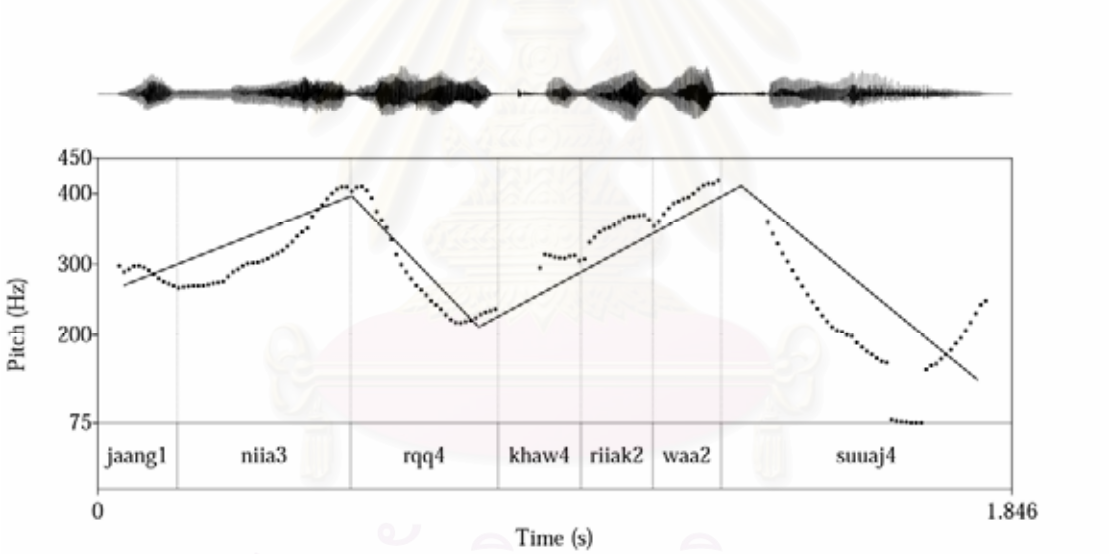
รูปที่ 12 ตัวอย่างของทำนองเสียงขึ้นแบบที่ 2: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค “แล้วพูดเรื่องอะไรกัน?” เมื่อผู้พูดเป็นผู้ชาย



รูปที่ 13 ตัวอย่างของทำนองเสียงขึ้นแบบที่ 2: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค “แล้วพูดเรื่องอะไรกัน?” เมื่อผู้พูดเป็นผู้หญิง



รูปที่ 14 ตัวอย่างของทำนองเสียงผสม: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค “อย่างเนี้ยเหรอเขาเรียกว่าสวย!” เมื่อผู้พูดเป็นผู้ชาย



รูปที่ 15 ตัวอย่างของทำนองเสียงผสม: รูปคลื่นเสียง (รูปบน) คอนทัวร์ F_0 (เส้นประในรูปล่าง) และเส้นแสดงทำนองเสียง (เส้นทึบในรูปล่าง) ของประโยค “อย่างเนี้ยเหรอเขาเรียกว่าสวย!” เมื่อผู้พูดเป็นผู้หญิง

การออกแบบคอนทัวร์

คอนทัวร์ F_0 ประกอบด้วยสารสนเทศ ทางภาษาศาสตร์ (linguistic information) สารสนเทศกึ่งภาษาศาสตร์ (paralinguistic information) และสารสนเทศที่ไม่ใช่ภาษาศาสตร์ (nonlinguistic information) ในภาษาไทย สารสนเทศทางภาษาศาสตร์ เป็นสารสนเทศที่ให้ข้อมูลเกี่ยวกับชนิดของเสียงวรรณยุกต์ในภาษาไทย สารสนเทศกึ่งภาษาศาสตร์เป็นสารสนเทศที่ให้ข้อมูลเกี่ยวกับทำนองเสียงพูดภาษาไทย สารสนเทศที่ไม่ใช่ภาษาศาสตร์ เป็นสารสนเทศที่ให้ข้อมูลเกี่ยวกับลักษณะทางความสูงต่ำของเสียงอื่น ๆ เช่น เพศ และ อายุ ของผู้พูด

จะเห็นได้ว่าการจะสร้างระบบรู้จำทำนองเสียงพูด จากคอนทัวร์ F_0 นั้น มีความจำเป็นต้อง ลดผลของสารสนเทศอื่น ๆ ที่ไม่ใช่สารสนเทศกึ่งภาษาศาสตร์ ออกไปให้มากที่สุด จากการศึกษาพบว่า เมื่อพิจารณาว่าคอนทัวร์ F_0 เป็นสัญญาณดิสครีตทางเวลา (discrete time signal) จะพบว่า สารสนเทศทางภาษาศาสตร์คือ รูปร่างขององค์ประกอบความถี่สูงของสัญญาณ (high frequency component) ซึ่งเป็นองค์ประกอบที่เปลี่ยนแปลงอย่างรวดเร็ว โดยรูปร่างของการเปลี่ยนแปลงของคอนทัวร์ F_0 ในแต่ละพยางค์จะแสดงให้เห็นถึงเสียงวรรณยุกต์ประเภทต่าง ๆ ของภาษาไทย ส่วนสารสนเทศกึ่งภาษาศาสตร์คือ รูปร่างขององค์ประกอบความถี่ต่ำของสัญญาณ (low frequency component) รวมทั้งช่วงกว้างในการแกว่งขององค์ประกอบความถี่สูงของสัญญาณ ส่วนสารสนเทศที่ไม่ใช่ภาษาศาสตร์คือระดับความสูงของคอนทัวร์ F_0 รวมทั้งช่วงกว้างในการแกว่งของคอนทัวร์ F_0 ด้วย

งานวิจัยนี้จึงได้นำเสนอคอนทัวร์ลักษณะ (feature contour) 2 คอนทัวร์ คือ คอนทัวร์ LFC (low frequency contour) และคอนทัวร์ FVC (F_0 variation contour) ซึ่งได้มาจากการนำคอนทัวร์ F_0 ไปลดผลของสารสนเทศทางภาษาศาสตร์ออกไป เพื่อให้เหมาะสมต่อการนำมาใช้รู้จำทำนองเสียงพูด ขั้นตอนในการหาคอนทัวร์ลักษณะมีดังนี้

การหาคอนทัวร์ F_0 จากสัญญาณเสียงพูด

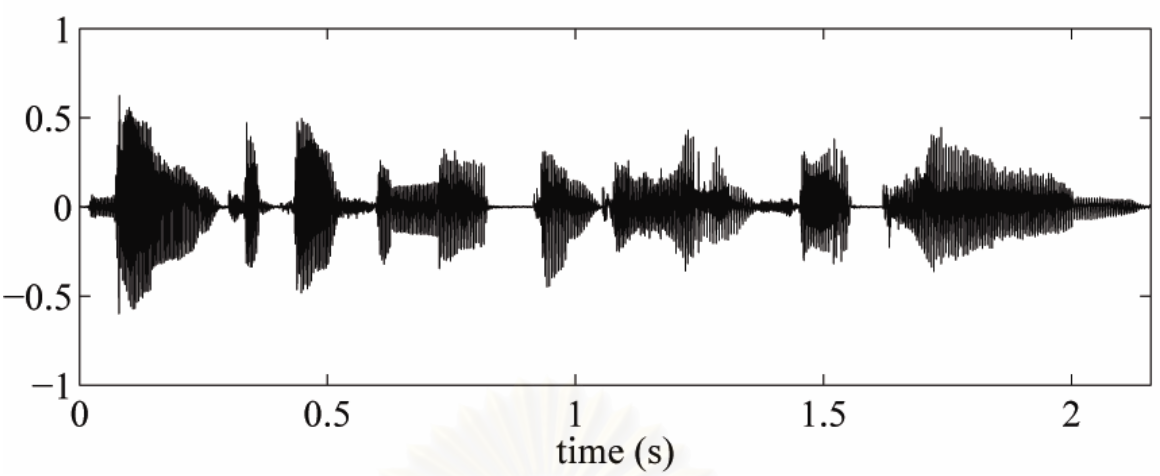
งานวิจัยนี้เลือกใช้โปรแกรม Praat (http://www.fon.hum.uva.nl/praat/) ในการหาคอนทัวร์ F_0 จากสัญญาณเสียงพูด ลักษณะของรูปคลื่นสัญญาณเสียง และคอนทัวร์ F_0 ที่หาได้โดยใช้โปรแกรม Praat แสดงดังรูปที่ 9

การทำให้คอนทัวร์ F_0 เรียบ

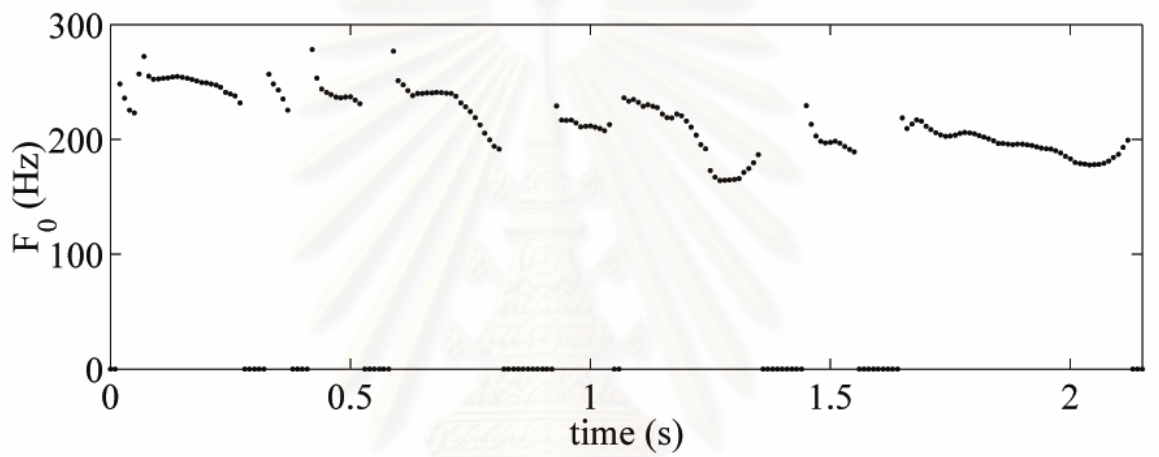
จากรูปที่ 16 จะเห็นได้ว่าคอนทัวร์ F_0 ที่หาได้ มีบางช่วงที่ไม่เรียบ ซึ่งเป็นผลมาจากกลไกในการให้กำเนิดเสียงพูด ทำให้สัญญาณพัลส์ที่ได้จากเส้นเสียงเกิดความไม่สม่ำเสมอที่เรียกว่า ปรากฏการณ์ไมโครโพรโซดิก (microprosodic effect)

วิธีที่นิยมใช้ลดความผิดพลาดดังกล่าวที่เป็นที่นิยมคือ การนำคอนทัวร์ F_0 ไปผ่านตัวกรองมัธยฐาน (median filter) จากการทดสอบกับคอนทัวร์ F_0 ที่ใช้ในงานวิจัยนี้ พบว่า การใช้ตัวกรองมัธยฐานขนาด 5 จุด (นำเอาค่า มัธยฐานของ F_0 ของF_0 ของเฟรมที่ต้องการหารวมทั้งเฟรมที่อยู่ข้างเคียง 4 เฟรม มาแทนที่ค่า F_0 ของเฟรมนั้น ๆ) สามารถกำจัดความไม่สม่ำเสมอของคอนทัวร์ F_0 ได้ดี ตัวอย่างของ คอนทัวร์ F_0 หลังจากผ่านตัวกรองมัธยฐานมีลักษณะดังรูปที่17 โดยจะเห็นได้ว่าคอนทัวร์ F_0 เรียบขึ้นเมื่อเทียบกับรูปที่ 16

หลังจากนั้นจะประมาณค่า F_0 ในช่วงที่เป็นเสียงไม่ก้องด้วยเส้นตรง เพื่อให้คอนทัวร์ F_0 มีความต่อเนื่องกันทั้งประโยค และสามารถนำไปผ่านตัวกรองได้ โดยเรียกคอนทัวร์ที่ได้จากกระบวนการนี้ว่า CF_0 (connected F_0) ดังที่แสดงในรูปที่ 18

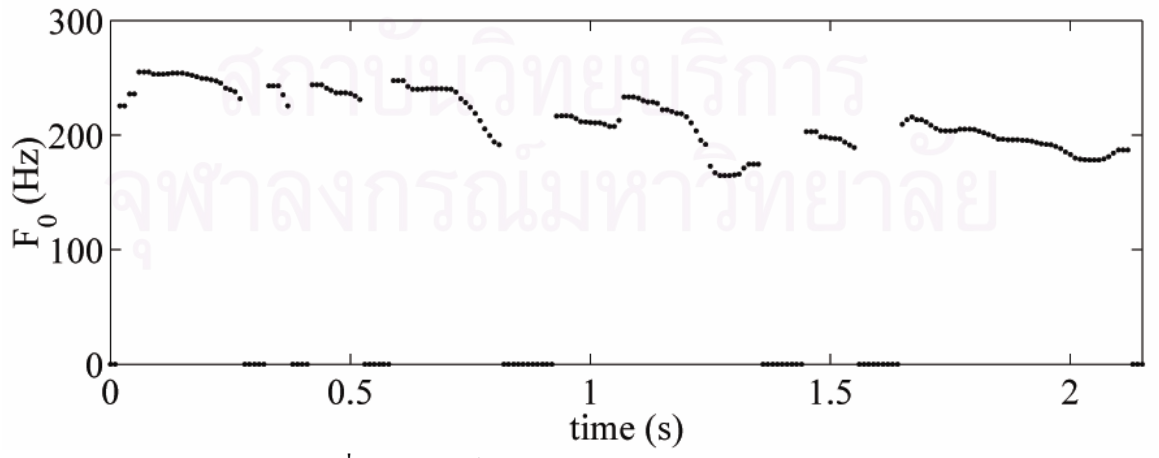


(ก)

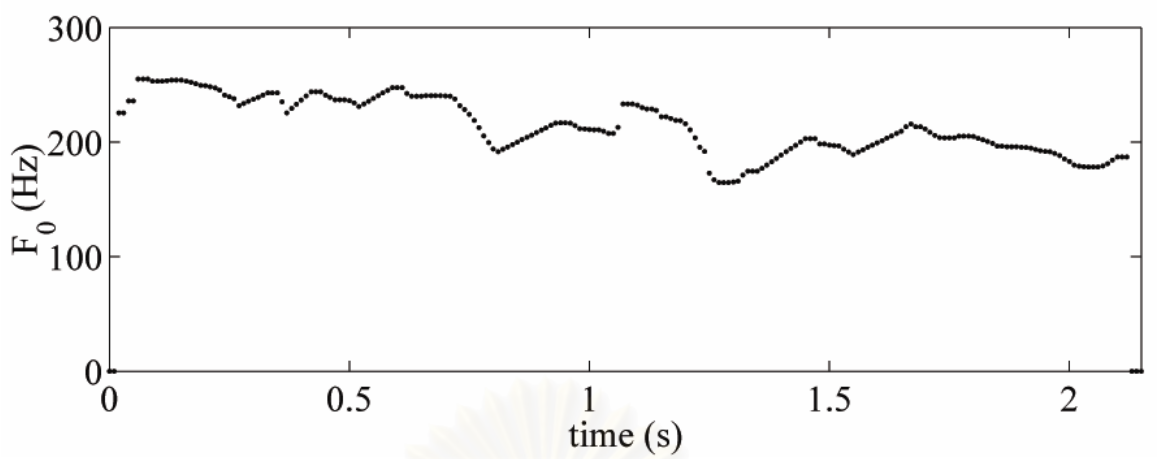


(ข)

รูปที่ 16 สัญญาณเสียงพูดที่ต้องการนำมาหาคอนทัวร์ลักษณะ:
(ก) รูปคลื่นของสัญญาณเสียงพูด (ข) คอนทัวร์ F_0 ที่หาโดยใช้โปรแกรม Praat



รูปที่ 17 คอนทัวร์ F_0 หลังจากผ่านตัวกรองมัธยฐาน

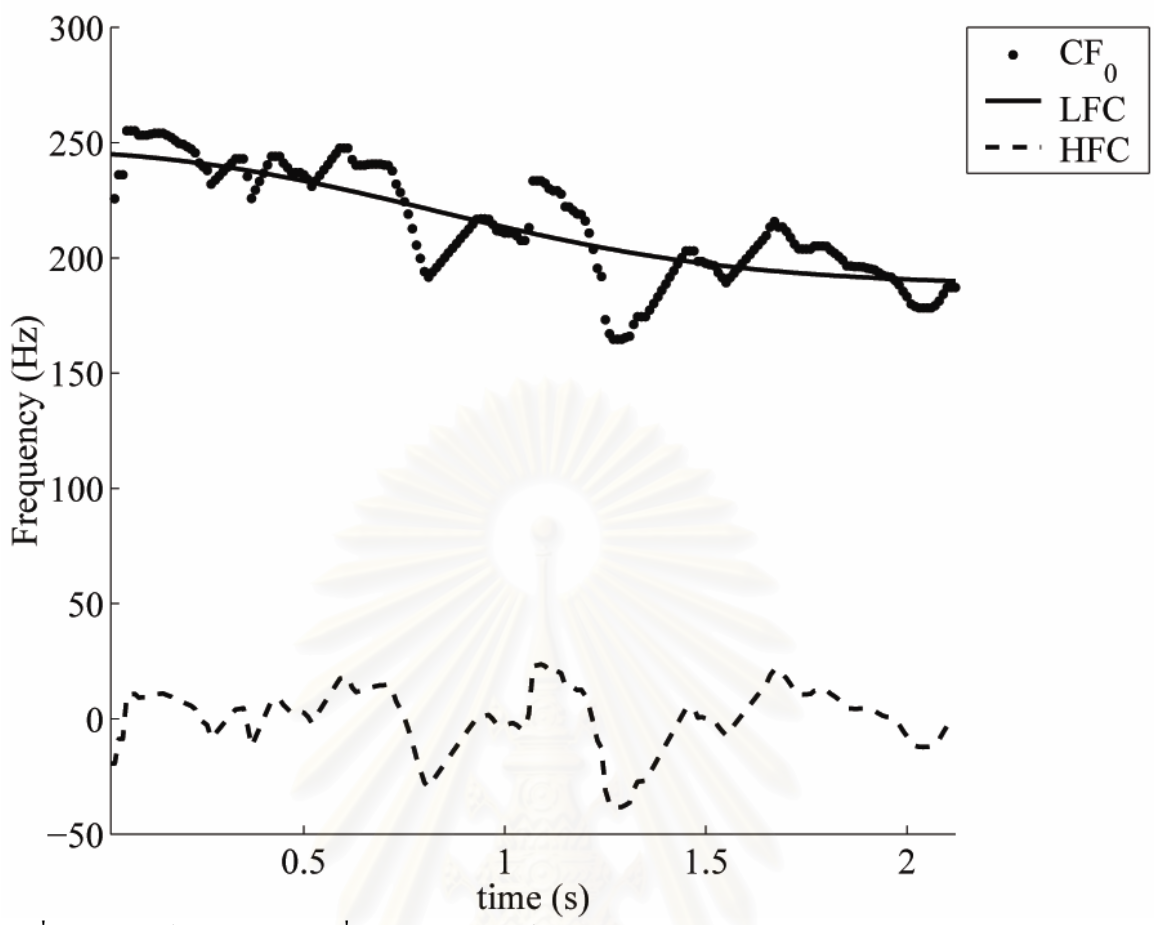


รูปที่ 18 คอนทัวร์ F_0 หลังจากผ่านการกำจัดช่วงที่เป็นเสียงไม่ก้อง (คอนทัวร์ CF_0)

การหาคอนทัวร์ LFC

สามารถหาคอนทัวร์ LFC ได้โดยการพิจารณาคอนทัวร์ CF_0 ว่าเป็นสัญญาณดิสครีตทางเวลาในลักษณะเดียวกับ โดยพิจารณาว่าองค์ประกอบวรรณยุกต์เป็นองค์ประกอบที่มีการเปลี่ยนแปลงอย่างรวดเร็ว เมื่อเทียบกับองค์ประกอบวลี จึงสามารถกำจัดองค์ประกอบวรรณยุกต์ออกไปได้โดยการใช้ตัวกรองผ่านต่ำ (low-pass filter) ซึ่งจะยอมให้สัญญาณในส่วนที่มีการเปลี่ยนแปลงอย่างช้า ๆ เท่านั้นที่สามารถผ่านไปได้ งานวิจัยนี้ได้เลือกใช้ตัวกรองแบบ FIR (finite impulse response) เนื่องจากจะทำให้สัญญาณขาออกมีการเลื่อนเฟสแบบเชิงเส้น จึงสามารถชดเชยผลของการเลื่อนเฟสของสัญญาณขาออกได้โดยง่าย โดยกำหนดให้สัญญาณขาเข้าที่ช่วงเวลาก่อนเริ่มคอนทัวร์ CF_0 และสัญญาณขาเข้าที่ช่วงเวลาหลังจากคอนทัวร์ CF_0 มีค่าเท่ากับค่าเฉลี่ยของคอนทัวร์ CF_0 เพื่อให้รูปร่างของสัญญาณขาออกไม่เพี้ยนที่ปลายทั้งสองข้าง โดยเรียกสัญญาณขาออก นี้ว่าคอนทัวร์ LFC ตามที่ได้กล่าวไปแล้วโดยกำหนดให้ Fc_{LFC} คือค่าความถี่ตัดของตัวกรองที่ใช้ในการหาคอนทัวร์ LFC ตัวอย่างของคอนทัวร์ LFC แสดงดังรูปที่ 19 (เส้นทึบ)

ในการทดลองหาค่า Fc_{LFC} ที่ให้อัตราการรู้จำทำนองเสียงพูดสูงที่สุด นอกจากจะใช้คอนทัวร์ LFC ที่ได้จากตัวกรองผ่านต่ำ งานวิจัยนี้ยังได้ใช้ LFC ที่เป็นเส้นตรงด้วย โดยเลือกใช้เส้นตรง 2 แบบ คือ เส้นตรงที่มีความชันเป็น 0 (เส้นแนวระดับ) และเส้นตรงที่มีความชัน โดยสามารถหาสมการเส้นตรงได้โดยใช้สมการถดถอย (regression equation) กับคอนทัวร์ CF_0



รูปที่ 19 คอนทัวร์ LFC และ HFC ที่หาได้จากคอนทัวร์ CF_0

การหาคอนทัวร์ FVC

จากที่กล่าวไปแล้วข้างต้นว่า องค์ประกอบเสียงวรรณยุกต์มีลักษณะที่แสดงให้เห็นถึงประเภทของทำนองเสียงได้ นั่นก็คือช่วงกว้างของการเปลี่ยนแปลงค่าของคอนทัวร์ F_0 เราจึงไม่สามารถทิ้งองค์ประกอบเสียงวรรณยุกต์ไปได้

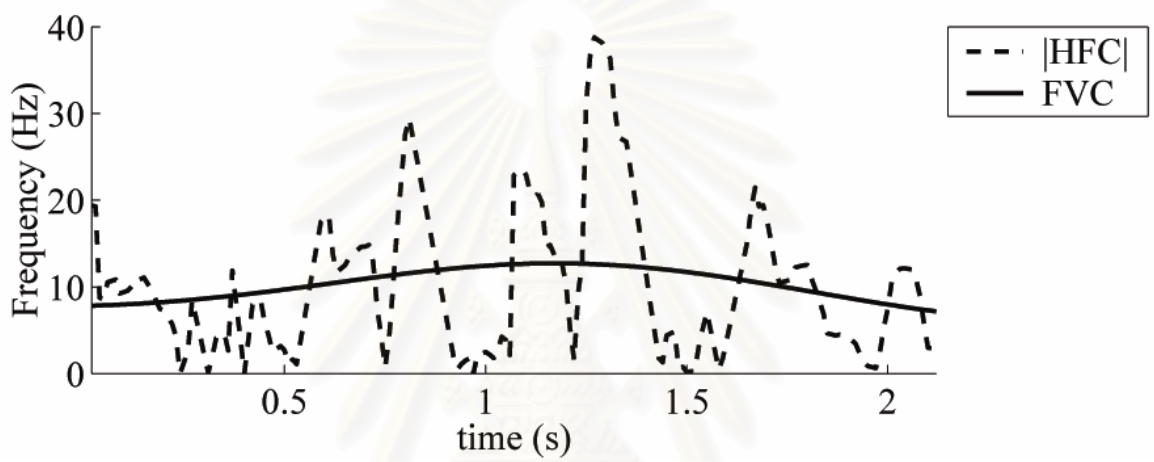
องค์ประกอบเสียงวรรณยุกต์เป็นองค์ประกอบที่มีการเปลี่ยนแปลงอย่างรวดเร็ว เราจึงสามารถสกัดเอาองค์ประกอบเสียงวรรณยุกต์จากคอนทัวร์ CF_0 ได้โดยใช้ตัวกรองผ่านสูง (high-pass filter) แต่เนื่องจากเรามีองค์ประกอบความถี่ต่ำ คือคอนทัวร์ LFC อยู่แล้ว เราจึงสามารถหาองค์ประกอบความถี่สูงได้โดยการนำคอนทัวร์ LFC ไปลบออกจากคอนทัวร์ CF_0 โดยเรียกคอนทัวร์ที่ได้ว่า คอนทัวร์ HFC (high frequency contour) แสดงดังรูปที่ 19 (เส้นประ)

สิ่งที่แสดงให้เห็นถึงลักษณะของทำนองเสียงในคอนทัวร์ HFC คือ ช่วงกว้างในการแกว่งของคอนทัวร์ HFC โดยไม่ต้องคำนึงถึงว่า เป็นการแกว่งขึ้น (HFC > 0) หรือการแกว่งลง (HFC < 0) จึงนำ HFC ไปใส่ค่าสัมบูรณ์ ลักษณะของ |HFC| แสดงดังรูปที่ 20

จากรูปที่ 20 จะเห็นว่า คอนทัวร์ |HFC| นั้นไม่เรียบ ซึ่งเป็นผลจากเสียงวรรณยุกต์ของแต่ละพยางค์ ที่ให้สารสนเทศภาษาศาสตร์ งานวิจัยนี้จึงได้นำ |HFC| ไปผ่านตัวกรองผ่านต่ำแบบ FIR แบบเดียวกับที่ใช้

ในการหาคอนทัวร์ LFC เพื่อกำจัดสารสนเทศทางภาษาศาสตร์ โดยจะได้ว่าสัญญาณขาออกของตัวกรอง แสดงให้เห็นถึงความมากน้อยในการแกว่งของ HFC ซึ่งก็คือความมากน้อยในการกวัดแกว่ง หรือการเปลี่ยนแปลงค่าของคอนทัวร์ F_0 นั่นเอง จึงเรียกสัญญาณขาออกนี้ว่า คอนทัวร์ FVC (F_0 variation contour) ดังแสดงในรูปที่ 20 โดยกำหนดให้ความถี่ตัดของตัวกรองผ่านต่ำ ที่ใช้หาคอนทัวร์ FVC มีค่าเป็น Fc_{LFC}

คอนทัวร์ FVC เป็นคอนทัวร์ที่แสดงให้เห็นถึงความมากน้อยในการเปลี่ยนแปลงค่าของ F_0 ถ้าคอนทัวร์ FVC มีระดับที่สูง แสดงว่าคอนทัวร์ F_0 มีช่วงการแกว่งที่กว้าง ซึ่งจะพบในทำนองเสียงทำนองเสียงพูดแบบผสม แต่ถ้าคอนทัวร์ FVC มีระดับที่ต่ำ แสดงว่าคอนทัวร์ F_0 มีช่วงการแกว่งที่แคบ ซึ่งมักจะพบได้ในทำนองเสียงพูดแบบตก หรือทำนองเสียงพูดแบบขึ้น



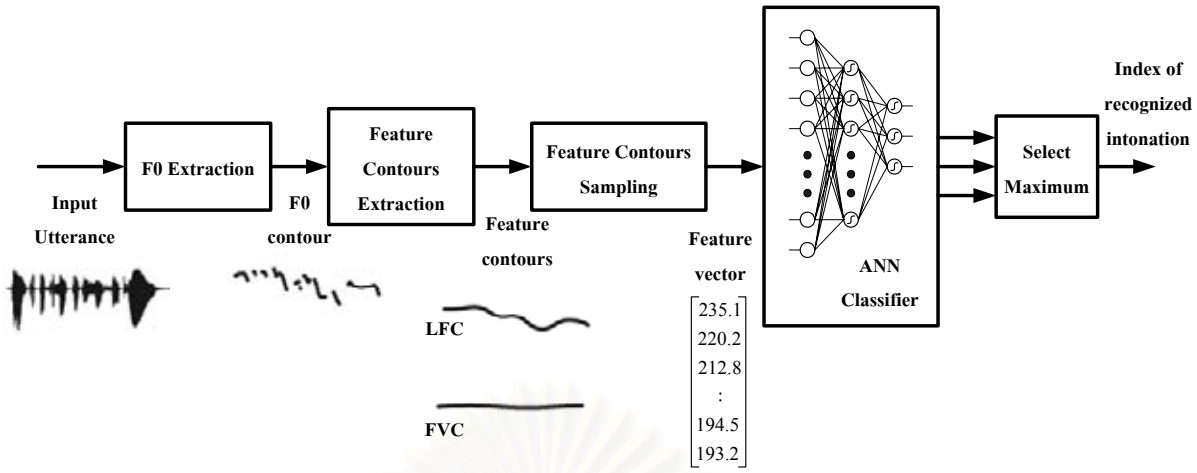
รูปที่ 20 คอนทัวร์ FVC ที่หาได้จากการนำ |HFC| ไปผ่านตัวกรองผ่านต่ำ

ระบบรู้จำทำนองเสียงพูด

ระบบรู้จำทำนองเสียงพูดภาษาไทยที่ออกแบบขึ้นเพื่อใช้ในการทดลองแสดงในรูปที่ 14 ระบบจะนำเสียงพูดที่ต้องการรู้จำทำนองเสียงไปหาค่า F_0 ในส่วน F_0 Extraction แล้วจึงนำคอนทัวร์ F_0 ที่ได้ไปหาคอนทัวร์ลักษณะ LFC และ FVC ในส่วน Feature Contours Extraction ตามวิธีที่แสดงในข้อ 2

จากนั้นจึงสุ่มตัวอย่าง (sampling) คอนทัวร์ลักษณะทั้งสอง โดยกำหนดให้จำนวนจุดที่ใช้สุ่มตัวอย่างขึ้นกับค่าความถี่ตัดของตัวกรองที่ใช้ในการหาคอนทัวร์ LFC และ FVC ตามทฤษฎีการสุ่มตัวอย่างของไนควิสต์ (Nyquist's sampling theory) คือสุ่มด้วยอัตรา 2 เท่าของความกว้างของช่วงความถี่ (bandwidth)

ค่าที่นำมาใช้เป็นเวกเตอร์ลักษณะ ได้จากความยาวของประโยค ค่าของ LFC และ FVC รวมทั้งค่าผลต่างอันดับหนึ่ง (first difference) ของคอนทัวร์ทั้งสอง (ΔLFC และ ΔFVC) ที่จุดที่สุ่มตัวอย่าง



รูปที่ 21 แผนภาพของระบบรู้จำทำนองเสียงพูด

การทดลอง

ข้อมูลเสียงพูดที่ใช้ในการวิจัย

ประโยคที่ใช้ทดสอบ มีทั้งสิ้น 61 ประโยค มาจากบทสนทนา 6 บท ดังที่ได้แสดงไว้ในภาคผนวก ข แต่ละบทมีผู้พูด 2 คนพูดคุยกัน โดยในแต่ละประโยคพูดได้เขียนกำกับไว้ว่าจะต้องพูดด้วยทำนองเสียงพูดแบบใด ความยาวของแต่ละประโยคจะอยู่ระหว่าง 1 – 11 พยางค์

ข้อมูลเสียงที่ใช้มาจากผู้พูด 12 คน ผู้ชาย 6 คน ผู้หญิง 6 คน ผู้พูดจะจับคู่กันและสนทนาตามบทพูด โดยผู้พูดทุกคนจะพูดแต่ละบทสนทนา 2 ครั้ง โดยสลับบทพูดกัน เพื่อให้ผู้พูดทุกคนได้พูดครบทั้ง 61 ประโยค จึงมีประโยคสำหรับทดสอบทั้งสิ้น $12 \times 61 = 732$ ประโยค

เนื่องจากผู้พูดแต่ละคนได้รับคำสั่งว่าให้พูดตามบทสนทนาให้เป็นธรรมชาติมากที่สุด จึงทำให้มีบางประโยค ที่ผู้พูดพูดด้วยทำนองเสียงที่ไม่ตรงกับทำนองเสียงที่ได้กำหนดไว้ให้ จึงได้กำหนดให้มีการนำข้อมูลเสียงทั้งหมดมาตรวจสอบประเภทของทำนองเสียงอีกครั้ง โดยการเขียนโปรแกรมเพื่อนำเสียงทั้ง 732 ประโยค มาสลับลำดับแบบสุ่ม แล้วเปิดให้ผู้ฟังทั้ง 2 คนฟัง แล้วให้ผู้ฟังตัดสินใจว่าเสียงที่ได้ยิน เป็นทำนองเสียงพูดแบบใด โดยให้ผู้ฟังแต่ละคนตรวจสอบเสียงพูดทุกเสียงคนละ 2 รอบ ดังนั้นจะได้ว่ามีการทดสอบการฟัง 4 ครั้งสำหรับประโยคเสียงพูดแต่ละประโยค งานวิจัยนี้จะนำเฉพาะประโยคที่มีทำนองเสียงที่ชัดเจนเท่านั้น มาทำการทดลอง โดยกำหนดว่าประโยคที่มีทำนองเสียงที่ชัดเจน คือประโยคที่ผู้ฟังเลือกให้มีทำนองเสียงประเภทเดียวกัน 3 ครั้งขึ้นไป จากการฟังทั้งหมด 4 ครั้ง

การทดลองรู้จำทำนองเสียง

การทดลองในงานวิจัยนี้จะแบ่งเป็น 2 ส่วนใหญ่ ๆ คือ การทดลองในชุดแรก จะแบ่งทำนองเสียงออกเป็น 3 ประเภท คือ ทำนองเสียงตก ทำนองเสียงขึ้น และทำนองเสียงผสม ส่วนการทดลองในชุดที่สอง

จะแบ่งทำนองเสียงออกเป็น 2 ประเภท คือ ทำนองเสียงตก และทำนองเสียงขึ้น โดยจะจัดกลุ่มทำนองเสียงผสมเข้าไปอยู่ในประเภทเดียวกับทำนองเสียงขึ้น

การทดลองในแต่ละชุด จะเริ่มจากการใช้ความยาวของประโยค และ LFC จากนั้นจึงเพิ่ม ΔLFC และ FVC เข้าไป เพื่อทดสอบผลของลักษณะต่าง ๆ ต่ออัตราการรู้จำ ในแต่ละการทดลองจะเปลี่ยนค่าความถี่ตัดของตัวกรองที่ใช้สร้างคอนทัวร์ LFC (Fc_{LFC}) และค่าความถี่ตัดของตัวกรองที่ใช้สร้างคอนทัวร์ FVC (Fc_{FVC}) โดยกำหนดให้ค่าทั้งสองมีค่าเป็น 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0 และ 4.5 Hz โดยจับคู่ Fc_{LFC} และ Fc_{FVC} ให้ครบทุกแบบเท่าที่เป็นไปได้ นอกจากนี้ยังได้ใช้ LFC และ FVC ที่เป็นเส้นตรง และเส้นตรงแนวระดับ (มีความชันเป็น 0) ซึ่งได้จากการหาสมการถดถอย (regression equation) ของคอนทัวร์ CF_0 และ คอนทัวร์ |HFC| อีกด้วย

การทดลองแต่ละประเภท จะเป็นลักษณะที่ขึ้นกับผู้พูด (speaker dependent) นั่นคือ ผู้พูดที่อยู่ในกลุ่มฝึกฝน และผู้พูดที่อยู่ในกลุ่มทดสอบเป็นกลุ่มเดียวกัน โดยในแต่ละการทดลองจะทดลองเสียงของผู้ชายและเสียงของผู้หญิงแยกจากกัน เนื่องจากคอนทัวร์ F_0 ของเสียงผู้ชาย และคอนทัวร์ F_0 ของเสียงผู้หญิงมีความแตกต่างกันมาก

ผลการทดลอง

กรณีที่แบ่งทำนองเสียงออกเป็น 3 ประเภท

ในกรณีของเสียงผู้ชาย การทดลองที่ให้อัตราการรู้จำที่ดีที่สุด (อัตราการรู้จำเฉลี่ยมีค่าสูงที่สุด และอัตราการรู้จำของทำนองเสียงที่มีอัตราการรู้จำต่ำที่สุดมีค่าสูงที่สุด) เกิดขึ้นเมื่อใช้ความยาวของประโยค และจุดสุ่มตัวอย่างของ LFC และ FVC เป็นเวกเตอร์ลักษณะ ซึ่งให้อัตราการรู้จำเฉลี่ยที่ร้อยละ 61.6 ผลการรู้จำแสดงในตารางที่ 5

ตารางที่ 5 ผลการรู้จำทำนองเสียง (%) ในกรณีที่ให้อัตราการรู้จำสูงที่สุดของเสียงผู้ชาย
เมื่อ Fc_{LFC} = 3.5 Hz และ Fc_{FVC} = 3.0 Hz

ประเภทของทำนองเสียงที่นำมารู้จำ	ประเภทของทำนองเสียงที่รู้จำได้			จำนวนประโยค
	ตก	ขึ้น	ผสม	
ตก	**84.7**	**10.2**	5.1	98
ขึ้น	**29.7**	**47.3**	23.1	91
ผสม	25.3	**25.3**	**49.4**	79
จำนวนประโยคทั้งหมด				268

ในกรณีของเสียงผู้หญิง การทดลองที่ให้อัตราการรู้จำที่ดีที่สุด เกิดขึ้นเมื่อใช้ความยาวของประโยค และจุดสุ่มตัวอย่างของ LFC ΔLFC และ FVC เป็นเวกเตอร์ลักษณะ ซึ่งให้อัตราการรู้จำเฉลี่ยอยู่ที่ร้อยละ 73.7 ผลการรู้จำแสดงในตารางที่ 6

ตารางที่ 6 ผลการรู้จำทำนองเสียง (%) ในกรณีที่ให้อัตราการรู้จำสูงที่สุด ของเสียงผู้หญิง
เมื่อ Fc_{LFC} = 3.0 Hz และ Fc_{FVC} = 1.5 Hz

ประเภทของทำนองเสียงที่นำมารู้จำ	ประเภทของทำนองเสียงที่รู้จำได้			จำนวนประโยค
	ตก	ขึ้น	ผสม	
ตก	**90.6**	0.9	**8.5**	106
ขึ้น	16.9	**43.1**	**40.0**	65
ผสม	11.5	**13.1**	**75.4**	122
	จำนวนประโยคทั้งหมด			293

กรณีที่แบ่งทำนองเสียงออกเป็น 2 ประเภท

ในกรณีของเสียงผู้ชาย การทดลองที่ให้อัตราการรู้จำที่ดีที่สุด เกิดขึ้นเมื่อใช้ความยาวของประโยค และจุดสุ่มตัวอย่างของ LFC และ FVC เป็นเวกเตอร์ลักษณะ ซึ่งให้อัตราการรู้จำเฉลี่ยที่ร้อยละ 81.7 ผลการรู้จำแสดงในตารางที่ 7

ตารางที่ 7 ผลการรู้จำทำนองเสียง (%) ในกรณีที่ให้อัตราการรู้จำสูงที่สุด ของเสียงผู้ชาย
เมื่อ Fc_{LFC} =1.5 Hz และใช้ FVC เป็นเส้นตรงที่มีความชันเป็น 0

ประเภทของทำนองเสียงที่นำมารู้จำ	ประเภทของทำนองเสียงที่รู้จำได้		จำนวนประโยค
	ตก	ขึ้น	
ตก	**77.6**	22.4	98
ขึ้น	15.9	**84.1**	170
	จำนวนประโยคทั้งหมด		268

ในกรณีของเสียงผู้หญิง การทดลองที่ให้อัตราการรู้จำที่ดีที่สุด เกิดขึ้นเมื่อใช้ความยาวของประโยค และจุดสุ่มตัวอย่างของ LFC ΔLFC และ FVC เป็นเวกเตอร์ลักษณะ ซึ่งให้อัตราการรู้จำเฉลี่ยที่ร้อยละ 90.8 ผลการรู้จำแสดงในตารางที่ 8

ตารางที่ 8 ผลการรู้จำทำนองเสียง (%) ในกรณีที่ให้อัตราการรู้จำสูงที่สุด ของเสียงผู้หญิง
เมื่อ Fc_{LFC} = 3.0 Hz และใช้ Fc_{FVC} = 1.0 Hz

ประเภทของทำนองเสียงที่นำมารู้จำ	ประเภทของทำนองเสียงที่รู้จำได้		จำนวนประโยค
	ตก	ขึ้น	
ตก	**90.6**	9.4	106
ขึ้น	9.1	**90.9**	187
จำนวนประโยคทั้งหมด			293

วิเคราะห์ผลการทดลอง

จากผลการทดลอง จะเห็นได้ว่า ทั้ง คอนทัวร์ LFC และคอนทัวร์ FVC ต่างก็ช่วยให้อัตราการรู้จำทำนองเสียงสูงขึ้นกว่าการใช้เพียงคอนทัวร์ F_0 (งานวิจัยนี้ไม่ได้ใช้คอนทัวร์ F_0 โดยตรง แต่สามารถเปรียบเทียบได้ว่าการสุ่มตัวอย่างคอนทัวร์ LFC ที่มีค่า Fc_{LFC} สูง ๆ มีค่าใกล้เคียงกับการสุ่มตัวอย่างคอนทัวร์ F_0) ทั้งในการทดลองที่แบ่งทำนองเสียงเป็น 3 ประเภท และในการทดลองที่แบ่งทำนองเสียงเป็น 2 ประเภท

ในการทดลองที่แบ่งทำนองเสียงเป็น 3 ประเภท จะเห็นได้ว่าตัวรู้จำมีความสับสนระหว่างทำนองเสียงขึ้น และทำนองเสียงผสมอยู่มาก เป็นผลทำให้อัตราการรู้จำเฉลี่ยมีค่าเพียงร้อยละ 61.6 สำหรับเสียงผู้ชาย และร้อยละ 73.7 สำหรับเสียงผู้หญิง

ในการทดลองที่แบ่งทำนองเสียงออกเป็น 2 ประเภท โดยกำหนดให้ทำนองเสียงผสมเป็นประเภทเดียวกับทำนองเสียงขึ้น พบว่าอัตราการรู้จำเฉลี่ยเพิ่มสูงขึ้นมาก โดยมีค่าเป็นร้อยละ 81.7 สำหรับเสียงผู้ชาย และร้อยละ 90.8 สำหรับเสียงผู้หญิง

สรุปผล

งานวิจัยนี้ได้นำเสนอคอนทัวร์ LFC และคอนทัวร์ FVC ขึ้นเพื่อนำมาใช้เป็นคอนทัวร์ลักษณะสำหรับระบบรู้จำทำนองเสียงพูดภาษาไทย ซึ่งจากการทดลองสามารถสรุปได้ว่าคอนทัวร์ทั้งสองสามารถทำให้อัตราการรู้จำทำนองเสียงพูดสูงขึ้นกว่าการใช้คอนทัวร์ F_0 ในการทดลองที่แบ่งทำนองเสียงเป็น 3 ประเภท คือ ทำนองเสียงตก ทำนองเสียงขึ้น และทำนองเสียงผสม จะเห็นได้ว่าตัวรู้จำมีความสับสนระหว่างทำนองเสียงขึ้นและทำนองเสียงผสมอยู่มาก ซึ่งจากการวิเคราะห์คอนทัวร์ LFC และคอนทัวร์ FVC โดยเฉลี่ย พบว่าคอนทัวร์ทั้งสอง ของสองทำนองเสียงนี้มีความใกล้เคียงกันมาก และจำเป็นต้องมีการศึกษาโดยละเอียดเพื่อหาลักษณะอื่น ๆ ของเสียงพูดมาช่วยในการรู้จำ เช่นอัตราเร็วในการพูด หรือระดับพลังงานของเสียงพูด โดยสามารถนำระบบรู้จำทำนองเสียงที่นำเสนอในงานวิจัย ไปใช้จำแนกทำนองเสียงตก ออกจากทำนองเสียงขึ้น และทำนองเสียงผสมก่อนแล้วจึงนำลักษณะอื่น ๆ ไปใช้จำแนกทำนองเสียง 2 ประเภทนี้ ซึ่งในการทดลองที่แบ่งทำนองเสียงออกเป็น 2 ประเภท (กำหนดให้ทำนองเสียงผสมเป็นทำนองเสียงประเภทเดียวกับทำนองเสียงขึ้น) ได้แสดงให้เห็นว่าระบบรู้จำสามารถจำแนกความแตกต่างระหว่างทำนองเสียงตกออกจากทำนองเสียงประเภทอื่น ๆ ได้ดี (ร้อยละ 81.7 สำหรับเสียงผู้ชาย และร้อยละ 90.8 สำหรับเสียงผู้หญิง)

5 งานวิจัยที่จะพัฒนาในปีถัดไป

- Telephone-Quality Speech Corpus (Monologue / Dialogue)
  ฐานข้อมูลเสียงพูดต่อเนื่องในระบบโทรศัพท์
  - ♦ Wired telephones
  - ♦ Wireless mobile phones—GSM900/GSM1800/CDMA
- Thai Continuous Speech Recognizer (Phase III)
  สร้างตัวระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยระยะที่ 3
  - ♦ Baseline Thai continuous speech recognition system
  - ♦ Thai CSR system development toolkit
  - ♦ Thai CSR baseline demonstration system
  - ♦ Language modeling development toolkit
- On-line Telephone-based Thai CSR System
  สร้างตัวระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยบนพื้นฐานของระบบโทรศัพท์
  - ♦ Monologue / Dialogue
  - ♦ On-line recording of speech/voice over telephone channel

6 ผลิตผลและหรือความสัมฤทธิผลของงาน

6.1 วิทยานิพนธ์จำนวน 3 เรื่องได้แก่

6.1.1 ระดับปริญญาดุษฎีบัณฑิต

- นายเอกฤทธิ์ มณีน้อย, "การศึกษาหน่วยตามของพยางค์เชิงกลสัทศาสตร์: พื้นฐานสำหรับระบบการรู้จำเสียงพูดต่อเนื่องภาษาไทย"

6.1.2 ระดับปริญญามหาบัณฑิต

- นายนัทธี งามเจตนารมย์, "การรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทยบนพื้นฐานแบบจำลอง ฟูจิซากิ"
- นายปฐวี ชาญไววิทย์, "ระบบรู้จำทำนองเสียงพูดสำหรับเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม"

6.2 บทความวิจัยจำนวน 4 เรื่อง ได้แก่

6.2.1. Maneenoi E., Ahkuputra V., Luksaneeyanawin S., and Jitapunkul S., "A Study on Acoustic Modeling for Speech Recognition of Predominantly Monosyllabic Languages", IEICE Transactions on Information and System, Vol.E87-D, No. 5, May 2004.

6.2.2. Jitapunkul S., Maneenoi E., Ahkuputra V, and Luksaneeyanawin S. "Performance Evaluation of Phonotactic and Contextual Onset-Rhyme Models for Speech Recognition of Thai Language", Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, pp. 1841-1844, 2003.

6.2.3. Charnvivit P., Thubthong N., Maneenoi E., Luksaneeyanawin S., and Jitapunkul S., "Recognition of Intonation Patterns in Thai Utterance", Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, pp. 137-140, 2003.

6.2.4. Ngarmchatetanarom N., Maneenoi E., Asdornwised W., and Jitapunkul S., "Tone Recognition of Thai Continuous Speech Using Fujisaki's Model", Proceedings of the 17th IEEE Canadian Conference on Electrical and Computer Engineering, Niagara Falls, Canada, May, 2-5, 2004.

6.2. ความตกลงร่วมมือกับภาคเอกชนและหรือรัฐวิสาหกิจจำนวน 1 ฉบับ ได้แก่

6.2.1. สัญญาความร่วมมือกับบริษัท SUN Systems Corporation Limited โดยผ่านทางศูนย์วิจัยประมวลผลภาษาและวัจนะ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย