การรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทยบนพื้นฐานแบบจำลองฟูจิซากิ

นายนัทธี งามเจตนรมย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2546
ISBN 974-17-4710-1

TONE RECOGNITION IN CONTINUOUS THAI SPEECH BASED ON FUJISAKI MODEL

Mr. Nutthee Ngarmchatetanarom

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Electrical Engineering

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2003

ISBN 974-17-4710-1

หัวข้อวิทยานิพนธ์	การรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทยบนพื้นฐานแบบจำลองฟูจิซากิ
โดย	นายนัทธี งามเจตนรมย์
สาขาวิชา	วิศวกรรมไฟฟ้า
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร.สมชาย จิตะพันธ์กุล
อาจารย์ที่ปรึกษาร่วม	อ.วิทยากร อัศดรวิเศษ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

…… ประธานกรรมการ
(อาจารย์ สุวิทย์ นาคพีระยุทธ)

…….. อาจารย์ที่ปรึกษา
(รศ.ดร.สมชาย จิตะพันธ์กุล)

……อาจารย์ที่ปรึกษาร่วม
(อ.วิทยากร อัศดรวิเศษ)

……กรรมการ
(อ.ดร.ณัฐกร ทับทอง)

นัทธี งามเจตนรมย์ : การรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทยบนพื้นฐานแบบจำลองฟูจิซากิ. (TONE RECOGNITION IN CONTINUOUS THAI SPEECH BASED ON FUJISAKI MODEL) อ. ที่ปรึกษา : รศ.ดร.สมชาย จิตะพันธ์กุล, อ.ที่ปรึกษาร่วม : อ.วิทยากร อัศดรวิเศษ 72 หน้า. ISBN 974-17-4710-1.

การรู้จำวรรณยุกต์ของคำพูดต่อเนื่องภาษาไทย เป็นส่วนเพิ่มเติมความสามารถในระบบการรู้จำเสียงพูดในคำพูดต่อเนื่องภาษาไทย วิทยานิพนธ์ฉบับนี้เสนอการรู้จำวรรณยุกต์ของคำพูดต่อเนื่องในภาษาไทยโดยประยุกต์ใช้พารามิเตอร์ของแบบจำลองฟูจิซากิเป็นค่าคุณลักษณะสำคัญเพื่อลดผลของการลดลงของเสียง และการควบร่วมของโทนเสียง อีกทั้งลดความซับซ้อนในการกระบวนการรู้จำโดยการลดขนาดของเวกเตอร์คุณลักษณะสำคัญ นอกจากนั้นยังนำเสนอกรรมวิธีหาค่าพารามิเตอร์ของแบบจำลองฟูจิซากิแบบอัตโนมัติ โดยผลการทดสอบการรู้จำให้ค่าเฉลี่ยความถูกต้องการรู้จำร้อยละ 96.35 และสำหรับกรณีการหาค่าพารามิเตอร์โดยอัตโนมัติให้ค่าเฉลี่ยความถูกต้องในการรู้จำร้อยละ 70.27

ภาควิชา.....วิศวกรรมไฟฟ้า..... ลายมือชื่อนิสิต.....
สาขาวิชา.....วิศวกรรมไฟฟ้า..... ลายมือชื่ออาจารย์ที่ปรึกษา.....
ปีการศึกษา 2546 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

4470372421 : MAJOR ELECTRICAL ENGINEERING

KEY WORD: THAI SPEECH / FUJISAKI MODEL / NEURAL NETWORK / TONE RECOGNITION /

NUTTHEE NGARMCHATETANAROM : TONE RECOGNITION IN CONTINUOUS THAI SPEECH BASED ON FUJISAKI MODEL. THESIS ADVISOR : ASSOC. PROF. SOMCHAI JITAPUNKUL, Dr.Ing., THESIS COADVISOR : WIDHYAKORN ASDORNWISED, 72 pp. ISBN 974-17-4710-1.

Thai language is a tonal language. Thus tone characteristic should be an essential characteristic in order to add up performance of recognition engine in Thai continuous speech recognition. This thesis purposes a tone recognition system in continuous Thai speech using parameters obtained from Fujisaki model as features. Its advantageous are those of reducing declination and tonal assimilation effects and also reducing the complexity of recognition system by decreasing the size of feature vectors. An automatic Fujisaki model parameters extraction process is proposed and the experimental results show that the recognition rates are 96.35% and 70.27% for manually process and automatically process, respectively.

Department.....ElectricalEngineering…….. Student's……

Field of study....ElectricalEngineering.....…… Advisor's……...

Academic year 2003..... Co-advisor's signature.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยการผลักดัน คำแนะนำและความช่วยเหลืออย่างดียิ่งของอาจารย์ที่ปรึกษาวิทยานิพนธ์ คือ รศ.ดร.สมชายจิตะพันธ์กุล พร้อมทั้งคำแนะนะของอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม คือ อาจารย์วิทยากร อัศดรวิเศษ ผู้วิจัยจึงขอกราบขอบพระคุณมา ณ ที่นี้

ขอขอบคุณอาจารย์ ดร.ณัฐกร ทับทอง ที่ได้ให้ความอนุเคราะห์ชุดข้อมูลเพื่อใช้สำหรับงานวิจัยในวิทยานิพนธ์ฉบับนี้

ขอขอบคุณพี่ ๆ เพื่อน ๆ และน้อง ๆ ในห้องปฏิบัติการวิจัยกรรมวิธีสัญญาณดิจิทัลทุก ๆ คน รวมทั้งบุคคลรอบตัวผู้วิจัยสำหรับความคิดเห็นและแรงบันดาลใจที่มีต่องานชิ้นนี้

สุดท้ายผู้วิจัยขอขอบคุณครอบครัวที่เป็นกำลังใจและคอยให้ความสนับสนุนอยู่เบื้องหลังผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

หน้า

บทที่ หน้า

บทที่ หน้า

บทที่ หน้า

สารบัญตาราง

สารบัญภาพ

ภาพประกอบ หน้า

บทที่ หน้า

บัญชีคำศัพท์

Accent command	คำสั่งสำเนียง
Accent component	องค์ประกอบสำเนียง
Accent control mechanism	กลไกควบคุมสำเนียง
Artificial neural network	โครงข่ายประสาทเทียม
Cepstrum	เซปสตรัม
Classification	การจำแนก
Command-line input	การป้อนบรรทัดคำสั่ง
Consonant	พยัญชนะ
Critical damp	หน่วงวิกฤติ
Declination	การลดระดับของเสียง
Discrete Fourier Transform	การแปลงฟูริเยร์แบบไม่ต่อเนื่อง
F0 contour	เส้นโค้งความถี่มูลฐาน
Falling tone	เสียงโท
Feature	ค่าคุณลักษณะสำคัญ
Feature extraction	การสกัดคุณลักษณะสำคัญ
Fujisaki model	แบบจำลองฟูจิซากิ
Fundamental frequency	ความถี่มูลฐาน
Global Phenomena	ผลในวงกว้าง
Glottal pulse	พัลส์เส้นเสียง
Hidden layer	ชั้นซ่อน
High frequency contour	เส้นโค้งความถี่สูง
High tone	เสียงตรี
Human-Machine Interface	การติดต่อระหว่างมนุษย์กับเครื่องจักร
Impulse response	ฟังก์ชันตอบสนองอิมพัลส์
Interpolation	การประมาณค่าในช่วง
Linear	เชิงเส้น
Linguistic information	สารสนเทศเชิงภาษาศาสตร์
Local maximum	ค่าสูงสุดเฉพาะที่
Local minimum	ค่าต่ำสุดเฉพาะที่

Local Phenomena	ผลเฉพาะจุด
Low frequency contour	เส้นโค้งความถี่ต่ำ
Low tone	เสียงเอก
Macroprosodic	สัทสัมพันธ์มหัพภาค
Microprosodic	จุลสัทสัมพันธ์
Mid tone	เสียงสามัญ
Non-linear	ไม่เป็นเชิงเส้น
Non-periodic	ไม่เป็นรายคาบ
Nonlinguistic information	สารสนเทศที่ไม่ใช่ภาษาศาสตร์
Paralinguistic information	สารสนเทศกึ่งภาษาศาสตร์
Pattern recognition	การรู้จำแบบรูป
Periodic	รายคาบ
Phrase command	คำสั่งวลี
Phrase component	องค์ประกอบวลี
Phrase control mechanism	กลไกควบคุมวลี
Post processing	การประมวลผลภายหลัง
Preprocessing	การประมวลผลเบื้องต้น
Programming	การป้อนชุดคำสั่ง
Quasi-periodic	คล้ายรายคาบ
Random signals	สัญญาณสุ่ม
Rising tone	เสียงจัตวา
Segmental	หน่วยเสียงเรียง
Segmentation	การแบ่งส่วน
Speech communication	การสื่อสารด้วยเสียงพูด
Speech production model	แบบจำลองการสร้างเสียงพูด
Speech recognition	การรู้จำเสียงพูด
Speech signals	สัญญาณเสียงพูด
Speaker dependent	ขึ้นกับผู้พูด
Speaker independent	ไม่ขึ้นกับผู้พูด
Supervised learning	การเรียนรู้แบบชี้นำ

Supra-segmental	หน่วยเสียงซ้อน
Syllable	พยางค์
Target Point	จุดเป้าหมาย
Time invariant	ไม่แปรผันตามเวลา
Time varying	แปรผันตามเวลา
Tonal assimilation	การควบร่วมของโทนเสียง
Tone	โทน, วรรณยุกต์
Tone command	คำสั่งวรรณยุกต์
Tone component	องค์ประกอบวรรณยุกต์
Tone control mechanism	กลไกควบคุมวรรณยุกต์
Tone language	ภาษาที่มีวรรณยุกต์
Transfer function	ฟังก์ชันส่งทอด
Transformation	การแปลง
Unsupervised learning	การเรียนรู้แบบไม่ชี้นำ
Unvoice	เสียงไม่ก้อง
Vocal tract	ช่องทางเดินเสียง
Voice	เสียงก้อง
Vowel	สระ

บทที่ 1
บทนำ

1.1 แนวเหตุผลและความเป็นมา

ปัจจุบัน มนุษย์ได้นำเครื่องจักรมาใช้ในการอำนวยความสะดวกและเพิ่มผลผลิตอย่างแพร่หลาย รวมทั้งการใช้เครื่องจักรที่มีความสามารถทางด้านการคำนวณสูงอย่างเช่นคอมพิวเตอร์เข้ามาใช้ในการวิเคราะห์ ช่วยในการตัดสินใจ และอื่น ๆ จากการพัฒนาการใช้งานคอมพิวเตอร์ในสาขาต่าง ๆ จนคอมพิวเตอร์เป็นที่นิยมแพร่หลายในวงกว้าง รวมไปถึงการนำคอมพิวเตอร์มาใช้ในครัวเรือน ทำให้เกิดแนวความคิดในการพัฒนาการติดต่อระหว่างมนุษย์กับเครื่องจักร (Human-Machine interface) ให้ใกล้เคียงกับธรรมชาติของมนุษย์มากขึ้น เริ่มต้นจากการป้อนรหัสไบนารีไปเป็นการป้อนชุดคำสั่ง (Programming) การป้อนบรรทัดคำสั่ง (Command-line input) ไปจนถึงความพยายามนำรูปภาพมาใช้ในการติดต่อกับผู้ใช้ (Graphical User Interface – GUI) ซึ่งทำให้ผู้ใช้ สามารถใช้งานเครื่องจักรนั้น ๆ ได้โดยแทบไม่ต้องผ่านการเรียนรู้และฝึกฝน อย่างไรก็ดี การพัฒนาการติดต่อระหว่างมนุษย์กับเครื่องจักรยังคงดำเนินไปอย่างต่อเนื่อง กล่าวกันว่าการติดต่อระหว่างมนุษย์และเครื่องจักรในอนาคตจะเหลือเพียง 2 รูปแบบคือการใช้จอสัมผัส (Touch screen) และการสั่งงานด้วยเสียง (Speech-controlled command) [1]

การติดต่อสื่อสารด้วยเสียง เป็นคุณสมบัติของสิ่งมีชีวิตชั้นสูงเพียงไม่กี่ชนิด และการสื่อสารด้วยเสียงพูด (Speech communication) เป็นคุณสมบัติเฉพาะตัวของมนุษย์ โดยสามารถสื่อความหมายที่ต้องการได้อย่างครบถ้วนด้วยเสียงพูด อีกทั้งยังสามารถใช้ในการแสดงอารมณ์และความรู้สึกได้เป็นอย่างดี ถึงกระนั้น เครื่องจักรดังเช่นคอมพิวเตอร์ สามารถจัดการกับสิ่งที่เป็นข้อมูลและตรรกะที่สามารถเข้าใจได้อย่างชัดเจนเท่านั้น จึงมีความจำเป็นในการเปลี่ยนรูปของเสียงพูด ให้อยู่ในรูปแบบของข้อมูลที่เครื่องจักรสามารถจัดการได้ อันเป็นหน้าที่ของการรู้จำแบบรูป (Pattern recognition) ที่จะเข้ามาจัดการเรื่องดังกล่าวในแง่ของการรู้จำเสียงพูด (Speech recognition)

ในภาษาอังกฤษซึ่งการรู้จำเสียงพูดได้พัฒนาไปมากนั้น โทนของเสียง (Tone) มีเพียงความหมายทางอารมณ์และความรู้สึก ขณะที่ภาษาไทยซึ่งเป็นภาษาที่โทนเสียงมีความหมายอย่างเด่นชัด (Explicit meaning) หรือเรียกว่าเป็นภาษาที่มีวรรณยุกต์ (Tone Language) เสียงพูดแต่ละคำที่มีเสียงวรรณยุกต์แตกต่างกันจะให้ความหมายต่างกัน เช่น คา (เสียงสามัญ) ข่า (เสียงเอก) ค่า (เสียงโท) ค้า (เสียงตรี) ขา (เสียงจัตวา) ล้วนมีความหมายแตกต่างกันโดยสิ้นเชิง ดังนั้น วิธีการรู้จำเสียงพูดสำหรับภาษาอื่นจึงไม่สามารถนำมาใช้กับภาษาไทยได้

งานด้านการรู้จำเสียงพูดภาษาไทยแบ่งออกเป็นส่วนตามองค์ประกอบทางภาษาคือ ส่วนพยัญชนะ (Consonants) สระ (Vowels) และโทนเสียง (Tone) หรือวรรณยุกต์ ส่วนพยัญชนะและสระมักทำการรู้จำร่วมกันเนื่องจากสามารถแบ่งส่วน (Segmentation) ออกจากกันได้อย่างค่อนข้างชัดเจน และจำแนก (Classified) ได้ด้วยค่าคุณลักษณะสำคัญแบบเดียวกัน ในขณะที่โทนเสียงมีลักษณะขึ้นกับค่าคุณลักษณะสำคัญอื่นค่อนข้างเด่นชัด อีกทั้งไม่มีอยู่ในภาษาที่ได้รับการพัฒนาระบบรู้จำเสียงพูดอย่างก้าวหน้าแล้ว จึงมีความน่าสนใจเป็นอย่างยิ่งในการศึกษาและพัฒนากระบวนการรู้จำวรรณยุกต์ในเสียงพูดภาษาไทย

การรู้จำเสียงพูดในภาษาที่ได้รับการพัฒนาไปมากอย่างภาษาอังกฤษนั้น จะทำการแบ่งย่อยเสียงพูดออกเป็นส่วนย่อยเพื่อขยายขอบเขตของศัพท์ที่สามารถรู้จำให้กว้างมากขึ้น โดยจะแบ่งจนถึงระดับหน่วยเสียงย่อยสุดคือ หน่วยเสียง Phone ซึ่งเป็นการแยกส่วนในระดับหน่วยเสียงเรียง (Segmental) ในขณะที่เสียงวรรณยุกต์เป็นคุณสมบัติของพยางค์ (Syllable) ซึ่งเป็นการแยกส่วนในระดับหน่วยเสียงซ้อน (Supra-Segmental) ด้วยเหตุนี้ กรรมวิธีการรู้จำเสียงพูดในแบบอย่างภาษาอังกฤษ ซึ่งทำการแยกส่วนเสียงพูดลงไปในระดับย่อยนั้น จึงไม่เหลือสารสนเทศเกี่ยวกับเสียงวรรณยุกต์พอที่จะใช้ในการรู้จำได้ ด้วยเหตุนี้การรู้จำวรรณยุกต์ของเสียงพูดจึงต้องเป็นกระบวนการที่กระทำแยกต่างหากจากการรู้จำเสียงพูดส่วนอื่น

ในการทำการรู้จำเสียงวรรณยุกต์ของเสียงพูดภาษาไทยใน [2, 3] ได้ใช้การวิเคราะห์หาจุดสูงสุดและต่ำสุดของเส้นโค้งความถี่มูลฐาน (F0 Contour) ร่วมกับกฎของรูปแบบเส้นโค้งความถี่มูลฐานของวรรณยุกต์ของเสียงพูด มาใช้ในการหาชุดของลำดับเสียงวรรณยุกต์ที่เป็นไปได้ของเสียงพูดที่ต้องการวิเคราะห์ เพื่อสังเคราะห์เสียงด้วยแบบจำลองฟูจิซากิ (Fujisaki Model) และนำไปเปรียบเทียบกับเสียงต้นแบบเพื่อหาชุดของลำดับเสียงวรรณยุกต์ที่น่าจะเป็นไปได้มากที่สุด โดยทำการทดลองกับชุดเสียงพูดที่กำหนดรูปแบบขึ้นเป็นประโยคที่ประกอบด้วย 4 พยางค์จำนวนทั้งสิ้น 11 ประโยคที่มีลักษณะโทนเสียงต่างกันไป ซึ่งให้ผลการทดลองที่มีระดับการรู้จำโดยเฉลี่ยร้อยละ 89.1

การนำแบบจำลองฟูจิซากิมาใช้ในการสังเคราะห์ใน [2, 3] เพื่อเป็นการตัดผลของการควบร่วมของโทนเสียง (Tonal assimilation) และผลของการลดระดับของเสียง (Declination) แต่โดยวิธีการเปรียบเทียบเส้นโค้งความถี่มูลฐานที่ถูกสังเคราะห์ขึ้น กับเส้นโค้งความถี่มูลฐานที่ต้องการวิเคราะห์นั้น จะเห็นได้ว่าปริมาณข้อมูลที่ต้องทำการเปรียบเทียบจะมีอยู่เป็นจำนวนมาก อีกทั้งเส้นโค้งความถี่มูลฐานของเสียงพูดที่เกิดจากการสังเคราะห์ด้วยลำดับเสียงวรรณยุกต์เดียวกัน จะมีลักษณะเหมือนกันเสมอเนื่องจากมีพารามิเตอร์เพียงชุดเดียวต่อหนึ่งชุดลำดับเสียงวรรณยุกต์ ในขณะที่เส้นโค้งความถี่มูลฐานของเสียงพูดจริงที่มีลำดับของเสียงวรรณยุกต์เดียวกัน อาจจะมีลักษณะต่างกันมาก เพียงแต่มีแนวโน้มของลักษณะไปในทางเดียวกัน ดังนั้นผลการเปรียบเทียบ

กับพารามิเตอร์มาตรฐานของแบบจำลองฟูจิซากิของชุดลำดับเสียงวรรณยุกต์แต่ละชุด จึงไม่อาจให้ผลที่ดีนัก นอกจากนี้ หากขยายความต้องการในการรู้จำวรรณยุกต์ให้เป็นไปได้หลายพยางค์ จะเห็นได้ว่ารูปแบบของเสียงที่เกิดขึ้นจะมีได้เป็นจำนวนมาก ซึ่งถือเป็นข้อจำกัดที่สำคัญของวิธีดังกล่าว

ในการรู้จำวรรณยุกต์ของเสียงพูดโดยใช้ Contextual tone features ร่วมกับการทำ Center-point intonation normalization และ Incorporated stress feature method [4] ซึ่งพิจารณาผลของการออกเสียงร่วม (Coarticulation) ทำนองเสียง (Intonation) และการเน้นเสียง (Stress) ผลที่ได้จากการทดลองกับเสียงพูดที่มีลักษณะเดียวกับ [3] ให้ผลการรู้จำถึงร้อยละ 93.6 อย่างไรก็ดี การตั้งสมมติฐานของการลดระดับของเสียง (Declination) เป็นเส้นตรงอาจไม่เหมาะสมนัก เนื่องจากการศึกษาของ [5] พบว่าการเปลี่ยนแปลงของความถี่พื้นฐานของเสียงโดยรวม (Global) ควรจะมีลักษณะเป็นเส้นโค้ง

ดังนั้นในวิทยานิพนธ์นี้ จึงมีความพยายามจะนำพารามิเตอร์ของแบบจำลองฟูจิซากิ มาใช้ในการรู้จำวรรณยุกต์ของเสียงพูด เนื่องจากการที่พารามิเตอร์นั้นสามารถสังเคราะห์เส้นโค้งความถี่มูลฐานของเสียงได้อย่างเหมาะสม แสดงให้เห็นว่าสารสนเทศที่เกี่ยวเนื่องกับเส้นโค้งความถี่มูลฐานของเสียง ซึ่งสามารถใช้เป็นตัวบ่งบอกถึงวรรณยุกต์ของเสียง ถูกบรรจุอยู่ในพารามิเตอร์ของแบบจำลองแล้วทั้งหมด อีกทั้งยังมีลักษณะของ Declination ที่เป็นธรรมชาติ โดยมีเป้าหมายที่เสียงพูดต่อเนื่อง และใช้พารามิเตอร์ของแบบจำลองเป็นค่าคุณลักษณะสำคัญของการรู้จำวรรณยุกต์เสียงพูดซึ่งมีปริมาณข้อมูลที่ใช้เปรียบเทียบน้อยกว่า อีกทั้งยังเป็นการเปรียบเทียบจากพารามิเตอร์ของเสียงจริง ซึ่งน่าจะให้ผลการทดลองที่ดีกว่า

1.2 วัตถุประสงค์

เพื่อพัฒนาวิธีการรู้จำวรรณยุกต์ในเสียงพูดภาษาไทยด้วยการนำแบบจำลองฟูจิซากิมาใช้

1.3 เป้าหมายและขอบเขตของงานวิจัย

1. สร้างระบบรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทยโดยใช้แบบจำลองฟูจิซากิ
2. ระบบรู้จำที่ได้มีความถูกต้องโดยเฉลี่ยไม่น้อยกว่าร้อยละ 80

1.4 ขั้นตอนและวิธีดำเนินงาน

1. ศึกษางานวิจัยที่เกี่ยวข้องกับการรู้จำโทนเสียงพูด
2. ศึกษางานวิจัยที่เกี่ยวข้องกับแบบจำลองฟูจิซากิ
3. ศึกษาการเขียนโปรแกรม Visual C++
4. เขียนโปรแกรมทดสอบระบบ
5. จัดเก็บตัวอย่างเสียงพูดภาษาไทย
6. วิเคราะห์ผลที่ได้จากการทดสอบ และแก้ไขข้อผิดพลาด
7. สรุปและรวบรวมข้อมูลทั้งหมดพร้อมทั้งจัดทำรูปเล่มวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ทราบถึงความเป็นไปได้ในการนำแบบจำลองฟูจิซากิมาใช้ในกระบวนการรู้จำวรรณยุกต์ในเสียงพูดต่อเนื่องภาษาไทย
2. สร้างแนวทางพัฒนาการรู้จำวรรณยุกต์ด้วยการใช้ค่าคุณลักษณะสำคัญแบบใหม่
3. สร้างกระบวนการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิที่ใช้กับเสียงพูดภาษาไทย
4. เป็นแนวทางในการเพิ่มความถูกต้องของการรู้จำเสียงพูดภาษาไทย

บทที่ 2
ความรู้พื้นฐาน

เนื้อหาในบทนี้กล่าวถึงทฤษฎีต่าง ๆ ที่เกี่ยวข้องกับวิทยานิพนธ์ คือ โครงสร้างภาษาไทย แบบจำลองการสร้างเสียงพูด ความถี่มูลฐาน แบบจำลองฟูจิซากิ และโครงข่ายประสาทเทียม

2.1 สารสนเทศในเสียงพูด

ในเสียงพูดของมนุษย์ประกอบด้วยสารสนเทศมากมาย โดยสามารถแบ่งออกได้เป็น 3 ประเภทคือ สารสนเทศเชิงภาษาศาสตร์ (Linguistic information) สารสนเทศกึ่งภาษาศาสตร์ (Paralinguistic information) และสารสนเทศที่ไม่ใช่ภาษาศาสตร์ (Nonlinguistic information)

2.1.1 สารสนเทศเชิงภาษาศาสตร์

สารสนเทศในเชิงภาษาศาสตร์ตาม [5] ที่กล่าวว่า “The symbolic information that is represented by a set of discrete symbols and rules for their combination” หมายความว่า “สารสนเทศเชิงสัญลักษณ์ซึ่งสามารถแสดงแทนได้ด้วยชุดของสัญลักษณ์และกฎของการรวมกันของสัญลักษณ์”

นั่นคือเป็นสารสนเทศที่สามารถแสดงเป็นตัวอักษรหรือภาษาเขียนได้ เช่นพยัญชนะ สระ หรือวรรณยุกต์ต่าง ๆ

2.1.2 สารสนเทศกึ่งภาษาศาสตร์

สารสนเทศกึ่งภาษาศาสตร์ตามที่ [5] ให้นิยามว่า “The information that is not inferable from the written counterpart but is deliberately added by the speaker to modify or supplement the linguistic information” หมายถึง “สารสนเทศที่ไม่สามารถแสดงได้โดยภาษาเขียน แต่ถูกผู้พูดเสริมเข้าไปเพื่อแก้ไขหรือเพิ่มเติมสารสนเทศทางภาษาศาสตร์”

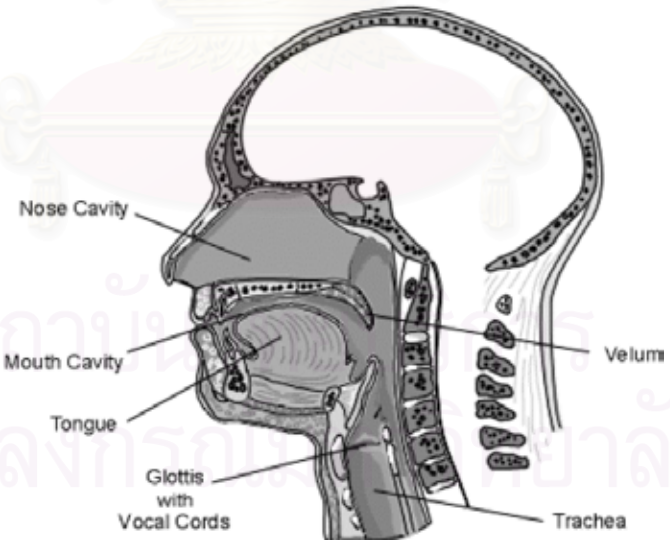
เนื่องจากประโยคในภาษาเขียน สามารถเปล่งเสียงได้หลายรูปแบบ ในการแสดงออกถึงความตั้งใจ และวิธีพูดของแต่ละบุคคล ซึ่งเป็นสิ่งที่สามารถควบคุมได้โดยผู้พูด ตัวอย่างเช่น ลักษณะเสียงพูดของประโยค “จะไปกินข้าว” สามารถเปล่งออกมาในลักษณะของประโยคบอกเล่า หรือสามารถเปล่งออกมาในลักษณะของประโยคคำถามได้โดยการเปลี่ยนแปลงทำนองเสียง

2.1.3 **สารสนเทศที่ไม่ใช่ภาษาศาสตร์**

สารสนเทศที่ไม่ใช่ภาษาศาสตร์หมายถึง สารสนเทศอื่นในเสียงพูดที่ไม่ใช่สารสนเทศเชิงภาษา และสารสนเทศกึ่งภาษา เช่น อายุ เพศ สถานะทางอารมณ์ หรือลักษณะทางร่างกายเป็นต้น สารสนเทศประเภทนี้ไม่สามารถควบคุมได้โดยผู้พูดในลักษณะการพูดปกติ แต่สามารถควบคุมได้ในลักษณะการควบคุมอารมณ์ในการพูด

2.2 **แบบจำลองการสร้างเสียงพูดและความถี่มูลฐาน**

สัญญาณเสียงพูด (Speech signals) เป็นสัญญาณสุ่ม (Random signals) มีองค์ประกอบที่มีลักษณะทั้งเป็นเชิงเส้น (Linear) และไม่เป็นเชิงเส้น (Non-linear) แปรผันตามเวลา (Time varying) และไม่แปรผันตามเวลา (Time invariant) เป็นรายคาบ (Periodic) และไม่เป็นรายคาบ (Non-periodic) อยู่รวมกัน โดยส่วนที่มีลักษณะคล้ายรายคาบ (Quasi-periodic) เรียกว่าเสียงก้อง (Voice) และส่วนที่ไม่เป็นรายคาบซึ่งมีลักษณะคล้ายสัญญาณรบกวนเรียกว่าเสียงไม่ก้อง (Unvoice)

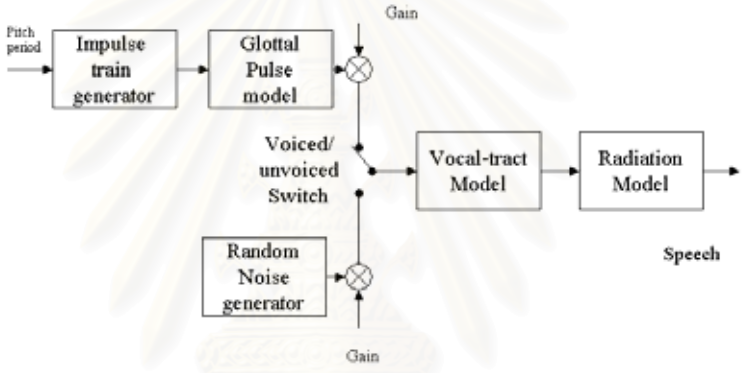


รูปที่ 2.1 อวัยวะที่เกี่ยวข้องกับการสร้างเสียงพูด [6]

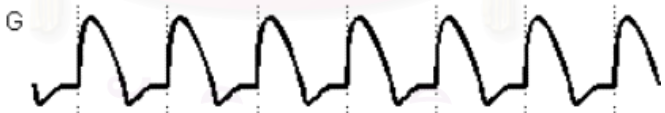
แบบจำลองการสร้างเสียงพูด (Speech production model) ซึ่งจำลองจากลักษณะของอวัยวะในการสร้างเสียงพูดดังแสดงในรูปที่ 2.1 สามารถแสดงได้ดังรูปที่ 2.2 โดยสัญญาณเสียงพูดในส่วนของเสียงก้องสามารถแสดงได้ดังสมการ (2.1)

$$S(n) = E(n) * \Theta(n) \tag{2.1}$$

เมื่อ $S(n)$ คือเสียงพูดที่สร้างขึ้นโดยเกิดจากการคอนโวลูชันของ $E(n)$ คือพัลส์เส้นเสียง (Glottal pulse) และ $\Theta(n)$ คือฟังก์ชันส่งทอด (Transfer function) ของช่องทางเดินเสียง (Vocal tract) [7] โดยลักษณะของพัลส์เส้นเสียงแสดงในรูปที่ 2.3



รูปที่ 2.2 แบบจำลองการสร้างเสียงพูด [7]

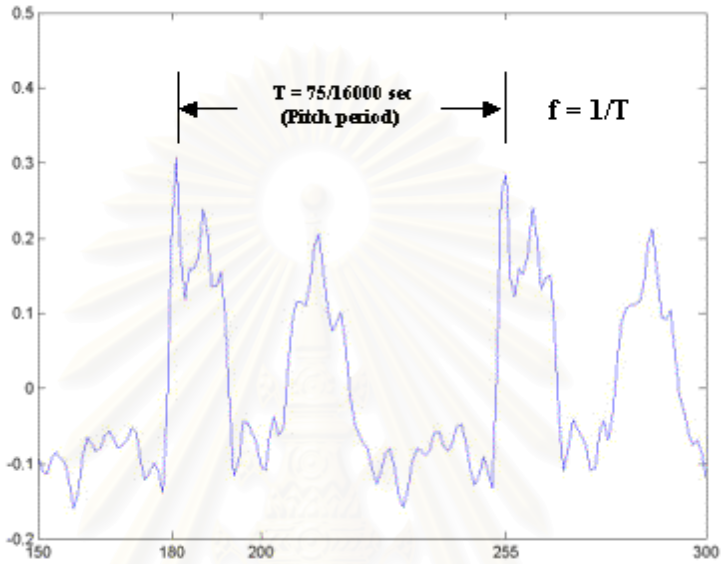


รูปที่ 2.3 พัลส์เส้นเสียง [6]

ในสัญญาณเสียงพูดส่วนที่เป็นเสียงก้อง ความคล้ายรายคาบสามารถแสดงได้ด้วยความถี่มูลฐาน (F0) ซึ่งเกิดขึ้นจากพัลส์เส้นเสียงและมีความถี่เท่ากับพัลส์เส้นเสียง โดยความถี่มูลฐานนิยามได้ด้วยส่วนกลับของคาบของสัญญาณเสียงดังแสดงได้ในรูปที่ 2.4 และสมการ (2.2)

$$F0 = 1/T \tag{2.2}$$

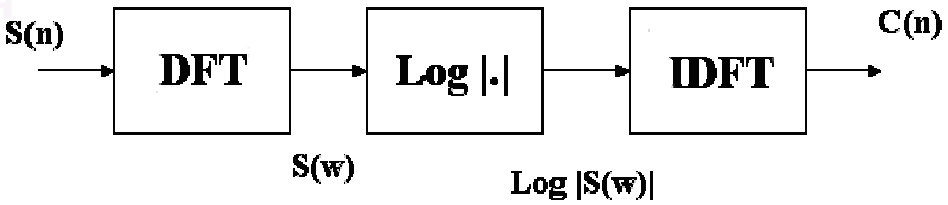
จากแบบจำลองการสร้างเสียงพูดข้างต้น การวิเคราะห์พัลส์เส้นเสียงเพื่อหาความถี่มูลฐานจากเสียงพูดที่บันทึกไว้ทำได้ยาก เนื่องจากสัญญาณที่บันทึกได้เกิดจากการคอนโวลูชันของพัลส์เส้นเสียงและฟังก์ชันส่งทอดของช่องทางเดินเสียง ซึ่งไม่สามารถแยกประมวลผลแบบเชิงเส้นได้ แนวคิดของเซปสตรัมจึงเกิดขึ้นเพื่อพยายามแยกองค์ประกอบทั้ง 2 ส่วนออกจากกันอย่างเป็นเชิงเส้น เซปสตรัมนิยามได้โดยสมการ (2.3) [8]



รูปที่ 2.4 ลักษณะคล้ายรายคาบและคาบของสัญญาณเสียงพูด

$$C(n) = F^{-1}\{\log|F\{S(n)\}|\} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|S(\omega)|e^{j\omega n}d\omega \qquad (2.3)$$

โดย $F\{\cdot\}$ ใช้แทนการแปลงฟูริเยร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform – DFT) และ $F^{-1}\{\cdot\}$ ใช้แทนการแปลงฟูริเยร์ผกผันแบบไม่ต่อเนื่อง (Inverse Discrete Fourier Transform – IDFT) ขั้นตอนการวิเคราะห์หาเซปสตรัมแสดงได้ดังรูปที่ 2.5



รูปที่ 2.5 ขั้นตอนการวิเคราะห์เซปสตรัม

จากสมการ (2.1) จะได้ว่า

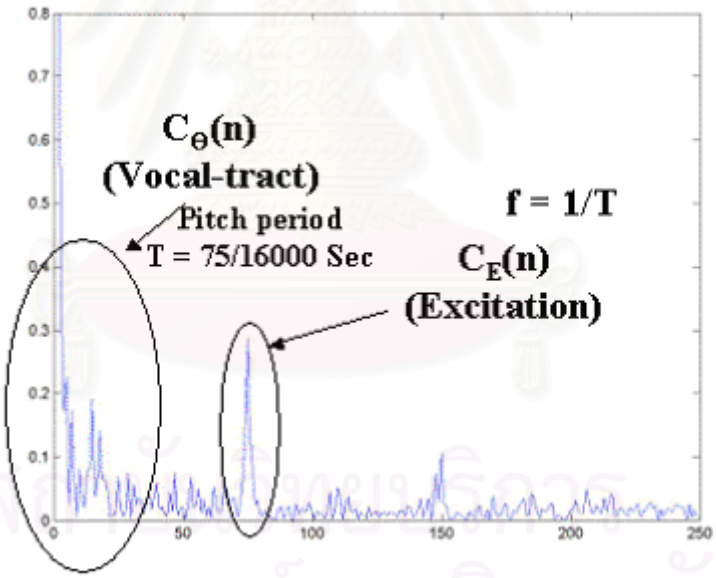
$$F\{S(n)\} = F\{E(n) * \Theta(n)\} = E(\omega) \cdot \Theta(\omega) \tag{2.4}$$

$$\log|F\{S(n)\}| = \log|E(\omega)| + \log|\Theta(\omega)| \tag{2.5}$$

$$F^{-1}\{\log|F\{S(n)\}|\} = F^{-1}\{\log|E(\omega)|\} + F^{-1}\{\log|\Theta(\omega)|\} \tag{2.6}$$

$$C(n) = C_E(n) + C_\Theta(n) \tag{2.7}$$

โดย $C_E(n)$ คือเซปสตรัมของพัลส์เส้นเสียงและ $C_\Theta(n)$ คือเซปสตรัมของฟังก์ชันส่งทอดของช่องทางเดินเสียง เมื่อพิจารณาลักษณะของสัญญาณเสียงพูด จะพบว่าเซปสตรัมของสัญญาณเสียงสามารถแบ่งส่วนที่เกิดจากช่องทางเดินเสียงและพัลส์เส้นเสียงได้อย่างชัดเจนดังรูปที่ 2.6 เวลาที่เกิดจุดสูงสุดในส่วนของพัลส์เส้นเสียงจะเป็นคาบของลักษณะคล้ายรายคาบของสัญญาณ ซึ่งสามารถนำมาหาส่วนกลับเป็นค่าความถี่มูลฐานต่อไป



รูปที่ 2.6 เซปสตรัมของสัญญาณเสียงพูด

2.3 โครงสร้างภาษาไทย

เสียงพูดของพยางค์ในภาษาไทยประกอบด้วยองค์ประกอบหลัก 3 ส่วนคือพยัญชนะ (Consonants) สระ (Vowel) และ วรรณยุกต์ (Tone) ซึ่งมีลักษณะโครงสร้างแสดงได้ดังรูปที่ 2.7

เสียงพยัญชนะในภาษาไทยประกอบด้วย 21 เสียง เสียงสระ 24 เสียง และ เสียงวรรณยุกต์ 5 เสียง ดังแสดงในตารางที่ 2.1 2.2 และ 2.3 ตามลำดับ

C : Consonants

V : Vowel

T : Tone

*สัญลักษณ์ในวงเล็บหมายถึงสามารถมีหรือไม่มีก็ได้

รูปที่ 2.7 โครงสร้างพยางค์ของเสียงพูดในภาษาไทย

ตารางที่ 2.1 เสียงพยัญชนะในภาษาไทย

เสียง	ตัวอักษร
กอ	ก
คอ	ข ค ฆ
งอ	ง หง-
จอ	จ
ชอ	ฉ ช ฌ
ซอ	ซ ศ ษ ส
ยอ	ญ ย หย- หญ- อย-
ดอ	ฎ ด
ตอ	ฏ ต
ทอ	ฐ ฑ ฒ ถ ท ธ
นอ	ณ น หน-

เสียง	ตัวอักษร
บอ	บ
ปอ	ป
พอ	ผ พ ภ
ฟอ	ฝ ฟ
มอ	ม หม-
รอ	ร หร-
ลอ	ล ฬ หล-
วอ	ว หว-
หอ	ห ฮ
ออ	อ

ตารางที่ 2.2 เสียงสระในภาษาไทย

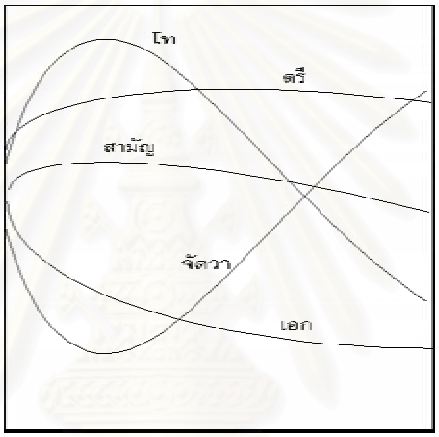
สระเสียงสั้น	สระเสียงยาว
อะ	อา
อิ	อี
อึ	อือ
อุ	อู
เอะ	เอ
แอะ	แอ
โอะ	โอ
เอาะ	ออ
เออะ	เออ
เอียะ	เอีย
เอือะ	เอือ
อัวะ	อัว

ตารางที่ 2.3 เสียงวรรณยุกต์ในภาษาไทย

เสียงสระ	ตัวอย่าง
สามัญ	คา แกง เตา
เอก	ข่า แกลบ ตบ
โท	ข้า ค่า แก้ม โคก
ตรี	ค้า โอ๊ย
จัตวา	ขา จ๋า

2.4 ความถี่มูลฐานกับวรรณยุกต์ในเสียงพูดภาษาไทย

ในภาษาพูดที่โทนเสียงไม่ได้สื่อความหมายเชิงภาษา ในโทนเสียงพูดนั้นจะยังคงมีสารสนเทศของลักษณะคำพูด อารมณ์ และ อื่น ๆ เช่น เพศและอายุของผู้พูด แต่ในภาษาที่โทนเสียงมีความหมายทางภาษา หรือภาษาที่มีวรรณยุกต์อย่างเช่นภาษาไทยนั้น โทนเสียงมีความเกี่ยวข้องอย่างมากกับความหมายต่าง ๆ ของเสียงพูด ซึ่งโทนเสียงนั้นเกิดจากการเปลี่ยนแปลงของความถี่มูลฐานในช่วงเวลาต่าง ๆ ของการออกเสียงพูด จากการศึกษาของ [9] แสดงให้เห็นถึงความสัมพันธ์ของความถี่มูลฐานกับวรรณยุกต์ในภาษาไทย โดยรูปที่ 2.8 แสดงเส้นความถี่มูลฐานของเสียงพูดภาษาไทยของคำโดดในการออกเสียงวรรณยุกต์ต่าง ๆ



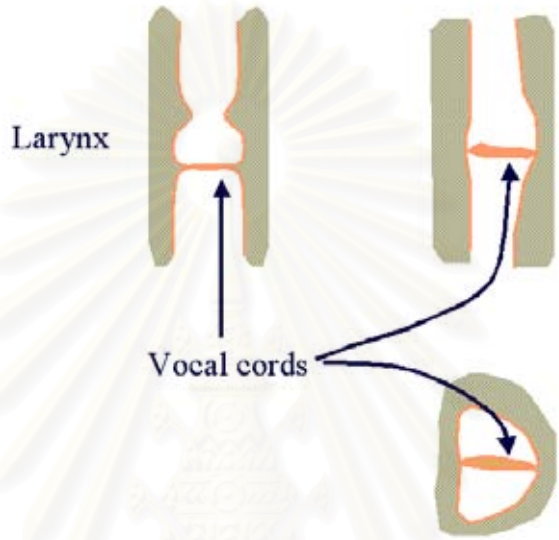
รูปที่ 2.8 เส้นความถี่มูลฐานของเสียงพูดภาษาไทยของคำโดดในการออกเสียงวรรณยุกต์ต่าง ๆ

จากรูปที่ 2.8 จะเห็นได้ว่าเสียงวรรณยุกต์กับลักษณะเส้นโค้งความถี่มูลฐานมีความสัมพันธ์กันโดยเสียงสามัญมีเส้นโค้งความถี่มูลฐานลักษณะเรียบและมีความถี่ระดับกลาง (Mid tone) เสียงเอกมีเส้นโค้งความถี่มูลฐานลักษณะเรียบและมีความถี่ระดับต่ำ (Low tone) เสียงโทมีเส้นโค้งความถี่มูลฐานที่สูงขึ้นและลดต่ำลง (Falling tone) เสียงตรีมีเส้นโค้งความถี่มูลฐานเรียบและมีความถี่ระดับสูง (High tone) และเสียงจัตวามีเส้นโค้งความถี่มูลฐานที่ต่ำลงและสูงขึ้น (Rising tone)

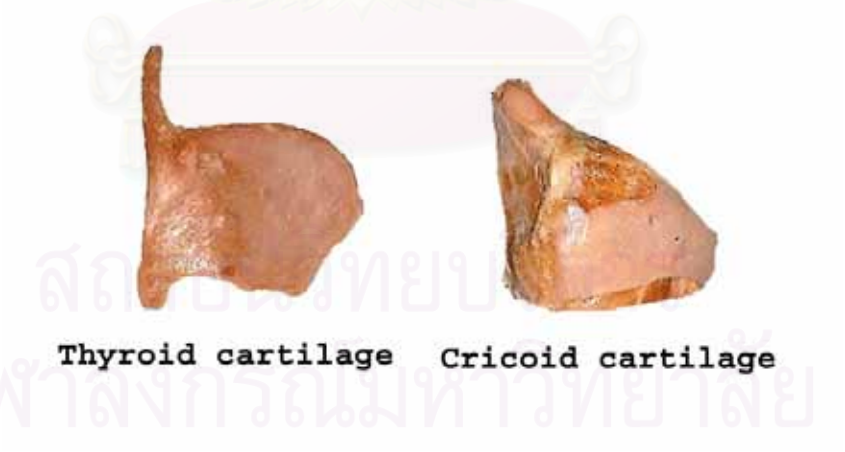
2.5 แบบจำลองฟูจิซากิ

ในการสร้างเสียงพูดให้ได้เหมือนเสียงพูดจริง จำเป็นอย่างยิ่งที่จะต้องทำการสร้างแบบจำลองสำหรับการสร้างความถี่พื้นฐานในเชิงปริมาณ จากแบบจำลองการสร้างเสียงพูด ความถี่

พื้นฐานของเสียงพูดเกิดจากความถี่ในการสั่นตัวของเส้นเสียง ซึ่งเปลี่ยนแปลงไปตามการขยับอวัยวะต่าง ๆ ที่มีส่วนเกี่ยวข้อง การขยับอวัยวะที่เกี่ยวข้องทำให้เส้นเสียงมีขนาดยาวขึ้นหรือสั้นลง ซึ่งมีผลโดยตรงกับความถี่ของการสั่นของเส้นเสียง รูปที่ 2.9 แสดงลักษณะของเส้นเสียง และรูปที่ 2.10 แสดงอวัยวะที่เกี่ยวข้องกับการเปลี่ยนแปลงความยาวของเส้นเสียงคือ Thyroid cartilage และ Cricoid cartilage โดยการเคลื่อนไหวของกล้ามเนื้อ Crico-Thyroid ก่อให้เกิดการขยับและแรงเค้นรวมทั้งความยาวที่เปลี่ยนไปของเส้นเสียง



รูปที่ 2.9 ภาพแสดงเส้นเสียงในภาพตัดขวางของ Larynx [5]



รูปที่ 2.10 Thyroid cartilage และ Cricoid cartilage

จาก [5] แรงดึงในเส้นเสียงมีความสัมพันธ์กับการเปลี่ยนแปลงความยาวของเส้นเสียงเป็น

$$T = T_0 \exp(bx) \tag{2.8}$$

โดย T คือแรงดึงในเส้นเสียง T_0 คือแรงดึงในเส้นเสียงในสภาวะปกติ b เป็นค่าคงที่ และ x คือความยาวที่เปลี่ยนแปลงไปของเส้นเสียง และความสัมพันธ์ของความถี่พื้นฐานในการสั่นของเส้นเสียงกับแรงดึงเป็น

$$F_0 = c_0\sqrt{T/\sigma} \tag{2.9}$$

โดย σ เป็นความหนาแน่นต่อหน่วยพื้นที่ของเส้นเสียง และ c_0 เป็นค่าคงที่ซึ่งแปรผกผันกับขนาดของเส้นเสียง จากสมการที่ 2.8 และ 2.9 เราจะได้ว่า

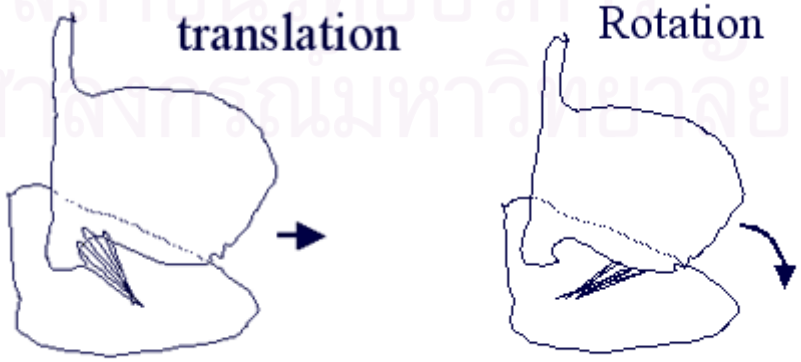
$$\log F_0 = \log(c_0\sqrt{T_0\exp(bx)/\sigma} \tag{2.10}$$

$$\log F_0 = \log(c_0\sqrt{T_0/\sigma}) + (b/2)x \tag{2.11}$$

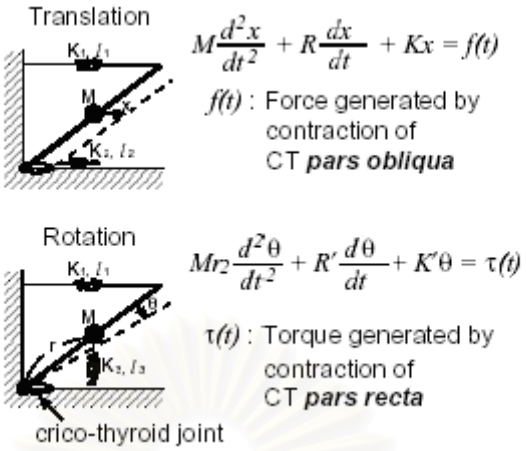
พจน์ $\log(c_0\sqrt{T_0/\sigma})$ มีค่าคงที่ขึ้นอยู่กับตัวบุคคล ในขณะที่ความยาวของเส้นเสียงจะมีการเปลี่ยนไปตามเวลา ดังนั้นหากให้ $c_0\sqrt{T_0/\sigma}$ แทนด้วย F_b ซึ่งเป็นความถี่ฐานของความถี่มูลฐาน F_0 จะได้สมการในเชิงเวลาเป็น

$$\log F_0(t) = \log F_b + (b/2)x(t) \tag{2.12}$$

การเปลี่ยนแปลงความยาวของเส้นเสียงเกิดเนื่องจากการเคลื่อนไหวสัมพัทธ์ของ Thyroid cartilage และ Cricoid cartilage ซึ่งการวิเคราะห์โครงสร้างของ Larynx แสดงให้เห็นว่าการเคลื่อนที่ดังกล่าวประกอบด้วย 2 องศาอิสระ (Degree of Freedom) โดยเกิดจากการเคลื่อนที่ในแนวนอน (Translation) และการหมุนรอบจุดยึด Crico-thyroid ดังรูปที่ 2.11



รูปที่ 2.11 การเคลื่อนตัวของ Crico-Thyroid [5]



รูปที่ 2.12 แบบจำลองกลศาสตร์ของ Crico-Thyroid [5]

การเคลื่อนที่ในแนวนอนและการหมุนของ Thyroid สามารถแสดงแทนได้ด้วยแบบจำลองทางกลศาสตร์สองระบบแยกออกจากกันดังรูปที่ 2.12 ซึ่งการเคลื่อนที่ทั้งสองแบบทำให้เกิดการเปลี่ยนแปลงความยาวของเส้นเสียง ดังนั้น $x(t)$ ในสมการ 2.11 ซึ่งใช้แสดงแทนการเปลี่ยนแปลงความยาวของเส้นเสียงจึงสามารถแยกออกเป็น $x_1(t)$ และ $x_2(t)$ ซึ่งแสดงแทนการเปลี่ยนแปลงความยาวของเส้นเสียงเนื่องจากการเคลื่อนที่ในแนวนอน และการหมุนตามลำดับ โดยแสดงได้เป็น

$$\log F_0(t) = \log F_b + (b/2)(x_1(t) + x_2(t)) \qquad (2.13)$$

สมการข้างต้นแสดงให้เห็นถึงองค์ประกอบที่แปรเปลี่ยนตามเวลาของ $\log F_0(t)$ ที่สามารถแสดงให้เห็นได้ว่าเกิดจากการรวมกันขององค์ประกอบที่แปรเปลี่ยนตามเวลา 2 องค์ประกอบ เนื่องจากการเคลื่อนที่ไปในแนวนอนของ Thyroid cartilage มีค่าคงที่ทางเวลามากกว่าค่าคงที่ทางเวลาของการหมุนของ Thyroid cartilage มาก ดังนั้นองค์ประกอบที่แปรเปลี่ยนตามเวลาที่เกิดจากการเคลื่อนที่ในแนวนอนจึงถูกใช้แสดงถึงผลในวงกว้าง (Global Phenomena) อย่างเช่นการเปลี่ยนแปลงของวลี ในขณะที่องค์ประกอบที่แปรเปลี่ยนตามเวลาที่เกิดจากการหมุนจะถูกใช้แสดงผลเฉพาะแห่ง (Local Phenomena) อย่างเช่นสำเนียงเสียง

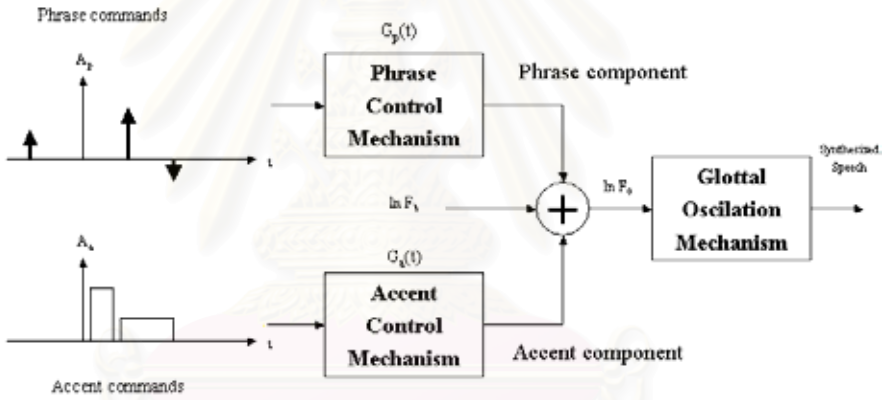
แบบจำลองฟูจิซากิเป็นแบบจำลองเชิงปริมาณของความถี่มูลฐานของเสียงพูด โดยใช้ความรู้พื้นฐานและความเข้าใจในระบบการเคลื่อนไหวอวัยวะที่มีผลกระทบต่อการสั่นของเส้นเสียง โดยแบบจำลองนี้ยึดถือสมมติฐาน 3 ข้อคือ

- คำสั่งวลี (Phrase Command) เป็นชุดของอิมพัลส์ และองค์ประกอบวลี (Phrase Component) เป็นผลตอบสนองของคำสั่งวลีต่อระบบเชิงเส้นอันดับที่สองแบบ critical-damped

- คำสั่งสำเนียง (Accent Command) เป็นชุดของฟังก์ชันขั้นบันได และองค์ประกอบสำเนียง (Accent Component) เป็นผลตอบสนองของคำสั่งสำเนียงต่อระบบเชิงเส้นอันดับที่สองแบบ critical-damped อีกระบบ

- องค์ประกอบวลี (Phrase Component) และองค์ประกอบสำเนียง (Accent Component) จะถูกรวมกันและเกิดเป็นการเปลี่ยนแปลงของความถี่มูลฐานในสเกลลอการิทึม และแม้ว่าระบบเชิงเส้นทั้งสองระบบจะไม่เป็น critical-damped อย่างแท้จริง แต่การวิเคราะห์เบื้องต้นของเส้นความถี่มูลฐานพบว่าสมมติฐานนี้ค่อนข้างเหมาะสม

บนพื้นฐานของสมมติฐานทั้ง 3 ข้อนี้ แบบจำลองสำหรับการสร้างเส้นความถี่มูลฐานของประโยคจึงแสดงได้ดังรูปที่ 2.13



รูปที่ 2.13 แบบจำลองกระบวนการสร้างเส้นความถี่มูลฐาน [5]

ซึ่งตามแบบจำลองนี้สามารถแสดงเส้นความถี่มูลฐานได้เป็น

$$\log F_0(t) = \log F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \left\{ G_a(t - T_{1j}) - G_a(t - T_{2j}) \right\} \quad (2.14)$$

$$G_p(t) = \begin{cases} \alpha^2 t \cdot \exp(-\alpha t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2.15)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t) \cdot \exp(-\beta t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2.16)$$

โดย $G_p(t)$ แทนฟังก์ชันตอบสนองอิมพัลส์ของกลไกควบคุมวลี (Phrase control mechanism) และ $G_a(t)$ แทนฟังก์ชันตอบสนองขั้นบันไดของกลไกควบคุมสำเนียง (Accent control mechanism) และตัวแปรต่าง ๆ มีความหมายดังนี้

t : เวลาที่จุดสังเกต

A_{pi} : แมกนิจูดของคำสั่งวลีที่ i

T_{0i} : เวลาของคำสั่งวลีที่ i

I : จำนวนคำสั่งวลีทั้งหมด

A_{aj} : แมกนิจูดของคำสั่งสำเนียงที่ j

T_{1j} : เวลาเริ่มต้นของคำสั่งสำเนียงที่ j

T_{2j} : เวลาสิ้นสุดของคำสั่งสำเนียงที่ j

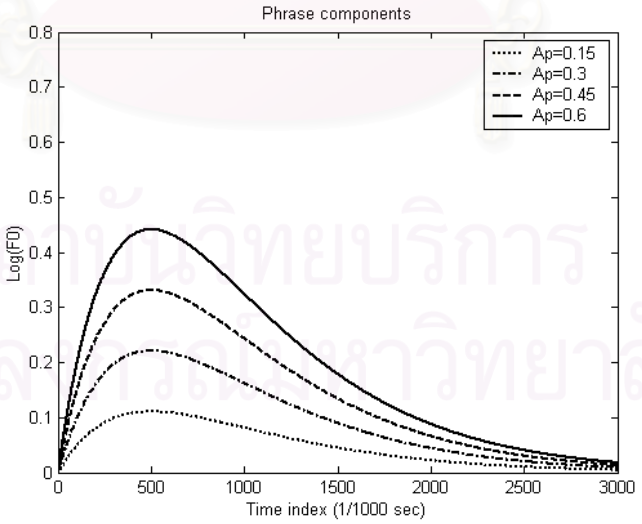
J : จำนวนคำสั่งสำเนียงทั้งหมด

α : ค่าคงที่เวลาของกลไกควบคุมวลีซึ่งเกิดจากการเคลื่อนที่ในแนวนอน

β : ค่าคงที่เวลาของกลไกควบคุมสำเนียงซึ่งเกิดจากการหมุนรอบจุดยึด

γ : ค่าจำกัดของผลตอบสนองของกลไกควบคุมสำเนียง

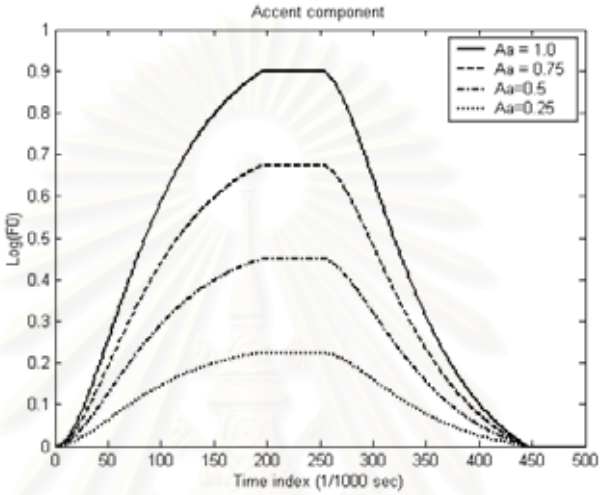
ค่า α และ β มีค่าคงที่ตลอดประโยค และจากการศึกษาของ [5] พบว่ามีค่าไม่แตกต่างกันมากนักในแต่ละบุคคล ส่วนค่า γ ให้ค่าไว้ที่ 0.9



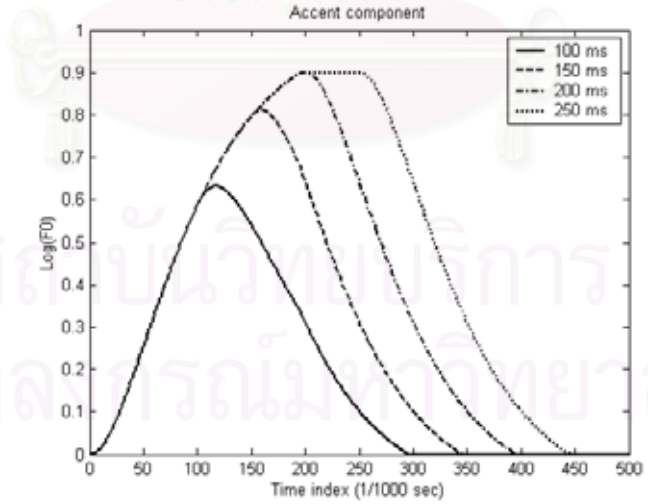
รูปที่ 2.14 องค์ประกอบวลีที่เวลา T_0 = 0 A_{pi} เป็น 0.15, 0.30, 0.45 และ 0.6

รูปที่ 2.14 แสดงลักษณะขององค์ประกอบวลีที่เวลา T_0 = 0 และ α = 2.0 โดยมีค่า A_{pi} เป็น 0.15, 0.30, 0.45 และ 0.6 ตามลำดับ รูปที่ 2.15 แสดงลักษณะขององค์ประกอบสำเนียงที่เกิดจาก

คำสั่งสำเนียงที่มีช่วงเวลา 250 มิลลิวินาที และมีค่า A_{aj} เป็น 1.0, 0.75, 0.5 และ 0.25 โดยมี ช่วงเวลา 250 มิลลิวินาที โดยมี β = 20.0 เห็นได้ว่า การเปลี่ยนแปลงของความถี่มูลฐานในส่วนของกลไกควบคุมสำเนียงจะแปรผันกับ A_{aj} รูปที่ 2.16 แสดงลักษณะขององค์ประกอบสำเนียงที่เกิดจากคำสั่งสำเนียงที่มีค่าช่วงเวลาเป็น 100, 150, 200 และ 250 มิลลิวินาทีโดยมีค่า A_{aj} = 1.0 และ β = 20.0



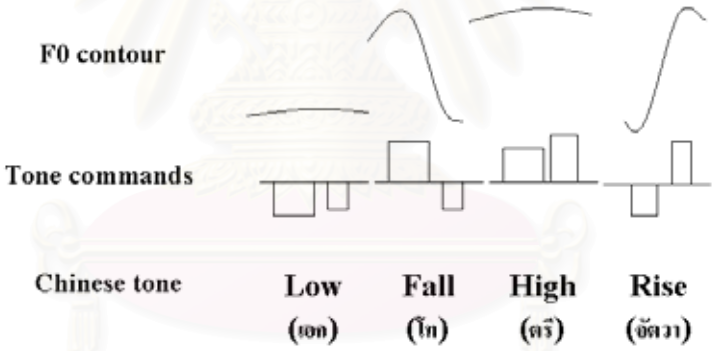
รูปที่ 2.15 องค์ประกอบสำเนียงที่ค่า A_{aj} เป็น 1.0, 0.75, 0.5 และ 0.25 ช่วงเวลา 250 มิลลิวินาที



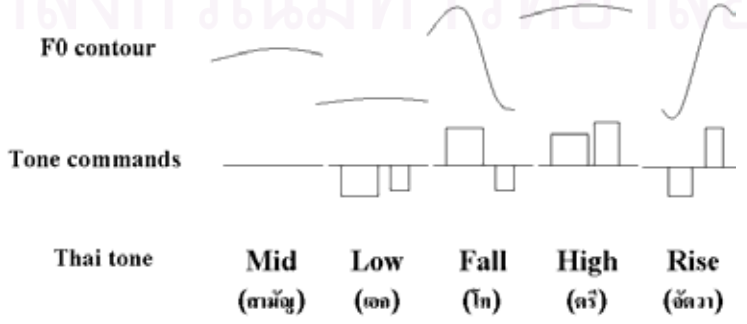
รูปที่ 2.16 องค์ประกอบสำเนียงที่ค่าช่วงเวลาเป็น 100, 150, 200 และ 250 มิลลิวินาที ที่ A_{aj} = 1.0

2.6 แบบจำลองฟูจิซากิกับวรรณยุกต์ภาษาไทย

วัตถุประสงค์แรกเริ่มในการพัฒนาแบบจำลองฟูจิซากิคือ การสังเคราะห์เสียงพูดภาษาญี่ปุ่น ซึ่งเป็นภาษาที่โทนเสียง (Tone) ไม่มีความหมายโดยเด่นชัด แต่ด้วยความสามารถในการสังเคราะห์เสียงพูดได้ใกล้เคียงกับเสียงพูดตามธรรมชาติ และจำนวนพารามิเตอร์ที่ใช้ในการสังเคราะห์เสียงพูดมีจำนวนน้อย ทำให้มีการนำแบบจำลองฟูจิซากิไปใช้กับหลายภาษา ซึ่งมีทั้งภาษาที่โทนเสียงไม่มีความหมายเด่นชัดเช่นภาษาอิตาลี และภาษาที่โทนเสียงมีความหมายเด่นชัดเช่นภาษาแมนดาริน สำหรับภาษาที่โทนเสียงมีความหมายเด่นชัด ความถี่มูลฐานของเสียงพูดจะเป็นตัวบ่งบอกถึงโทนเสียงแทนสำเนียง (Accent) องค์ประกอบสำเนียง (Accent components) ถูกเรียกแทนที่ด้วยองค์ประกอบโทน (Tone components) กลไกการควบคุมสำเนียง (Accent control mechanism) ถูกเรียกแทนที่ด้วยกลไกการควบคุมโทน (Tone control mechanism) และคำสั่งสำเนียง (Accent command) ถูกเรียกแทนที่ด้วยคำสั่งโทน (Tone commands) ซึ่งโทนและสำเนียงมีความหมายในเชิงสัญญาณเช่นเดียวกัน

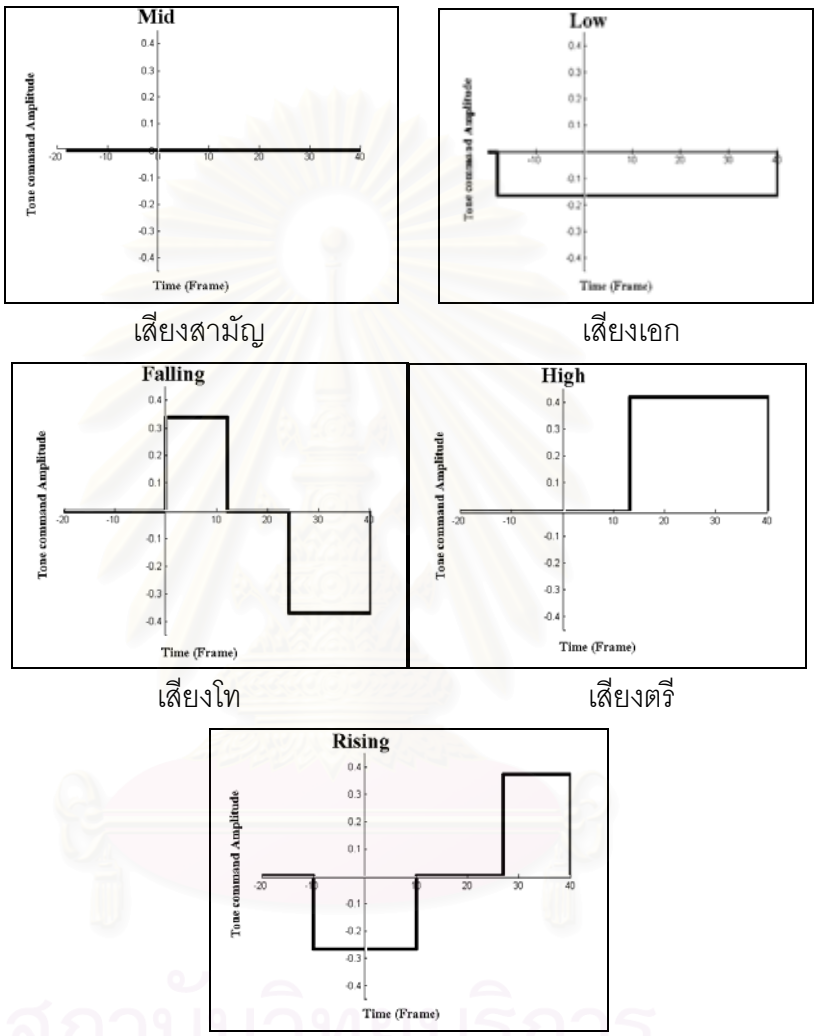


รูปที่ 2.17 คำสั่งวรรณยุกต์ในภาษาแมนดาริน [10]



รูปที่ 2.18 คำสั่งวรรณยุกต์ในภาษาไทย [9]

การวิเคราะห์สัญญาณเสียงพูดโดยแบบจำลองฟูจิซากิ สามารถทำได้โดยการสังเคราะห์เสียงพูดเลียนแบบเสียงที่ต้องการวิเคราะห์ และนำพารามิเตอร์ที่ใช้ในการสังเคราะห์นั้นมาทำการวิเคราะห์ จึงเรียกกระบวนการวิเคราะห์นี้เป็น “การวิเคราะห์โดยการสังเคราะห์”



เสียงสามัญ เสียงเอก

เสียงโท เสียงตรี

เสียงจัตวา

รูปที่ 2.19 คำสั่งวรรณยุกต์ในภาษาไทย [11]

ในการวิเคราะห์พารามิเตอร์ของแบบจำลองฟูจิซากิ เพื่อการแบ่งแยกสำหรับภาษาที่มีเสียงวรรณยุกต์ เช่นภาษาแมนดาริน [10] แบบจำลองฟูจิซากิแบบปกติไม่สามารถใช้แบ่งแยกเสียงวรรณยุกต์ได้ จึงได้มีการปรับแก้แบบจำลองฟูจิซากิ โดยปรับให้องค์ประกอบวลี ซึ่งมีค่าความถี่ในช่วงต่ำกว่าเส้นโค้งความถี่มูลฐาน มีความถี่อยู่ในช่วงกึ่งกลางของเส้นความถี่มูลฐานของเสียงพูด ซึ่งมีผลให้มีรูปแบบของคำสั่งวรรณยุกต์มากขึ้น โดยเสียงตรี (High tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีค่าแมกนิจูดเป็นบวก-บวก เสียงเอก (Low tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มี

ค่าแมกนิจูดเป็นลบ-ลบ เสียงโท (Fall tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีค่าแมกนิจูดเป็นบวก-ลบ และเสียงจัตวา (Rise tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีค่าแมกนิจูดเป็นลบ-บวก ดังแสดงได้ในรูปที่ 2.17

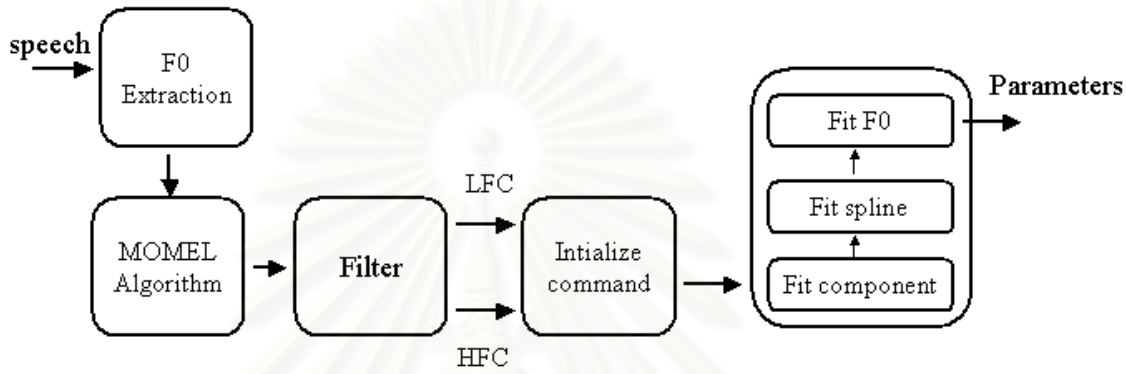
สำหรับภาษาไทย [9] จากการทดลองโดยควบคุมการควบร่วมของโทนเสียง (Tonal Assimilation) ด้วยการให้พยางค์ก่อนหน้าและหลังพยางค์ที่ต้องการวิเคราะห์มีเสียงสามัญ พบว่าแบบจำลองฟูจิซากิที่ถูกปรับแก้สามารถใช้แบ่งแยกเสียงวรรณยุกต์ได้ โดยเสียงสามัญ (Mid tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีแมกนิจูดเป็นศูนย์-ศูนย์ เสียงเอก (Low tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีแมกนิจูดเป็นลบ-ลบ เสียงโท (Falling tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีแมกนิจูดเป็นบวก-ลบ เสียงตรี (High tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีแมกนิจูดเป็นบวก-บวก และเสียงจัตวา (Rising tone) แทนได้ด้วยคำสั่งวรรณยุกต์ที่มีแมกนิจูดเป็นลบ-บวก ดังแสดงได้ในรูปที่ 2.18

ผลการศึกษาของ [9] สอดคล้องกับการศึกษาของ [11] ซึ่งสามารถแสดงลักษณะของคำสั่งโทนในแต่ละวรรณยุกต์ได้ดังรูปที่ 2.19

2.7 การหาค่าพารามิเตอร์ของแบบจำลองฟูจิซากิ

แบบจำลองฟูจิซากิ เป็นแบบจำลองที่ใช้ในการสร้างเส้นโค้งความถี่มูลฐาน จากพารามิเตอร์จำนวนน้อย ซึ่งการสร้างเส้นโค้งความถี่มูลฐานจากพารามิเตอร์สามารถทำได้โดยง่ายเพียงการแทนค่าลงในสมการของแบบจำลองเพื่อให้ได้เส้นโค้งที่มีลักษณะที่ต้องการ ในทางกลับกันการหาค่าพารามิเตอร์สำหรับแบบจำลองฟูจิซากิ สำหรับเส้นโค้งความถี่มูลฐานที่กำหนดกลับเป็นเรื่องที่ค่อนข้างยากลำบาก และยังไม่มีวิธีที่แน่นอน กระบวนการแยกพารามิเตอร์ที่ใช้กันอยู่ในปัจจุบันมักเป็นการแยกพารามิเตอร์โดยอาศัยการตัดสินใจของมนุษย์ อย่างไรก็ดี ในการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิสำหรับสัญญาณเสียงตัวอย่างจำนวนมาก และรวมไปถึงการแยกพารามิเตอร์ในการใช้งานจริงโดยเฉพาะในการนำไปใช้กับการรู้จำนั้น การแยกพารามิเตอร์โดยอาศัยการตัดสินใจของมนุษย์ย่อมไม่สามารถทำได้จริง ดังจะเห็นว่าได้มีความพยายามในการพัฒนากระบวนการแยกพารามิเตอร์แบบอัตโนมัติแบบใหม่ และนำเสนอออกมาอย่างต่อเนื่อง [12,13] นั่นคือยังไม่มีกระบวนการแยกพารามิเตอร์แบบอัตโนมัติ ที่เป็นที่ยอมรับกันในปัจจุบัน กระบวนการแยกพารามิเตอร์สำหรับแบบจำลองฟูจิซากิแบบอัตโนมัติที่ถูกอ้างว่ามีความใกล้เคียงกับกระบวนการแยกพารามิเตอร์โดยอาศัยการตัดสินใจของมนุษย์ในปัจจุบันคือวิธีของ Hansjorg Mixdorff [14] ซึ่งมีรายละเอียดดังต่อไปนี้

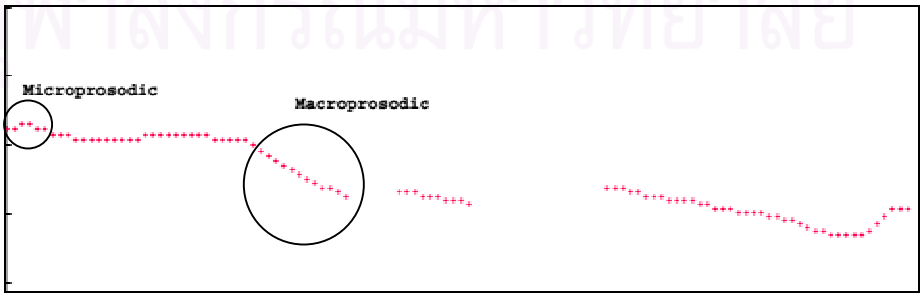
ตามวิธีของ Mixdorff การแยกพารามิเตอร์สำหรับแบบจำลองฟูจิซากิจากเส้นโค้งความถี่มูลฐานที่ได้ประกอบด้วย 4 ขั้นตอนหลักคือ การประมาณด้วยฟังก์ชันเสมือนพหุนามกำลังสอง (Quadratic Spline Stylisation) การกรองและแยกองค์ประกอบ (Filtering and Component Separation) การให้คำสั่งเริ่มต้นของแบบจำลอง (Fujisaki Model Command Initialization) และการวิเคราะห์โดยสังเคราะห์ (Analysis-by-Synthesis) ซึ่งสามารถแสดงได้ด้วยแผนภาพดังรูปที่ 2.20



รูปที่ 2.20 แผนภาพแสดงการแยกพารามิเตอร์ตามกรรมวิธีของ Mixdorff

2.7.1 การประมาณด้วยฟังก์ชันเสมือนพหุนามกำลังสอง

ขั้นตอนของการประมาณด้วยฟังก์ชันเสมือนพหุนามกำลังสองประกอบด้วย 2 ส่วนคือการประมาณค่าของเส้นโค้งความถี่มูลฐานที่หายไปในช่วงเสียงไม่ก้อง และการทำให้ไมโครโปรโซดิก (Microprosodic) เรียบ ทั้งนี้เนื่องจากแบบจำลองฟูจิซากิพิจารณาแต่เพียงมาโครโปรโซดิก (Macroprosodic) ซึ่งมีการเปลี่ยนแปลงที่ช้ากว่าเท่านั้น โดยวิธีของ Mixdorff ได้ใช้แบบจำลอง MOMEL ในการกำจัดไมโครโปรโซดิก



รูปที่ 2.21 Microprosodic และ Macroprosodic ของเสียง “กินอยู่กับปาก”

2.7.1.1 **แบบจำลอง** MOMEL [15]

MOMEL มาจาก MELodic MOdelisation ซึ่งถูกเสนอขึ้นโดย Daniel Hirst และ Robert Espesser ในปี 1991 เพื่อใช้ในการแสดงแทนเส้นโค้งเมโลดิก (Melodic) หรือเส้นโค้งความถี่มูลฐาน โดยแปลงจุดต่าง ๆ ของเส้นโค้งความถี่มูลฐานให้กลายเป็นจุดเป้าหมาย (Target Point) เพื่อใช้เป็นจุดอ้างอิงในการทำเส้นฟังก์ชันเสมือนพหุนาม การสร้างจุดเป้าหมายตามอัลกอริทึมของ MOMEL ประกอบด้วย 4 ขั้นตอน

2.7.1.1.1 **การประมวลผลเบื้องต้น**

ค่าความถี่มูลฐานทุกค่าที่มีค่าสูงหรือต่ำกว่าค่าข้างเคียงเกินกว่าร้อยละ 5 จะถูกตัดทิ้งหรือให้ค่าเป็น 0 กรณีที่ให้ค่าของความถี่มูลฐานสำหรับส่วนเสียงไม่ก้องเป็น 0

2.7.1.1.2 **ประมาณค่าทาร์เกตแคนดิเดต** (Target-candidate)

ขั้นตอนนี้เป็นการกระทำในทุกจุดเวลา x บนเส้นโค้งความถี่มูลฐาน โดยขั้นตอนประกอบด้วย

1. กำหนดหน้าต่างสำหรับการวิเคราะห์ที่มีความกว้าง A ซึ่งปกติจะมีค่าเป็น 300 มิลลิวินาทีโดยมีจุดศูนย์กลางของหน้าต่างที่ x ค่า F_0 ภายในหน้าต่างที่มีค่าน้อยกว่า $hz\min$ หรือมากกว่า $hz\max$ ที่กำหนดไว้จะถูกตัดทิ้ง โดยค่าที่กำหนดไว้โดยปกติจะเป็น $hz\min$ = 50 เฮิร์ตซและ $hz\max$ = 500 เฮิร์ตซ
2. ทำการหาพารามิเตอร์ของสมการถดถอยแบบกำลังสองจากจุดที่เหลือในหน้าต่าง
3. ค่า F_0 ที่มีค่าแตกต่างจาก F_0 ที่ได้จากการประมาณด้วยสมการถดถอยแบบกำลังสองเกินกว่าค่าที่กำหนดจะถูกตัดทิ้ง ซึ่งโดยปกติจะกำหนดค่าไว้ที่ร้อยละ 5 ของค่าที่ได้จากการประมาณ
4. ทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งไม่มีค่าใดถูกตัดทิ้ง
5. สำหรับทุกจุดเวลา x จะทำการคำนวณหาทาร์เกตแคนดิเดต $<t,h>$ จากสัมประสิทธิ์ถดถอย

$$\hat{y} = a + bx + cx^2 \qquad (2.17)$$

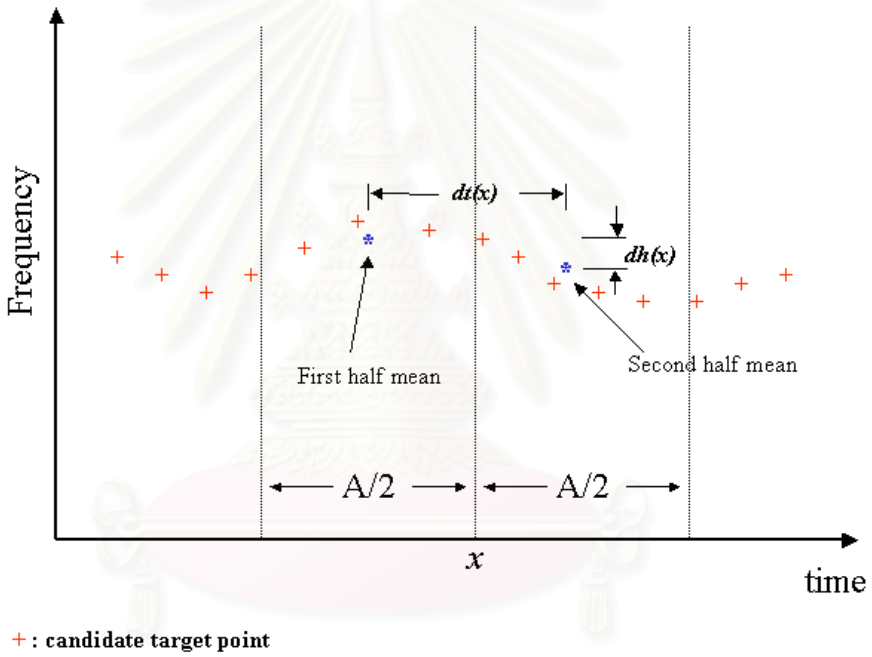
โดย

$$t = -b/(2c)$$
$$h = a + bt + ct^2 \qquad (2.18)$$

หากค่า $t < x - (A/2)$ หรือ $t > x + (A/2)$ หรือค่า $h < hz\min$ หรือ $h > hz\max$ ให้ตัดค่า t และ h ออกไปจากการเป็นทาร์เกตแคนดิเดต

6. ขั้นตอนที่ 2 ถึง 5 จะถูกทำซ้ำในทุกค่าของ x ซึ่งจะทำให้ได้ทาร์เกตแคนดิเดต $< t, h >$ หนึ่งจุด หรือจุดเป้าหมายที่ถูกตัดทิ้งสำหรับทุกค่า F_0

2.7.1.1.3 แบ่งทาร์เกตแคนดิเดต



รูปที่ 2.22 ทาร์เกตแคนดิเดต

กำหนดให้ความกว้างหน้าต่างเคลื่อนได้ R มีค่าปกติเป็น 200 มิลลิวินาทีและมีจุดกึ่งกลางที่ x ทำการคำนวณค่าระยะทางเฉลี่ยสัมบูรณ์ของค่า t และ h ระหว่างครึ่งแรกและครึ่งหลังของหน้าต่างได้เป็น $dt(x)$ และ $dh(x)$ ตามลำดับ จากนั้นคำนวณค่าระยะทางรวมถ่วงน้ำหนักได้

$$d(x) = \frac{dt(x) \cdot wd + dh(x) \cdot wh}{wd + wh} \qquad (2.19)$$

โดย

$$wd = \frac{1}{mean(dt(x))}$$
$$wh = \frac{1}{mean(dh(x))} \quad (2.20)$$

กำหนดขอบเขตของช่วงแบ่ง (partition) ที่ทุกค่า x ที่มีคุณสมบัติ 3 ข้อดังนี้

1. $d(x) > d(x-1)$
2. $d(x) > d(x+1)$
3. $d(x) > mean(d(x))$

2.7.1.1.4 ลดทาร์เกตแคนดิเดต

ในแต่ละช่วงแบ่ง ทาร์เกตแคนดิเดตที่มีค่า $dt(x)$ หรือ $dh(x)$ มากกว่าค่าเฉลี่ยภายในช่วงเกินกว่าหนึ่งส่วนเบี่ยงเบนมาตรฐานจะถูกตัดทิ้งไป และค่าเฉลี่ยของทาร์เกตแคนดิเดตที่เหลือในช่วงแบ่งนั้นจะถูกใช้เป็นจุดเป้าหมาย $< t, h >$ ของช่วงแบ่งนั้น

2.7.1.2 ประมาณเส้นความถี่มูลฐานจากจุดเป้าหมายด้วยฟังก์ชันเสมือนพหุนาม

การประมาณเส้นความถี่มูลฐานจากจุดเป้าหมาย ที่ได้จากขั้นตอนของการใช้อัลกอริทึม MOMEL จะทำการประมาณด้วยฟังก์ชันพหุนามกำลังสอง ซึ่งเป็นการประมาณค่าที่ไม่ทราบในข้อมูลแบบไม่ต่อเนื่องวิธีหนึ่ง

การประมาณค่าในช่วง (Interpolation) คือการประมาณค่าที่เกิดขึ้นระหว่างข้อมูลไม่ต่อเนื่องสองจุดด้วยสมการ โดยใช้ข้อมูลที่มีอยู่ ณ จุดต่าง ๆ ในการประมาณ ซึ่งปกติแล้วสมการที่ใช้มักเป็นสมการพหุนามเนื่องจากความง่ายในประเด็นต่าง ๆ ดังนี้

- หาแก้สมการ
- การหาค่าอนุพันธ์
- การหาค่าอินทิเกรต

การประมาณค่าในช่วงด้วยสมการพหุนาม คือการหาสมการพหุนามอันดับที่ n ที่ผ่านจุดทั้งหมดจำนวน $n+1$ จุดข้อมูล แต่การที่อันดับของสมการที่ใช้สอดคล้องกับจำนวนชุดข้อมูลมีค่ามากเกินไป และส่งผลให้อาจเกิดการออสซิเลตได้ ดังนั้นจึงได้มีแนวคิดสำหรับการประมาณค่าในช่วงที่มีจำนวนจุดข้อมูลมากโดยใช้ฟังก์ชันเสมือนพหุนาม และฟังก์ชันเสมือนพหุนามที่ใช้กันโดยทั่วไปได้แก่ฟังก์ชันพหุนามเชิงเส้น ฟังก์ชันพหุนามกำลังสอง และฟังก์ชันพหุนามกำลังสาม

สำหรับจุดข้อมูลที่มีค่าพิกัด $(x_0, y_0),(x_1, y_1),\text{K},(x_{n-1}, y_{n-1}),(x_n, y_n)$ การประมาณค่าด้วยฟังก์ชันเสมือนพหุนามกำลังสองคือ การหาฟังก์ชันเสมือนพหุนามกำลังสองที่เหมาะสมกับจุดข้อมูลเหล่านี้ โดยกำหนดให้ฟังก์ชันเสมือนพหุนามเป็น

$$f(x) = a_1x^2 + b_1x + c_1 \qquad , x_0 \le x \le x_1$$
$$f(x) = a_2x^2 + b_2x + c_2 \qquad , x_1 \le x \le x_2$$
$$f(x) = a_3x^2 + b_3x + c_3 \qquad , x_2 \le x \le x_3$$
$$\text{M}$$
$$f(x) = a_nx^2 + b_nx + c_n \qquad , x_{n-1} \le x \le x_n$$

การประมาณด้วยฟังก์ชันเสมือนพหุนามกำลังสองคือ การหาค่าสัมประสิทธิ์เพื่อที่จะหาฟังก์ชันที่เหมาะสมจำนวน $3n$ ตัวคือ

$$a_i, i = 1,2,..,n$$
$$b_i, i = 1,2,..,n$$
$$c_i, i = 1,2,..,n$$

ในการหาค่าตัวแปรจำนวน $3n$ ตัวจะต้องใช้สมการจำนวน $3n$ สมการ และต้องแก้สมการทั้งหมดเพื่อหาคำตอบ โดยสมการทั้งหมดหาได้จาก

- ทุกเส้นของฟังก์ชันเสมือนพหุนาม จะต้องผ่านจุดข้อมูลสองจุดที่ติดกัน ดังนั้นจะได้สมการจำนวน $2n$ สมการคือ

$$a_1x_0^2 + b_1x_0 + c_1 = f(x_0)$$
$$a_1x_1^2 + b_1x_1 + c_1 = f(x_1)$$
$$\text{M}$$
$$a_ix_{i-1}^2 + b_ix_{i-1} + c_i = f(x_{i-1})$$
$$a_ix_i^2 + b_ix_i + c_i = f(x_i)$$
$$\text{M}$$
$$a_nx_{n-1}^2 + b_nx_{n-1} + c_n = f(x_{n-1})$$
$$a_nx_n^2 + b_nx_n + c_n = f(x_n)$$

- อนุพันธ์อันดับที่หนึ่งของฟังก์ชันเสมือนพหุนามสองเส้นที่ติดกันจะต้องต่อเนื่อง

ตัวอย่างเช่นฟังก์ชันเสมือนพหุนามเส้นแรก

$$\frac{d(a_1x^2 + b_1x + c_1)}{dx} = 2a_1x + b_1$$

ฟังก์ชันเสมือนพหุนามเส้นที่สอง

$$\frac{d(a_2x^2 + b_2x + c_2)}{dx} = 2a_2x + b_2$$

และอนุพันธ์อันดับที่หนึ่งของทั้งสองฟังก์ชันจะมีค่าเท่ากันที่ $x = x_1$ จะได้เป็น

$$2a_1x_1 + b_1 = 2a_2x1 + b_2$$
$$2a_1x_1 + b_1 - 2a_2x_1 - b_2 = 0$$

และจะได้สมการในลักษณะเดียวกันของจุดข้อมูลอื่นเป็น

$$2a_2x_2 + b_2 - 2a_3x_2 - b_3 = 0$$
$$\text{M}$$
$$2a_ix_i + b_i - 2a_{i+1}x_i - b_{i+1} = 0$$
$$\text{M}$$
$$2a_{n-1}x_{n-1} + b_{n-1} - 2a_nx_{n-1} - b_n = 0$$

ดังนั้นจะได้สมการจำนวน $n-1$ สมการ

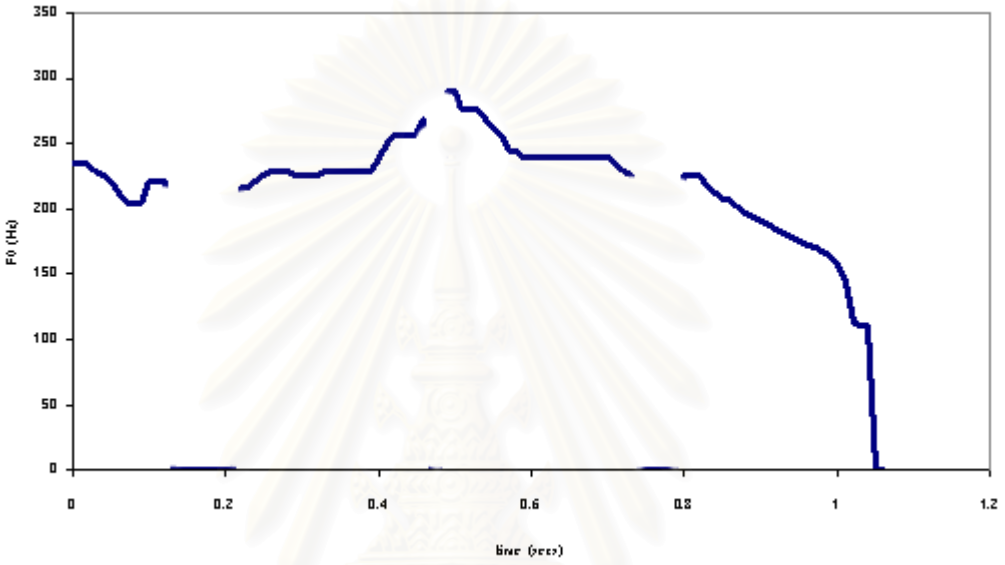
- ตั้งสมมติฐานให้ฟังก์ชันเสมือนพหุนามฟังก์ชันแรกเป็นเส้นตรง นั่นคือ

$$a_1 = 0$$

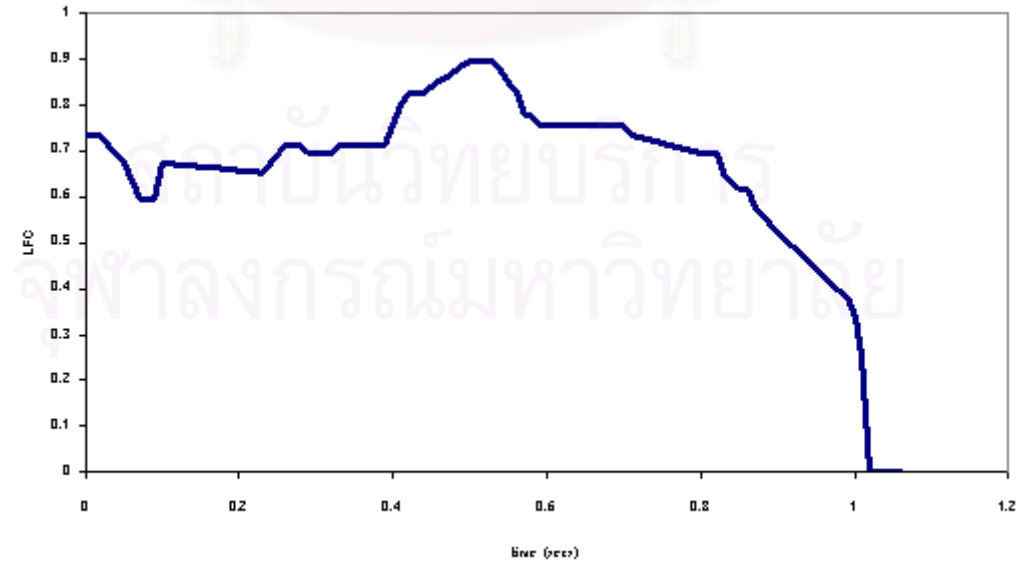
จะได้จำนวนสมการทั้งสิ้นเป็น $3n$ สมการ เมื่อทำการแก้สมการหาค่าสัมประสิทธิ์ทั้งหมดจะทำให้ได้ฟังก์ชัน $f(x)$ ซึ่งใช้แทนเส้นที่ต้องการประมาณตลอดช่วงข้อมูลตั้งแต่ (x_0, y_0) จนถึง (x_n, y_n)

2.7.2 การกรองและแยกองค์ประกอบ

จากการที่แบบจำลองฟูจิซากิทำการสร้างเส้นโค้งความถี่มูลฐานในสเกลลอการิทึม ด้วยการรวมกันขององค์ประกอบสามส่วนคือ องค์ประกอบวลีซึ่งเป็นส่วนที่เกี่ยวข้องกับวลีและการลดระดับอย่างช้า ๆ (declination) โดยรวมของความถี่มูลฐาน องค์ประกอบสำเนียงซึ่งเป็นการเปลี่ยนแปลงที่รวดเร็วกว่าในเส้นโค้งความถี่มูลฐาน และความถี่ฐานที่เป็นค่าคงที่ขึ้นอยู่กับตัวบุคคล

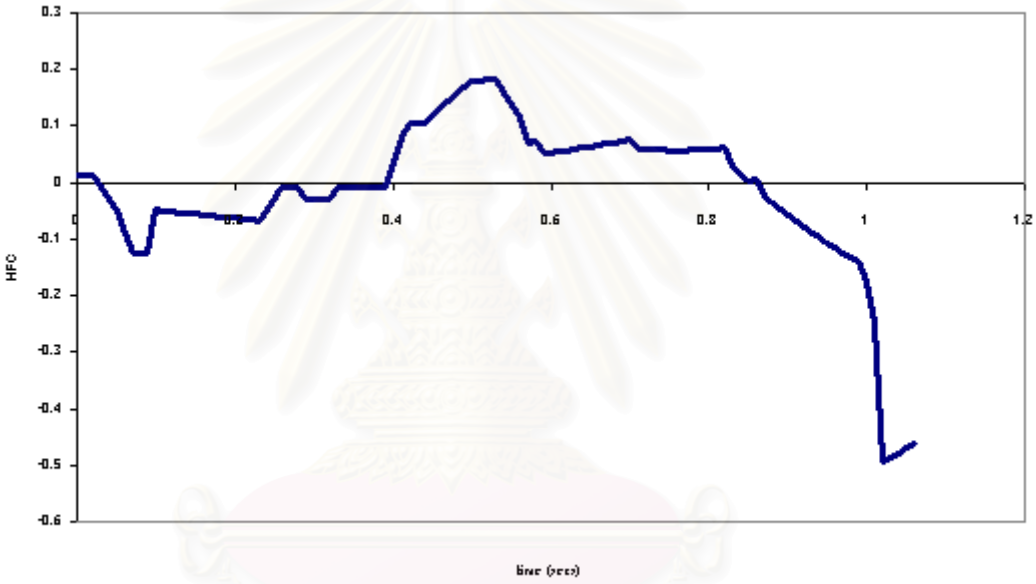


รูปที่ 2.23 เส้นโค้งความถี่มูลฐานของเสียง “จับแพะชนแกะ”



รูปที่ 2.24 เส้นโค้งความถี่ต่ำของเสียง “จับแพะชนแกะ”

ในการแยกองค์ประกอบสำเนียงออกจากองค์ประกอบวลีและความถี่ฐาน เส้นโค้งจากการประมาณค่าด้วยฟังก์ชันเสมือนพหุนามกำลังสอง จะผ่านตัวกรองผ่านสูง (Highpass) ที่มีความถี่ตัดที่ 0.5 เฮิร์ตซ ผลลัพธ์ที่ได้จากตัวกรองเรียกเป็นเส้นโค้งความถี่สูง (High Frequency Contour-HFC) จะถูกหักออกจากเส้นโค้งฟังก์ชันเสมือนพหุนาม ทำให้ได้ผลลัพธ์เป็นเส้นโค้งความถี่ต่ำ (Low Frequency Contour-LFC) ซึ่งเป็นส่วนขององค์ประกอบวลีและความถี่ฐาน โดยกำหนดให้ค่าความถี่ฐานคือค่าความถี่ที่ต่ำที่สุดของเส้นโค้งความถี่ต่ำ ดังนั้นด้วยวิธีดังกล่าว ทำให้สามารถแยกองค์ประกอบวลี องค์ประกอบสำเนียง และความถี่ฐานออกจากกันได้อย่างหยาบ ๆ ดังแสดงในรูปที่ 2.23 2.24 และ 2.25



รูปที่ 2.25 เส้นโค้งความถี่สูงของเสียง “จับแพะชนแกะ”

2.7.3 การให้คำสั่งเริ่มต้นของแบบจำลอง

กระบวนการให้ค่าเริ่มต้นของคำสั่ง จะใช้คุณลักษณะของผลตอบสนองต่อคำสั่งวลีและคำสั่งสำเนียง ที่ทำให้เกิดเป็นองค์ประกอบวลีและองค์ประกอบสำเนียง

จากการที่ผลตอบสนองต่อคำสั่งวลี เริ่มมีค่าสูงขึ้นที่จุดที่เกิดคำสั่งวลีซึ่งมีลักษณะเป็นอิมพัลส์ และสูงขึ้นจนถึงจุดสูงสุดแล้วลดลงอย่างช้า ๆ โดยขึ้นอยู่กับค่าคงตัวเวลา α ดังนั้น สำหรับการกำหนดคำสั่งวลี จะได้ว่าค่าเวลาของคำสั่งวลีจะหาได้จากค่าต่ำสุดท้องถิ่น (Local minimum) ขององค์ประกอบวลี โดยในกระบวนการนี้จะทำการหาค่าต่ำสุดท้องถิ่นโดยมีระยะห่างจากันระหว่างคำสั่งวลีที่ติดกันไม่น้อยกว่า 1 วินาทีซึ่งเป็นค่าที่ได้จากการทดลองโดย [14] สำหรับการตั้ง

ค่าเริ่มต้นของค่าแมกนิจูด A_p จะทำการหาค่าจุดสูงสุดท้องถิ่นหลังจากเวลาเริ่มต้นของคำสั่งวลีนั้น โดย A_p จะถูกคำนวณได้จากการให้เป็นสัดส่วนกับค่าความถี่ที่พบที่จุดดังกล่าว จากการที่สามารถมีหลายคำสั่งวลีได้ในหนึ่งประโยค ทำให้การหาค่าเริ่มต้นของ A_p จะต้องคำนึงถึงผลของการเกิดขึ้นของคำสั่งวลีก่อนหน้านี้ด้วยในการคำนวณ ซึ่งมีผลให้ค่า A_p ที่กำหนดมีค่าลดลง สำหรับค่า α จากการทดลองของ [14] พบว่ามีค่าเหมาะสมที่ 1.0

ผลตอบสนองต่อคำสั่งสำเนียงเริ่มต้นจากค่า 0 ที่เวลาออนเซ็ทของคำสั่ง T_1 ไปจนมีค่าสูงสุดที่เวลาออฟเซ็ทของคำสั่ง T_2 ซึ่งทำให้ผลตอบสนองเริ่มลดลง ในการให้ค่าตั้งต้นที่เหมาะสมของ T_1 และ T_2 ของคำสั่งสำเนียง จะทำการหาค่าต่ำสุดเฉพาะแห่งของเส้นโค้งความถี่สูงซึ่งต้องเป็นค่าต่ำที่สุดในช่วงเวลาก่อนหน้าและหลังจากนั้นเป็นระยะเวลลา 100 มิลลิวินาทีเพื่อมิให้เกิดการกำหนดค่าที่จุดอานม้า (Saddle point) และภายในระหว่างสองจุดที่ได้จะถูกกำหนดให้มีคำสั่งสำเนียงขึ้นโดยมี T_1 ที่จุดที่เริ่มต้นช่วง

จากการที่ผลตอบสนองของคำสั่งสำเนียงจะต้องใช้เวลาส่วนหนึ่งเพื่อกลับสู่ค่า 0 หลังจากการเกิด T_2 ดังนั้นจึงตั้งค่า T_2 ไว้ก่อนการเกิดขึ้นของจุดต่ำสุดท้องถิ่นถัดไปเป็นเวลา 200 มิลลิวินาทีโดย ค่าคงตัวเวลา β ถูกตั้งค่าไว้ที่ 20 ส่วนค่าเริ่มต้นของแอมปลิจูดของคำสั่งสำเนียง A_a จะถูกตั้งให้มีค่าเป็นสัดส่วนกับจุดที่ความถี่สูงสุดในช่วง T_1 ถึง T_2 ตามความสัมพันธ์ของผลตอบสนองของกลไกควบคุมสำเนียงกับค่าแมกนิจูดของคำสั่งสำเนียง โดยค่าคงที่ทั้งหมดได้จากการทดลองของ [14]

2.7.4 การวิเคราะห์โดยสังเคราะห์

การวิเคราะห์โดยสังเคราะห์ ประกอบด้วยสามขั้นตอนในการปรับค่าพารามิเตอร์ที่ได้จากค่าเริ่มต้นโดยใช้วิธีเปลี่ยนค่าพารามิเตอร์ให้มากขึ้นหรือน้อยลงเรื่อย ๆ เพื่อลดค่าความผิดพลาดกำลังสองเฉลี่ยโดยรวมทั้งหมดในสเกลลอการิทึม

ในขั้นตอนที่ 1 องค์ประกอบวลีและองค์ประกอบสำเนียงจะถูกแยกพิจารณาออกจากกันโดยใช้เส้นโค้งความถี่ต่ำและเส้นโค้งความถี่สูงเป็นเป้าหมายในการปรับค่าพารามิเตอร์

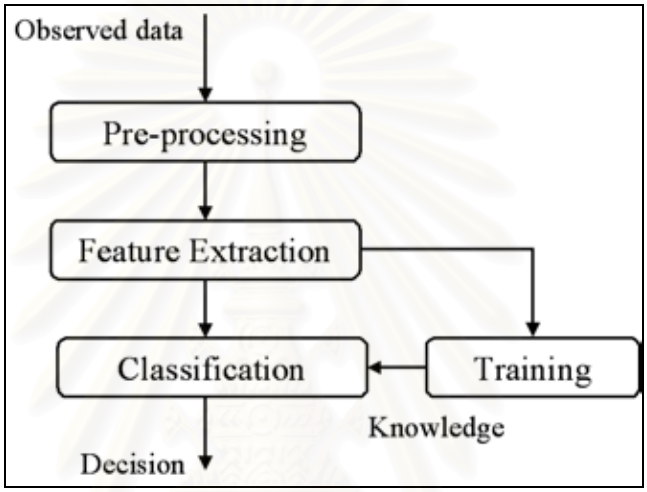
ขั้นตอนที่ 2 องค์ประกอบวลี องค์ประกอบสำเนียงและความถี่พื้นฐานจะถูกพิจารณาร่วมกันในการปรับค่าให้ใกล้เคียงกับเป้าหมาย คือเส้นโค้งที่ได้จากการประมาณด้วยฟังก์ชันเสมือนพหุนามกำลังสอง

ขั้นตอนที่ 3 พารามิเตอร์ทั้งหมดจะถูกปรับแต่งอีกครั้งโดยใช้เส้นโค้งความถี่มูลฐานเป็นเป้าหมายโดยมีการถ่วงน้ำหนัก ซึ่งค่าที่ใช้ในการถ่วงน้ำหนักจะได้จากผลคูณของระดับความเป็น

เสียงก้องกับพลังงานในเฟรมนั้น ๆ ของทุกค่าความถี่มูลฐาน โดยระดับความเป็นเสียงก้องมีค่าเป็น 0 เมื่อช่วงดังกล่าวไม่ได้เป็นเสียงก้อง และมีค่าเป็น 1 เมื่อช่วงดังกล่าวเป็นเสียงก้อง

2.8 **การรู้จำแบบรูป** (Pattern Recognition)

การรู้จำแบบรูปเป็นกระบวนการที่ใช้ในการตัดสินใจของระบบจากข้อมูลเชิงปริมาณ เช่น ระบบคอมพิวเตอร์ โดยการรู้จำแบบรูปมีขั้นตอนพื้นฐานดังนี้



รูปที่ 2.26 ขั้นตอนของการรู้จำแบบรูป

2.8.1 **การประมวลผลเบื้องต้น** (Pre-processing)

การประมวลผลเบื้องต้น เป็นกระบวนการปรับข้อมูลที่สังเกต (Observe) หรือวัดได้ เพื่อให้อยู่ในรูปแบบที่เหมาะสมกับการดำเนินการขั้นต่อไป รวมทั้งเป็นการลดข้อมูลที่ไม่ถูกต้องบางส่วนอันอาจเกิดจากการรบกวนหรือความผิดพลาดของการสังเกต ซึ่งมีส่วนทำให้ประสิทธิผลของการรู้จำดีขึ้นด้วย

2.8.2 **การสกัดคุณลักษณะสำคัญ** (Feature Extraction)

เป็นการประมวลผล เพื่อให้ได้มาซึ่งค่าที่ใช้เป็นตัวแทนของข้อมูลที่สังเกตได้สำหรับใช้ในการแยกประเภทต่อไป ประสิทธิภาพของการรู้จำขึ้นอยู่กับความเหมาะสมของค่าคุณลักษณะสำคัญมาก เนื่องจากคุณลักษณะสำคัญที่ดี จะทำให้การแยกประเภททำได้ง่ายและถูกต้อง ใน

ขณะที่ค่าคุณลักษณะสำคัญบางอย่างมีความใกล้เคียงกันมากในแต่ละประเภทของข้อมูลที่สังเกต ทำให้การแยกประเภทเป็นไปได้ยาก และให้ผลที่ไม่ดี และการใช้จำนวนค่าคุณลักษณะที่มากขึ้นมักจะมีผลทำให้การคำนวณในส่วนของการแยกประเภทมากขึ้นตามไปด้วย สำหรับเสียงพูดแล้วค่าคุณลักษณะที่ใช้กันทั่วไปอย่างกว้างขวางได้แก่ ค่าพลังงาน ค่าในเชิงสเปกตรัม และค่าที่ได้จากการแปลง (Transform) ต่าง ๆ

2.8.3 **การจำแนก** (Classification)

การจำแนกคือกระบวนการตัดสินใจ โดยข้อมูลที่ใช้ในการตัดสินใจคือค่าคุณลักษณะสำคัญที่ได้จากขั้นตอนการแยกคุณลักษณะสำคัญ แต่มักจะต้องมีความรู้ (Knowledge) ที่มีบันทึกไว้ล่วงหน้า โดยความรู้นั้นอาจได้มาจากการตั้งกฎ เช่นการใช้ Decision Tree เป็นตัวแยกประเภท (Classifier) หรือได้มาจากการฝึกให้ตัวแยกประเภทมีความรู้ ซึ่งได้จากการนำค่าสังเกตที่แปลงเป็นค่าคุณลักษณะสำคัญมาทำการฝึกฝนระบบ ตัวอย่างของตัวแยกประเภทชนิดนี้ได้แก่ แบบจำลองฮิดเดนมาร์คอฟ และโครงข่ายประสาทเทียม ซึ่งตัวแยกประเภทแต่ละชนิดจะมีข้อดีและข้อเสียต่างกันไป อีกทั้งความเหมาะสมของตัวแยกประเภทยังขึ้นอยู่กับลักษณะของค่าคุณลักษณะสำคัญที่ใช้อีกด้วย

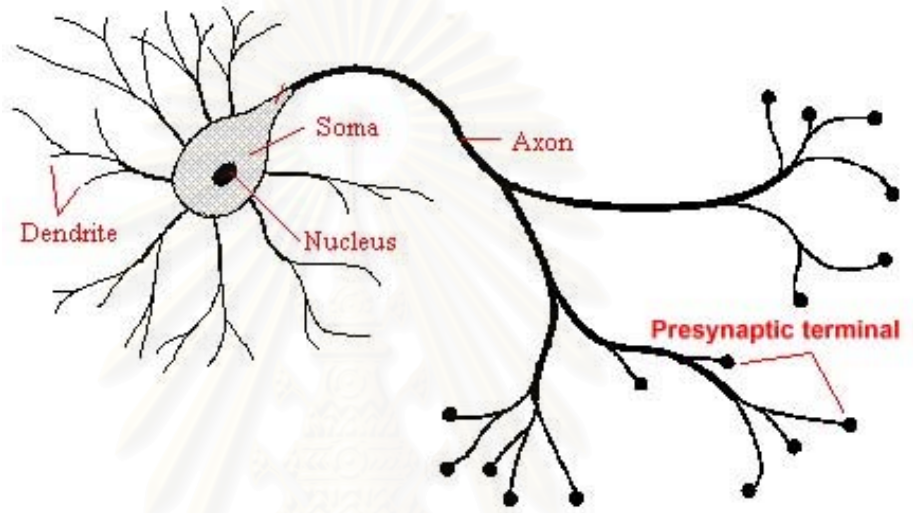
2.8.4 **การประมวลผลภายหลัง** (Post-processing)

การประมวลผลภายหลัง คือกระบวนการที่ทำหลังจากการแยกประเภทได้ทำการตัดสินใจแล้ว โดยกระบวนการภายหลังนี้ จะนำความรู้ทางด้านอื่นเข้ามาประมวลผลร่วมเพื่อทำให้ผลการตัดสินใจที่ได้ถูกต้องมากขึ้น เช่นการนำความรู้ทางด้านภาษาเช่นโครงสร้างทางด้านภาษา ศัพท์ที่เป็นไปได้ เข้ามาร่วมพิจารณา

สำหรับวิทยานิพนธ์นี้ จะใช้โครงข่ายประสาทเทียมเป็นตัวแยกประเภท เนื่องจากมีลักษณะเหมาะสมกับค่าคุณลักษณะที่ได้จากพารามิเตอร์ของแบบจำลองฟูจิซากิ และจะไม่กล่าวถึงตัวแยกประเภทชนิดอื่นเช่นแบบจำลองฮิดเดนมาร์คอฟ หรือ ซัพพอร์ตเวกเตอร์แมชชีน ในรายละเอียด

2.9 **โครงข่ายประสาทเทียม** (Artificial Neural Network)

โครงข่ายประสาทเทียม เป็นโครงข่ายที่สร้างขึ้นเพื่อเลียนแบบการทำงานของโครงข่ายประสาทจริง เนื่องจากการทำงานของโครงข่ายประสาทจะทำงานพร้อม ๆ กันไปในแต่ละเซลล์ประสาท จึงทำให้โครงข่ายประสาทสามารถทำการประมวลผลได้อย่างรวดเร็ว โดยในเซลล์ประสาทแต่ละเซลล์ประกอบด้วย

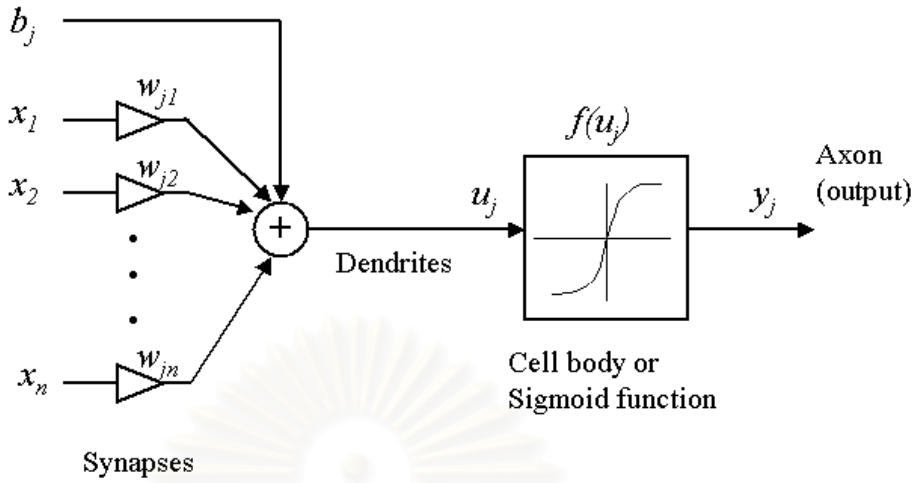


รูปที่ 2.27 โครงสร้างของเซลล์ประสาททางชีวภาพ

- Soma เป็นตัวเซลล์ประสาท
- Axon เป็นเส้นใยประสาทที่ต่อออกจากเซลล์ประสาทโดยทำหน้าที่เป็นด้านออกของเซลล์ประสาทเพื่อส่งต่อไปยังเซลล์ประสาทอื่น
- Dendrites ทำหน้าที่เป็นด้านเข้าของเซลล์ประสาทเพื่อรับสัญญาณจากเซลล์ประสาทตัวอื่น โดยผ่าน Axon ซึ่งต่อกับ Synapses เพื่อส่งไปยัง Soma
- Synapses เป็นตัวต่อ Dendrite กับ Axon จากเซลล์ประสาทอื่น

โครงสร้างของเซลล์ประสาทแสดงได้ดังรูปที่ 2.27

กล่าวโดยทั่วไปแล้ว โครงข่ายประสาทเทียมคือระบบประมวลสัญญาณที่ประกอบด้วยตัวประมวลผลอย่างง่าย ๆ จำนวนมาก ซึ่งเรียกหน่วยประมวลผลเหล่านี้ว่านิวรอน (Neuron) มาต่อเข้าด้วยกันเป็นโครงข่าย ซึ่งการทำงานของโครงข่ายนี้จะกระทำขนานไปพร้อม ๆ กันในแต่ละนิวรอนเพื่อแก้ปัญหาที่ต้องการ โดยแบบจำลองพื้นฐานของนิวรอนแสดงได้ดังรูปที่ 2.28



รูปที่ 2.28 แบบจำลองพื้นฐานของเซลล์ประสาทเทียม

จากรูปที่ 2.28 จะเห็นว่าค่าถ่วงน้ำหนัก w_{ji} ทำหน้าที่เสมือนการกำหนดความสำคัญของ Axon จากเซลล์ก่อนหน้าที่จะเชื่อมต่อกับ Synapses เพื่อต่อด้านเข้า x_i เข้าสู่ตัวรวมซึ่งทำหน้าที่เหมือน Dendrite เพื่อส่งสัญญาณเข้าสู่ตัว activation function $f(u_j)$ ซึ่งเปรียบเสมือน Soma ส่วน bias b_j เป็นค่าระดับอ้างอิงที่ป้อนจากภายนอก ผลของการประมวลผล y_j จะถูกส่งออกที่ด้านออกของตัวขยายแบบไม่เป็นเชิงเส้นซึ่งเปรียบเสมือนกับ Axon โดยสามารถเขียนสมการได้เป็น

$$y_j = f(\sum_{i=1}^{n} w_{ji} x_i + b_j) \tag{2.21}$$

ค่าระดับอ้างอิงจากภายนอก b_j สามารถเขียนในรูปของค่าถ่วงน้ำหนักได้เป็น $w_{j0} = b_j$ และ $x_0 = 1$ จึงทำให้สมการที่ 2.21 เขียนใหม่ได้เป็น

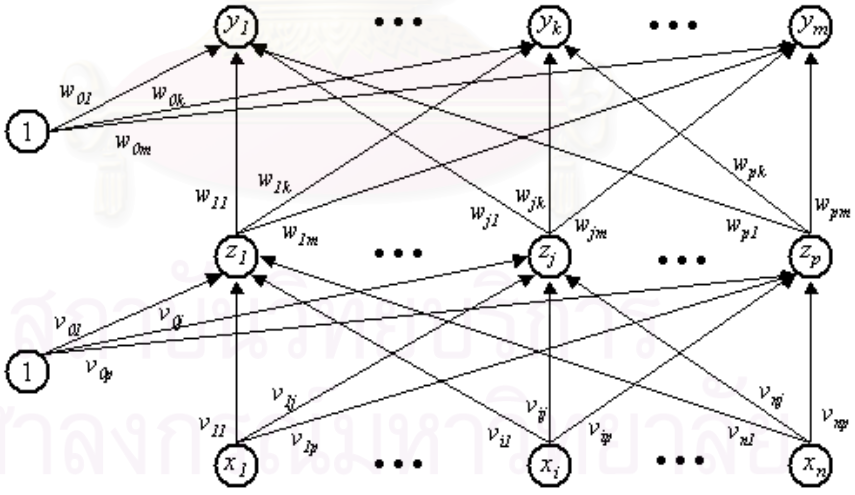
$$y_j = f(\sum_{i=0}^{n} w_{ji} x_i) \tag{2.22}$$

การใช้โครงข่ายประสาทเทียมในการแก้ปัญหา คือการสร้างโครงข่ายของนิวรอนที่มีค่าถ่วงน้ำหนักที่เหมาะสม ซึ่งค่าถ่วงน้ำหนักที่เหมาะสมจะหาได้จากกระบวนการฝึกฝนโครงข่าย การฝึก (Training) โครงข่ายมีด้วยกัน 2 แบบคือการเรียนรู้แบบชี้นำ (Supervised Learning) และการเรียนรู้แบบไม่มีการชี้นำ (Unsupervised Learning) โดยการเรียนรู้แบบชี้นำจะทำการกำหนดเซต

ของการฝึกให้กับโครงข่าย ซึ่งเซตนี้ประกอบด้วยอินพุตและเอาต์พุตที่ต้องการ เมื่อป้อนอินพุตให้กับโครงข่าย โครงข่ายจะทำการประมวลผลจนได้คำตอบและค่าถ่วงน้ำหนักออกมาชุดหนึ่ง สำหรับคำตอบที่ได้จะถูกนำมาคำนวณค่าความผิดพลาด โดยวัดเป็นระยะทางว่ามีความห่างจากคำตอบที่ต้องการของอินพุตในชุดเดียวกันมากน้อยเพียงใด ถ้ายังมีความผิดพลาดสูงอยู่ก็จะมีการปรับค่าถ่วงน้ำหนักและทำการฝึกต่อไปจนกว่าค่าความผิดพลาดมีค่าน้อยพอที่จะยอมรับได้จึงหยุดการฝึก

การเรียนรู้แบบไม่มีการชี้นำ จะทำการป้อนอินพุตเข้าสู่โครงข่าย และภายในโครงข่ายจะมีเอาต์พุตโนดอยู่หลายโนดด้วยกัน โดยแต่ละโนดจะแทนกลุ่มของข้อมูลที่มีคุณสมบัติเหมือนกัน เมื่อป้อนอินพุตเข้าสู่โครงข่าย โครงข่ายจะคำนวณค่าความสัมพันธ์ที่มีภายในเซตของอินพุต โดยอาศัยค่าถ่วงน้ำหนักเป็นตัวแยกความแตกต่างของอินพุตไปเก็บไว้ในโนดเอาต์พุตของโครงข่าย การเรียนรู้โดยวิธีนี้จะไม่สามารถระบุได้ว่าเอาต์พุตโนดใดเป็นของข้อมูลกลุ่มไหน ซึ่งผู้ใช้จะต้องกำหนดเอง

ในวิทยานิพนธ์ฉบับนี้ ได้นำโครงข่ายประสาทเทียมแบบหลายชั้นป้อนไปข้างหน้า (Feedforward Multi-layer Neural Network) มาใช้ และใช้การฝึกโครงข่ายโดยใช้อัลกอริทึมการแพร่ย้อนกลับ (Backpropagation Algorithm) ซึ่งเป็นการฝึกแบบมีการชี้นำ เนื่องจากเข้าใจง่าย มีความซับซ้อนน้อยและมีความสามารถในการจำแนกได้ดี



รูปที่ 2.29 แสดงภาพโครงข่ายประสาทเทียมที่มีโครงสร้างแบบ Feedforward

จากรูปที่ 2.29 เป็นโครงข่ายประสาทเทียมที่มีชั้นซ่อน (Hidden layer) 1 ชั้น โดยที่ตัวแปร x แทนอินพุตโนด ตัวแปร z แทนโนดซ่อน และตัวแปร y แทนเอาต์พุตโนด bias ที่เอาต์พุต

โนด y_k กำหนดเป็น w_{0k} และ bias ที่โนดซ่อน z_j กำหนดเป็น v_{0j} ตามลำดับ ซึ่งกำหนดด้วยตัวแปรที่แตกต่างกันเพื่อแยกแสดงค่าถ่วงน้ำหนักและ bias ในแต่ละชั้นโดยกำหนดตัวแปรดังนี้

X : Input training vector $X = (x_1, \mathrm{K}, x_i, \mathrm{K}, x_n)$

T : Output target vector $T = (t_1, \mathrm{K}, t_k, \mathrm{K}, t_m)$

δ_k : Error term เอาต์พุตโนดเพื่อนำไปปรับค่าถ่วงน้ำหนักระหว่างชั้นซ่อนกับชั้นเอาต์พุต

δ_j : Error term ของโนดซ่อน เพื่อนำไปปรับค่าถ่วงน้ำหนักระหว่างชั้นอินพุตกับชั้นซ่อน

λ : อัตราการเรียนรู้ (Learning rate)

v_{0j} : bias ของโนดซ่อน z_j

z_j : โนดซ่อนที่ j โดยมีอินพุตเป็น

$$z\_in_j = v_{0j} + \sum_i x_i v_{ij}$$

และมีเอาต์พุตเมื่อผ่านฟังก์ชันกระตุ้นเป็น

$$z_j = f(z\_in_j)$$

w_{0k} : bias ของเอาต์พุตโนด y_k

y_k : เอาต์พุตโนดที่ k โดยกำหนดอินพุตเป็น

$$y\_in_k = w_{0k} + \sum_j z_j w_{jk}$$

และมีเอาต์พุตเมื่อผ่านฟังก์ชันกระตุ้นเป็น

$$y_k = f(y\_in_k)$$

จะได้พจน์ผิดพลาดของเอาต์พุตโนด y_k เป็น

$$\delta_k = (t_k - y_k) f'(y\_in_k)$$

ได้ค่าที่ใช้ปรับค่าถ่วงน้ำหนักเป็น

$$\Delta w_{jk} = \lambda \delta_k z_j$$

$$\Delta w_{0k} = \lambda \delta_k$$

และได้พจน์ผิดพลาดของโนดซ่อน z_j เป็น

$$\delta\_in_j = \sum_{k=1}^{m} \delta_k w_{jk}$$

$$\delta_j = \delta\_in_j f'(z\_in_j)$$

ได้ค่าที่ใช้ปรับค่าถ่วงน้ำหนักเป็น

$$\Delta v_{ij} = \lambda \delta_j x_i$$

$$\Delta v_{0j} = \lambda \delta_j$$

จะได้ค่าถ่วงน้ำหนักใหม่เป็น

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk}$$

$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij}$$

การฝึกจะกระทำซ้ำจนกระทั่งค่าความผิดพลาดมีค่าต่ำกว่าที่กำหนดโดยค่าความผิดพลาดหาได้จาก

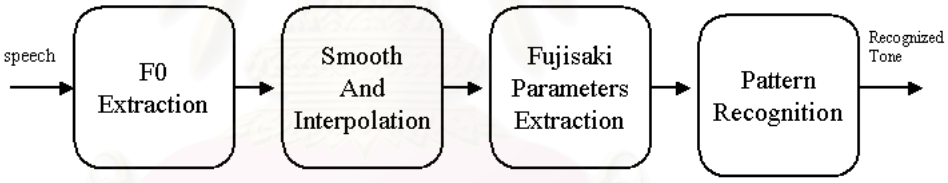
$$E = \frac{1}{2} \sum_p \sum_k (t_k - y_k)^2$$

ผลลัพธ์ที่ได้จากการใช้โครงข่ายประสาทเทียมในการแก้ปัญหาจะอยู่ในรูปของเวกเตอร์ผลลัพธ์ที่ได้จากเอาต์พุตโนด โดยผลลัพธ์ที่ได้หากมีค่าใกล้เคียงกับเอาต์พุตเวกเตอร์เป้าหมายใด ก็จะถือว่าโครงข่ายนั้นตัดสินใจให้ผลลัพธ์อยู่ในประเภทเดียวกับเอาต์พุตเวกเตอร์เป้าหมายนั้น

บทที่ 3
แนวคิดที่นำเสนอ

จากแนวคิดเรื่องแบบจำลองฟูจิซากิที่ทำการสร้างเส้นโค้งความถี่มูลฐานเลียนแบบเส้นโค้งความถี่มูลฐานที่ได้จากเสียงที่บันทึกไว้ แสดงให้เห็นว่า พารามิเตอร์ของแบบจำลองฟูจิซากิได้ทำการเก็บสารสนเทศทั้งหมดของเส้นโค้งความถี่มูลฐานนั้นเอาไว้ และจากความสัมพันธ์ของเส้นโค้งความถี่มูลฐานกับวรรณยุกต์ในเสียงพูดภาษาไทย จึงน่าจะนำพารามิเตอร์ของแบบจำลองฟูจิซากิมาใช้เป็นคุณลักษณะสำคัญในการรู้จำวรรณยุกต์ของเสียงพูดภาษาไทยได้ วิทยานิพนธ์ฉบับนี้ได้นำเสนอการนำพารามิเตอร์ของแบบจำลองฟูจิซากิ มาใช้เป็นคุณลักษณะสำคัญในการรู้จำวรรณยุกต์ของเสียงพูดภาษาไทยแบบต่อเนื่องโดยมีการปรับเปลี่ยนชื่อของคำสั่งจากคำสั่งสำเนียง (Accent Command) เป็นคำสั่งวรรณยุกต์ (Tone Command) เนื่องจากเป็นการนำมาใช้กับวรรณยุกต์เป็นสำคัญ

3.1 โครงสร้างระบบรู้จำวรรณยุกต์ของเสียงพูด



รูปที่ 3.1 โครงสร้างระบบรู้จำวรรณยุกต์ของเสียงพูด

โครงสร้างของระบบรู้จำวรรณยุกต์ของเสียงพูดที่นำเสนอแสดงได้ดังรูปที่ 3.1 โดยเสียงพูดจะถูกนำไปคำนวณหาเส้นโค้งความถี่มูลฐาน จากนั้นจึงผ่านการ smoothing และประมาณค่าในช่วง (Interpolation) เพื่อลดสัญญาณที่อาจเกิดจากสัญญาณรบกวน ซึ่งถือเป็นกระบวนการประมวลผลก่อนหน้า จากนั้นทำการสกัดค่าคุณลักษณะสำคัญด้วยการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิ และขั้นตอนสุดท้ายเป็นการจำแนกแบบรูปโดยใช้โครงข่ายประสาทเทียมเป็นตัวจำแนกแบบรูป

3.2 การ smoothing และการประมาณค่าในช่วง

ขั้นตอนการทำ smoothing ประกอบด้วย 2 ขั้นตอนคือการ Neutralization และการทำ Median filtering

3.2.1 การ Neutralization

การ Neutralization คือกำจัดค่าที่แตกต่างจากค่าข้างเคียงมาก โดยทำการตรวจสอบค่าความถี่มูลฐานแต่ละจุดสังเกต และเปรียบเทียบกับค่าข้างเคียง หากค่าความถี่มูลฐานที่จุดสังเกตมีค่าแตกต่างจากค่าข้างเคียงเกินกว่าร้อยละ 5 จะถือว่าค่าความถี่มูลฐานที่จุดสังเกตนั้นไม่มีอยู่จริง เนื่องจากค่าความถี่มูลฐานที่เกิดจากเสียงพูดที่ติดกัน จะไม่มีการเปลี่ยนแปลงอย่างรวดเร็ว

3.2.2 การทำ Median filtering

ขั้นตอนการทำ Median filtering เป็นการทำให้ค่าความถี่มูลฐานที่จุดสังเกตต่างๆ มีความเรียบมากขึ้น โดยการกำหนดค่าความกว้างของจุดสังเกตที่จะใช้ในการหาค่ามัธยฐาน (Median) และให้นำค่าของช่วงความถี่มูลฐานที่กว้างเท่ากับค่าที่กำหนด โดยมีจุดสังเกตเป็นจุดกึ่งกลางมาทำการหาค่ามัธยฐาน และใช้เป็นตัวแทนของจุดสังเกตนั้นต่อไป วิธีนี้จะทำให้เส้นโค้งความถี่มูลฐานที่ได้มีความเรียบมากขึ้น

การประมาณค่าในช่วง จะใช้การประมาณค่าในช่วงแบบเชิงเส้น (Linear Interpolation) เนื่องจากสามารถทำได้ง่ายและรวดเร็ว อีกทั้งช่วงของค่าเส้นโค้งความถี่มูลฐานในส่วนที่มีข้อมูลเกี่ยวกับโทนเสียง จะเป็นช่วงที่เป็นเสียงก้อง และมีค่าความถี่มูลฐาน จึงมีผลของความผิดเพี้ยนจากการประมาณค่าในช่วงไม่มากนัก

3.3 การปรับเส้นโค้งความถี่มูลฐานให้อยู่ในสเกลลอการิทึม

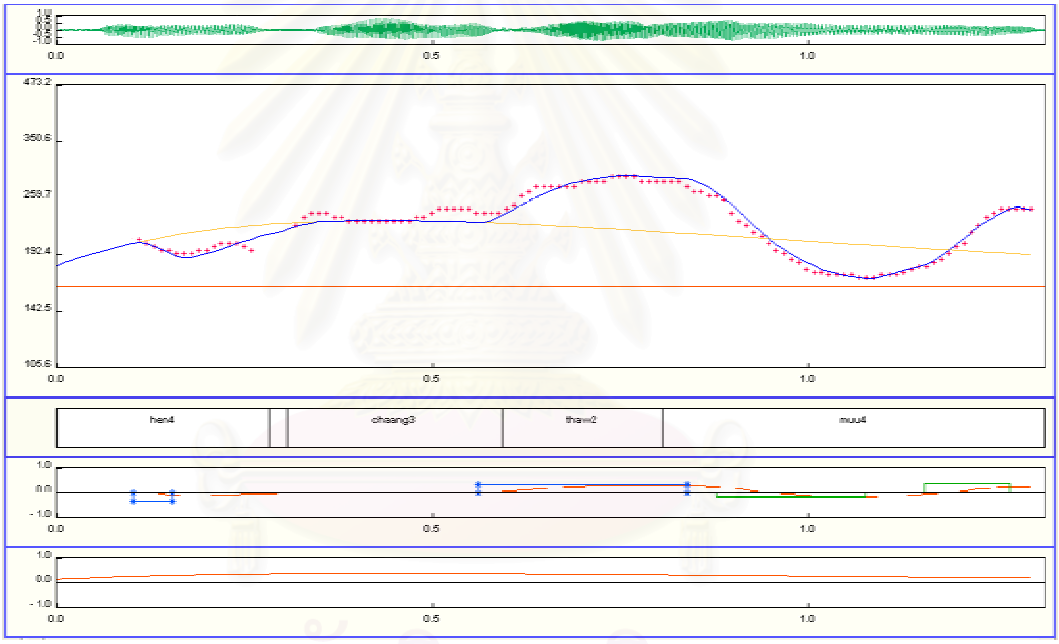
ทำการปรับค่าความถี่มูลฐานทุกจุดให้อยู่ในสเกลลอกริทึมตามลักษณะของแบบจำลองฟูจิซากิ โดยใช้สมการ

$$\log f_0(t) = \begin{cases} \log(f_0(t)), & f_0(t) > 0 \\ 0 & , f_0(t) \le 0 \end{cases} \qquad (3.1)$$

และจากนี้ไปจะพิจารณาเส้นโค้งความถี่มูลฐานในสเกลลอการิทึมโดยตลอด

3.4 การแยกพารามิเตอร์ของแบบจำลองฟูจิซากิ

วิทยานิพนธ์นี้ทำการทดสอบความเป็นไปได้ ในการนำพารามิเตอร์ของแบบจำลองฟูจิซากิมาใช้เป็นค่าคุณลักษณะสำคัญในการรู้จำวรรณยุกต์ภาษาไทย โดยได้ทำการทดลองกับแนวคิดดังกล่าวด้วยการหาพารามิเตอร์ของแบบจำลองทั้งสิ้น 4 วิธีคือ การหาพารามิเตอร์โดยการใช้มนุษย์และการหาพารามิเตอร์แบบอัตโนมัติ 3 วิธีคือ การหาพารามิเตอร์ตามวิธีของ Hansjorg Mixdorff [14] การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ และการหาพารามิเตอร์โดยใช้ขอบเขตพยางค์



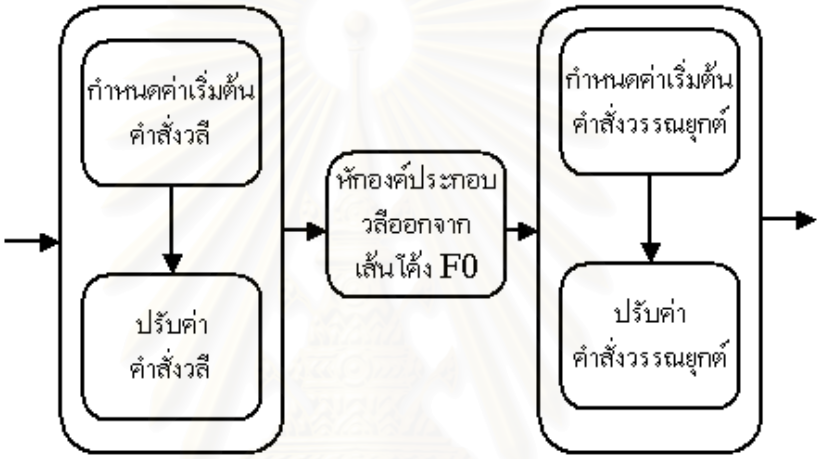
รูปที่ 3.2 ตัวอย่างผลที่ได้โดยใช้วิธีของ Mixdorff ของเสียง “เห็นช้างเท่าหมู”

การหาพารามิเตอร์โดยใช้มนุษย์ อาศัยการตัดสินใจและปรับค่าพารามิเตอร์ต่าง ๆ โดยมนุษย์ ซึ่งถือเป็นพารามิเตอร์ที่ถูกต้องที่สุด การหาพารามิเตอร์ตามแบบของ Hansjorg Mixdorff เป็นวิธีที่ถูกอ้างว่าให้ค่าใกล้เคียงกับการหาพารามิเตอร์โดยใช้มนุษย์มากที่สุดในปัจจุบัน [14] การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ เป็นการหาค่าพารามิเตอร์โดยไม่มีข้อมูลเกี่ยวกับจุดเริ่มต้น และจุดสิ้นสุดของพยางค์ในการกำหนดค่าเริ่มต้นของคำสั่งวลีและคำสั่งวรรณยุกต์ ทำให้การหาค่าพารามิเตอร์ทำได้อย่างอัตโนมัติทุกขั้นตอน ส่วนการหาค่าพารามิเตอร์โดยใช้ขอบเขตพยางค์

เป็นการหาค่าพารามิเตอร์โดยนำข้อมูลของจุดเริ่มต้นและจุดสิ้นสุดของพยางค์ และความรู้เกี่ยวกับรูปแบบที่ควรจะเป็นของพารามิเตอร์ในแต่ละพยางค์ มาใช้ในการกำหนดค่าตั้งต้นของคำสั่งวลีและคำสั่งวรรณยุกต์

โดยในส่วนนี้จะไม่กล่าวถึงการหาพารามิเตอร์ตามวิธีของ Hansjorg Mixdorff เนื่องจากได้กล่าวไว้แล้วในบทที่ 2

3.4.1 การหาพารามิเตอร์โดยใช้มนุษย์

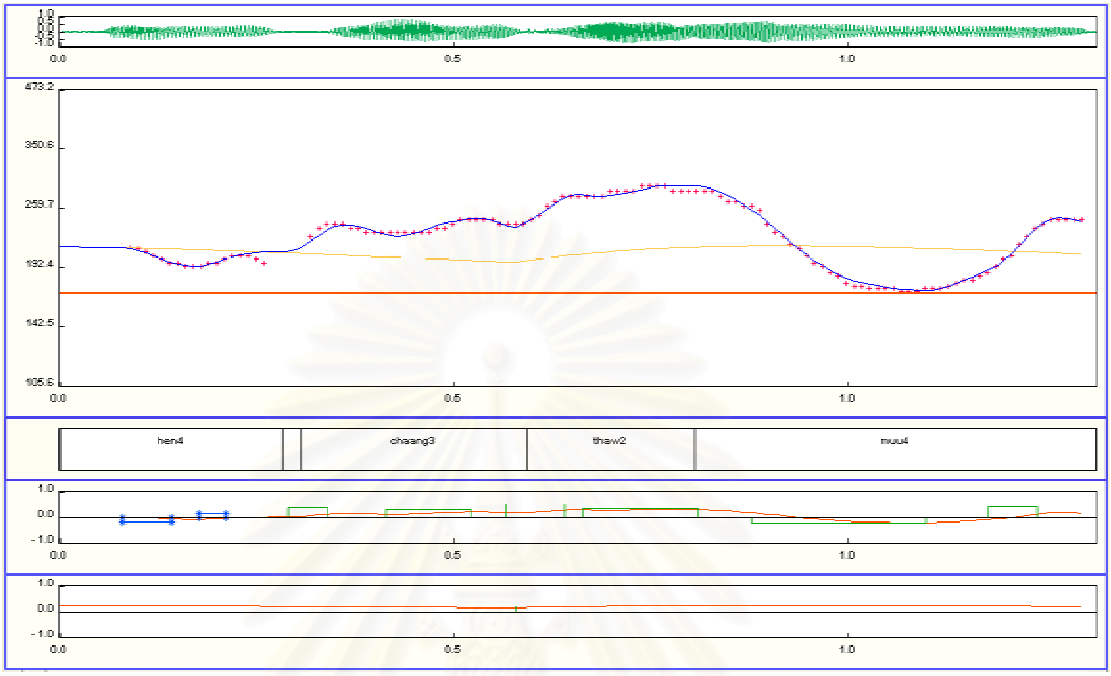


รูปที่ 3.3 ขั้นตอนการหาพารามิเตอร์โดยใช้มนุษย์

การหาพารามิเตอร์โดยอาศัยการพิจารณาและตัดสินใจจากมนุษย์ประกอบด้วยขั้นตอนต่างๆ ดังนี้

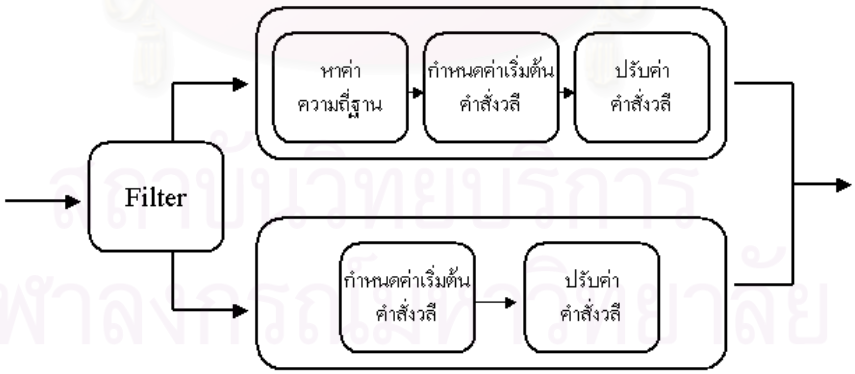
- พิจารณาการเปลี่ยนแปลงโดยรวมของเส้นโค้งความถี่มูลฐานที่ได้จากเสียงพูด และกำหนดค่าเริ่มต้นสำหรับคำสั่งวลี
- ปรับค่าคำสั่งวลีทั้งในส่วนของเวลา และแมกนิจูดให้องค์ประกอบวลีที่สร้างขึ้นมีค่าใกล้เคียงกับการเปลี่ยนแปลงโดยรวมของเส้นโค้งความถี่มูลฐานที่ได้จากเสียงพูด
- นำองค์ประกอบวลีที่ได้จากคำสั่งวลีที่ปรับแล้วไปหักออกจากเส้นโค้งความถี่มูลฐาน
- กำหนดค่าเริ่มต้นสำหรับคำสั่งวรรณยุกต์ โดยพิจารณาจากเส้นโค้งส่วนที่เหลือจากการหักคำสั่งวลีออกจากเส้นโค้งความถี่มูลฐาน
- ปรับค่าคำสั่งวรรณยุกต์ทั้งเวลาเริ่มต้น เวลาสิ้นสุด และแมกนิจูดจนมีเส้นโค้งความถี่มูลฐาน ที่สร้างขึ้นจากการรวมกันขององค์ประกอบวลีและองค์ประกอบวรรณยุกต์ใกล้เคียงกับเส้นโค้งความถี่มูลฐานที่ได้จากเสียงพูดมากที่สุด

ขั้นตอนการหาพามิเตอร์โดยใช้มนุษย์แสดงได้ในรูปที่ 3.3 และผลที่ได้จากการหาค่าพารามิเตอร์ แสดงได้ในรูปที่ 3.4



รูปที่ 3.4 ตัวอย่างผลที่ได้จากการหาพารามิเตอร์โดยใช้มนุษย์ของเสียง “เห็นช้างเท่าหมู”

3.4.2 การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์



รูปที่ 3.5 ขั้นตอนการหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์

การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ เป็นการหาค่าพารามิเตอร์โดยอัตโนมัติ ซึ่งข้อมูลที่ใช้ในการหาพารามิเตอร์มีเพียงเส้นโค้งความถี่มูลฐานเพียงอย่างเดียว โดยเป็นการดัดแปลงวิธีจากการหาค่าพารามิเตอร์ตามวิธีของ Hansjorg Mixdorff [14] ซึ่งข้อแตกต่างหลักคือการไม่ใช้

อัลกอริทึม MOMEL เนื่องจากพบว่า การนำอัลกอริทึม MOMEL มาใช้กับเสียงในภาษาไทย จะทำให้การเปลี่ยนแปลงของเส้นโค้งความถี่ถูกทำให้เรียบจนข้อมูลเกี่ยวกับวรรณยุกต์หายไปมาก

การหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์แสดงได้ดังรูปที่ 3.7 โดยมีขั้นตอนดังนี้

3.4.2.1 แยกองค์ประกอบโดยใช้การกรอง

การแยกองค์ประกอบโดยใช้การกรอง คือการนำเส้นโค้งความถี่มูลฐานที่ได้จากเสียงพูดและผ่านการทำให้เรียบด้วยการทำ Neutralization และทำ Median Filtering มาผ่านวงจรกรองที่มีความถี่ตัดที่ 0.5 เฮิร์ตซ ซึ่งองค์ประกอบที่มีความถี่ต่ำกว่าความถี่ตัด จะถูกใช้เป็นตัวแทนขององค์ประกอบวลี เรียกว่าเส้นโค้งความถี่ต่ำ (LFC-Low Frequency Contour) และ องค์ประกอบที่มีความถี่สูงกว่าจะถูกใช้เป็นตัวแทนขององค์ประกอบวรรณยุกต์ เรียกว่าเส้นโค้งความถี่สูง (HFC-High Frequency Contour)

3.4.2.2 หาค่าความถี่ฐาน

จากเส้นโค้งความถี่ต่ำที่ได้ ทำการหาค่าต่ำสุดของค่าทั้งหมด และใช้ค่าดังกล่าวเป็นค่าความถี่ฐาน

$$\log F_b = \min(LFC) \tag{3.2}$$

3.4.2.3 กำหนดค่าตั้งต้นสำหรับคำสั่งวลี

ทำการหาค่าต่ำสุดเฉพาะแห่งและสูงสุดเฉพาะแห่งขององค์ประกอบความถี่ต่ำ และกำหนดให้เป็นค่าเวลาเริ่มต้นของคำสั่งวลีในทุกจุดต่ำสุด และจากสมการผลตอบสนองของกลไกควบคุมวลี

$$G_p(t) = \alpha^2 t \cdot \exp(-\alpha t) \tag{3.3}$$

จะได้

$$\frac{d}{dt} G_p(t) = \alpha^2 \cdot \exp(-\alpha t) - \alpha^3 t \cdot \exp(-\alpha t) \tag{3.4}$$

พิจารณาจุดที่ผลตอบลัพธ์สูงสุด t^* จะได้ว่า

$$\frac{d}{dt}G_p(t) = 0 \tag{3.5}$$

$$t^* = 1/\alpha \tag{3.6}$$

$$G_{p\_\max} = \alpha \cdot \exp(-1) \tag{3.7}$$

ดังนั้น ที่จุดสูงสุดเฉพาะแห่งขององค์ประกอบความถี่ต่ำที่ i มีค่าเป็น $f_{\max\_i}$ จะได้ว่า

$$f_{\max\_i} = A_{pi} \cdot G_{p\_\max} \tag{3.8}$$

$$A_{pi} = f_{\max\_i} \cdot \exp(1)/\alpha \tag{3.9}$$

และใช้ค่า A_{pi} ดังกล่าวเป็นค่าตั้งต้นของแมกนิจูดสำหรับคำสั่งวลี

3.4.2.4 ปรับค่าพารามิเตอร์สำหรับคำสั่งวลี

หาค่าความผิดพลาดกำลังสองเฉลี่ย (Mean Square Error) ขององค์ประกอบวลีที่ได้จากค่าเริ่มต้นของคำสั่งวลีเทียบกับองค์ประกอบความถี่ต่ำ

เพิ่มหรือลดค่าพารามิเตอร์ครั้งละขั้นจนกว่าค่าความผิดพลาดกำลังสองเฉลี่ยจะไม่ลดลง

3.4.2.5 กำหนดค่าเริ่มต้นสำหรับคำสั่งวรรณยุกต์

หาจุดตัดข้ามศูนย์ (Zero crossing) ค่าสูงสุดเฉพาะแห่ง (Local maximum) ที่เป็นบวก และค่าต่ำสุดเฉพาะแห่ง (Local minimum) ที่เป็นลบ กำหนดให้เวลาที่เกิดจุดตัดข้ามศูนย์เป็นเวลาเริ่มต้น (Onset) ของคำสั่งวรรณยุกต์ และให้จุดที่มีค่าสูงสุดและต่ำสุดเฉพาะแห่งเป็นเวลาสิ้นสุดของคำสั่งวรรณยุกต์

โดยการประมาณแล้ว จุดสูงสุดหรือต่ำสุดขององค์ประกอบวรรณยุกต์จะอยู่ที่เวลาสิ้นสุดของคำสั่งวรรณยุกต์ หรือ $t = T_{2j}$ และจากสมการผลตอบสนองของกลไกควบคุมวรรณยุกต์

$$G_a(t) = 1 - (1 + \beta t) \cdot \exp(-\beta t) \tag{3.10}$$

และองค์ประกอบวรรณยุกต์

$$A_{aj}\left\{G_a(t - T_{1j}) - G_a(t - T_{2j})\right\} \tag{3.11}$$

ที่จุดสูงสุดหรือต่ำสุดเฉพาะแห่งที่ j ขององค์ประกอบความถี่สูงมีค่าเป็น $f_{\max\_j}$ และ $t = T_{2j}$ จะได้ว่า

$$f_{\max\_j} = A_{aj}\left\{G_a(T_{2j} - T_{1j}) - G_a(T_{2j} - T_{2j})\right\} \quad (3.12)$$

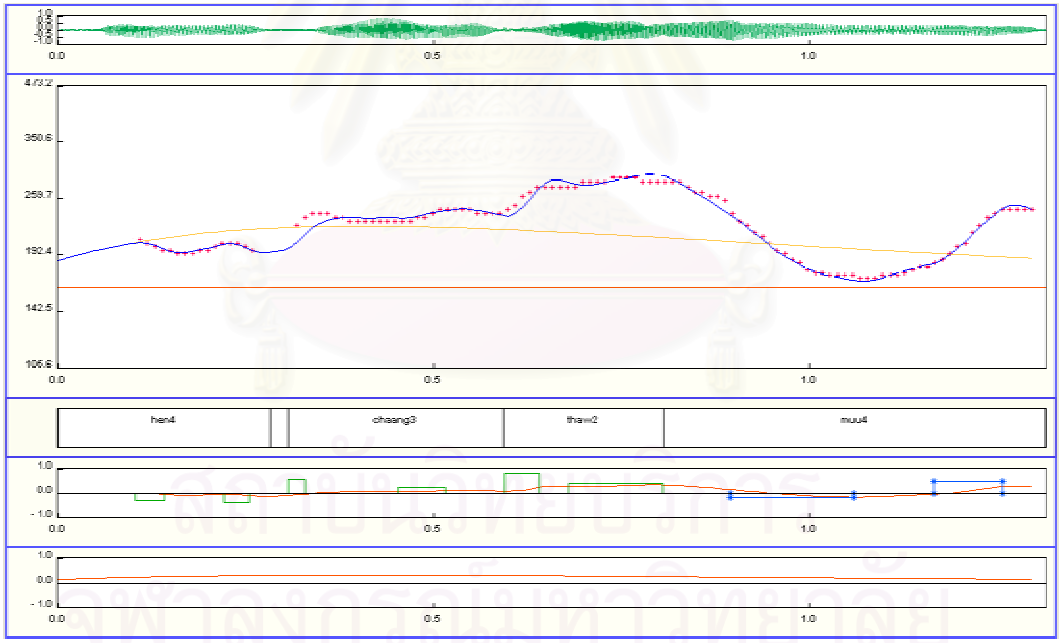
$$f_{\max\_j} = A_{aj}(1-(1+\beta(T_{2j} - T_{1j}))\cdot\exp(-\beta(T_{2j} - T_{1j}))) \quad (3.13)$$

ให้ผลต่างระหว่างเวลาของจุดสูงสุดหรือต่ำสุดเฉพาะแห่งกับเวลาที่เกิดจุดตัดข้ามศูนย์ก่อนหน้าเป็น Δt เป็นค่าแทน $T_{2j} - T_{1j}$ จะได้ว่า

$$f_{\max\_j} = A_{aj}(1-(1+\Delta t\cdot\beta)\cdot\exp(-\Delta t\cdot\beta)) \quad (3.14)$$

$$A_{aj} = f_{\max\_j}/(1-(1+\Delta t\cdot\beta)\cdot\exp(-\Delta t\cdot\beta)) \quad (3.15)$$

และใช้ค่า A_{aj} แทนค่าเริ่มต้นของแมกนิจูดสำหรับคำสั่งวรรณยุกต์



รูปที่ 3.6 ตัวอย่างผลที่ได้จากการหาพารามิเตอร์แบบไม่ใช้ขอบเขตพยางค์ของเสียง “เห็นช้างเท่าหมู”

3.4.2.6 ปรับค่าพารามิเตอร์สำหรับคำสั่งวรรณยุกต์

หาค่าความผิดพลาดกำลังสองเฉลี่ย (Mean Square Error) ขององค์ประกอบวรรณยุกต์ที่ได้จากค่าเริ่มต้นของคำสั่งวรรณยุกต์เทียบกับองค์ประกอบความถี่สูง

เพิ่มหรือลดค่าพารามิเตอร์ครั้งละขั้นจนกว่าค่าความผิดพลาดกำลังสองเฉลี่ยจะไม่ลดลง

3.4.3 การหาพารามิเตอร์โดยใช้ขอบเขตพยางค์

การหาพารามิเตอร์โดยใช้ขอบเขตพยางค์ เป็นการนำข้อมูลเกี่ยวกับขอบเขตของพยางค์เข้ามาประกอบในการแยกพารามิเตอร์ โดยอาศัยความรู้เกี่ยวเนื่องกับพารามิเตอร์ในแต่ละพยางค์ของเสียงพูดวรรณยุกต์ต่าง ๆ กันมาใช้ในขั้นตอนการกำหนดค่าเริ่มต้นให้กับคำสั่งวรรณยุกต์ซึ่งมีขั้นตอนอื่นเหมือนกับการหาพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ การกำหนดค่าเริ่มต้นให้กับคำสั่งวรรณยุกต์มีขั้นตอนดังนี้

3.4.3.1 กำหนดค่า $f_{\max\_j}$

แบ่งช่วงที่พิจารณาในแต่ละพยางค์ออกเป็นสองส่วนตามเวลา โดยทำการหาค่าเฉลี่ยขององค์ประกอบความถี่สูงในแต่ละช่วงที่พิจารณา หากมีค่าเฉลี่ยมากกว่า 0 ให้กำหนดค่า $f_{\max\_j}$ เป็นค่าสูงสุดในช่วงที่พิจารณา หากมีค่าเฉลี่ยน้อยกว่า 0 ให้กำหนดค่า $f_{\max\_j}$ เป็นค่าต่ำสุดในช่วงที่พิจารณา

3.4.3.2 กำหนดเวลาเริ่มต้นและเวลาสิ้นสุดของคำสั่งวรรณยุกต์

หากค่าเฉลี่ยขององค์ประกอบความถี่สูงในช่วงที่พิจารณามีค่ามากกว่า 0 กำหนดให้ค่าเวลาเริ่มต้น T_{1j} เป็นเวลาที่จุดที่มีค่าต่ำสุด และมีค่าเวลา t น้อยกว่าจุดที่เกิด $f_{\max\_j}$ และให้เวลาสิ้นสุดของคำสั่งวรรณยุกต์ T_{2j} คือจุดที่เกิด $f_{\max\_j}$

หากค่าเฉลี่ยขององค์ประกอบความถี่สูงในช่วงที่พิจารณามีค่าน้อยกว่า 0 กำหนดให้เวลาเริ่มต้น T_{1j} เป็นเวลาที่จุดที่มีค่าสูงสุดและมีค่าเวลา t น้อยกว่าจุดที่เกิด $f_{\max\_j}$ และให้เวลาสิ้นสุดของคำสั่งวรรณยุกต์ T_{2j} คือจุดที่เกิด $f_{\max\_j}$

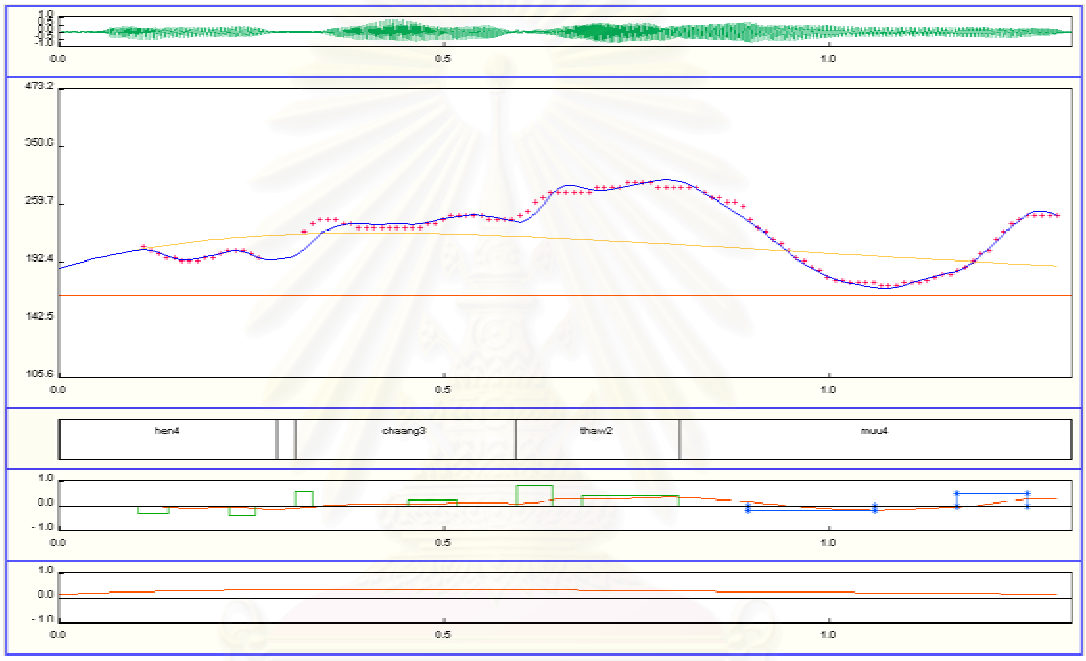
3.4.3.3 กำหนดค่าเริ่มต้นของแมกนิจูดของคำสั่งวรรณยุกต์

ให้

$$\Delta t = T_{2j} - T_{1j} \tag{3.16}$$

กำหนดให้ค่าตั้งต้นของแมกนิจูดของคำสั่งวรรณยุกต์ A_{aj} เป็น

$$A_{aj} = f_{\max\_j} / (1 - (1 + \Delta t \cdot \beta) \cdot \exp(-\Delta t \cdot \beta)) \tag{3.17}$$



รูปที่ 3.7 ตัวอย่างผลที่ได้จากการหาพารามิเตอร์แบบใช้ขอบเขตพยางค์ของเสียง “เห็นช้างเท่าหมู”

3.5 แยกพารามิเตอร์ตามพยางค์ (Syllable segmentation)

พารามิเตอร์ของแบบจำลองฟูจิซากิที่ได้จากกระบวนการในขั้นตอน 3.4 จะเป็นพารามิเตอร์สำหรับทั้งประโยคของเสียงพูด จึงต้องทำการแยกพารามิเตอร์สำหรับแต่ละพยางค์ เพื่อใช้ในขั้นตอนการจำแนกแบบรูปต่อไป

การแยกพารามิเตอร์ตามพยางค์คือการพิจารณาเฉพาะพารามิเตอร์สำหรับพยางค์นั้น โดยนำข้อมูลของขอบเขตพยางค์มาใช้ในการแบ่งพารามิเตอร์ โดยคำสั่งวรรณยุกต์ที่อยู่ภายใน

ขอบเขตพยางค์ หรือมีบางส่วนอยู่ในขอบเขตพยางค์จะถูกนำมาพิจารณาว่าเป็นคำสั่งวรรณยุกต์ของพยางค์นั้น ๆ

เวลาเริ่มต้นและเวลาสิ้นสุดของคำสั่งวรรณยุกต์จะถูกนอร์แมลไลซ์ด้วยขอบเขตพยางค์ โดยจุดเริ่มต้นของพยางค์จะมีค่าเวลาเป็น 0 และจุดสิ้นสุดของพยางค์จะมีค่าเวลาเป็น 1 ค่าเวลาเริ่มต้นของคำสั่งวรรณยุกต์ที่ถูกนอร์แมลไลซ์แล้วมีค่าต่ำกว่า 0 จะถูกให้ค่าเป็น 0 และค่าเวลาสิ้นสุดของคำสั่งวรรณยุกต์ที่ถูกนอร์แมลไลซ์แล้วมีค่ามากกว่า 1 จะถูกให้ค่าเป็น 1 ค่า Magnitude ของคำสั่งวรรณยุกต์จะถูกนอร์แมลไลซ์ให้มีค่าอยู่ระหว่าง –1 ถึง 1

3.6 **ตัวจำแนกแบบรูป** (Pattern Classifier)

สำหรับวิทยานิพนธ์นี้เลือกใช้ตัวจำแนกแบบรูปเป็นโครงข่ายประสาทเทียม เนื่องจากมีรูปแบบที่เหมาะสมกับค่าคุณลักษณะที่ได้จากการแยกพารามิเตอร์ และสามารถทำงานได้รวดเร็ว

3.6.1 **โครงข่ายประสาทเทียมสำหรับการแยกพารามิเตอร์แบบไม่ใช้ขอบเขตพยางค์**

การแยกพารามิเตอร์ของแบบจำลองฟูจิซากิโดยไม่ใช้ขอบเขตพยางค์ จะทำให้ได้จำนวนคำสั่งวรรณยุกต์ 0 – 3 ชุดต่อพยางค์ ในแต่ละคำสั่งวรรณยุกต์ประกอบด้วยพารามิเตอร์ที่จะนำมาใช้เป็นค่าคุณลักษณะสำคัญ 3 ตัวคือเวลาเริ่มต้นของคำสั่งวรรณยุกต์ เวลาสิ้นสุดของคำสั่งวรรณยุกต์ และแมกนิจูดของคำสั่งวรรณยุกต์ ดังนั้นโครงข่ายประสาทเทียมสำหรับใช้กับพารามิเตอร์ที่แยกแบบไม่ใช้ขอบเขตพยางค์จึงมีอินพุตโนดจำนวน 9 โนด และมีค่าเป็น 0 ในกรณีที่มีคำสั่งวรรณยุกต์ไม่ครบทั้ง 3 ชุด โครงข่ายประสาทเทียมมีเอาต์พุตโนดจำนวน 5 โนดแทนวรรณยุกต์ทั้ง 5 และมีโนดซ่อนจำนวน 40 โนด ซึ่งเป็นจำนวนโนดที่ให้ค่าการรู้จำสูงที่สุดในการทดลอง

3.6.2 **โครงข่ายประสาทเทียมสำหรับการแยกพารามิเตอร์แบบใช้ขอบเขตพยางค์**

การใช้ขอบเขตพยางค์ในการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิจะทำให้ได้คำสั่งวรรณยุกต์ 2 ชุดต่อพยางค์เสมอ ดังนั้นโครงข่ายประสาทเทียมจะมีอินพุตโนดจำนวน 6 โนด และมีเอาต์พุตโนดจำนวน 5 โนดตามจำนวนวรรณยุกต์ทั้ง 5 แบบ และมีโนดซ่อนจำนวน 40 โนด ซึ่งเป็นจำนวนโนดที่ให้ค่าการรู้จำที่สูงที่สุดในการทดลอง

บทที่ 4
การทดสอบ

ในบทนี้จะกล่าวถึงการนำพารามิเตอร์ที่ได้จากการแยกโดยวิธีต่าง ๆ ทั้ง 4 วิธี ได้แก่

- การแยกพารามิเตอร์โดยใช้มนุษย์
- การแยกพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์
- การแยกพารามิเตอร์โดยใช้ขอบเขตพยางค์
- การแยกพารามิเตอร์ตามวิธีของ Hansjorg Mixdorff [14]

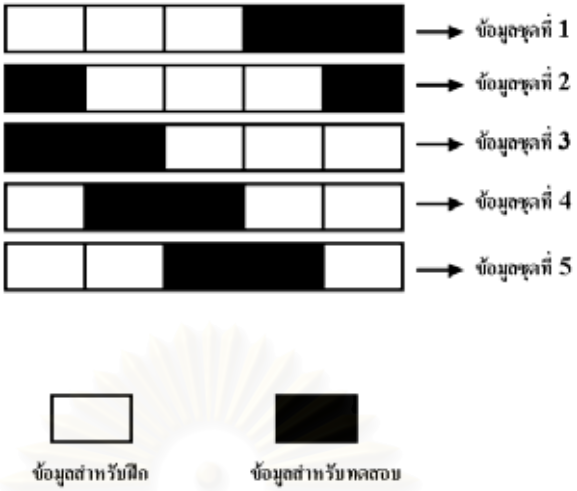
และทำการทดสอบผลที่เกิดจากการเปลี่ยนแปลงความถี่ตัด ในขั้นตอนการแยกองค์ประกอบโดยใช้วงจรกรอง ที่มีผลต่อการรู้จำ

4.1 ข้อมูลเสียงที่ใช้ในการทดสอบ

ข้อมูลเสียงที่ใช้ในการทดสอบสำหรับวิทยานิพนธ์ เป็นข้อมูลเสียง Thai Proverb Corpus [16] ที่ได้รับความอนุเคราะห์จาก อ.ดร.ณัฐกร ทับทอง ประกอบด้วยคำพังเพย 30 ประโยค ความยาว 4 พยางค์ 10 ประโยค ความยาว 5 พยางค์ 10 ประโยคและความยาว 6 พยางค์ 10 ประโยค ซึ่งเป็นเสียงพูดที่ได้จากการบันทึกจากผู้พูด 40 คน เป็นชาย 20 คนหญิง 20 คน อายุ 17 ถึง 29 ปี โดยมีค่าเฉลี่ยอายุเป็น 20.78 ปี และส่วนเบี่ยงเบนมาตรฐานเป็น 2.35 ปี ผู้พูดทุกคนพูดคำพังเพยทั้งหมดหนึ่งครั้งด้วยความเร็วระดับการสนทนา ดังนั้นชุดข้อมูลนี้จึงประกอบด้วยเสียงพูดทั้งหมด 1,200 ประโยค

4.2 วิธีการทดสอบ

การทดสอบจะกระทำทั้งสิ้น 5 ชุด โดยจะใช้วิธีแบ่งข้อมูลออกเป็น 5 ส่วนเท่าๆ กัน จากนั้นใช้ข้อมูล 3 ส่วนในการฝึกโครงข่ายประสาทเทียม และใช้ 2 ส่วนที่เหลือในการทดสอบผลรวมนับเป็น 1 ชุด จากนั้นจะทำการสลับชุดข้อมูลสำหรับฝึกโครงข่ายและชุดข้อมูลสำหรับทดสอบเป็นข้อมูลชุดใหม่จนครบทั้งสิ้น 5 ชุด เช่นชุดที่ 1 ใช้ข้อมูลส่วนที่ 1, 2 และ 3 ในการฝึก และใช้ข้อมูลส่วนที่ 4 และ 5 ในการทดสอบ ชุดที่ 2 ใช้ข้อมูลส่วนที่ 2, 3 และ 4 ในการฝึก และใช้ข้อมูลส่วนที่ 5 และ 1 ในการทดสอบเป็นต้น ทั้งนี้ เพื่อให้มีความน่าเชื่อถือในการทดสอบมากขึ้น โดยผลการทดสอบจะไม่โน้มเอียงไปตามชุดของข้อมูลชุดใดชุดหนึ่ง ลักษณะการเวียนสลับข้อมูลแสดงได้ในรูปที่ 4.1



รูปที่ 4.1 ลักษณะการเวียนข้อมูลสำหรับทดสอบ

ข้อมูลแต่ละชุดจะถูกทำการฝึกและทดสอบ 5 ครั้งและนำค่าที่ได้มาทำการเฉลี่ยผลลัพธ์และผลของข้อมูลแต่ละชุดจะถูกนำมาเฉลี่ยผลลัพธ์อีกครั้ง

4.3 ผลการทดสอบ

หัวข้อนี้เป็นการแสดงผลการรู้จำ ของการนำพารามิเตอร์ของแบบจำลองฟูจิซากิมาใช้ในการรู้จำวรรณยุกต์ของเสียงพูด โดยกระบวนการแยกความถี่มูลฐานทำโดยวิธีอัตสหสัมพันธ์ (Autocorrelation) โดยใช้ความกว้างของหน้าต่างเป็น 40 มิลลิวินาที และมีการเลื่อนของหน้าต่างเป็น 10 มิลลิวินาที ผลการทดสอบแสดงได้ด้วย confusion matrix

4.3.1 ผลการทดสอบด้วยพารามิเตอร์ที่ได้จากการแยกโดยใช้มนุษย์

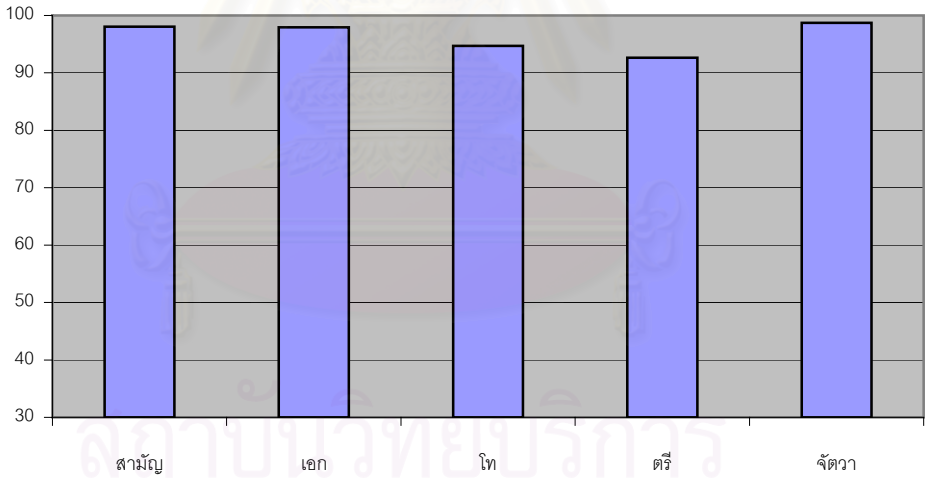
การแยกพารามิเตอร์โดยใช้มนุษย์ ทำโดยบุคคลผู้มีประสบการณ์ในการพิจารณา และแยกพารามิเตอร์ของแบบจำลองฟูจิซากิ 1 คน ผลการทดสอบการรู้จำโดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยใช้มนุษย์ ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 98.06 และมีความผิดพลาดเป็นเสียงเอกและเสียงตรีร้อยละ 0.69 และ 1.25 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 97.94 มีความผิดพลาดเป็นเสียงสามัญ เสียงตรี และเสียงจัตวาร้อยละ 0.72 0.10 และ 1.24 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 94.70 มีความผิดพลาดเป็นเสียงตรีร้อยละ 5.3 ผลความถูกต้องของการรู้

จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 92.61 มีความผิดพลาดเป็นเสียงสามัญและเสียงโทร้อยละ 1.32 และ 6.07 ตามลำดับ และมีผลความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 98.73 มีความผิดพลาดเป็นเสียงเอกร้อยละ 1.27

ตารางที่ 4.1 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้มนุษย์

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**98.06**	0.69	0.00	1.25	0.00
เอก	0.72	**97.94**	0.00	0.10	1.24
โท	0.00	0.00	**94.70**	5.30	0.00
ตรี	1.32	0.00	6.07	**92.61**	0.00
จัตวา	0.00	1.27	0.00	0.00	**98.73**

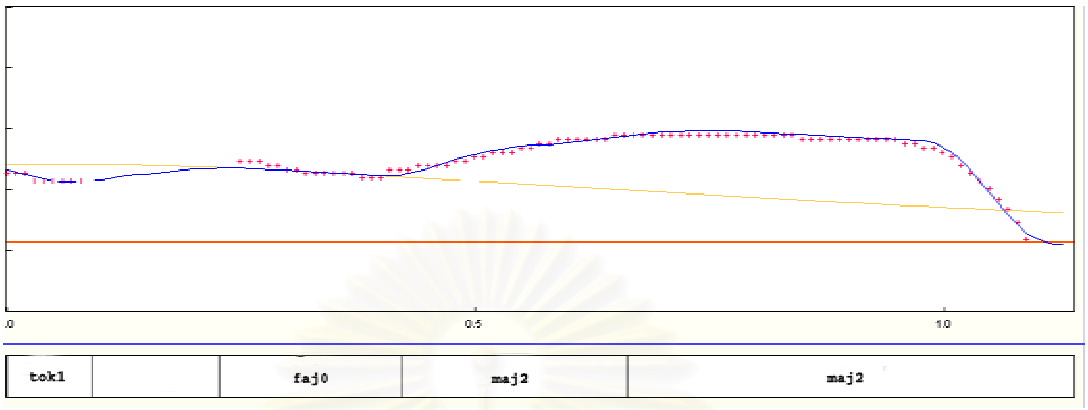
อัตราการรู้จำเฉลี่ยร้อยละ 96.35



รูปที่ 4.2 ความถูกต้องของการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้มนุษย์

พารามิเตอร์ของแบบจำลองฟูจิซากิที่แยกโดยใช้มนุษย์ให้ผลการรู้จำถูกต้องเฉลี่ยร้อยละ 96.35 โดยความผิดพลาดสูงสุดเกิดจากการรู้จำผิดพลาดระหว่างเสียงวรรณยุกต์โท และเสียงวรรณยุกต์ตรีที่ร้อยละ 5.30 และ 6.07 ตามลำดับ สาเหตุเนื่องมาจากการที่ไม่สามารถตัดผลของการควบรวมของเสียงได้หมด ดังเช่นในประโยค “ตกไฟไม่ไหม้” ที่มีพยางค์ที่ 3 และพยางค์ที่ 4 ที่

ควรจะมีลักษณะของเส้นโค้งความถี่มูลฐานสูงขึ้น-ลดลง-สูงขึ้น-ลดลง แต่ด้วยการควบรวมของเสียงทำให้เส้นโค้งความถี่มูลฐานเป็นสูงขึ้น-สูงขึ้น-สูงขึ้น-ลดลง ดังแสดงในรูปที่ 4.3



รูปที่ 4.3 เส้นโค้งความถี่มูลฐานของประโยค “ตกไฟไม่ไหม้”

อย่างไรก็ดี การแยกพารามิเตอร์โดยใช้มนุษย์นั้น ค่าพารามิเตอร์ที่ได้จะมีความโน้มเอียงไปยังค่าที่ถูกคาดหมายไว้ล่วงหน้าของผู้แยกพารามิเตอร์ อีกทั้งการแยกพารามิเตอร์โดยบุคคลเดียวกันในแต่ละครั้ง อาจได้ค่าพารามิเตอร์ที่แตกต่างจากเดิมด้วย

4.3.2 ผลการทดสอบด้วยพารามิเตอร์ที่ได้จากการแยกโดยไม่ใช้ขอบเขตพยางค์

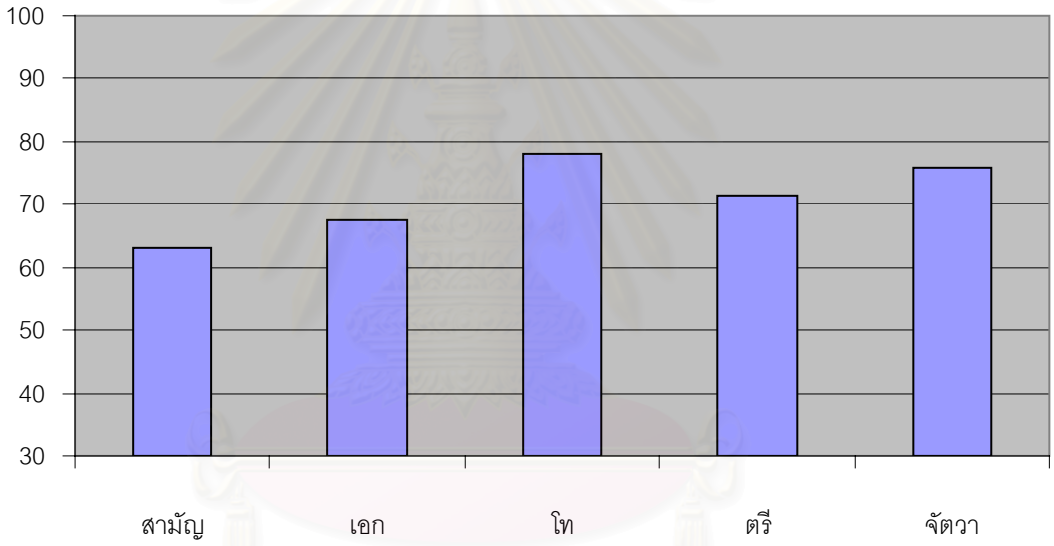
ตารางที่ 4.2 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยไม่ใช้ขอบเขตพยางค์

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**63.17**	18.59	5.98	9.57	2.69
เอก	17.29	**67.50**	1.50	1.43	12.28
โท	10.26	1.65	**78.14**	9.32	0.62
ตรี	10.86	6.61	7.40	**71.27**	3.85
จัตวา	3.68	15.89	0.66	4.07	**75.71**

อัตราการรู้จำเฉลี่ยร้อยละ 70.27

ผลการทดสอบการรู้จำโดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยไม่ใช้ขอบเขตพยางค์ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 63.17 และมีความผิดพลาด

เป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 18.59, 5.98, 9.57 และ 2.69 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 67.50 มีความผิดพลาดเป็นเสียงสามัญ เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 17.29, 1.50, 1.43 และ 12.28 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 78.14 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 10.26, 1.65, 9.32 และ 0.62 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 71.27 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 10.86, 6.61, 7.40 และ 3.85 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 75.71 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงตรีร้อยละ 3.68, 15.89, 0.66 และ 4.07 ตามลำดับ โดยมีอัตราการรู้จำเฉลี่ยเป็นร้อยละ 70.27



รูปที่ 4.4 ความถูกต้องของการรู้จำด้วยพารามิเตอร์ที่แยกโดยไม่ใช้ขอบเขตพยางค์

การรู้จำโดยใช้พารามิเตอร์ที่ได้จากการแยกแบบอัตโนมัติ โดยไม่ใช้ขอบเขตพยางค์ มีความผิดพลาดสูงสุดในการรู้จำเสียงวรรณยุกต์สามัญและเอก โดยเป็นการรู้จำสลับกันมากที่สุด ทั้งนี้เนื่องมาจากลักษณะของเส้นโค้งความถี่มูลฐานที่ใกล้เคียงกัน การแยกเสียงวรรณยุกต์ของทั้งสองเสียงนี้ออกจากกันจึงแยกโดยอาศัยความแตกต่างของระดับของความถี่ ซึ่งขึ้นอยู่กับการแยกองค์ประกอบวลีเป็นอย่างมาก นั่นคือ เกิดจากกระบวนการแยกองค์ประกอบวลีออกจากองค์ประกอบวรรณยุกต์ที่ไม่เหมาะสม ความผิดพลาดที่เกิดขึ้นสูงอีกแห่งคือการรู้จำผิดพลาดจากวรรณยุกต์จัตวาไปเป็นวรรณยุกต์เอก ซึ่งมีความผิดพลาดถึงร้อยละ 15.89 โดยสาเหตุเกิดจากการแยกองค์ประกอบวลีที่ไม่เหมาะสมเช่นกัน รวมทั้งเกิดจากการที่ไม่สามารถลดผลของการควบร่วม

ของเสียงได้อย่างเพียงพออีกด้วย ทำให้ช่วงท้ายของพยางค์ที่ควรจะมีค่าสั่งวรรณยุกต์ค่าเป็นบวกถูกควบร่วมกับเสียงที่ตามมา ทำให้กลับมีค่าสั่งวรรณยุกต์ที่มีค่าเป็นลบ

จาก confusion matrix แสดงให้เห็นว่า การรู้จำผิดพลาด จะเป็นการรู้จำผิดเป็นวรรณยุกต์เสียงสามัญมากที่สุด ยกเว้นการรู้จำเสียงวรรณยุกต์จัตวา เมื่อพิจารณาจากเส้นโค้งความถี่สูงในแต่ละช่วงพยางค์ ที่ได้จากการสร้างด้วยพารามิเตอร์ที่แยกได้ พบว่าในส่วนของเสียงสามัญซึ่งควรมีค่าใกล้เคียงค่า 0 ตลอดช่วงพยางค์ กลับมีเส้นโค้งความถี่สูงที่มีลักษณะหลายรูปแบบ ซึ่งคล้ายกับเส้นโค้งความถี่สูงในส่วนของเสียงวรรณยุกต์อื่น ก่อให้เกิดความสับสนในการจำแนกวรรณยุกต์ออกจากกัน

สาเหตุจากการเกิดปรากฏการณ์ข้างต้น น่าจะเกิดจากการแยกองค์ประกอบวลีออกจากองค์ประกอบวรรณยุกต์ที่ไม่เหมาะสม หรือกระบวนการในการกรองเพื่อแยกเส้นโค้งความถี่มูลฐานมีลักษณะที่ไม่เหมาะสมนั่นเอง จึงได้ทำการทดลองปรับค่าความถี่ตัดของวงจรกรองเพื่อทดสอบหาความถี่ตัดที่เหมาะสม

4.3.2.1 **ผลทดสอบการเปลี่ยนแปลงค่าความถี่ตัด**

หัวข้อนี้เป็นการแสดงผลทดสอบ การเปลี่ยนแปลงความถี่ผ่านของการกรองเพื่อแยกสัญญาณเป็นองค์ประกอบความถี่ต่ำและองค์ประกอบความถี่สูง ในการแยกพารามิเตอร์แบบไม่ใช้ขอบเขตพยางค์ที่มีผลต่ออัตราการรู้จำ โดยได้มีการทดสอบที่ความถี่ 0.5, 1, 1.5 และ 2 เฮิร์ตซ

4.3.2.1.1 **ผลการรู้จำที่ความถี่ตัด 0.5 เฮิร์ตซ**

ตารางที่ 4.3 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้ความถี่ตัด 0.5 เฮิร์ตซ

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**63.17**	18.59	5.98	9.57	2.69
เอก	17.29	**67.50**	1.50	1.43	12.28
โท	10.26	1.65	**78.14**	9.32	0.62
ตรี	10.86	6.61	7.40	**71.27**	3.85
จัตวา	3.68	15.89	0.66	4.07	**75.71**

อัตราการรู้จำเฉลี่ยร้อยละ 70.27

ผลการทดสอบการรู้จำโดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยไม่ใช้ขอบเขตพยางค์และใช้ความถี่ตัดในขั้นตอนการแยกองค์ประกอบความถี่สูงและองค์ประกอบความถี่ต่ำที่ 0.5 เฮิร์ตซ ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 63.17 และมีความผิดพลาดเป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 18.59, 5.98, 9.57 และ 2.69 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 67.50 มีความผิดพลาดเป็นเสียงสามัญ เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 17.29, 1.50, 1.43 และ 12.28 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 78.14 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 10.26, 1.65, 9.32 และ 0.62 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 71.27 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 10.86, 6.61, 7.40 และ 3.85 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 75.71 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงตรีร้อยละ 3.68, 15.89, 0.66 และ 4.07 ตามลำดับ โดยมีอัตราการรู้จำเฉลี่ยเป็นร้อยละ 70.27

4.3.2.1.2 **ผลการรู้จำที่ความถี่ตัด 1 เฮิร์ตซ**

ตารางที่ 4.4 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้ความถี่ตัด 1 เฮิร์ตซ

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**56.72**	28.39	7.39	5.08	2.42
เอก	17.58	**71.29**	3.81	2.52	4.79
โท	4.69	4.25	**88.39**	2.06	0.61
ตรี	13.05	9.00	27.65	**44.80**	5.50
จัตวา	5.70	23.65	2.30	1.85	**66.50**

อัตราการรู้จำเฉลี่ยร้อยละ 67.95

ผลการทดสอบการรู้จำโดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยไม่ใช้ขอบเขตพยางค์และใช้ความถี่ตัดในขั้นตอนการแยกองค์ประกอบความถี่สูงและองค์ประกอบความถี่ต่ำที่ 1 เฮิร์ตซ ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 56.72 และมีความผิดพลาดเป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 28.39, 7.39, 5.08 และ 2.42 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 71.29 มีความผิดพลาดเป็นเสียงสามัญ

เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 17.58, 3.81, 2.52 และ 4.79 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 88.39 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 4.69, 4.25, 2.06 และ 0.61 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 44.80 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 13.05, 9.00, 27.65 และ 5.50 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 66.50 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงตรีร้อยละ 5.70, 23.65, 2.30 และ 1.85 ตามลำดับ โดยมีอัตราการรู้จำเฉลี่ยเป็นร้อยละ 67.95

4.3.2.1.3 **ผลการรู้จำที่ความถี่ตัด 1.5 เฮิร์ตซ**

ตารางที่ 4.5 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้ความถี่ตัด 1.5 เฮิร์ตซ

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**51.00**	26.94	12.67	5.08	4.31
เอก	13.52	**74.90**	5.83	2.17	3.58
โท	8.50	8.69	**79.86**	2.00	0.94
ตรี	21.55	13.20	16.30	**40.60**	8.35
จัตวา	7.75	20.45	2.40	4.20	**65.20**

อัตราการรู้จำเฉลี่ยร้อยละ 65.14

ผลการทดสอบการรู้จำ โดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยไม่ใช้ขอบเขตพยางค์และใช้ความถี่ตัดในขั้นตอนการแยกองค์ประกอบความถี่สูงและองค์ประกอบความถี่ต่ำที่ 1.5 เฮิร์ตซ ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 51.00 และมีความผิดพลาดเป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 26.94, 12.67, 5.08 และ 4.31 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 74.90 มีความผิดพลาดเป็นเสียงสามัญ เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 13.52, 5.83, 2.17 และ 3.58 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 79.86 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 8.50, 8.69, 2.00 และ 0.94 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 40.60 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 21.55, 13.20, 16.30 และ 8.35 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 65.20 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและ

เสียงตรีร้อยละ 7.75, 20.45, 2.40 และ 4.20 ตามลำดับ โดยมีอัตราการรู้จำเฉลี่ยเป็นร้อยละ 65.14

4.3.2.1.4 ผลการรู้จำที่ความถี่ตัด 2 เฮิร์ตซ

ผลการทดสอบการรู้จำ โดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยไม่ใช้ขอบเขตพยางค์และใช้ความถี่ตัดในขั้นตอนการแยกองค์ประกอบความถี่สูงและองค์ประกอบความถี่ต่ำที่ 2 เฮิร์ตซ ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 54.14 และมีความผิดพลาดเป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 23.22, 13.50, 5.42 และ 3.72 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 76.21 มีความผิดพลาดเป็นเสียงสามัญ เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 11.75, 6.17, 1.85 และ 4.02 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 77.47 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 9.89, 7.69, 3.22 และ 1.72 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 35.00 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 23.55, 18.45, 15.10 และ 7.90 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 65.15 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงตรีร้อยละ 9.10, 17.60, 3.10 และ 5.05 ตามลำดับ โดยมีอัตราการรู้จำเฉลี่ยเป็นร้อยละ 64.99

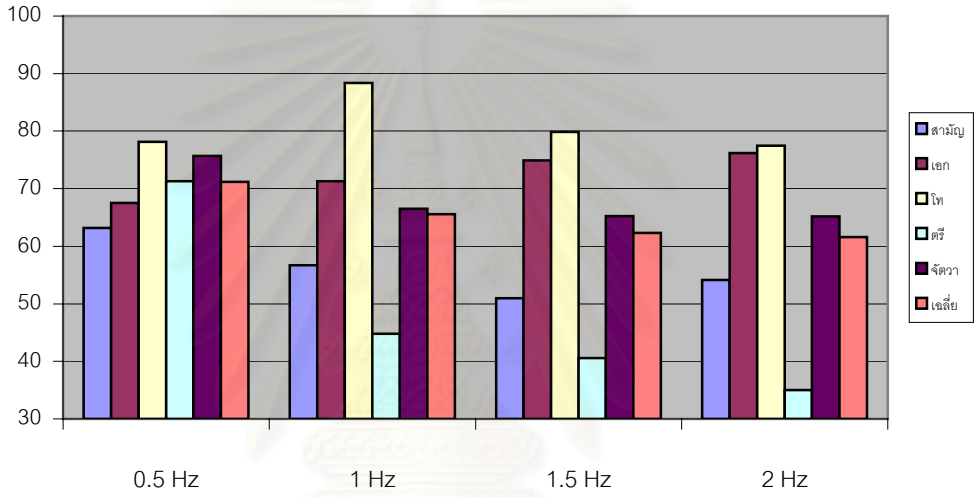
ตารางที่ 4.6 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้ความถี่ตัด 2 เฮิร์ตซ

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**54.14**	23.22	13.50	5.42	3.72
เอก	11.75	**76.21**	6.17	1.85	4.02
โท	9.89	7.69	**77.47**	3.22	1.72
ตรี	23.55	18.45	15.10	**35.00**	7.90
จัตวา	9.10	17.60	3.10	5.05	**65.15**

อัตราการรู้จำเฉลี่ยร้อยละ 64.99

ผลการรู้จำเมื่อใช้พารามิเตอร์ที่แยกโดยไม่ใช้ขอบเขตพยางค์ที่เปลี่ยนแปลงไปตามความถี่ตัดของวงจรกรองสำหรับแยกองค์ประกอบความถี่สูงและองค์ประกอบความถี่ต่ำ แสดงให้เห็นว่า

ความถี่ตัดที่ให้ความถูกต้องในการรู้จำสูงสุดคือ 0.5 เฮิร์ตซ ซึ่งเป็นค่าที่เหมาะสมและแนะนำโดย [14] เมื่อความถี่ตัดของวงจรกรองสูงขึ้น หมายถึงองค์ประกอบความถี่สูงจะเข้าไปอยู่ในเส้นโค้งความถี่ต่ำมากขึ้น หรือกล่าวได้ว่า เส้นโค้งความถี่ต่ำจะมีอัตราการเปลี่ยนแปลงที่มากขึ้น ทำให้ลักษณะของเส้นโค้งความถี่สูง ในส่วนของวรรณยุกต์ที่มีการเปลี่ยนแปลงมาก เช่นวรรณยุกต์จัตวามีการเปลี่ยนแปลงน้อย ก่อให้เกิดความถี่สับสนในการจำแนกวรรณยุกต์ และเส้นโค้งความถี่สูงของเสียงตรี จะมีลักษณะเรียบ เนื่องจากการเปลี่ยนแปลงของเส้นโค้งถูกกรองทิ้งไป ก่อให้เกิดความคล้ายคลึงกับลักษณะเส้นโค้งความถี่สูงของเสียงวรรณยุกต์สามัญมากขึ้น ดังจะเห็นได้จากความผิดพลาดในการรู้จำของเสียงวรรณยุกต์ตรีที่เพิ่มขึ้นมาก เมื่อความถี่ตัดของวงจรกรองสูงขึ้น



รูปที่ 4.5 แสดงผลเปรียบเทียบการรู้จำที่ความถี่ตัดต่างๆ

4.3.3 ผลการทดสอบด้วยพารามิเตอร์ที่ได้จากการแยกโดยใช้ขอบเขตพยางค์

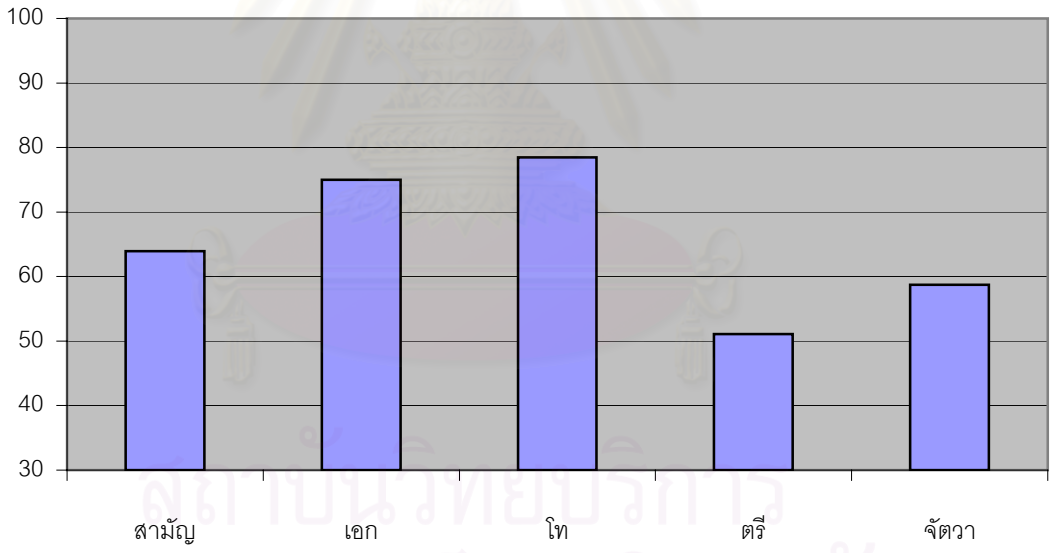
ผลการทดสอบการรู้จำโดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยใช้ขอบเขตพยางค์ ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 63.92 และมีความผิดพลาดเป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 18.44, 9.28, 6.11 และ 2.25 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 75.02 มีความผิดพลาดเป็นเสียงสามัญ เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 15.67, 2.29, 1.90 และ 5.12 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 78.47 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 11.06, 3.28, 6.69 และ 0.50 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 51.10 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 23.00, 12.35, 11.25 และ 2.30 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์

จัตวาเป็นร้อยละ 58.70 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงตรีร้อยละ 3.75, 34.80, 0.65 และ 2.10 ตามลำดับ โดยมีอัตราการรู้จำเฉลี่ยเป็นร้อยละ 68.27

ตารางที่ 4.7 ผลการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้ขอบเขตพยางค์

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**63.92**	18.44	9.28	6.11	2.25
เอก	15.67	**75.02**	2.29	1.90	5.12
โท	11.06	3.28	**78.47**	6.69	0.50
ตรี	23.00	12.35	11.25	**51.10**	2.30
จัตวา	3.75	34.80	0.65	2.10	**58.70**

อัตราการรู้จำเฉลี่ยร้อยละ 68.27



รูปที่ 4.6 ความถูกต้องของการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้ขอบเขตพยางค์

การแยกพารามิเตอร์โดยใช้ขอบเขตพยางค์ ไม่ได้ทำให้ผลการรู้จำดีขึ้น เทียบกับการรู้จำโดยไม่ใช้ขอบเขตพยางค์ และมีค่าความถูกต้องใกล้เคียงกับการแยกพารามิเตอร์โดยไม่ใช้ขอบเขตพยางค์ โดยมีค่าความถูกต้องในการรู้จำเสียงสามัญ เอก และโทดีกว่าการไม่ใช้ขอบเขตพยางค์ แต่ให้ความถูกต้องในการรู้จำเสียงตรีและจัตวาต่ำกว่าการไม่ใช้ขอบเขตพยางค์ถึงร้อยละ 20.17 และ 17.01 ตามลำดับ ความผิดพลาดในการรู้จำจากเสียงตรีเป็นเสียงสามัญมีร้อยละ 23.00 และความ

ผิดพลาดในการรู้จำจากเสียงจัตวาเป็นเสียงเอกมีร้อยละ 34.80 สาเหตุเนื่องมากจากการบังคับให้มีคำสั่งวรรณยุกต์ในช่วงที่มีค่าองค์ประกอบความถี่สูงที่มีค่ามากหรือน้อยกว่า 0 เพียงเล็กน้อยก่อให้เกิดความสับสนในการแยกเสียงวรรณยุกต์ออกจากกัน

4.3.4 ผลการทดสอบด้วยพารามิเตอร์ที่ได้จากการแยกโดยวิธีของ Hansjorg Mixdorff [13]

ผลการทดสอบการรู้จำโดยการใช้พารามิเตอร์ที่ได้จากการแยกโดยใช้วิธีของ Hansjorg Mixdorff ให้ผลความถูกต้องของการรู้จำในเสียงวรรณยุกต์สามัญเป็นร้อยละ 45.72 และมีความผิดพลาดเป็นเสียงเอก เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 29.83, 13.89, 7.14 และ 3.42 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์เอกเป็นร้อยละ 65.02 มีความผิดพลาดเป็นเสียงสามัญ เสียงโท เสียงตรีและเสียงจัตวาร้อยละ 17.42, 7.98, 4.27 และ 5.31 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์โทเป็นร้อยละ 67.42 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงตรีและเสียงจัตวาร้อยละ 13.33, 9.39, 7.89 และ 1.97 ตามลำดับ ผลความถูกต้องของการรู้จำเสียงวรรณยุกต์ตรีเป็นร้อยละ 40.80 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโทและเสียงจัตวาร้อยละ 17.35, 12.30, 19.20 และ 10.35 ตามลำดับ และมีความถูกต้องของการรู้จำเสียงวรรณยุกต์จัตวาเป็นร้อยละ 53.75 มีความผิดพลาดเป็นเสียงสามัญ เสียงเอก เสียงโท และเสียงตรีร้อยละ 11.70, 26.90, 4.55 และ 3.10 ตามลำดับ

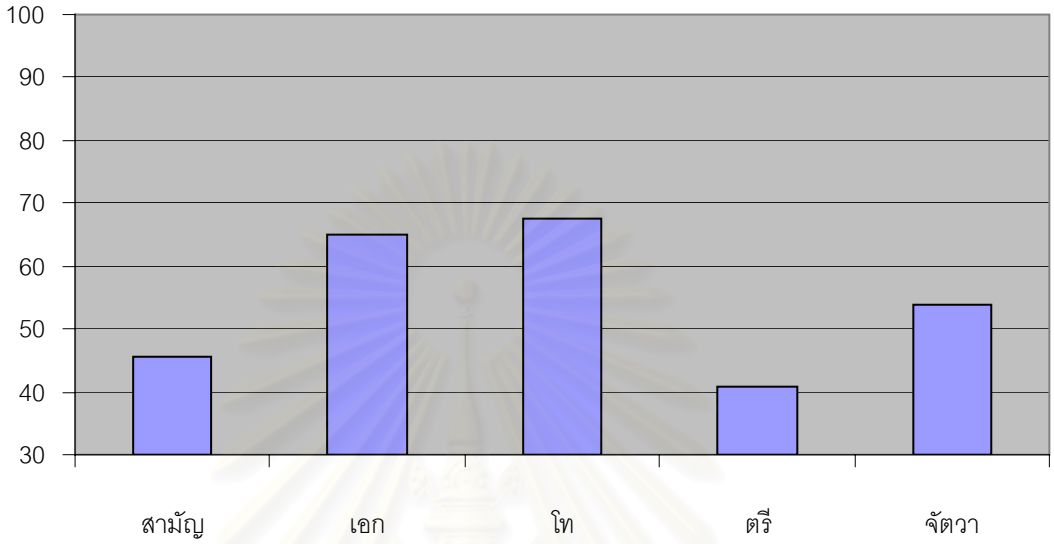
ตารางที่ 4.8 ผลการรู้จำด้วยพารามิเตอร์ที่แยกตามวิธีของ Mixdorff

วรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
สามัญ	**45.72**	29.83	13.89	7.14	3.42
เอก	17.42	**65.02**	7.98	4.27	5.31
โท	13.33	9.39	**67.42**	7.89	1.97
ตรี	17.35	12.30	19.20	**40.80**	10.35
จัตวา	11.70	26.90	4.55	3.10	**53.75**

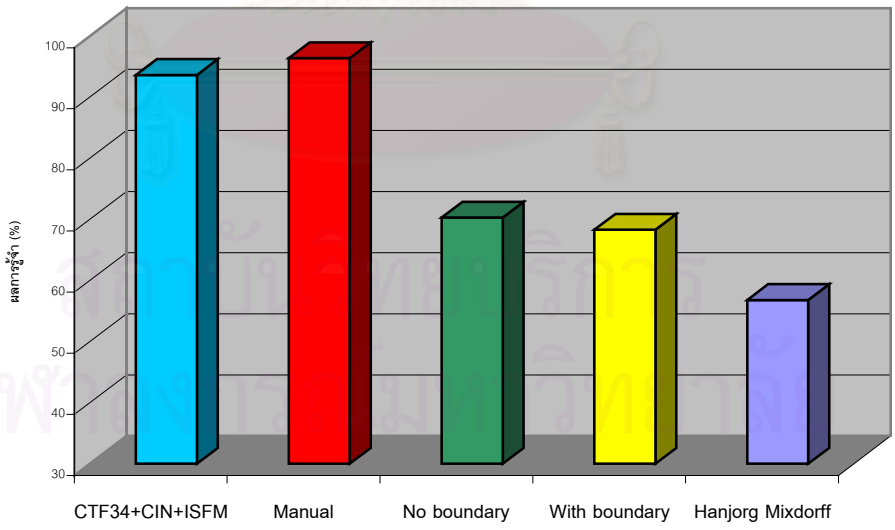
อัตราการรู้จำเฉลี่ยร้อยละ 56.78

สำหรับการรู้จำโดยใช้พารามิเตอร์ของแบบจำลองฟูจิซากิที่ทำการแยกโดยอัตโนมัติตามวิธีของ Hansjorg Mixdorff ได้ความถูกต้องเฉลี่ยร้อยละ 56.78 และมีความผิดพลาดในการรู้จำไปเป็นเสียงวรรณยุกต์สามัญ และเสียงวรรณยุกต์เอกมากที่สุด เนื่องมาจากการทำ smoothing โดย

ใช้อัลกอริทึม MOMEL ทำให้ข้อมูลด้านการเปลี่ยนแปลงของเส้นโค้งความถี่มูลฐานสูญหายไป หรือกล่าวได้ว่ามีการทำ smoothing มากเกินและส่งผลโดยตรงกับการตั้งค่าเริ่มต้นของคำสั่งวลี และคำสั่งวรรณยุกต์



รูปที่ 4.7 ความถูกต้องของการรู้จำด้วยพารามิเตอร์ที่แยกโดยใช้วิธีของ Mixdorff



รูปที่ 4.8 แสดงผลการรู้จำวรรณยุกต์ด้วยวิธีต่างๆ

พารามิเตอร์ของแบบจำลองฟูจิซากิที่แยกโดยใช้มนุษย์ ให้ผลการรู้จำถูกต้องร้อยละ 96.35 ซึ่งมีค่าสูงกว่าการรู้จำวรรณยุกต์ด้วย Contextual Tone Features ร่วมกับการทำ Center-

point intonation normalization และ Incorporated stress feature method [4] ที่ร้อยละ 93.60 โดยความผิดพลาดเกิดจากการที่ไม่สามารถตัดผลของการควบรวมของเสียงได้หมด สำหรับการรู้จำโดยใช้พารามิเตอร์ของแบบจำลองฟูจิซากิที่ทำการแยกโดยอัตโนมัติ ผลการทดสอบสำหรับวิธีการแยกของ Hansjorg Mixdorff ได้ความถูกต้องเพียงร้อยละ 56.78 ซึ่งมีความผิดพลาดในการรู้จำเนื่องมาจากการทำ smoothing โดยใช้อัลกอริทึม MOMEL ที่ทำให้สูญเสียข้อมูลด้านการเปลี่ยนแปลงของเส้นโค้งความถี่มูลฐาน ทั้งนี้อาจเนื่องจากอัลกอริทึม MOMEL มิได้ถูกพัฒนาเพื่อภาษาไทยโดยเฉพาะ อีกทั้งกรรมวิธีของ Mixdorff ยังใช้เวลาในการแยกพารามิเตอร์มาก เนื่องจากการใช้อัลกอริทึม MOMEL จะต้องทำการวนซ้ำ (Iteration) จำนวนมาก และการปรับค่าถึง 3 ครั้งโดยวิธีการเพิ่มค่าครั้งละขั้น จะใช้เวลาในการประมวลผลสูง

ข้อดีของการแยกพารามิเตอร์ตามวิธีของ Mixdorff คือ เส้นโค้งความถี่มูลฐานที่ได้จากการสังเคราะห์ด้วยพารามิเตอร์ที่ได้ จะมีความใกล้เคียงกับเส้นโค้งความถี่มูลฐานต้นแบบมากกว่า เนื่องจากทำการปรับค่าเทียบกับเส้นโค้งความถี่มูลฐานต้นแบบในขั้นตอนสุดท้าย ทำให้ได้ค่าที่ถูกต้องมากกว่าในกรณีที่นำไปใช้กับการสังเคราะห์เสียงพูด

สำหรับวิธีการแยกพารามิเตอร์ตามวิธีที่นำเสนอ ทำการ smoothing โดยใช้การ Neutralize และ Median Filtering ซึ่งการ smoothing ด้วยวิธีนี้ จะให้ผลการ smooth ที่ไม่เรียบมากนัก ทำให้ได้ผลการรู้จำที่ดีกว่าวิธีของ Hansjorg Mixdorff และการปรับค่าคำสั่งวลีและคำสั่งวรรณยุกต์จะกระทำเพียงครั้งเดียว โดยเปรียบเทียบองค์ประกอบวลีและองค์ประกอบวรรณยุกต์ที่สร้างขึ้นจากคำสั่งวลีและคำสั่งวรรณยุกต์ กับเส้นโค้งความถี่ต่ำและเส้นโค้งความถี่สูงที่ได้จากเสียงที่นำมาเปรียบเทียบ ซึ่งใช้เวลาน้อยกว่าวิธีของ Hansjorg Mixdorff ที่ทำการปรับค่าคำสั่งถึง 3 ครั้ง และพบว่าความถี่ตัดที่เหมาะสมในการแยกเส้นโค้งความถี่มูลฐานออกเป็น เส้นโค้งความถี่ต่ำ และเส้นโค้งความถี่สูงคือ 0.5 เฮิร์ตซตามคำแนะนำของ [14] การแยกพารามิเตอร์โดยใช้ขอบเขตพยางค์ไม่ได้ทำให้ผลการรู้จำโดยรวมดีขึ้น สาเหตุเนื่องมากจากการบังคับให้มีคำสั่งวรรณยุกต์ในช่วงที่มีค่าองค์ประกอบความถี่สูงที่มีค่ามากหรือน้อยกว่า 0 เพียงเล็กน้อยก่อให้เกิดความสับสนในการแยกเสียงวรรณยุกต์ออกจากกัน แต่การแยกพารามิเตอร์โดยใช้ของเขตพยางค์จะทำให้ได้จำนวนพารามิเตอร์ต่อพยางค์ที่แน่นอน เท่ากันทุกครั้ง ซึ่งเหมาะสมกับการใช้กับตัวจำแนกที่ต้องการจำนวนค่าสังเกตที่คงที่

ถึงแม้ว่าการรู้จำโดยใช้พารามิเตอร์ของแบบจำลองฟูจิซากิที่แยกโดยใช้มนุษย์จะให้ผลที่ดีที่สุด และเป็นที่น่าพอใจ อย่างไรก็ดี การแยกพารามิเตอร์วิธีนี้ ไม่สามารถนำไปประยุกต์ใช้งานจริงในระบบรู้จำวรรณยุกต์ได้ เนื่องจากการรู้จำวรรณยุกต์ และการรู้จำเสียงพูด ต้องมีการทำงานและตอบสนองได้ในทันที นั่นคือขั้นตอนการทำงานทั้งหมดต้องเป็นไปโดยอัตโนมัติ ถึงกระนั้นก็ยังได้

ทำให้เห็นว่า การนำพารามิเตอร์ของแบบจำลองฟูจิซากิไปใช้เป็นค่าคุณลักษณะสำคัญในการรู้จำวรรณยุกต์ของเสียงพูดต่อเนื่องนั้น สามารถกระทำได้จริง และให้ผลที่ค่อนข้างน่าพอใจ

การพัฒนาการรู้จำวรรณยุกต์ของเสียงพูดต่อเนื่องในอนาคต โดยใช้แบบจำลองฟูจิซากินี้สมควรที่จะหาวิธีที่เหมาะสมและรวดเร็ว ในการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิโดยอัตโนมัติ เนื่องจากวิธีหลักในการหาค่าพารามิเตอร์ดังเช่นใน [11] ยังคงไม่เป็นการหาค่าพารามิเตอร์แบบอัตโนมัติ หรือวิธีที่เสนอโดยวิทยานิพนธ์ฉบับนี้และวิธีที่เสนอโดย [14] ยังให้ผลการรู้จำที่ไม่ดีเพียงพอที่จะนำไปใช้งานได้จริง อีกทั้งยังใช้เวลาในการแยกพารามิเตอร์ในขั้นตอนการปรับค่ามาก แม้วิธีที่นำเสนอจะใช้เวลาน้อยกว่าวิธีที่เสนอโดย [14] มาก แต่ยังไม่สามารถแยกพารามิเตอร์ได้รวดเร็วเพียงพอที่จะนำไปใช้งานในการรู้จำเสียงพูดได้จริง เนื่องจากการรู้จำเสียงพูดนั้นต้องการการทำงานที่รวดเร็ว และตอบสนองในทันที นอกจากนั้น ยังมีการพิจารณาองค์ประกอบอื่นที่มีผลต่อวรรณยุกต์ของเสียงพูด ที่มิได้นำมาพิจารณาในที่นี้อีกด้วย ตัวอย่างเช่นการเน้นเสียง (Stress) ซึ่งนับว่ามีผลอย่างมากในการรู้จำวรรณยุกต์

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์นี้นำเสนอการรู้วรรณยุกต์ของเสียงพูดต่อเนื่อง โดยใช้พารามิเตอร์ของแบบจำลองฟูจิซากิเป็นคุณลักษณะสำคัญ เพื่อนำไปประยุกต์ใช้ร่วมกับการรู้จำเสียงพูดต่อเนื่องในระดับหน่วยเสียงเรียง (Segmental) จากผลการทดสอบกับชุดข้อมูลที่ใช้จะเห็นได้ว่า วิธีที่นำเสนอสามารถให้ผลการรู้จำได้ถูกต้องเฉลี่ยถึงร้อยละ 96.35 ในกรณีที่แยกพารามิเตอร์โดยมนุษย์ ซึ่งสูงกว่ากรณีการใช้ Half-tone model ร้อยละ 2.53 ในขณะที่มีขนาดของโครงข่ายประสาทเทียมที่ใช้เป็นตัวจำแนกประเภท (Classifer) ที่เล็กกว่า ซึ่งลดระยะเวลาการคำนวณในส่วนของโครงข่ายประสาทเทียม

อย่างไรก็ตาม กรรมวิธีการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิโดยใช้มนุษย์ยังไม่สามารถนำไปประยุกต์ใช้ได้จริง และสำหรับการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิแบบอัตโนมัติซึ่งยังไม่มีกรรมวิธีที่เป็นที่ยอมรับกันโดยทั่วไปนั้น วิทยานิพนธ์นี้ได้ทำการทดสอบกับระบบการรู้จำวรรณยุกต์โดยใช้กรรมวิธีที่ถูกอ้างว่าใกล้เคียงกับการแยกพารามิเตอร์โดยมนุษย์มากที่สุด รวมทั้งได้เสนอกรรมวิธีการแยกพารามิเตอร์ที่ง่ายและซับซ้อนน้อยกว่า โดยผลการทดสอบแสดงให้เห็นว่าการใช้พารามิเตอร์ที่แยกด้วยกรรมวิธีที่วิทยานิพนธ์นี้นำเสนอ สามารถให้ผลการรู้จำเฉลี่ยที่ 70.27 ซึ่งสูงกว่าการใช้พารามิเตอร์ที่แยกโดยวิธีที่ถูกอ้างดังกล่าวร้อยละ 13.49

5.2 ข้อเสนอแนะสำหรับงานวิจัยในอนาคต

ปัญหาที่เกิดขึ้นในการทำงานวิจัยในวิทยานิพนธ์ ได้แก่

1. ยังมีผู้สนใจในการใช้งานแบบจำลองฟูจิซากิกับภาษาไทยไม่มาก ทำให้ข้อมูลของแบบจำลองฟูจิซากิกับภาษาไทยมีอยู่จำกัด
2. การแยกพารามิเตอร์ของแบบจำลองฟูจิซากิโดยใช้มนุษย์จะเสียเวลามาก ทำให้การทดสอบต่างๆ ทำได้ช้า

งามที่สมควรได้รับการศึกษาหรือพัฒนาต่อไปในอนาคตคือ

1. ปรับปรุงกรรมวิธีการแยกพารามิเตอร์ของแบบจำลองฟูจิซากิโดยอัตโนมัติให้ได้ผลดี และเป็นที่ยอมรับโดยทั่วไป เพื่อเพิ่มประสิทธิภาพการรู้จำแบบอัตโนมัติ
2. ปรับปรุงโครงสร้างการรู้จำวรรณยุกต์ของเสียงพูดให้สามารถทำงานได้โดยไม่ต้องใช้ข้อมูลของขอบเขตพยางค์ เพื่อให้ระบบการรู้จำสามารถทำงานอย่างอัตโนมัติได้โดยสมบูรณ์
3. พัฒนากระบวนการรู้จำที่รวมการรู้จำวรรณยุกต์ของเสียงพูดและการรู้จำเสียงพูดเข้าด้วยกัน

รายการอ้างอิง

1. Brad A. Myers. A Brief History of Human Computer Interaction Technology. ACM interactions. Vol. 5 (March 1998) : 44-54
2. S.Potisuk, and M.P.Harper. Speaker-Independent Automatic Classification of Thai Tones in Connected Speech by Analysis-Synthesis Method. ICASSP-95 International Conference on Acoustics, Speech, and Signal Processing. Vol. 1 (1995) : 632-635
3. S.Potisuk, M.P.Harper and J.Gandour. Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method. IEEE Transactions on Speech and Audio Processing. Vol. 7, No. 1 (Jan 1999) : 95-102
4. N. Thubthong, B. Kijsirikul, S. Luksaneeyanawin. Tone Recognition in Thai Continuous Speech Based on Coarticulation, Intonation and Stress Effects. Proceedings of the 7th International Conference in Spoken Language Processing-ICSLP 2002, and Interspeech 2002. Vol. 2 (September 2002) : 1169-1172.
5. H. Fujisaki. Modeling in the Study of Tonal Feature of Speech with Application to Multilingual Speech Synthesis. Proceeding of the 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop. (2002)
6 . K.Fellbaum and J.Richter. Human Speech Production Using Interactive Modules and the Internet – a Tutorial for the Virtual University.[Online]. Available from: http://www.kt.tu-cottbus.de/speech-analysis/tech.html, [2002, June 20]
7. David P. Morgan and Christopher L. Scofield. Neural Networks and Speech Processing. Kluwer Academic Publishers, 1991.
8. J.R.Deller Jr., J.G.Proakis and J.H.L.Hansen. Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, 1993.
9. H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki and P. Charnvivit. Perception of Tone and Vowel Quantity in Thai. Proceedings of the 7th International Conference in Spoken Language Processing-ICSLP 2002 and Interspeech 2002. (2002)
10. C.-F. Wang, H. Fujisaki, and K. Hirose, Chinese Four Tone Recognition based on the Model for Process of Generating Contours of Sentences. ICSLP'90 Internation Conference on Spoken Language Precessing. (1990) : 221-224

11. P.Seresangtakul and T.Takara. Analysis of Pitch Contour of Thai Tone using Fujisaki's Model. International Conference on Acoustics, Speech, and Signal Processing. (2002)

12. Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose and Hiroya Fujisaki. A Method for Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Vol 1. (May 2002) : 509-512.

13. Pierluigi Salvo Rossi, Francesco Palmieri and Francesco Cutugno. A Method for Automatic Extraction of Fujisaki-Model Parameters. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Vol 1 (April 2003) : 520-523

14. Hansjorg Mixdorff. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. ICASSP'00.IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 3 (2000) : 1281-1284

15. Daniel Hirst , Robert Espesser. Automatic Modelling of Fundamental Frequency using a Quadratic Spline Function. Travaux de l'Institut de Phonetique d'Aix. Vol. 15 (1993) : 75-85

16. Thubthong, N., Kijsirikul, B. and Luksaneeyanawin, S. An empirical study for constructing Thai tone models. The 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop. (2002) : 179-186

บทความทางวิชาการที่ได้รับการเผยแพร่

1. ได้รับการตอบรับงานประชุมวิชาการ IEEE Canadian Conference on Electrical and Computer Engineering 2004 (CCECE2004) สำหรับบทความ “Tone Recognition of Thai Continuous Speech using Fujisaki Model” กำหนดการจัดประชุมโดย IEEE Canada ในระหว่างวันที่ 2-5 พฤษภาคม 2547

TONE RECOGNITION OF THAI CONTINUOUS SPEECH USING FUJISAKI'S MODEL

Nutthee Ngarmchatetanarom, Ekkarit Maneenoi, Widhyakorn Asdornwised, and Somchai Jitapunkul
Digital Signal Processing Research Laboratory, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Somchai.j@chula.ac.th

Abstract

This paper proposes the recognition algorithm using Fujisaki's model parameters as feature for tone recognition of predominantly tonal language without limitation of number of syllables in the utterance. The algorithm is composed of 4 steps. Initially, the F0 contour of speech is extracted. Fujisaki's model parameters of whole utterance are extracted manually. The utterance-base tone commands are then segmented into syllable-base tone commands and are used as classification feature and fed into MLP neural network recognition system. The experimented results show that the proposed system when Thai continuous speech was selected to test achieves average recognition rate of 96.35%.

Keywords: *Speech recognition, Tone, Neural Network.*

1. INTRODUCTION

For tonal language (e.g. Chinese, Thai), tone is an important part of speech understanding system because the reference meaning of an utterance is dependent on the lexical tones. In Thai, there are five different lexical tones traditionally labeled mid (M), low (L), falling (F), high (H) and rising (R). The following examples show the effect of tones on the meaning of utterances: M/khāa/ ("a kind of grass"); L/khàa/("galangale"); F/khâa/("to kill"); H/kháa/("to trade"); and R/khăaa/("a leg"). For Thai language the classification of a tone relies on the shape of the fundamental frequency (F0) contour which is a characteristic of voice portion of each syllable. Figure 1 shows the average of F0 contours of five different Thai tones when syllables are spoken in isolation by a speaker [1].

Thai language can be classified as a predominantly tonal language. Thai tone recognition for monosyllable using polynomial regression function to extract the feature from F0 contour works well on single syllable utterance [2]. But on the continuous speech, the phrase declination and tonal assimilation from adjacent syllable in the utterance degrade the recognition rate of the system.

Thai tone classification using Analysis-by-Synthesis method analyzes the pattern of peaks and valleys sequence of F0 contours of speech, synthesizes the contours with trained Fujisaki's model parameters for possible sequences, and compares the synthesized F0 contours with the exact one to find out the most matching tone sequence [3, 4]. This method takes the advantage of tonal assimilation and phrase declination handling of Fujisaki's model. But there are disadvantages on limitation of the number of syllables in utterance which is combined to the possible tone sequences for matching of the fixed synthesized fundamental frequency contour by trained parameter with the variation of raw fundamental frequency.

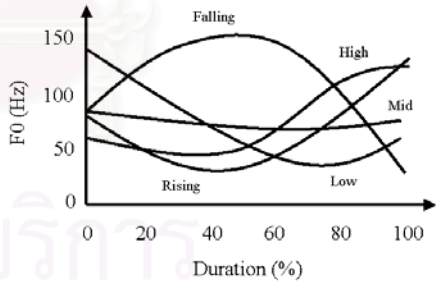


Figure 1: F0 contours of the five Thai tone [1].

Tone recognition technique in [1] used half-tone model which employed the 5 points of ERB-scale frequency value and slope of F0 contour, 2 points of preceding syllable and 3 points of following syllable to be 10 points feature. The model used declination normalization in order to reduce its effect. The normalization assumed the steady downdrift of the mean of F0 contour [1]. This assumption is not true for the long utterances. The technique in [1] is the most effective for

CCECE 2004- CCGEI 2004, Niagara Falls, May/mai 2004
0-7803-8253-6/04/$17.00 ©2004 IEEE

Thai continuous speech which yield the recognition rate of 93.82%

In this paper, the input speech is re-synthesized using Fujisaki's model-based synthesis and uses the tone commands of the synthesis as the new feature for tone recognition. The attraction of using Fujisaki's model to extract the feature is that the model already considers the declination effect and tonal assimilation effect of continuous speech. This yields that the parameter from synthesis already handles the declination and tonal assimilation effect.

This paper is organized as follows. The next section describes the detail of Fujisaki's model and relation of Thai tones with Fujisaki's model tone commands. The recognition system configuration of proposed system and baseline system are described in section 3. The results of experiment are presented in section 4 and finally concluded in section 5.

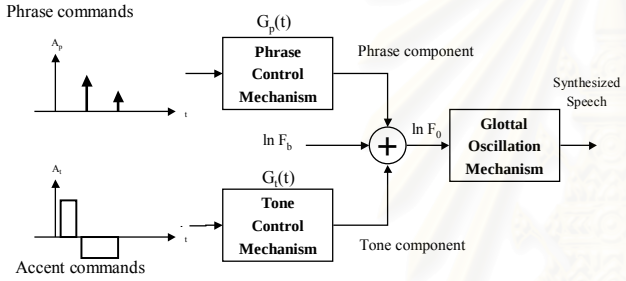


Figure 2: Block diagram of Fujisaki's model.

2. FUJISAKI'S MODEL

Fujisaki's model was proposed by Fujisaki H. and primarily used in natural speech synthesis which based on dynamical synthesis of fundamental frequency contour to yield the naturalness of speech utterances. The model is adopted to describe the lexical tone in many languages and also Thai [3, 4]. The model for tonal languages consists of phrase and tone control mechanism. Block diagram of Fujisaki's model is illustrated in Figure 2. The phrase commands are assumed to be impulse which are applied to the phrase control mechanism and generate the phrase components. Tone commands are step function in both positive and negative polarities that are applied to tone control mechanism to generate the tone components. The mathematical expression of the model is as follow:

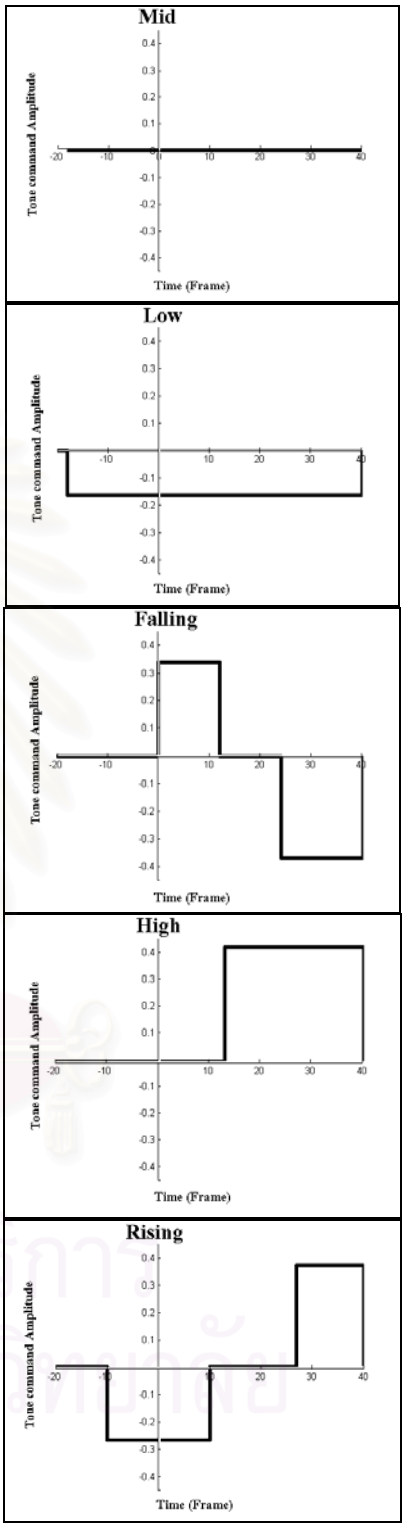


Figure 3: Tone commands for each tone [4].

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{tj}[G_t(t - T_{1j}) - G_t(t - T_{2j})] \quad (1)$$

$$G_p(t) = \alpha^2 t \cdot \exp(-\alpha t), \quad for \quad t \geq 0 \qquad = 0 \quad for \quad t < 0 \quad (2)$$

$$G_t(t) = \min[1 - (1 + \beta t) \cdot \exp(-\beta t), \gamma], \quad for \quad t \geq 0 \qquad = 0 \quad for \quad t < 0 \quad (3)$$

$G_p(t)$ represents the impulse response function of phrase control mechanism. α is time constant parameter in phrase generation. T_{0i} and A_{pi} denote the i[th] phrase command time and its amplitude respectively. $G_t(t)$ represents the step response function of tone control mechanism. β is the time constant parameter in tone generation. γ is the limitation parameter of tone control mechanism. T_{1j}, T_{2j} and A_{tj} are onset time, offset time and amplitude of the j[th] tone command respectively. I and J are number of phrase and tone commands respectively. And F_b is the base frequency of F0 contour which is depended on speaker.

F_0 is the synthesized F0 contour that is superposition in logarithmic scale of base frequency, phrase components and tone components. The synthesized F0 contour is passed to glottal oscillation mechanism to generate the naturalness of utterances in continuous speech synthesis system.

From [4], the onset and offset time and amplitude of tone commands are able to describe the difference of 5 Thai tones as illustrated in Figure 3. It can be clearly seen that the polarities of tone commands for mid, low, falling, high and rising tone are zero, negative, positive/negative, positive, and negative/positive respectively.

3. RECOGNITION SYSTEM CONFIGURATION

Block diagram of recognition system is shown in Figure 4. F0 contour is extracted from input speech and then used as a reference signal for Fujisaki's model parameter extraction. Tone commands of the utterance are then segmented manually to syllable-based tone commands and used as a feature for recognition using MLP neural network classifier.

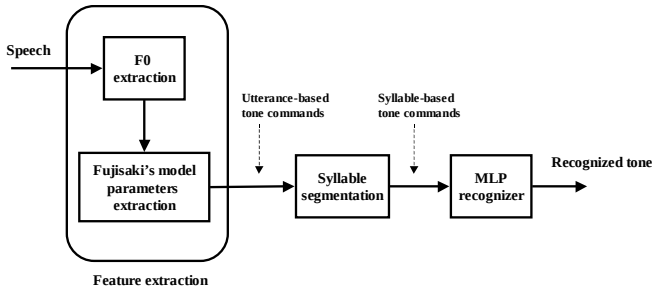


Figure 4: Block diagram of recognition system.

3.1 Speech corpus

Tone recognition experiment is based on Thai Proverb Corpus which is the same corpus as [1]. The corpus composed of 10 sentences of Thai proverb. Each sentence composed of 4 syllables. There are 20 male and 20 female speakers uttered each sentences in continuous manner with reading style. The total of speech data become 400 utterances which composed of 3600, 4800, 3600, 2000 and 2000 words of mid, low, falling, high and rising tone respectively.

3.2 Fujisaki's parameter extraction for Thai tones

The Fujisaki's model parameter extraction procedure is described below.

3.2.1 The F0 extraction uses 40 milliseconds frame size with 10 milliseconds step. The logarithm of each extracted F0 contours are then used as reference for perfect Fujisaki's model parameter extraction.

3.2.2 The $\ln F_b$ is set to the lowest frequency of the contour and subtracted out, then the rest portion is used to estimate the phrase command time and amplitude. The phrase components calculated from phrase commands are subtracted out of the contour.

3.2.3 The last portion is used as reference for tone command feature extraction. Tone command then is set to best fit with the reference. The tone commands of utterance then are segmented into syllable-based tone command token and used as feature parameters.

3.3 Tone recognition

The time parameters of tone commands are normalized by the duration of syllable to lie between 0

and 1. The amplitude parameters of tone commands are normalized to lie between –1 and 1.

This paper employs MLP neural network, the NICO toolkit, to train and test the system. This MLP neural network is composed of 9 units in input layer, 40 units in hidden layer, and 5 units in output layer. The network is trained by error back-propagation with random initial weights from –1 to 1.

4. EXPERIMENTAL RESULTS

The experiment uses 5-fold cross-validation approach [1]. All tones are divided into five disjoint sets. Five training sets are derived by overlapping three disjoint sets systematically and the rest two sets of each fold are combined to use as test sets.

The experimental results are shown in confusion matrix in Table 1. The results show the recognition rate of 97.75%, 97.42%, 95.19%, 94.10% and 95.60% in mid, low, falling, high and rising tone respectively. From the recognition result, the average recognition rate is 96.35% which is better than 93.82% of [1].

Table 1: The experimental results.

	Mid	Low	Falling	High	Rising
Mid	97.75	1.06	0	1.17	0.03
Low	1.19	97.42	0	0.15	1.25
Falling	0	0.08	95.19	4.50	0.22
High	1.10	0.10	4.30	94.10	0.40
Rising	0	2.65	0.05	1.70	95.60

5. CONCLUSIONS

The continuous Thai speech tone recognition using Fujisaki's model has been proposed in this paper. The system uses tone commands of Fujisaki's model that is extracted from F0 contour of input speech as feature. The tone recognition processes use MLP-based neural network to classify 5 different Thai tones. The recognition results are average to 96.35% which is 2.53% improved from 93.82% of [1].

Acknowledgements

The authors would like to acknowledge Government Research Grant in Research and Development Cooperative Project between EE Dept and Private Sector for financial support for this research.

References

[1] Thubthong N., Pusittrakul A., and Kijsirikul B., "Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model," *Proc. of International Conference on Intelligent Technologies,* pp. 229-234, 2000.

[2] Charnvivit P., Jitapunkul S., Ahkuputra V., and Maneenoi E., "F0 Feature Extraction by Polynomial Regression Function for Monosyllabic Thai Tone Recognition," *Proc. of European Conference on Speech,* 2001.

[3] Potisuk S., and Harper M.P., "Speaker-Independent Automatic Classification of Thai Tones in Connected Speech by Analysis-Synthesis Method," *Proc. of International Conference on Acoustics, Speech, and Signal Processing,* Vol. 1, pp.632-635, 1995.

[4] Potisuk S., Harper M.P., and Gandour J., "Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method," *IEEE Trans. Speech and Audio Processing,* Vol. 7, pp. 95-102, 1999.

[5] Seresangtakul P., and Takara T., "Analysis of Pitch Contour of Thai Tone Using Fujisaki's Model," *IEEE Trans. Speech and Audio Processing,* Vol. 1, pp.505-508, 2002.

ประวัติผู้เขียนวิทยานิพนธ์

นายนัทธี งามเจตนรมย์ เกิดวันที่ 19 เมษายน พ.ศ.2515 ที่จังหวัดสมุทรปราการ เข้าศึกษาในหลักสูตรวิศวกรรมศาสตรบัณฑิต คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2532 และเข้าศึกษาต่อในหลักสูตรวิศกรรมศาสตรมหาบัณฑิต ที่ห้องปฏิบัติการวิจัยกรรมวิธีสัญญาณดิจิทัล ภาควิชาวิศวกรรมไฟฟ้า จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2544