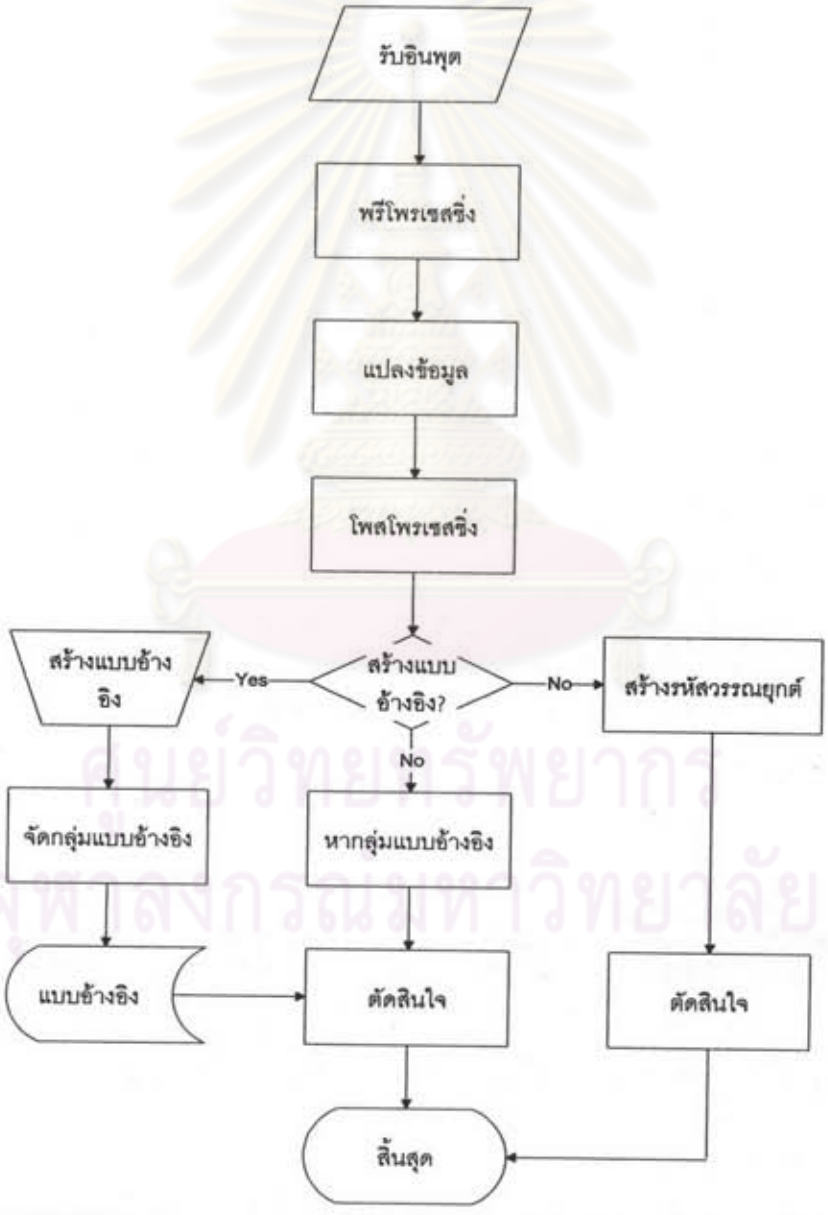
บทที่ 4

ขั้นตอนการทำงานของระบบ

การทำงานของระบบรู้จำเสียงพูด แบ่งออกเป็นส่วนต่าง ๆ ดังแผนภาพข้างล่าง

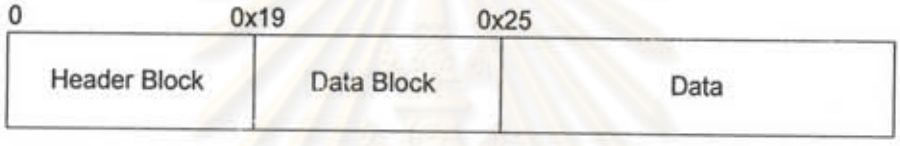


รูปที่ 4.1 แผนผังการทำงานของระบบ

ส่วนรับอินพุต

ส่วนรับอินพุตนี้จะทำหน้าที่รับไฟล์เสียง ซึ่งใช้ฟอร์แมตไฟล์เสียงของ Sound Blaster ที่ใช้ไฟล์นามสกุล VOC จะบันทึกระดับความดังของสัญญาณเสียง ซึ่งจะถูกแทนด้วยรหัสไบนารี ขนาด 8 บิต ต่อหนึ่ง Sample จึงมีระดับความดังสัญญาณได้ทั้งหมด 256 ระดับ โดยใช้อัตราการสุ่มสัญญาณ (Sampling Rate) ที่ 8 kHz สำหรับข้อมูลที่จะนำไปประมวลผล จะต้องตัดส่วนที่ไม่ต้องการทิ้งไป เหลือเฉพาะข้อมูลสัญญาณเสียงแต่เพียงอย่างเดียว เพื่อส่งไปยังส่วนถัดไป ส่วนรายละเอียดของไฟล์เสียงประกอบด้วยส่วนต่าง ๆ ดังนี้

ส่วนที่ 1 Header Block
ส่วนที่ 2 Data Block
ส่วนที่ 3 Data



รูปที่ 4.2 ฟอร์แมตของไฟล์เสียง

ส่วนพรีโพรเซสชิง

ส่วนพรีโพรเซสชิงนี้จะเตรียมและปรับข้อมูลก่อนที่จะแปลงสัญญาณเสียงให้อยู่ในเชิงความถี่ ซึ่งมีส่วนที่ทำหน้าที่ต่าง ๆ ดังนี้

1 ปรับระดับสัญญาณจาก DC ให้เป็น AC เนื่องจากข้อมูลที่ได้มาจาก Sound Blaster จะถูกบันทึกไว้แบบ Character คือมีค่าเป็นบวกอยู่ระหว่าง 0 ถึง 256 ดังนั้นจะต้องทำการปรับให้สัญญาณมีค่าบวกและลบ แกว่งอยู่รอบแกนที่ระดับ 128 เพื่อลดสเปคตรัมของสัญญาณไฟตรงบริเวณ 0 Hz ลง

2 หาขอบเขตของสัญญาณ (End Point Detection) โดยยึดเอาระดับสัญญาณที่ค่า ๆ หนึ่งเป็นเกณฑ์ (Threshold) เมื่อมีสัญญาณที่มีระดับความดัง (Amplitude) ของสัญญาณเกินค่าที่ตั้งเอาไว้ จำนวน 3 ครั้ง [2] กำหนดให้จุดนั้นเป็นจุดเริ่มต้นของสัญญาณ และทำเช่นเดียวกันกับในส่วนท้ายของสัญญาณเสียงเพื่อหาจุดสิ้นสุด การตรวจสอบความดังของสัญญาณที่มีค่าเกินค่าที่กำหนดไว้ 3 ครั้งติดต่อกัน เพื่อยืนยันว่าสัญญาณนั้นเป็นสัญญาณเสียงจริง ๆ ไม่ใช่สัญญาณรบกวนที่มีระดับความแรงของสัญญาณค่าสูง ๆ ชั่วขณะ แต่ถ้าระดับสัญญาณของสัญญาณรบกวนมีค่าสูงกว่าค่าที่ตั้งกำหนดเอาไว้ตลอด จะไม่สามารถหาขอบเขตของสัญญาณเสียงที่แท้จริงได้ อาจต้องทำการเลือกกำหนดค่าตรวจระดับความแรงของสัญญาณ หรือปรับระดับความดังของสัญญาณทั้งหมดใหม่ ถ้าสัญญาณทั้งหมดจะถูกส่งต่อไปยังขั้นตอนต่อไปโดยไม่ได้ตัดสัญญาณส่วนที่ไม่ต้องการทิ้ง จะทำให้เสียเวลาในการคำนวณโดยไม่จำเป็น

ส่วนการแปลงข้อมูล (Transformation)

ในส่วนการแปลงข้อมูลนี้ จะนำข้อมูลที่ผ่านการหาขอบเขตของสัญญาณจนได้เฉพาะสัญญาณเสียง แล้วจึงนำเอาสัญญาณเสียงในเชิงเวลานี้มาแบ่งเป็นช่วง ๆ เพื่อนำไปทรานส์ฟอร์มโดยใช้ FFT (Fast Fourier Transform) ซึ่งข้อมูลในแต่ละเฟรมจะถูกเลือกให้มีจำนวน Sample เท่ากับ 200 หรือ 25 มิลลิวินาที ซึ่งช่วงเวลา 20 - 40 มิลลิวินาที จะเป็นช่วงที่เหมาะกับการวิเคราะสัญญาณเสียงพูดทั้งเพศชายและเพศหญิง [15] ความยาวข้อมูลแต่ละเฟรมที่ทรานส์ฟอร์มมีขนาด 512 Sample เท่ากันทั้งหมด แต่ส่วนที่เป็นข้อมูลจริงเพียง 200 Sample เท่านั้น อีก 312 Sample จะถูกกำหนดให้เป็นศูนย์ (Zero Padding) และเพื่อลดปัญหาของความไม่ต่อเนื่องของสเปคตรัมเสียงที่เกิดขึ้นระหว่างเฟรม จึงได้มีการนำเอาข้อมูลในเฟรมก่อนหน้านี้มาใช้ในเฟรมปัจจุบันด้วย ทำให้เกิดการซ้อนทับกัน (Overlap) ของข้อมูลในแต่ละเฟรม ถ้าข้อมูลมีการซ้อนทับกันมากจะมีผลให้สเปคตรัมเสียงมีความต่อเนื่องกัน แต่จะเสียเวลาในการคำนวณเพิ่มขึ้น ซึ่งในที่นี้จะใช้การซ้อนทับกันของข้อมูล ครึ่งหนึ่งของจำนวนข้อมูลในแต่ละเฟรม คือ 100 Sample (สุนิสา จันทวีกูล, 2536) และสัญญาณเสียงในเชิงเวลาแต่ละเฟรมจะถูกคูณด้วย Hamming Window ฟังก์ชัน [17] เพื่อลดการเปลี่ยนแปลงของสเปคตรัมที่เกิดขึ้นอย่างรวดเร็ว ทำให้สเปคตรัมที่ได้มีลักษณะใกล้เคียงกับ Power Envelop ของสัญญาณเสียงนั้น

ส่วนโพสโพรเซสซิง

1 การลดสัญญาณรบกวน เนื่องจากอาจมีสัญญาณรบกวนในขั้นตอนพรีโพรเซสซิงหลงเหลืออยู่ ซึ่งจะทำให้เกิดปัญหาในขั้นตอนของการรู้จำและการสร้างแบบอ้างอิงได้ จึงได้มีการตรวจและลดสัญญาณรบกวนที่ไม่ต้องการในเชิงความถี่ที่สามารถทำได้ง่ายกว่า เช่น การลดสัญญาณรบกวนที่ความถี่ต่ำ 0 - 50 Hz สามารถตัดออกไปโดยการให้ค่า Magnitude ที่ความถี่ 0 - 50 Hz เป็นศูนย์ได้โดยตรง แล้วหาขอบเขตของสัญญาณเสียงใหม่อีกครั้ง โดยตรวจพลังงานรวมในแต่ละเฟรมจากเฟรมที่มีพลังงานรวมมากที่สุดไปหาจุดเริ่มต้นและจุดสุดท้ายตาม threshold ที่ได้กำหนดไว้

2 ส่วนลดขนาดข้อมูล ในส่วนนี้จะทำการนอร์มอลไลซ์ข้อมูลให้ระดับสัญญาณมีความใกล้เคียงกัน และแปลง Magnitude ของสเปคตรัมให้อยู่ใน Log Scale เพื่อปรับให้สเปคตรัมที่ช่วงความถี่ต่ำ และที่ช่วงความถี่สูงมี Magnitude ใกล้เคียงกัน ข้อมูลจะเก็บอยู่ในฟอร์แมตของคาร์เรคเตอร์ (Character) ซึ่งทำให้ขนาดของไฟล์ลดลงจากเดิมครึ่งหนึ่ง

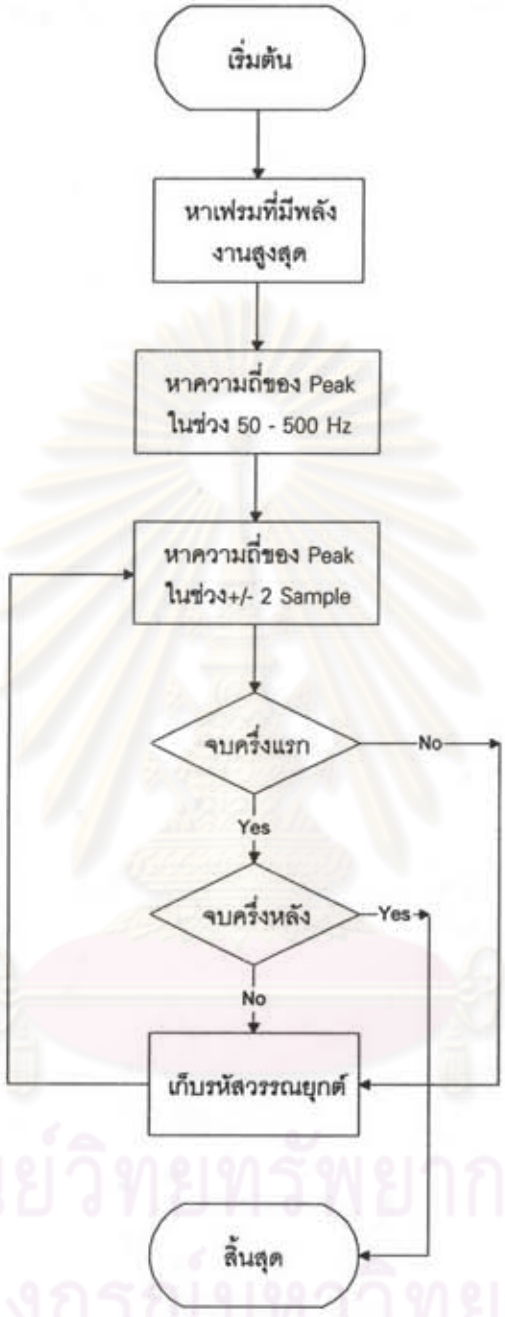
ส่วนการสร้างแบบอ้างอิง

แบบอ้างอิงที่ใช้ในวิทยานิพนธ์นี้จะใช้หนึ่งแบบต่อหนึ่งเสียง โดยแบบอ้างอิงจะได้จากการนำไฟล์เสียงแต่ละเสียงที่จะนำมาสร้างแบบอ้างอิงมาหาค่าเฉลี่ย ซึ่งแบบอ้างอิงที่ได้จะปรากฏส่วน Magnitude ของสเปคตรัมเสียงที่ซ้ำกันมีค่ามากกว่าส่วนที่มีสเปคตรัมไม่เหมือนกัน การสร้างแบบอ้างอิงนั้น จะนำเอาไฟล์ข้อมูลของเสียงที่ต้องการจะสร้างแบบอ้างอิง โดยการนำไฟล์เสียงทั้งหมดมารวมกันในลักษณะเฟรมต่อเฟรม คือ เฟรมที่มีตำแหน่งเหมือนกันจะถูกนำมารวมกัน โดยกำหนดให้ต้นไฟล์ของทุกไฟล์คือเฟรมที่ 1 และถัดมาตามลำดับ สำหรับไฟล์ซึ่งมีจำนวนเฟรมน้อยกว่า จะทำการเพิ่มจำนวนเฟรมให้เท่ากันโดยเฟรมที่เพิ่มเข้าไปนั้นจะมีค่าเป็น 0 ทั้งหมด และทุก ๆ ค่าในแต่ละเฟรมจะถูกหารด้วยจำนวนไฟล์ทั้งหมด ซึ่งไฟล์ที่ได้จะเป็นแบบอ้างอิงของเสียงนั้นต่อไป

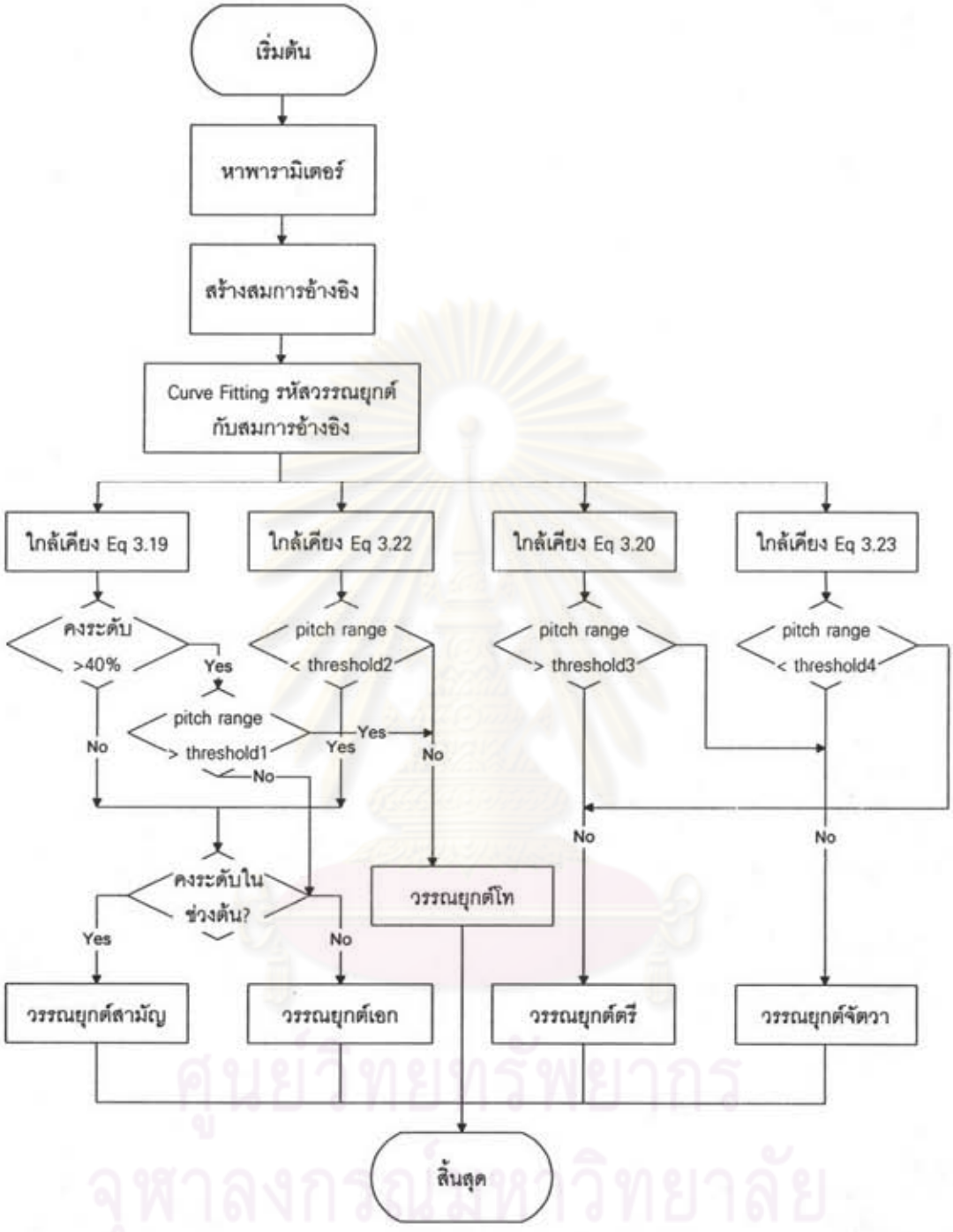
ส่วนการสร้างรหัสเสียงวรรณยุกต์

จากไฟล์นามสกุล .COD ซึ่งเป็นไฟล์ข้อมูลเสียงในเชิงความถี่ ไฟล์เดียวกันกับที่ใช้ทดสอบการรู้จำเสียงสระ นำมาตรวจหา Peak ของสเปคตรัมในช่วง 50-500 Hz เพื่อนำมาเป็นรหัสข้อมูลของเสียงวรรณยุกต์ มีขั้นตอนการทำงานตามรูปที่ 4.4 ดังนี้

1. ทำการหาเฟรมที่มีพลังงานสูงสุดจากจำนวนเฟรมที่มีอยู่ทั้งหมดในไฟล์เสียงนั้น
2. ทำการตรวจหาความถี่ของ Peak ของสเปคตรัมในช่วงความถี่ 50 - 500 Hz ของแฟรมที่มีพลังงานสูงสุด
3. ทำการตรวจหาความถี่ของ Peak ของสเปคตรัมในช่วงครึ่งเวลาแรกของเฟรมถัดไป โดยช่วงพิสัยการตรวจจะไม่เกินไปจากช่วงความถี่ที่พบ Peak ไม่เกิน 2 Sample นับจากจุดนั้น หรือประมาณ 30 Hz จนกว่าค่า Magnitude ที่ได้จะมีค่าต่ำกว่าค่าที่กำหนด จึงทำการหา Peak ในช่วงครึ่งหลังต่อไป ซึ่งรหัสวรรณยุกต์จะเก็บตำแหน่งของเวลา และความถี่เอาไว้ เพื่อวิเคราะห์ลักษณะของวรรณยุกต์ต่อไป
4. ทำการตรวจหาในครึ่งหลังเช่นเดียวกับข้อ 3. เพื่อเก็บรหัสวรรณยุกต์ จนกว่าค่า Magnitude จะมีค่าต่ำกว่าค่าที่กำหนด



รูปที่ 4.3 ขั้นตอนการทำงานของการหารหัสวรรณยุกต์

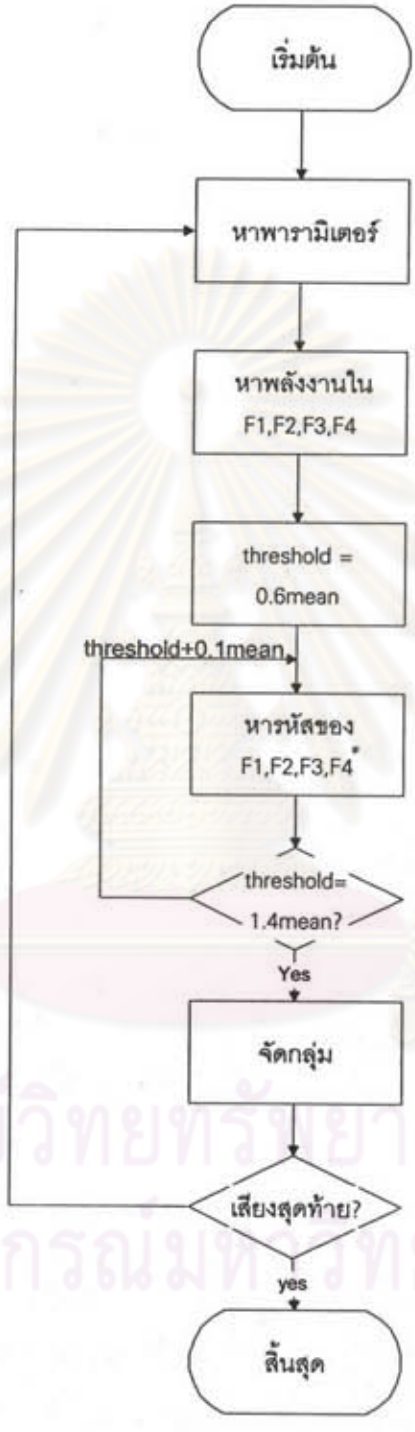


รูปที่ 4.4 ขั้นตอนการทำงานของการจำแนกเสียงวรรณยุกต์

ส่วนระบบรู้จำเสียงวรรณยุกต์

ในส่วนของระบบรู้จำเสียงวรรณยุกต์นั้นจะใช้วิธีการทำ Curve Fitting ระหว่างรหัสวรรณยุกต์ที่จะทดสอบกับสมการแบบอ้างอิงของเสียงวรรณยุกต์ 4 สมการ ซึ่งค่า MSE ที่ได้จากการทำ Curve Fitting จะถูกนำไปตรวจสอบลักษณะของวรรณยุกต์เพื่อยืนยันอีกครั้ง ดังขั้นตอนในรูปที่ 4.4

1. หาค่าพารามิเตอร์จากแบบทดสอบเพื่อนำมาสร้างเป็นสมการของแบบอ้างอิง
2. นำพารามิเตอร์มาสร้างสมการตามสมการที่ 3.19, 3.22, 3.20 และ 3.23
3. นำรหัสวรรณยุกต์ที่ได้ไปทำ Curve Fitting กับสมการอ้างอิงที่ 3.19, 3.20, 3.22 และ 3.23
4. ถ้ารหัสวรรณยุกต์ที่นำมาทดสอบใกล้เคียงกับสมการ 3.19 ที่สุด จึงทำการตรวจว่ามีการคงระดับหรือมีความถี่คงที่มากกว่า 40% ของช่วงเวลาทั้งหมด และเกิดตั้งแต่ช่วงต้น ๆ ของระยะเวลาทั้งหมดแบบทดสอบจะเป็นวรรณยุกต์สามัญ ถ้าเกิดในช่วงท้าย หรือ มี Pitch Range น้อยกว่าค่าที่กำหนด จะเป็นวรรณยุกต์เอก แต่ถ้ามี Pitch Range มากกว่าค่าที่กำหนด จะเป็นวรรณยุกต์โท
5. ถ้ารหัสวรรณยุกต์ที่นำมาทดสอบใกล้เคียงกับสมการ 3.20 ที่สุด จึงจะทำการตรวจว่ามี Pitch Range มากกว่าค่าที่กำหนดหรือไม่ ถ้าใช่จะเป็นวรรณยุกต์จัตวา ถ้าไม่ใช่จะเป็นวรรณยุกต์ตรี
6. ถ้ารหัสวรรณยุกต์ที่นำมาทดสอบใกล้เคียงกับสมการ 3.22 ที่สุด จะทำการตรวจว่ามี Pitch Range น้อยกว่าค่าที่กำหนดหรือไม่ ถ้าใช่จะไปตรวจว่าเป็นวรรณยุกต์เอกหรือสามัญ ถ้าไม่ใช่จะเป็นวรรณยุกต์โท
7. ถ้ารหัสวรรณยุกต์ที่นำมาทดสอบใกล้เคียงกับสมการ 3.23 ที่สุด จะทำการตรวจว่ามี Pitch Range น้อยกว่าค่าที่กำหนดหรือไม่ ถ้าไม่ใช่จะเป็นวรรณยุกต์จัตวา ถ้าใช่จะเป็นวรรณยุกต์ตรี



รูปที่ 4.5 ขั้นตอนการทำงานของการจัดกลุ่มอ้างอิง

ส่วนการจัดกลุ่มแบบอ้างอิง

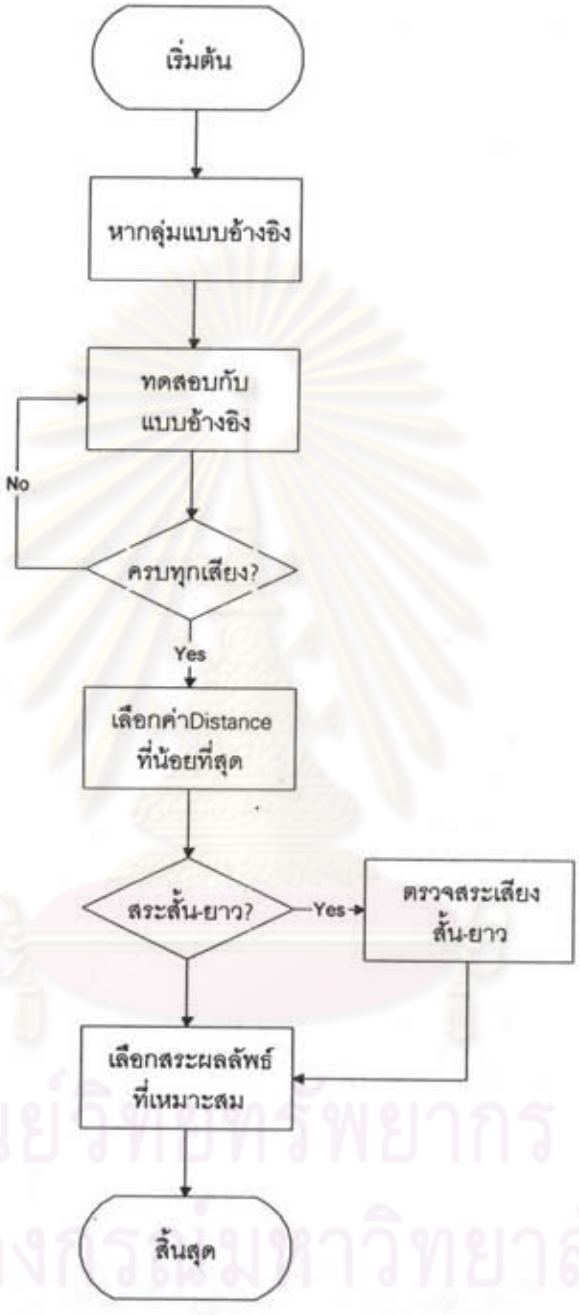
ในส่วนนี้จะเป็นการช่วยลดจำนวนแบบอ้างอิงที่นำมาทดสอบให้น้อยลง ใช้เวลาในการทดสอบน้อยลง โดยจะทำการแบ่งช่องความถี่ของข้อมูลที่มีความกว้าง 4 kHz ออกเป็น 4 ช่วง โดยจะทำการตรวจหาว่าในแต่ละ ช่วงความถี่ F_1 = 0 - 1 kHz, F_2 = 1 - 2 kHz, F_3 = 2 - 3 kHz และช่วง F_4 = 3 - 4 kHz ว่ามีพลังงานในแถบความถี่ย่อยเกินค่าที่กำหนดในแต่ละช่วงหรือไม่ ซึ่งในแต่ละช่วงจะถูกแบ่งออกเป็นแถบความถี่ย่อย ๆ 8 แถบ ถ้าแถบความถี่ย่อยใด ๆ ในแต่ละช่วงความถี่ F_1, F_2, F_3, และ F_4 มีพลังงานเกินค่าที่กำหนด ช่วงนั้นจะมีค่าเป็น 1 และถ้ามีค่าต่ำกว่า จะเป็น 0 ดังรายละเอียดขั้นตอนต่อไปนี้

1. หาค่าพารามิเตอร์ต่าง ๆ เช่น ค่าพลังงานเฉลี่ยทั้งหมด
2. หาพลังงานในแต่ละแถบความถี่ย่อย ๆ ในช่วงความถี่ F_1, F_2, F_3, และ F_4
3. กำหนดค่า threshold โดยเริ่มจาก 0.6 mean ของพลังงานทั้งหมด แล้วตรวจค่าพลังงานแต่ละแถบความถี่ย่อยว่าเกิน threshold หรือไม่ ตั้งแต่ 1 -2 kHz, 2 - 3 kHz และ 3 - 4 kHz แล้วเพิ่มค่า threshold ครั้งละ 0.1 จนมีค่าเท่ากับ 1.4 mean
4. ค่าผลลัพธ์ที่ได้จากการตรวจจะเป็น 1 เมื่อมีค่าเกิน threshold และเป็น 0 มีค่าต่ำกว่า threshold
5. นำเสียงที่ใช้เป็นแบบอ้างอิงถัดมาจัดเข้ากลุ่ม จนครบทุกเสียง

4.10 ส่วนการทดสอบการรู้จำเสียงสระ

เป็นขั้นตอนการตัดสินใจสำหรับระบบรู้จำเสียงสระ โดยใช้การวัดค่าระยะห่าง (Distance) ของเสียงแบบทดสอบกับแบบอ้างอิง มีขั้นตอนดังนี้

1. หากลุ่มแบบอ้างอิงเพื่อใช้ทดสอบกับแบบทดสอบ
2. ทดสอบแบบทดสอบกับแบบอ้างอิงต่าง ๆ ในกลุ่มที่เลือกไว้ทั้งหมด โดยการหาค่า Distance ที่น้อยที่สุด
3. เมื่อได้สระที่มีค่า Distance น้อยที่สุดแล้ว ทำการตรวจว่าเป็นสระที่อยู่ในกลุ่มสระเสียงสั้น - ยาวหรือไม่ โดยการตรวจดูจำนวนเฟรมของแบบทดสอบว่ามีค่าอยู่ในช่วงของสระเสียงสั้นหรือเสียงยาว เพื่อใช้พิจารณาผลลัพธ์อีกครั้ง
4. แบบอ้างอิงที่มีค่าระยะห่างน้อยที่สุดเป็นจะถูกพิจารณาร่วมกับข้อ 3. ในการตัดสินใจให้เป็นเสียงสระที่รู้จำได้



รูปที่ 4.6 ขั้นตอนตัดสินใจการรู้จำเสียงสระ