การหาแบบจำลองที่เหมาะสมสำหรับการรู้จำเสียงวรรณยุกต์ภาษาไทย โดยใช้วิธีการออโตคอร์รีเลชั่นแบบเซ็นเตอร์คลิปปิ้ง สำหรับการหาค่าพิทช์ในสภาวะที่มีสัญญาณรบกวนเกาส์เซียนขาว

OPTIMIZATION OF THAI TONE RECOGNITION USING AUTOCORRELATION WITH CENTER CLIPPING METHOD FOR PITCH EXTRACTION IN THE PRESENCE OF WHITE GAUSSIAN NOISE

กฤษกร สุนทรมนทกานติ

KITSAKORN SOONTORNMONTAKANTI

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ.2548

ISBN 974-15-1613-4

OPTIMIZATION OF THAI TONE RECOGNITION USING AUTOCORRELATION WITH CENTER CLIPPING METHOD FOR PITCH EXTRACTION IN THE PRESENCE OF WHITE GAUSSIAN NOISE

KITSAKORN SOONTORNMONTAKANTI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING
SCHOOL OF GRADUATE STUDIES
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2005

ISBN 974-15-1613-4

หัวข้อวิทยานิพนธ์	การหาแบบจำลองที่เหมาะสมสำหรับการรู้จำเสียงวรรณยุกต์ภาษาไทย โดยใช้วิธีการออโตคอร์รีเลชั่นแบบเซ็นเตอร์คลิปปิ้งสำหรับการหาค่าพิทช์ในสภาวะที่มีสัญญาณรบกวนเกาส์เซียนขาว
นักศึกษา	นายกฤษกร สุนทรมนทกานติ
รหัสประจำตัว	43061071
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
พ.ศ.	2548
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร. ไกรสิน ส่งวัฒนา

บทคัดย่อ

วิทยานิพนธ์นี้เสนอวิธีการ Pitch Extraction ในการรู้จำหน่วยเสียงวรรณยุกต์สำหรับภาษาไทยในสภาวะแวดล้อมที่มีเสียงรบกวน โดยใช้ออโตคอร์รีเลชั่นที่มีขั้นตอนการคลิปยอดของสัญญาณ (Autocorrelation Method using Center Clipping : AUTOC) ในการคำนวณหาค่าคาบเวลาพิทช์ ซึ่งจะทำการคลิปยอดของสัญญาณในระดับที่ต่างกัน ค่าคาบเวลาพิทช์ที่ได้จากการคลิปยอดของสัญญาณในระดับที่ต่างกันนี้ จะถูกแปลงเป็นค่าความถี่มูลฐานซึ่งเป็นตัวบ่งชี้ระดับสูง-ต่ำของเสียง ซึ่งลำดับของความถี่มูลฐานที่ได้จะถูกปรับปรุงให้มีความต่อเนื่องของข้อมูลโดยใช้มีเดียนฟิลเตอร์ จากนั้นทำการหาค่าการเปลี่ยนแปลงของความถี่มูลฐานของพิทช์นั้นๆเทียบกับเวลา โดยทำการควอนไตซ์การเบี่ยงเบนออกเป็น 3 ระดับตามทิศทางการเพิ่มขึ้นหรือลดลงของความถี่มูลฐาน ซึ่งค่าที่ได้จากการควอนไตซ์นี้จะถูกนำไปใช้เป็นข้อมูลฝึกสอนให้กับการสร้างแบบจำลองการรู้จำของหน่วยเสียงวรรณยุกต์ทั้ง 5 ระดับด้วยวิธี Hidden Markov Model (HMM)

งานวิจัยนี้ได้ทำการสอดแทรกสัญญาณรบกวนที่เป็นสัญญาณรบกวนเกาส์เซียนขาวเข้าไปในเสียงพูดวรรณยุกต์ในภาษาไทยพยางค์เดียวทั้ง 5 ระดับ เพื่อทดสอบความทนทานในการรู้จำเสียงวรรณยุกต์ในภาษาไทยต่อเสียงรบกวนและหาค่าการคลิปยอดของสัญญาณที่เหมาะสมสำหรับการหาค่าคาบเวลาพิทช์ในสภาวะที่มีเสียงรบกวน โดยใช้ข้อมูลเสียงที่ได้จากเพศชาย 5 คนและเพศหญิง 5 คน ผลที่ได้คือการคลิปยอดของสัญญาณที่ระดับ 30 เปอร์เซ็นต์ให้ผลการรู้จำเสียงวรรณยุกต์ภาษาไทยแม่นยำถึง 90.00 เปอร์เซ็นต์ ในระดับสัญญาณต่อสัญญาณรบกวนที่ 10 dB

Thesis Title	Optimization of Thai Tone Recognition using Autocorrelation with Center Clipping Method for Pitch Extraction in the Presence of White Gaussian Noise
Student	Mr. Kitsakorn Soontornmontakanti
Student ID.	43061071
Degree	Master of Engineering
Programme	Electrical Engineering
Year	2005
Thesis Advisor	Assoc.Prof.Dr. Kraisin Songwatana

ABSTRACT

This thesis presents pitch extraction method for recognition of Thai tone in a noisy environment. The Autocorrelation using Center Clipping (AUTOC) is used to find the pitch period with different levels of clipping. The sequence of fundamental frequency from each word is smooth out by using median filter. The observed sequence of pitch levels are preprocessed to find the pitch differences and the sequence of pitch differences are then grouped into three quantized levels. The resultant sequence is used as bases for training a Hidden Markov Model and recognition of 5 tones.

The white gaussian noise is added into the in isolated monosyllable Thai word to find optimal estimation clipping level of signal for pitch period calculation. The tests are carried out on 5-male and 5-female Thai subject. It shows that clipping level at 30 percent is accuracy of recognition to 90 percent at signal to noise ratio of 10 dB.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยดีจากการให้คำแนะนำ คำปรึกษาและความช่วยเหลือจาก รศ.ดร.ไกรสิน ส่งวัฒนา ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้เขียนจึงขอกราบขอบพระคุณเป็นอย่างสูง

- ขอขอบคุณ พ่อ แม่ ที่ให้การสนับสนุน ส่งเสริม คอยให้กำลังใจ เป็นห่วงเป็นใย ช่วยเหลือดูแลกับลูกด้วยดีตลอดมา
- ขอขอบคุณอาจารย์ทุกท่านที่ให้คำแนะนำและให้คำปรึกษา ตลอดระยะเวลาที่ศึกษาอยู่
- ขอขอบคุณ พี่ๆ เพื่อนๆ ทุกท่านทั้งที่ห้องวิจัย ที่ทำงาน และที่เรียนด้วยกัน ที่คอยให้กำลังใจ ดูแลเอาใจใส่อย่างสม่ำเสมอ
- ขอขอบคุณ คุณสมศักดิ์ ตรียากิจที่ให้ยืม notebook ตลอดการทำวิทยานิพนธ์ฉบับนี้
- ขอขอบคุณเจ้าของเสียงทุกๆท่าน ที่ได้เสียสละเวลามาบันทึกเสียงให้
- สุดท้ายนี้ขอขอบคุณบัณฑิตวิทยาลัย ที่ได้ให้การสนับสนุนการทำวิทยานิพนธ์ครั้งนี้

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ขอขอบคุณด้วยความจริงใจ
กฤษกร สุนทรมนทกานติ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

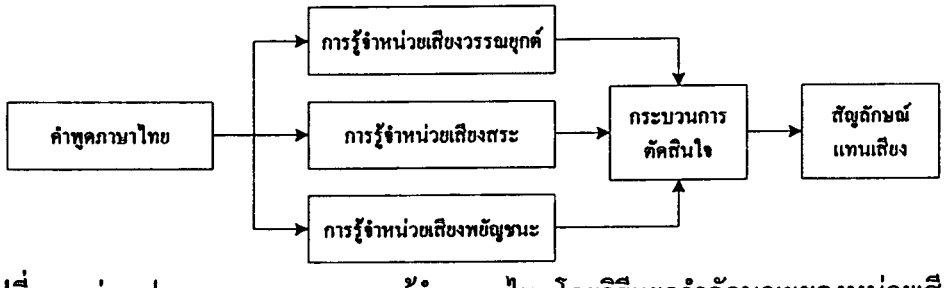
1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเทคโนโลยีทางด้านการคอมพิวเตอร์ได้ถูกพัฒนาอย่างต่อเนื่องและให้มีขีดความสามารถมากขึ้น และยังสามารถนำมาใช้ร่วมกับเทคโนโลยีในด้านต่างๆ เพื่อทำให้การประมวลผลเป็นไปอย่างรวดเร็วและถูกต้อง การพัฒนาให้คอมพิวเตอร์สามารถรับรู้คำสั่งจากเสียงพูดของมนุษย์เป็นอีกเทคโนโลยีหนึ่งที่น่าสนใจซึ่งจะทำให้การติดต่อสื่อสารระหว่างมนุษย์กับคอมพิวเตอร์ทำได้ง่ายและสะดวกขึ้น

จากความต้องการให้เครื่องคอมพิวเตอร์สามารถรับรู้เสียงพูดได้ จึงทำให้เกิดศาสตร์แขนงหนึ่งเรียกว่า "Speech Recognition" แต่เนื่องจากการพูดของมนุษย์มีความซับซ้อน และมีความแตกต่างกันในแต่ละบุคคลจึงทำให้การพัฒนาเป็นไปอย่างล่าช้า โดยสามารถแบ่งวิธีการรู้จำเสียงพูดออกได้เป็น 2 วิธี คือ

1. พิจารณาทั้งหน่วยภาษาที่เปล่งเสียงออกมาทั้งชุด มีทั้งระบบการรู้จำคำเดี่ยว[1]-[2] (Isolated word Recognition) และระบบรู้จำคำพูดต่อเนื่อง (Continuous word Recognition) ซึ่งข้อดีของระบบเหล่านี้คือ ง่าย เนื่องจากมีการหลีกเลี่ยงผลกระทบอันเนื่องจากฐานของเสียงภายในคำหรือกลุ่มคำนั้น แต่ข้อเสีย คือสามารถรู้จำคำได้ในจำนวนคำที่จำกัด เนื่องจากต้องใช้เนื้อที่จำนวนมากในการจัดเก็บแบบจำลองอ้างอิง และต้องใช้เวลาในการคำนวณเพื่อเปรียบเทียบมากตามจำนวนของแบบจำลองอ้างอิงที่มีอยู่

2. พิจารณาโดยการแยกแยะรายละเอียดของหน่วยเสียง (Phonetic Recognition)[3]-[5] วิธีนี้จะพิจารณาลักษณะของหน่วยเสียงที่มีขนาดเล็กลงไป เช่น หน่วยเสียงพยัญชนะ หน่วยเสียงสระ และหน่วยเสียงวรรณยุกต์ ดังแสดงในรูปที่ 1.1 โดยจะใช้หน่วยเสียงย่อยเหล่านี้เป็นหลักในการรู้จำเสียงพูด ซึ่งวิธีนี้เหมาะสำหรับการพัฒนาไปสู่ระบบการรู้จำคำจำนวนมาก



รูปที่ 1.1 ส่วนประกอบของระบบการรู้จำภาษาไทยโดยวิธีแยกจำลักษณะของหน่วยเสียง

พิจารณาภาษาไทยจะเห็นว่าหน่วยเสียงวรรณยุกต์ หน่วยเสียงสระและหน่วยเสียงพยัญชนะล้วนแต่เป็นส่วนประกอบที่สำคัญที่ทำให้เกิดคำและความหมายขึ้นมาใช้ในภาษาไทย แต่ด้วยลักษณะของหน่วยเสียงวรรณยุกต์มีหน้าที่ๆทำให้เกิดคำขึ้นใหม่ในภาษาไทยมากขึ้น และถ้าเปลี่ยนเสียงวรรณยุกต์ก็จะทำให้เกิดคำที่มีความหมายใหม่เพิ่มขึ้น จึงเลือกพิจารณาหน่วยเสียงวรรณยุกต์

จากการศึกษาระบบการรู้จำเสียงพูดสำหรับภาษาไทย ปัญหาที่ไม่สามารถมองข้ามไปอีกอย่างหนึ่งคือ สัญญาณรบกวน (Noise) เพราะในการใช้งานจริงย่อมเกิดสัญญาณรบกวนได้ทุกที่และลักษณะของสัญญาณรบกวนมีความแตกต่างกัน ปัญหานี้จึงเป็นปัญหาที่ไม่สามารถหลีกเลี่ยงได้ซึ่งมีผลกับระบบการรู้จำเสียงพูดมีประสิทธิภาพที่ด้อยลง

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

วิทยานิพนธ์ฉบับนี้ มีวัตถุประสงค์ที่จะศึกษาผลกระทบของสัญญาณรบกวนที่มีต่อระบบการรู้จำเสียงภาษาไทยและศึกษาวิธีการที่จะลดผลกระทบอันเนื่องมาจากสัญญาณรบกวนโดยมุ่งเน้นไปที่ระบบการรู้จำเสียงหน่วยวรรณยุกต์ในภาษาไทย ในลักษณะคำโดด(Monosyllabic)หรือคำพยางค์เดียว

1.3 สมมุติฐานของการศึกษา

จากปัญหาและวัตถุประสงค์ข้างต้น ที่ต้องการศึกษาผลกระทบของสัญญาณรบกวนที่มีต่อระบบการรู้จำเสียง ทำให้คิดว่าจะมีขั้นตอนหรือวิธีไหนบ้างที่จะลดผลกระทบอันเนื่องมาจากสัญญาณรบกวนในระบบการรู้จำเสียง โดยขั้นตอนแรกในการวิเคราะห์สัญญาณเสียงคือการหาค่าพิทช์ ซึ่งในขั้นตอนนี้มีผลอย่างมากทั้งการแยกแยะคำและนำมาสร้างแบบจำลองอ้างอิง โดยถ้าสามารถลดผลกระทบของสัญญาณรบกวนในขั้นตอนนี้ได้ ระบบการรู้จำเสียงน่าจะมีประสิทธิภาพที่ดีในสภาวะที่มีเสียงรบกวน

1.4 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

เสียงวรรณยุกต์ภาษาไทยมีระดับเสียงสูง-ต่ำ(Pitch)ที่แตกต่างกัน ซึ่งเกิดจากการสั่นสะบัดของเส้นเสียง จากคุณลักษณะดังกล่าวเป็นวิธีที่จะช่วยแยกแยะคำในภาษาไทย โดยในการหาค่าพิชท์ (Pitch Extraction) มีวิธีในการหาได้หลายวิธี ในวิทยานิพนธ์ฉบับนี้ได้ใช้วิธีออโตคอร์รีเลชั่นฟังก์ชั่นโดยใช้เทคนิคการคลิปยอดของสัญญาณ (Autocorrelation Method using Center Clipping : AUTOC) โดยทำการกำหนดระดับการคลิปยอดของสัญญาณในระดับต่างกันเพื่อหาค่าการคลิปยอดของสัญญาณที่เหมาะสมและสามารถทำให้ระบบการรู้จำเสียงมีประสิทธิภาพที่ดีในสภาวะที่มีเสียงรบกวน

1.5 ขอบเขตการวิจัย

1. งานวิจัยนี้มุ่งศึกษาพัฒนาอัลกอริธึมบนเครื่องคอมพิวเตอร์ส่วนบุคคล โดยใช้โปรแกรม Microsoft Visual C++ 6.0
2. ข้อมูลเสียงที่ใช้ในวิทยานิพนธ์ฉบับนี้ ได้ใช้คำพยางค์เดียวที่มีพยัญชนะต้น สระ และพยัญชนะสะกดที่แตกต่างกันจำนวนทั้งสิ้น 100 คำ ใช้ผู้ออกเสียงผู้ชาย 5 คนและผู้หญิง 5 คน เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี โดยมีการสอดแทรกสัญญาณรบกวนที่เป็นสัญญาณรบกวนเกาส์เซียนขาว (White Gaussian Noise) ในระดับสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio:SNR) ที่ 40dB, 35dB, 30dB, 25dB, 20dB, 15dB, 10dB และ 5dB
3. การทดลองได้ทำการคำนวณหาค่าพิทช์ใช้วิธีออโตคอร์รีเลชั่นฟังก์ชั่นโดยใช้เทคนิคการคลิปยอดของสัญญาณในระดับที่ 80%, 70%, 60%, 50%, 40%, 30% และ 20% ของแอมปลิจูดสูงสุดของสัญญาณ แล้วนำมาสร้างแบบจำลองอ้างอิงการรู้จำเสียงวรรณยุกต์ในแต่ละระดับของการคลิปยอดของสัญญาณ
4. ทำการทดสอบกับแบบจำลองอ้างอิงการรู้จำเสียงวรรณยุกต์ในแต่ละระดับของการคลิปยอดของสัญญาณ โดยใช้ข้อมูลเสียงในการทดสอบจากผู้ออกเสียงผู้ชาย 5 คนและผู้หญิง 5 คน เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี ที่มีการสอดแทรกสัญญาณรบกวนขาวที่ระดับสัญญาณต่อสัญญาณรบกวนตามหัวข้อข้างต้น

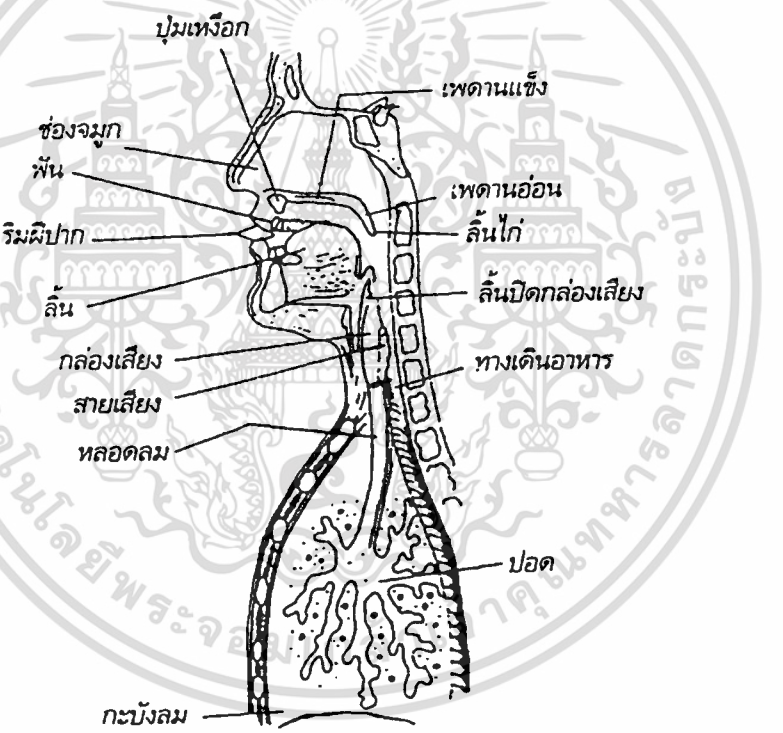
1.6 ขั้นตอนของการศึกษา

1. ศึกษาระบบการรู้จำภาษาในลักษณะต่างๆที่ได้มีการศึกษามาแล้ว
2. ศึกษาทฤษฎีทางด้านภาษาศาสตร์ และอัลกอลิธึมต่างๆที่ใช้ในการวิเคราะห์เสียงพูด
3. กำหนดขอบเขตการวิจัย
4. บันทึกข้อมูลเสียง โดยใช้ผู้พูดทั้งหมด 10 คน ผู้ชาย 5 คนและผู้หญิง 5 คน เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี
5. เก็บข้อมูลวิเคราะห์เสียง และทดสอบความถูกต้อง
6. สรุปผลการวิจัย

บทที่ 2
ระบบเสียงในภาษาไทย

2.1 ทฤษฎีการสร้างเสียงพูด

การพูดของมนุษย์มิใช่อาการที่เกิดเฉพาะที่ปากเท่านั้น หากเริ่มจากลมหายใจเข้าของมนุษย์เองที่นำลมเข้าสู่ปอด จากนั้นจะใช้ลมจากปอดซึ่งก็คือลมหายใจออก มาทำให้เกิดเสียงพูด โดยลมจะถูกบังคับให้ผ่านอวัยวะต่างๆที่สำคัญ คือ เส้นเสียงซึ่งอยู่ในช่องของหลอดลม หรือบริเวณที่เรียกว่าลูกกระเดือก แล้วลมนั้นจะผ่านต่อออกมาทางช่องปากหรือช่องจมูก อวัยวะในช่องปากก็จะดัดแปลงให้เป็นเสียงต่างๆตามที่ผู้พูดต้องการ [7]



รูปที่ 2.1 ภาพตัดขวางแสดงอวัยวะในระบบการพูดของมนุษย์

2.1.1 อวัยวะที่ใช้ในการออกเสียงพูด

อวัยวะส่วนที่มีหน้าที่โดยตรงในการออกเสียงพูด ดังแสดงในรูปที่ 2.1 มีดังนี้คือ

1. ริมฝีปาก เป็นอวัยวะส่วนที่เคลื่อนไหวได้มาก และทำให้เสียงแตกต่างกันได้มาก เราอาจจะบังคับริมฝีปากให้อยู่ชิดกัน ห่างกัน ยื่นออก หรือห่อกลม ฯลฯ ลักษณะริมฝีปากต่างๆนี้ล้วนแต่มีอิทธิพลต่อการออกเสียง และการทำให้เสียงแตกต่างกันไปทั้งสิ้น

2. ฟัน เป็นอวัยวะที่เกิดของเสียงหลายชนิด เช่นเมื่อฟันบนกดลงบนริมฝีปากล่าง หรือฟันล่าง ลมที่ผ่านออกมาโดยแรงจะลอดช่องที่พอผ่านได้ออกมา ทำให้เกิดเป็นเสียงชนิดที่เรียกว่าเสียงเสียดแทรก เป็นต้น

3. ปุ่มเหงือก เป็นส่วนนูนออกมาอยู่หลัง ฟันด้านบน ถ้าเอาลิ้นแตะดูจะรู้สึกว่ามีลักษณะเป็นคลื่น

4. เพดานแข็ง หรือ เพดานปาก คือ ส่วนเฉพาะที่โค้งเป็นกระดูกแข็ง

5. เพดานอ่อน คือ ส่วนของเพดานที่อยู่ต่อจากเพดานแข็งไปข้างในมีลักษณะเป็นกระดูกอ่อนที่ขยับขึ้น-ลงได้ เวลาหายใจเพดานอ่อนและลิ้นไก่ซึ่งอยู่ตอนปลายจะลดระดับลงมาเปิดช่องให้ลมออกไปทางจมูก เวลาพูดส่วนใหญ่ปลายเพดานอ่อนและลิ้นไก่จะถูกยกขึ้นไปจรดกับหลังคอ นอกจากเวลาออกเสียงนาสิกเท่านั้นที่เพดานอ่อนจะลดระดับลงมา เพื่อให้ลมออกทางช่องจมูก

6. ลิ้นไก่ เป็นก้อนเนื้อเล็กๆอยู่ต่อปลายเพดานตรงกลางปาก อวัยวะส่วนนี้สั่นรัวได้

7. ลิ้น เป็นส่วนที่เคลื่อนไหวมากที่สุดในการออกเสียงพูด จึงต้องแบ่งออกเป็น 3 ส่วนตามหน้าที่ในการออกเสียง คือ

7.1) ปลายลิ้น คือ ส่วนปลายลิ้นซึ่งสามารถยกขึ้นไปแตะอวัยวะส่วนต่างๆในปากตอนบนได้โดยง่าย

7.2) หน้าลิ้น คือ ลิ้นที่อยู่ตรงข้ามกับเพดานแข็ง

7.3) หลังลิ้น คือ ส่วนของลิ้นที่อยู่ตรงข้ามกับเพดานอ่อน

8. แผ่นเนื้อปากหลอดลม เป็นก้อนเนื้อเล็กๆคล้ายลิ้นไก่ อยู่ต่อโคนลิ้นลงไปในคอ มีหน้าที่ปิดช่องลมเมื่อรับประทานอาหาร และเปิดช่องลมเมื่อพูด

9. กรวยคอ หมายถึง โพรงคอที่อยู่ถัดจากปากลงไปจนถึงเส้นเสียง

10. เส้นเสียง หรือ สายเสียง เป็นอวัยวะสำคัญที่เกิดของเสียง เส้นเสียงมีลักษณะเป็นกล้ามเนื้อ 2 แผ่นปิดขวาง อยู่บริเวณปากช่องหลอดลมจากด้านหลังมาด้านหน้า ระหว่างเส้นเสียงจะมีช่องว่าง ซึ่งเป็นทางผ่านให้ลมเข้าถึงปอดและออกมาจากปอดได้ ช่องว่างนี้เรียกว่า ช่องระหว่างเส้นเสียง (Glottis) เส้นเสียงทั้งสองสามารถดึงออกให้ห่างจากกันหรือดึงเข้าหากันได้ ซึ่งเส้นเสียงนี้เป็นส่วนสำคัญที่ทำให้เกิดเสียงพูดขึ้นในภาษา

11. ช่องจมูก หมายถึง โพรงในช่องจมูก ซึ่งอยู่เหนือลิ้นไก่ขึ้นไป เป็นช่องที่ลมซึ่งผ่านเส้นเสียงขึ้นมาจะผ่านออกไปทางจมูกได้เมื่อเวลาหายใจและเวลาออกเสียงนาสิก

12. เส้นเสียงปลอม เป็นอวัยวะที่มีลักษณะเหมือนเส้นเสียงแต่อยู่เหนือเส้นเสียงขึ้นไป เส้นเสียงปลอมนี้เข้าใจกันว่าจะดึงเข้าหากันเมื่อเวลาพูดเสียงกระซิบ

2.1.2 ลักษณะร่วมของเสียงพูด

เสียงที่ใช้ในภาษาพูดนั้นจะมีลักษณะที่สำคัญบางประการร่วมกัน ซึ่งเรียกได้ว่าเป็นลักษณะร่วมของเสียงพูด ลักษณะที่กล่าวถึงนี้มีอยู่หลายประการ [8] คือ

1. ความก้อง หรือ ไม่ก้องของเสียง

เสียงก้อง หรือ เสียงโฆษะ (Voice)

คือเสียงที่เกิดในขณะที่เส้นเสียงเกิดการตึงตัวหรือเรียกว่าเส้นเสียงปิด เมื่อมีแรงดันให้อากาศไหลผ่านกล่องเสียงในขณะที่เส้นเสียงปิดจะเกิดการสั่นสะบัดของเส้นเสียง เป็นผลทำให้สัญญาณเสียงที่ได้ (speech waveform) มีลักษณะเป็นคาบ (quasi-periodic) ซึ่งสามารถเรียกความถี่ในการปิด-เปิดของเส้นเสียงนี้ว่า "ความถี่มูลฐาน" (Fundamental Frequency: F_0) ตัวอย่างของเสียงก้องได้แก่ เสียงสระต่างๆ และเสียงพยัญชนะเช่น บ ด ที่เกิดจากการเปล่งเสียงออกทางปาก หรือเสียงพยัญชนะ ม น ง ที่เกิดจากการเปล่งเสียงออกทางจมูก

เสียงไม่ก้อง หรือเสียงอโฆษะ (Unvoice หรือ voiceless)

คือเสียงที่เกิดในขณะที่เส้นเสียงคลายจากการตึงหรือเรียกว่าเส้นเสียงเปิด เมื่อมีแรงดันให้อากาศไหลผ่านกล่องเสียงในขณะที่เส้นเสียงเปิด อากาศที่ไหลผ่านอย่างรวดเร็วจะเกิดการไหลวนและปั่นป่วนทำให้เกิดเสียงที่มีลักษณะเป็นเสียงของสัญญาณรบกวน (Noise) ซึ่งไม่เป็นคาบ ตัวอย่างของเสียงไม่ก้องได้แก่ เสียงพยัญชนะ ฟ ซ ส ฯลฯ หรือเกิดจากการสร้างแรงดันอากาศหลังตำแหน่งปิดกั้นของช่องทางเดินเสียง และเมื่อการปิดกั้นนี้ถูกเปิดออก อากาศจะถูกปล่อยออกมาอย่างทันทีทันใดเกิดเป็นเสียงที่เรียกว่าเสียงระเบิด (Plosive Sound) เช่น การเปล่งเสียงเริ่มแรกของพยัญชนะต้นของคำต่างๆ

2. ความยาวของเสียง (Length)

หมายถึง การที่เสียงใดเสียงหนึ่งเปล่งออกมาได้นานเท่าใด เสียงพูดบางเสียงอาจจะเปล่งออกมาได้ติดต่อกันได้นาน เช่น เสียงสระ เสียงพยัญชนะนาสิก หรือ เสียงพยัญชนะเสียดแทรก

ในภาษาไทย เสียงพูดที่มีความยาว-สั้น ก็มีเพียงเสียงสระเท่านั้น เช่น อะ อิ อุ เป็นเสียงสั้น อา อี อู เป็นเสียงยาวเป็นต้น

3. ระดับเสียงสูง-ต่ำ (Pitch)

เสียงพูดจะมีระดับ สูง หรือ ต่ำ อยู่ที่ความถี่ของเสียง (Fundamental frequency) ถ้าความถี่ต่ำเสียงก็จะต่ำ อวัยวะส่วนที่ทำให้เสียงมีระดับ สูง-ต่ำ คือเส้นเสียง ดังนั้นระดับเสียงสูง-ต่ำก็คืออัตราการสั่นสะบัดของเส้นเสียงนั่นเอง

ในการพูดเสียงที่มีระดับสูง-ต่ำได้คือเสียงก้องเท่านั้นเพราะมีการสั่นสะเทือนของเส้นเสียงที่ทำให้เกิดมีความถี่ระดับต่างๆได้ ในภาษาไทยระดับเสียง สูง-ต่ำ ของคำเราเรียกว่า "วรรณยุกต์"

4. ความดัง (Loundness)

ความดังขึ้นอยู่กับปริมาณของลม ที่ผู้พูดเปล่งเสียงออกมาในช่วงเวลาหนึ่งๆ

5. การลงน้ำหนัก (Stress)

หมายถึง การออกเสียงพยางค์ใดพยางค์หนึ่งให้ดังเน้นมากหรือน้อยกว่าพยางค์อื่นที่อยู่ข้างเคียง (เพื่อต้องการเรียกร้องความสนใจเป็นพิเศษ หรือแสดงอารมณ์อย่างใดอย่างหนึ่ง)

6. ช่วงต่อของเสียง (Juncture)

หมายถึงช่วงระยะที่ผู้พูดเปล่งเสียงหนึ่งแล้วต่อไปเปล่งอีกเสียงหนึ่งซึ่งเรียงกันมาเป็นลำดับ เสียงที่ประกอบกันเข้าเป็นพยางค์จะมีช่วงต่อของเสียงแนบสนิทจนไม่เห็นร่องรอย (close juncture) แต่ถ้าเสียงปรากฏอยู่คนละพยางค์หรือคนละคำ จะมีช่วงต่อ "ห่าง" จนสังเกตได้ชัด (open juncture) ดังนั้นช่วงต่อของเสียง โดยเฉพาะช่วงต่อห่างจะมีความสำคัญมากในการแบ่งคำในภาษา

2.2 หน่วยเสียงสำคัญในภาษาไทย

"หน่วยเสียง" (phoneme) เป็นหน่วยเล็กที่สุดของภาษา หน่วยดังกล่าวได้แก่เสียงสำคัญๆ ในภาษาใดภาษาหนึ่ง ซึ่งทำหน้าที่ให้ความหมายของคำที่ใช้ในภาษานั้น และทำให้ความหมายของคำนั้นๆมีความหมายแตกต่างจากคำอื่นๆ หน่วยเสียงสำคัญในภาษาไทยมี 3 ประเภทใหญ่ๆคือ เสียงพยัญชนะ เสียงสระ และเสียงวรรณยุกต์ หน่วยเสียงทั้ง 3 นี่เองที่ประกอบกันเข้าเป็นคำที่ใช้ในภาษาไทย

เสียงพูดของมนุษย์ซึ่งมีความแตกต่างกันมากมายนั้นถ้าเราพิจารณาอย่างกว้างๆจะพบว่าสามารถแบ่งออกเป็น 2 ประเภทใหญ่ คือ

1. **เสียงเรียง** (segmental sound) เป็นหน่วยเสียงที่สามารถแยกออกจากเสียงอื่นได้โดยเด็ดขาดเพราะมีลักษณะเด่นเฉพาะตัว ในภาษาไทยได้แก่เสียงสระ และเสียงพยัญชนะ
2. **เสียงซ้อน** (supra-segmental feature) เป็นเสียงที่ทำหน้าที่เป็นส่วนประกอบของเสียงอื่นเพราะไม่สามารถแยกเปล่งเสียงได้ตามลำพัง ในภาษาไทยได้แก่เสียงวรรณยุกต์และทำนองเสียง เป็นต้น

2.3 หน่วยเสียงสระ

2.3.1 ลักษณะของเสียงสระ

ลักษณะสำคัญของเสียงสระก็คือ "เป็นเสียงก้องที่เปล่งเสียงออกมาโดยให้ลมออกทางช่องปากโดยไม่ถูกลิ้นกักหรือขัดขวาง" ดังนั้นเวลาเราออกเสียงสระจะออกเสียงได้สะดวกและออกเสียงได้นาน ทั้งนี้เพราะคุณสมบัติของเสียงสระมีความดังเด่นกว่าเสียงอื่นๆที่เรียงอยู่ข้างเสมอ อวัยวะที่เกี่ยวข้องกับการออกเสียงสระได้แก่ ลิ้น กับริมฝีปาก ถ้าลิ้นส่วนใดทำหน้าที่เพียงส่วนเดียว เสียงที่เกิดขึ้นก็จะมีเพียงเสียงเดียว เสียงเช่นนี้เรียกว่า "สระเดี่ยว" แต่ถ้าลิ้นส่วนอื่นทำหน้าที่ร่วมด้วยเสียงสระนั้นเรียงว่า "สระประสม"

สำหรับภาษาไทยมีหน่วยเสียงสระทั้งหมด 24 หน่วยเสียง แยกออกเป็นสระเดี่ยว 18 หน่วยเสียง และสระประสม 6 หน่วยเสียง [9]

สระเดี่ยว

เสียงสระเดี่ยว 18 หน่วยเสียง พิจารณาการเกิดเสียงได้เป็น 2 กรณีใหญ่ๆ คือ

1. พิจารณาการเกิดจากส่วนต่างๆของลิ้น หมายถึง ลมผ่านส่วนหน้า ส่วนกลาง หรือส่วนหลังของลิ้น
2. พิจารณาการเกิดจากลมผ่านลิ้นในขณะที่ลิ้นอยู่ในระดับ สูง กลาง หรือ ต่ำ

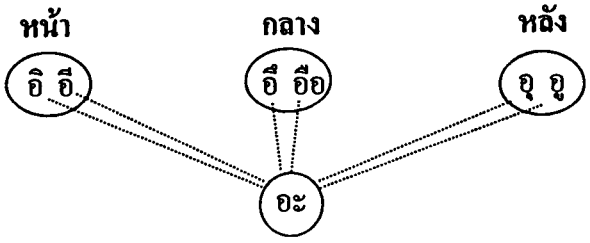
สระ	หน้า	กลาง	หลัง
สูง	อิ อี	อึ อือ	อุ อู
กลาง	เอะ เอ	เออะ เออ	โอะ โอ
ต่ำ	แอะ แอ	อะ อา	เอาะ ออ

นอกจากนี้ หน่วยเสียงสระเดี่ยว 18 หน่วย สามารถแบ่งตามความสั้น-ยาวของการออกเสียงได้เป็น

- สระเดี่ยวเสียงสั้น 9 หน่วย ได้แก่ อะ อิ อึ อุ เอะ แอะ โอะ เอาะ เออะ
- สระเดี่ยวเสียงยาว 9 หน่วย ได้แก่ อา อี อือ อู เอ แอ โอ ออ เออ

สระประสม

เสียงสระประสม 6 หน่วยเสียง เกิดจากลมผ่านกระทบลิ้น 2 ส่วนคือส่วนบนและส่วนล่าง ซึ่งในขณะออกเสียงลิ้นจะอยู่ในระดับสูงแล้วลดลงต่ำ โดยเสียงหลังเป็นเสียงสระ อะ เสมอ ดังแผนผังดังนี้



เสียงสระประสม 6 หน่วยเสียงได้แก่ เอียะ (อิ+อะ) เอีย (อี+อะ) เอือะ (อึ+อะ) เอือ (อือ+อะ) อัวะ (อุ+อะ) อัว (อู+อะ)

2.3.2 หน้าที่ของหน่วยเสียงสระในภาษาไทย

หน่วยเสียงสระในภาษาไทยทั้ง 24 หน่วยเสียงนี้ ทำหน้าที่เป็นแกนกลางของพยางค์หรือคำ กล่าวคือ คำ ทุกคำในภาษาไทยจะต้องมีเสียงสระอยู่ด้วย และเสียงสระในภาษาไทยจะสามารถเกิดกับเสียงพยัญชนะต้นได้ทุกเสียงและสามารถเกิดกับหน่วยเสียงวรรณยุกต์ได้ทุกหน่วย แต่ไม่สามารถเกิดกับหน่วยเสียงพยัญชนะสะกดได้ทุกหน่วย หน่วยเสียงสระที่ทำให้เกิดคำหรือพยางค์ใช้ได้มากที่สุดในภาษามักเป็นหน่วยเสียงสระยาว

2.4 หน่วยเสียงพยัญชนะ

เสียงพยัญชนะในภาษาไทยมีทั้งหมด 21 หน่วยเสียง(44 รูป) ดังแสดงในตารางที่ 2.1 หน่วยเสียงพยัญชนะออกเสียงได้ไม่สะดวกเท่าหน่วยเสียงสระ เพราะเวลาออกเสียงลมหายใจที่พุ่งออกมาจากหลอดลมจะถูกขัดขวางตามส่วนต่างๆของปาก เสียงพยัญชนะจึงออกเสียงให้ยาวนานอย่างเสียงสระไม่ได้ และเสียงพยัญชนะก็ไม่ใช่เสียงก้องเสมอไป

2.4.1 ลักษณะของเสียงพยัญชนะ

หน่วยเสียงพยัญชนะ 21 หน่วยเสียงนี้จำแนกเป็น เสียงก้อง เสียงไม่ก้อง และลักษณะการเกิดเสียง ดังนี้

เสียงก้อง (โฆษะ)มี 9 หน่วยเสียง คือ /ง/ /ย/ /บ/ /ด/ /ม/ /น/ /ร/ /ล/ /ว/

เสียงไม่ก้อง (อโฆษะ) มี 12 หน่วยเสียง คือ /ก/ /ค/ /จ/ /ช/ /ซ/ /ท/ /ต/ /ป/ /พ/ /ฟ/ /อ/ /ฮ/

ตารางที่ 2.1 เสียงพยัญชนะในภาษาไทย

ลำดับที่	อักษรไทยใช้แทนหน่วยเสียง	แทนสัญลักษณ์หน่วยเสียง	
		แบบสากล	แบบไทย
1.	ก	k	ก
2.	ข ฃ ค ฅ ฆ	kh	ค
3.	ง	ŋ	ง
4.	จ (จร_*)	c	จ
5.	ฉ ช ฌ	ch	ช
6.	ญ ย (หย_*) (หญ_*)	j	ย
7.	ซ ศ ษ ส (ทร_*)	s	ซ
8.	ฐ ฑ ฒ ถ ท ธ (ทร_*)	th	ท
9.	บ	b	บ
10.	ฎ ด (ฑ_*)	d	ด
11.	ฏ ต	t	ต
12.	ป	p	ป
13.	ผ พ ภ	ph	พ
14.	ฝ ฟ	f	ฟ
15.	ม (หม_*)	m	ม
16.	น ณ (หน_*)	n	น
17.	ร	r	ร
18.	ล ฬ (หล_*)	l	ล
19.	ว (หว_*)	w	ว
20.	อ	?	อ
21.	ฮ ห	h	ฮ
อักษร 44 รูป		21 หน่วยเสียง	

หมายเหตุ (*) หมายถึง พยัญชนะที่อยู่ในตำแหน่งพยัญชนะต้นแล้วออกเสียงเช่นเดียวกับเสียงพยัญชนะที่อยู่ในลำดับนั้นๆ

ลักษณะของเสียงพยัญชนะ จำแนกตามลักษณะรูปเสียง (classification by form)

1. **เสียงกัก หรือ เสียงหยุด** (Stop)

เป็นเสียงที่เมื่อผ่านกล่องเสียงเข้ามาถึงช่องปากแล้ว ในปากจะมีฐานกรณ์แห่งใดแห่งหนึ่งกั้นเสียงนี้ไว้ไม่ให้ออกจากปากแต่การกั้นเป็นเพียงชั่วระยะเวลาอันสั้นเท่านั้น แล้วฐานกรณ์ที่กั้นนั้นจะเปิดออก อากาศที่ถูกกักไว้จะถูกปล่อยออกมา เนื่องจากอากาศถูกกั้นไว้เมื่อถูกปล่อยออกมาจึงออกมาในลักษณะระเบิด บางทีจึงเรียกเสียงประเภทนี้ว่าเป็นเสียงระเบิด (plosive sound) มี 9 หน่วยเสียง คือ /ป/ /พ/ /บ/ /ต/ /ท/ /ด/ /ก/ /ค/ /อ/ เสียงพยัญชนะสะกดทุกเสียงในภาษาจะมีลักษณะเป็นเสียงระเบิด หรือเสียงกัก

2. เสียงเสียดแทรก (Fricative)

เป็นเสียงที่เมื่ออากาศผ่านขึ้นมาจากปอดผ่านกล่องเสียงเข้ามาถึงช่องปากแล้วในปากจะมีฐานกรณ์แห่งใดแห่งหนึ่งกั้นอากาศนี้ไว้ แต่การกั้นนี้ไม่สนิทมิดชิดเหมือนเสียงหยุด ยังมีช่องให้อากาศเล็ดลอดแทรกออกมาได้ทำให้เกิดเสียงขณะแทรกออกมา มี 3 หน่วยเสียงคือ /ซ/ /ฟ/ /ฮ/

3. เสียงกึ่งเสียดแทรก (Affricate)

เป็นเสียงในช่องปากที่มีคุณสมบัติเหมือนกับเริ่มต้นด้วยเสียงหยุดและตามด้วยเสียงแทรก มี 2 หน่วยเสียง คือ /จ/ และ /ช/

4. เสียงนาสิก (Nasal)

เป็นเสียงที่เมื่ออากาศผ่านกล่องเสียง ผ่านช่องคอแล้วก็เข้าสู่ช่องจมูก โดยที่ช่องปากมีฐานกรณ์กั้นไว้สนิทไม่ให้อากาศออกทางช่องปาก เสียงที่อากาศผ่านออกมาทางช่องจมูก มี 3 หน่วยเสียง คือ /ม/ /น/ /ง/

5. เสียงข้าง (Lateral)

เป็นเสียงที่อากาศในช่องปากออกสู่ภายนอกปากโดยผ่านทางข้างๆลิ้น มีหน่วยเสียงเดียว คือ /ล/

6. เสียงรัว (Trill)

คือเสียงที่เมื่ออากาศเข้ามาอยู่ในช่องปากแล้วมีการกระดกปลายลิ้นรัวเพดานหลายๆครั้ง มีหน่วยเสียงเดียว คือ /ร/

7. เสียงครึ่งสระ (Semi-Vowel)

คำอธิบายทั่วๆไปของเสียงประเภทนี้ไม่ค่อยชัดแจ้งนัก มักกล่าวว่าตำแหน่งลิ้นเมื่อเริ่มต้นเสียงต่างไปจากตำแหน่งในตอนท้ายๆเสียง เสียงประเภทนี้บางครั้งก็เรียกว่าเสียงครึ่งสระ เพราะมีตำแหน่งลิ้นและปากคล้ายเสียงสระ มี 2 หน่วยเสียง คือ เสียง /ว/ และ /ย/

2.4.2 หน้าที่ของหน่วยเสียงพยัญชนะในภาษาไทย

เสียงพยัญชนะในภาษาไทย 21 หน่วยเสียงนี้สามารถทำหน้าที่ได้ดังนี้

1. เป็นพยัญชนะต้นของพยางค์ คือสามารถนำหน้าเสียงสระในพยางค์หนึ่งๆได้ ในตำแหน่งนี้เสียงพยัญชนะสามารถเกิดได้หน่วยเดียว หรือ สองหน่วยดังนี้
   - เกิดได้ หน่วยเดียว คือ ทำหน้าที่เป็นพยัญชนะต้นเดี่ยว หน่วยเสียงทั้ง 21 หน่วยเสียงนี้สามารถทำหน้าที่เป็นพยัญชนะต้นเดี่ยวได้ทั้งสิ้น
   - เกิดได้ สองหน่วย คือ ทำหน้าที่เป็นพยัญชนะต้นควบ โดยหน่วยเสียงแรกเป็น /ก/ /ค/ /ต/ /ป/ และ /พ/ กับหน่วยเสียงที่สองเป็น /ร/ /ล/ หรือ/ว/
2. เป็นพยัญชนะสะกดของพยางค์ ในตำแหน่งนี้เสียงพยัญชนะในภาษาไทยสามารถเกิดได้ 9 หน่วยเสียง คือ /ป/ (แม่กบ) /ต/ (แม่กด) /ก/ (แม่กก) /ม/ (แม่กม) /ง/ (แม่กง) /น/ (แม่กน) /ย/ (แม่เกย) /ว/ (แม่เกอว) และไม่มีเสียงพยัญชนะสะกด (แม่กา)

2.5 หน่วยเสียงวรรณยุกต์

เสียงวรรณยุกต์ คือ ระดับเสียงสูง-ต่ำ ของคำในภาษาไทย เช่นเดียวกับภาษาจีน และภาษาอื่นๆ ที่เป็นภาษาคำโดดซึ่งมีการกำหนดเสียงสูงต่ำไว้ตายตัวในคำแต่ละคำ ถ้าออกเสียงสูง-ต่ำผิดไปความหมายย่อมผิดตามไปด้วย

ในภาษาไทยหน่วยเสียงวรรณยุกต์เป็นหน่วยเสียงสำคัญ ที่ทำให้คำที่มีส่วนประกอบแวดล้อมอื่นๆเหมือนกัน คือมี เสียงพยัญชนะต้น สระ และพยัญชนะสะกดอย่างเดียวกันมีความหมายต่างกัน ดังนั้นอาจกล่าวได้ว่าหน้าที่ของหน่วยเสียงวรรณยุกต์ก็คือ การทำให้เกิดคำขึ้นใช้ในภาษามากขึ้นและเป็นวิธีการสร้างคำขึ้นใช้เพิ่มขึ้นในภาษาเป็นวิธีแรก ทั้งนี้เพราะถ้าเราเปลี่ยนเสียงวรรณยุกต์ก็จะทำให้คำเกิดความหมายเพิ่มขึ้นใหม่นั่นเอง

เสียง สูง-ต่ำ ในภาษาพูด เกิดจากการสั่นสะเทือนของเส้นเสียงในอัตราต่างๆกัน โดยเสียงที่เปล่งออกมาในขณะที่เส้นเสียงสั่นนั้นจะต้องเป็นเสียงก้อง ดังนั้นหน่วยเสียงวรรณยุกต์ในภาษาไทยจึงจัดเป็นหน่วยเสียงซ้อน สัทอักษรที่ใช้จึงเป็นรูปเครื่องหมายเขียนซ้อนข้างบนหน่วยเสียงสระ(ซึ่งเป็นเสียงก้อง) ซึ่งมีรูปวรรณยุกต์อยู่ 4 รูป แทนเสียงวรรณยุกต์ทั้งหมด 5 หน่วยเสียง โดยเสียงสามัญไม่มีรูปวรรณยุกต์

2.5.1 ลักษณะของเสียงวรรณยุกต์

สามารถแบ่งออกตามลักษณะระดับเสียงได้เป็น 2 กลุ่มใหญ่ๆ คือ

1. กลุ่มวรรณยุกต์ระดับ (Level tone) มี 3 หน่วยเสียง คือ

1.1 หน่วยเสียงวรรณยุกต์ระดับต่ำ (Low tone) แทนด้วยสัญลักษณ์ /˧/

คือ เสียงวรรณยุกต์เอก หน่วยเสียงนี้จะปรากฏในพยางค์ของภาษาไทยได้ทุกแบบ

1.2 หน่วยเสียงวรรณยุกต์ระดับกลาง (Mid tone) ไม่มีสัญลักษณ์

คือ เสียงวรรณยุกต์สามัญ หน่วยเสียงนี้จะไม่ปรากฏในพยางค์ที่มีตัวสะกดเป็นพยัญชนะกัก (พยางค์คำตาย)

1.3 หน่วยเสียงวรรณยุกต์ระดับสูง (High tone) แทนด้วยสัญลักษณ์ /ˊ/

คือ เสียงวรรณยุกต์ตรี หน่วยเสียงนี้จะไม่ปรากฏในพยางค์ที่ประสมด้วยสระเสียงยาว ซึ่งมีตัวสะกดเป็นเสียงกัก

2. กลุ่มวรรณยุกต์เปลี่ยนระดับ (Contour tone) มี 2 หน่วยเสียง คือ

2.1 หน่วยเสียงวรรณยุกต์เปลี่ยนตก (Falling tone) แทนด้วยสัญลักษณ์ /ˆ/

คือ เสียงวรรณยุกต์โท หน่วยเสียงนี้จะไม่ปรากฏในพยางค์ที่มีสระเสียงสั้น และมีเสียงพยัญชนะสะกดเป็นพยัญชนะกัก

2.2 หน่วยเสียงวรรณยุกต์เปลี่ยนขึ้น (Rising tone) แทนด้วยสัญลักษณ์ /ˇ/

คือ เสียงวรรณยุกต์จัตวา หน่วยเสียงนี้จะไม่ปรากฏในพยางค์ที่มีเสียงพยัญชนะสะกดเป็นเสียงกักเลย

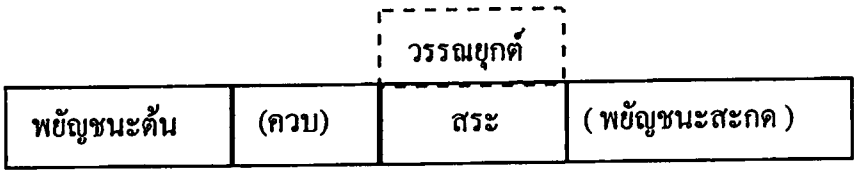
2.6 ลักษณะพยางค์ของคำไทย

2.6.1 คำจำกัดความของพยางค์และคำในภาษาไทย

กาญจนา นาคสกุล (2520:104) ได้ให้ความหมายของพยางค์ในระบบเสียงภาษาไทยว่า "พยางค์ หมายถึง จำนวนเสียงที่ดังเด่นซึ่งปรากฏในกลุ่มเสียงที่เรียงกันเป็นคำพูด ส่วนเสียงอื่นๆที่อยู่ข้างเคียงก็จะประกอบกันเข้าเป็นส่วนของพยางค์" เสียงที่ดังเด่นในกลุ่มเสียงก็คือเสียงสระ ซึ่งมีลักษณะประจำตัวก็คือเป็นเสียงก้องซึ่งดังเด่นกว่าเสียงอื่นๆ ดังนั้นเสียงสระจึงมักเป็นเสียงที่ทำให้เกิดพยางค์ ถ้ามีเสียงสระเด่นอยู่กี่เสียง พยางค์ก็จะมีจำนวนเท่านั้นด้วย

พยางค์ที่เปล่งออกมาครั้งหนึ่งๆ อาจมีความหมายหรือไม่ก็ได้ แต่เมื่อใดพยางค์ที่ประกอบขึ้นจาก เสียงสระ พยัญชนะ และวรรณยุกต์ เป็นอย่างน้อยที่สุด และกลุ่มเสียงเหล่านี้มีความหมายและสามารถปรากฏได้โดยลำพัง พยางค์นั้นๆก็จะกลายเป็นคำในภาษา

คำในภาษาไทยส่วนใหญ่จะเป็นคำพยางค์เดียว ซึ่งเป็นคำพื้นฐาน (Base words) ของภาษา ภาษาไทยจึงจัดอยู่ในตระกูลภาษาคำโดด หรือ คำพยางค์เดียว (Monosyllabic language) หน่วยเสียงที่ประกอบกันเข้าเป็นพยางค์จะต้องมีอย่างน้อย 3 หน่วย คือ หน่วยเสียงพยัญชนะต้น 1 หน่วย หน่วยเสียงสระ 1 หน่วย และ หน่วยเสียงวรรณยุกต์ 1 หน่วย และมีหน่วยเสียงอย่างมากไม่เกิน 5 หน่วย คือเพิ่มหน่วยเสียงพยัญชนะต้นที่เป็นเสียงควบกล้ำอีก 1 หน่วย และหน่วยเสียงพยัญชนะสะกดอีก 1 หน่วย โดยมีองค์ประกอบของหน่วยเสียงต่างๆในพยางค์ แสดงได้ดังรูปที่ 2.2



รูปที่ 2.2 องค์ประกอบของพยางค์ในภาษาไทย

2.6.2 ลักษณะโครงสร้างของคำพยางค์เดียวต่อการผันของเสียงวรรณยุกต์

เราต้องตระหนักเสมอว่ารูปวรรณยุกต์ในคำภาษาไทยบางครั้งไม่แสดงเสียงให้เห็นในการเขียนเสมอไป ทั้งนี้การกำหนดเสียงวรรณยุกต์ขึ้นอยู่กับลักษณะของพยางค์ว่าเป็น "คำเป็น" หรือ "คำตาย"

ลักษณะโครงสร้างของคำพยางค์เดียวในภาษาไทยมี 5 แบบ ซึ่งลักษณะโครงสร้างที่ต่างกันของพยางค์จะมีผลต่อการผันของเสียงวรรณยุกต์ ดังแสดงในตารางที่ 2.2

ตารางที่ 2.2 แสดงลักษณะของคำพยางค์เดียวในภาษาไทย

เสียงวรรณยุกต์ / โครงสร้างพยางค์	สามัญ	เอก	โท	ตรี	จัตวา
1. พ (พ) ส ส $^{0\text{-}4}$	+	+	+	+	+
2. พ (พ) ส น $^{0\text{-}4}$	+	+	+	+	+
3. พ (พ) ส ส น $^{0\text{-}4}$	+	+	+	+	+
4. พ (พ) ส ก ^{1,3}	−	+	−	+	−
5. พ (พ) ส ส ก ^{1,2}	−	+	+	−	−

<u>หมายเหตุ</u> + หมายถึงโครงสร้างพยางค์สามารถผันระดับเสียงวรรณยุกต์นั้นได้

− หมายถึงโครงสร้างพยางค์ไม่สามารถผันระดับเสียงวรรณยุกต์นั้นได้

เมื่อกำหนดให้ พ แทนหน่วยเสียงพยัญชนะต้น 1 หน่วย

พพ แทนหน่วยเสียงพยัญชนะต้น 2 หน่วยควบกัน หรือพยัญชนะต้นควบ โดยหน่วยเสียงที่ 2 คือ ร /r/ ล /l/ หรือ ว /w/

ส แทนหน่วยเสียงสระเดี่ยวสั้น

สส แทนหน่วยเสียงสระเดี่ยวยาว และหน่วยเสียงสระประสม

น แทนหน่วยเสียงพยัญชนะสะกดที่เป็นพยัญชนะนาสิก / m, n, ŋ/ และครึ่งสระ / j, w /

ก แทนหน่วยเสียงสะกดที่เป็นพยัญชนะกัก / p, t, k, ? /

0 แทนหน่วยเสียงวรรณยุกต์ สามัญ

1 แทนหน่วยเสียงวรรณยุกต์ เอก

2 แทนหน่วยเสียงวรรณยุกต์ โท

3 แทนหน่วยเสียงวรรณยุกต์ ตรี

4 แทนหน่วยเสียงวรรณยุกต์ จัตวา

และจากจำนวนอักษร 44 รูป ในภาษาไทยได้แบ่งเพื่อสะดวกต่อการผันเสียงวรรณยุกต์ เป็นอักษรไตรยางค์ ดังได้แสดงไว้ในตารางที่ 2.3

ตารางที่ 2.3 อักษรไตรยางค์

อักษรไตรยางค์		รูปวรรณยุกต์			
		เอก	โท	ตรี	จัตวา
อักษร สูง	ข ฃ ฉ ถ ฐ ผ ฝ ศ ษ ส ห	+	+	−	−
อักษร กลาง	ก จ ด ฎ ต ฏ บ ป อ	+	+	+	+
อักษร ต่ำ- คู่	ค ฆ ฅ ช ฌ ฑ ท ธ ฒ พ ภ ฟ ซ ฮ	+	+	−	−
ต่ำ-เดี่ยว	ม น ง ณ ย ญ ร ล ฬ ว	+	+	−	−

อักษรสูง 11 ตัว ผันวรรณยุกต์ได้ 3 เสียง เช่น ขา ข่า ข้า

อักษรกลาง 9 ตัว ผันวรรณยุกต์ได้ครบทั้ง 5 เสียง เช่น จา จ่า จ้า จ๊า จ๋า

อักษรต่ำ 24 ตัว ผันวรรณยุกต์ได้ 3 เสียง เช่น ทา ท่า ท้า

การผันอักษรต่ำนี้มีข้อสังเกต คือถ้ามีรูปวรรณยุกต์เอกจะเป็นเสียงวรรณยุกต์โท ถ้ารูปวรรณยุกต์โทจะเป็นเสียงวรรณยุกต์ตรี นอกจากนี้อักษรต่ำยังแบ่งออกเป็นอักษรต่ำคู่ 14 ตัวและอักษรต่ำเดี่ยวอีก 10 ตัว เพื่อประโยชน์ในการผันเสียงวรรณยุกต์คือ เมื่อนำคำที่เป็นอักษรต่ำคู่มาผันร่วมกับคำที่เป็นอักษรสูงจะเกื้อกูลกันทำให้การผันเสียงวรรณยุกต์ทำได้ครบทั้ง 5 เสียง ดังตัวอย่างในตารางที่ 2.4

ตารางที่ 2.4 ตัวอย่างการผันเสียงอักษรต่ำคู่ กับอักษรสูง

เสียงวรรณยุกต์				
สามัญ	เอก	โท	ตรี	จัตวา
คา	ข่า	ค่า ข้า	ค้า	ขา

อักษรต่ำที่ใช้คู่กับอักษรสูง แล้วทำให้ผันเสียงวรรณยุกต์ได้ครบทั้ง 5 เสียง มี 7 คู่ ดังตารางที่ 2.5

ตารางที่ 2.5 การจับคู่ในการผันเสียงวรรณยุกต์

คู่ที่	อักษรสูง	อักษรต่ำ
1	ข ฃ	ค ฆ ฅ
2	ฉ	ช ฌ
3	ถ ฐ	ฑ ท ธ ฒ
4	ผ	พ ภ
5	ฝ	ฟ
6	ศ ษ ส	ซ
7	ห	ฮ
	11 ตัว	14 ตัว

ส่วนอักษรต่ำเดี่ยวอีก 10 ตัวนั้นการที่จะทำให้ผันเสียงได้ครบนั้นจะนำตัว "ห" มาช่วยก็จะทำให้ผันเสียงวรรณยุกต์ได้ครบ เช่น นา หน่า น่า น้า หนา

บทที่ 3
สัญญาณรบกวน

3.1 คุณสมบัติทั่วไปของสัญญาณรบกวน

ในธรรมชาติจะมีปริมาณทางกายภาพหลายอย่างที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา โดยหากฎเกณฑ์ที่แน่นอนไม่ได้ ในเรื่องของสัญญาณไฟฟ้าก็เช่นกัน สัญญาณบางชนิดเกิดเปลี่ยนแปลงอยู่ตลอดเวลาตามธรรมชาติและอาจจะมารบกวนสัญญาณข่าวสารที่ต้องการให้เกิดความผิดเพี้ยนหรือไม่ชัดเจนไปได้ เราเรียกสัญญาณเช่นนี้ว่า "สัญญาณรบกวน" (Noise) ถ้าจะกล่าวโดยทั่วไป สัญญาณรบกวน ก็คือสัญญาณที่เราไม่พึงปรารถนา ไม่ว่าจะเกิดจากสาเหตุที่มีกฎเกณฑ์หรือไร้กฎเกณฑ์ก็ตาม จะเกิดขึ้นตามธรรมชาติหรือจะเกิดขึ้นจากการกระทำของมนุษย์ โดยความจงใจหรือไม่จงใจก็ตาม ถ้าหากเป็นสัญญาณที่เราไม่พึงต้องการแล้วเราจะจัดว่ามันเป็นสัญญาณรบกวนทั้งสิ้น ตัวอย่างของสัญญาณรบกวนอันเกิดจากการกระทำของมนุษย์ (man made noise) ได้แก่ สัญญาณจากการจุดระเบิดของหัวเทียนเครื่องยนต์ สัญญาณที่เกิดจากเครื่องใช้ไฟฟ้า เช่น สว่าน หรือ เลื่อยไฟฟ้า เป็นต้น สัญญาณรบกวนตามธรรมชาติ (natural noise) ได้แก่ สัญญาณรบกวนเนื่องจากอุณหภูมิ (thermal noise) และสัญญาณรบกวนจากระบบสุริยะ (solar noise) เป็นต้น

เนื่องจากสัญญาณรบกวนนั้นมีคุณสมบัติโดยทั่วไปที่คาดเดาอะไรกับมันโดยแน่นอนไม่ได้ (random) กล่าวคือ คุณสมบัติของสัญญาณรบกวนโดยทั่วไปมักจะเป็นสัญญาณสุ่ม ดังนั้นการที่จะบอกถึงคุณสมบัติอะไรเกี่ยวกับสัญญาณรบกวนนั้น จึงบอกได้แต่คุณสมบัติที่เป็นค่าเชิงสถิติเท่านั้น จากนั้นจึงอาศัยคุณสมบัติเหล่านั้นมาเป็นเครื่องเชื่อมโยงเกี่ยวเนื่อง เพื่ออธิบายคุณสมบัติทางกายภาพของสัญญาณรบกวนนั้น คุณสมบัติทางสถิติที่เราควรรู้ไว้ เมื่อเริ่มศึกษาถึงคุณสมบัติของสัญญาณรบกวนนั้น ได้แก่ ค่าเฉลี่ยในลักษณะต่างๆ เช่น ค่าเฉลี่ย (average value) หรือค่ามัธยฐาน (mean value) ค่าเฉลี่ยของกำลังสอง (mean square value) และ ค่าความแปรปรวน (variance) เป็นต้น แม้ปริมาณเหล่านี้จะไม่สามารถใช้บอกคุณสมบัติของสัญญาณรบกวนได้อย่างสมบูรณ์

ถ้าสมมุติให้ $n(t)$ เป็นฟังก์ชันของเวลาที่ใช้แทนสัญญาณรบกวน ค่าเฉลี่ยของสัญญาณรบกวนในเชิงสถิติ ตามที่ได้อ้างถึงนั้น มีการนิยามดังต่อไปนี้ คือ

3.1.1 ค่าเฉลี่ยของสัญญาณรบกวน

ค่าเฉลี่ยของสัญญาณรบกวน $n(t)$ เขียนแทนด้วยสัญลักษณ์ $\overline{n(t)}$ มีการนิยามดังนี้คือ

$$\overline{n(t)} \equiv \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} n(t)\ dt \qquad (3.1)$$

ค่าเฉลี่ย $\overline{n(t)}$ นี้บอกให้เรารู้ถึงคุณสมบัติทางกายภาพคือ บอกว่าระดับไฟตรงของสัญญาณรบกวน

ในการหาค่าเฉลี่ยทางปฏิบัติ ค่าระยะเวลา T ในสมการที่ 3.1 นั้นจะใช้เพียงค่า ระยะเวลาที่ยาวนานค่าหนึ่งเท่านั้น ซึ่งถ้าค่าระยะเวลานี้ยาวนานพอสมควรแล้ว ค่า $\overline{n(t)}$ ที่หามาได้ก็จะใช้ประมาณบอกถึงค่าเฉลี่ย ตามสมการที่ 3.1 ได้

3.1.2 ค่าเฉลี่ยของกำลังสองของสัญญาณรบกวน

ค่าเฉลี่ยของกำลังสองของสัญญาณรบกวน $\overline{n^2(t)}$ หาได้ดังนี้คือ

$$\overline{n^2(t)} \equiv \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \left| n(t) \right|^2 dt \qquad (3.2)$$

ค่าเฉลี่ยตามสมการที่ 3.2 นี้ แสดงถึงปริมาณทางกายภาพ คือ ค่ากำลังเฉลี่ย (average power) หรือค่ากำลังประสิทธิผล (effective power) ทั้งหมดของสัญญาณรบกวน

3.1.3 ค่าความแปรปรวนของสัญญาณรบกวน

ค่าความแปรปรวนของสัญญาณรบกวน σ_n^2 หาได้จากนิยามดังต่อไปนี้ คือ

$$\sigma_n^2 = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \{\overline{n(t)} - n(t)\}^2\ dt \qquad (3.3)$$

เนื่องจาก $\overline{n(t)} - n(t)$ นั้นแสดงถึงส่วนของ $n(t)$ ที่เบี่ยงเบนไปจากค่าเฉลี่ย $\overline{n(t)}$ ดังนั้นค่า $\{\overline{n(t)} - n(t)\}$ จึงมีความหมายแทนปริมาณส่วนที่เป็นไฟสลับของสัญญาณรบกวน $n(t)$ ด้วยเหตุนี้ ค่าความแปรปรวนของสัญญาณรบกวน σ_n^2 จึงแสดงให้เรารู้ถึงค่ากำลังเฉลี่ยของส่วนที่เป็นไฟสลับของสัญญาณรบกวน $n(t)$ และค่าความเบี่ยงเบนมาตรฐาน σ_n จะบอกให้เรารู้ถึงค่าอาร์เอ็มเอส (RMS) ของสัญญาณรบกวนนั้น

ค่าทางสถิติสมการที่ 3.1 ~ 3.3 นั้นมีความสัมพันธ์ดังต่อไปนี้ คือ

$$\overline{n^2(t)} = \sigma_n^2 + \{\overline{n(t)}\}^2 \tag{3.4}$$

ความสัมพันธ์สมการที่ 3.4 นี้สามารถพิสูจน์ได้โดยอาศัยสมการที่ 3.2 ดังต่อไปนี้

$$\begin{aligned}\overline{n^2(t)} &= \overline{[\{n(t)-\overline{n(t)}\}+\overline{n(t)}]^2} \\ &= \overline{\{n(t)-\overline{n(t)}\}^2 + 2\overline{n(t)}\{n(t)-\overline{n(t)}\} + \{\overline{n(t)}\}^2} \\ &= \overline{\{n(t)-\overline{n(t)}\}^2} + 2\overline{n(t)}\,\overline{\{n(t)-\overline{n(t)}\}} + \overline{\{n(t)\}}^2\end{aligned}$$

เนื่องจาก $\overline{\{n(t)-\overline{n(t)}\}}$ มีค่าเท่ากับศูนย์ เพราะฉะนั้นโดยอาศัยคำจำกัดความของ σ_n^2 ตามสมการที่ 3.3 จะทำให้สามารถสรุปผลจากสมการบนได้ว่า คือสมการที่ 3.4 มีความหมายในทางกายภาพว่า กำลังเฉลี่ยทั้งหมดของสัญญาณรบกวนมีค่าเท่ากับกำลังเฉลี่ยของส่วนที่เป็นไฟสลับของสัญญาณรบกวน รวมกับกำลังของส่วนที่เป็นไฟตรงของสัญญาณรบกวนนั้น

3.2 อัตราส่วนสัญญาณต่อสัญญาณรบกวน

เพราะเราไม่สามารถจะคาดการณ์ว่าสัญญาณรบกวนนั้น น่าจะมีค่าเฉลี่ยตามสมการที่ 3.1 ~ 3.3 เป็นเท่าไร ดังนั้นเมื่อเราต้องการจะวิเคราะห์เกี่ยวกับสัญญาณรบกวน จึงสมควรที่จะพิจารณาใช้ค่ากำลังเฉลี่ยของสัญญาณรบกวนเหล่านั้นมาเป็นเกณฑ์ในการวิเคราะห์ เรื่องสำคัญที่ควรพิจารณา เมื่อทำปฏิบัติการเกี่ยวกับสัญญาณที่มีความแรงในระดับต่ำ เราจะพบว่าสัญญาณรบกวนมักจะเข้ามามีอิทธิพลครอบงำสัญญาณนั้น ซึ่งตัวดัชนีที่ใช้ช่วยบอกว่าสัญญาณนั้นถูกรบกวนโดยสัญญาณรบกวนมากน้อยเท่าใดนั้นอย่างหนึ่งก็คือ "ค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน " (signal to noise ratio) ซึ่งนิยมเขียนแทนด้วยสัญลักษณ์ $\frac{S}{N}$ หรือ SNR

ค่า เอสเอ็นอาร์ นั้นเป็นค่าอัตราส่วนของกำลังเฉลี่ย ของสัญญาณ $\overline{s^2(t)}$ ต่อค่ากำลังเฉลี่ยของสัญญาณรบกวน $\overline{n^2(t)}$ ซึ่งเขียนเป็นสมการได้ดังนี้ คือ

$$\frac{S}{N} = \frac{\overline{s^2(t)}}{\overline{n^2(t)}} \tag{3.5}$$

ค่าเอสเอ็นอาร์นี้ ปรกติมักนิยมแสดงค่าในหน่วย เดซิเบล (decibel) ซึ่งเขียนย่อว่า dB การแปลงค่าเอสเอ็นอาร์ ตามสมการที่ 3.5 ให้มีหน่วยเป็นเดซิเบลทำได้ดังนี้

$$\left.\frac{S}{N}\right|_{dB} = 10\log\left(\frac{\overline{s^2(t)}}{\overline{n^2(t)}}\right) \tag{3.6}$$

ค่าเอสเอ็นอาร์นี้ เป็นค่าซึ่งแสดงถึงคุณภาพของสัญญาณที่กำลังพิจารณาว่า มีระดับกำลังสูงกว่าระดับกำลังของสัญญาณรบกวนที่ปนอยู่กับสัญญาณนั้นมากน้อยเท่าไร แต่อย่างไรก็ดี ค่าเอสเอ็นอาร์นี้ ไม่สามารถใช้แสดงให้รู้ถึงความดีเลวของระบบ ในแง่ที่จะแสดงให้รู้ว่าระบบนั้นก่อกำเนิดสัญญาณรบกวนจากภายในตัวระบบ ออกมาปนกันสัญญาณที่ต้องการมากหรือน้อยอย่างไร เพื่อความสะดวกในการแสดงคุณสมบัติดังกล่าว จึงได้มีการนิยามใช้ค่าดัชนีตัวใหม่ที่สามารถจะบอกให้เรารู้ถึงคุณสมบัติของระบบ ในทัศนะดังกล่าวได้ ค่าที่ถูกนิยามขึ้นนี้มีชื่อว่า ค่าตัวเลขสัญญาณรบกวน (noise figure) ของระบบ ซึ่งเรียกย่อว่าค่า เอ็นเอฟ (NF) โดยมีคำนิยามดังนี้ คือค่า เอ็นเอฟ หมายถึงค่าอัตราส่วนของกำลังเฉลี่ยทั้งหมดของสัญญาณรบกวนที่ปรากฏที่เอาต์พุตของระบบ $\overline{n_{t_o}^2(t)}$ ต่อค่ากำลังเฉลี่ยของสัญญาณรบกวนในส่วนของเอาต์พุตอันเป็นผลจากปฏิบัติการของระบบ (เช่น การขยายสัญญาณ) ที่มีโดยตรงต่อสัญญาณรบกวนที่มีมาจากอินพุต $\overline{n_{i_o}^2(t)}$ โดยไม่คำนึงถึงส่วนที่เกิดจากสัญญาณรบกวนที่เกิดภายในระบบเลยกล่าวคือ

$$NF = \frac{\overline{n_{t_o}^2(t)}}{\overline{n_{i_o}^2(t)}} \tag{3.7}$$

ถ้ากำหนดให้ $\overline{n_{s_o}^2(t)}$ คือค่ากำลังเฉลี่ยของสัญญาณรบกวนในส่วนของเอาต์พุตของระบบอันเนื่องมาจากสัญญาณรบกวนที่เกิดขึ้นมาในตัวระบบนั้นเอง แล้วจะพบความสัมพันธ์ดังนี้ คือ

$$\overline{n_{t_o}^2(t)} = \overline{n_{i_o}^2(t)} + \overline{n_{s_o}^2(t)} \tag{3.8}$$

จากสมการที่ 3.7 และ 3.8 จะได้

$$NF = \frac{\overline{n_{i_o}^2(t)} + \overline{n_{s_o}^2(t)}}{\overline{n_{i_o}^2(t)}} \tag{3.9}$$

หรือ

$$NF = 1+\frac{\overline{n_{s_o}^2(t)}}{n_{i_o}^2(t)} \tag{3.10}$$

สมการที่ 3.10 บอกให้เรารู้ว่า เมื่อระบบที่กำลังพิจารณานั้นเป็นระบบที่ไม่ก่อให้เกิดสัญญาณรบกวนขึ้นในตัวเองคือ $\overline{n_{s_o}^2(t)}=0$ จะได้ค่า $NF = 1$ แต่ถ้าระบบนั้นมีคุณภาพต่ำลง คือ $\overline{n_{s_o}^2(t)}\neq 0$ แล้ว ค่าเอ็นเอฟของระบบนั้นจะมีค่ามากกว่า 1

เมื่อพิจารณาสมการที่ 3.7 อีกครั้ง โดยพิจารณา $\overline{n_{i_o}^2(t)}$ เกิดมาจากการปฏิบัติการของระบบต่อสัญญาณรบกวนที่เข้ามาที่อินพุตเท่านั้น ดังนั้นถ้าให้ $\overline{n_{i_i}^2(t)}$ คือ กำลังเฉลี่ยของสัญญาณรบกวนที่อินพุตของระบบและให้การปฏิบัติการของระบบ คือ การขยายกำลังสัญญาณ k เท่า เราก็จะจัดรูปสมการที่ 3.7 ได้ใหม่ เป็น

$$NF=\frac{\overline{n_{t_o}^2(t)}}{kn_{i_i}^2(t)} \tag{3.11}$$

และโดยการตั้งข้อสังเกตว่า ค่าอัตราการขยายกำลังสัญญาณ k เท่า ได้มาจากอัตราส่วนของกำลังเฉลี่ยสัญญาณ $\overline{s_o^2(t)}$ ที่เอาต์พุต ต่อกำลังเฉลี่ยสัญญาณ $\overline{s_i^2(t)}$ ที่อินพุตกล่าวคือ

$$NF = \frac{\overline{s_o^2(t)}}{\overline{s_i^2(t)}} \tag{3.12}$$

เราจะเขียนสมการที่ 3.11 ใหม่ได้เป็น

$$NF = \frac{\dfrac{\overline{s_i^2(t)}}{n_{i_i}^2(t)}}{\dfrac{\overline{s_o^2(t)}}{n_{t_o}^2(t)}} \tag{3.13}$$

หรือ

$$NF = \frac{\left[\frac{S}{N}\right]_i}{\left[\frac{S}{N}\right]_o} \tag{3.14}$$

เมื่อ $\left[\frac{S}{N}\right]_i$ และ $\left[\frac{S}{N}\right]_o$ คือ ค่าเอสเอ็นอาร์ที่อินพุต และเอาต์พุต ของระบบตามลำดับ

ดังนั้นจึงกล่าวนิยามค่าเอ็นเอฟในอีกนัยหนึ่งได้ว่า คือ " อัตราส่วนของค่าเอสเอ็นอาร์ที่อินพุตต่อค่าเอสเอ็นอาร์ที่เอาต์พุตของระบบ" ซึ่งจะเห็นได้ว่าการกำหนดนิยาม ค่าเอ็นเอฟตามสมการที่ 3.7 หรือ 3.10 นั้นทำให้เห็นชัดถึงความหมายที่ซ่อนลึกอยู่ในตัวคำจำกัดความของเอ็นเอฟได้อย่างดี แต่ค่าเอ็นเอฟที่ถูกดัดแปลงให้อยู่ในรูปสมการที่ 3.14 นั้นทำให้เรารู้ถึงวิธีการที่จะวัดค่าเอ็นเอฟในทางปฏิบัติได้อย่างสะดวก

ในบางครั้ง ค่าเอ็นเอฟนั้น จะแสดงค่าในหน่วยเดซิเบลซึ่งสามารถคำนวณได้ดังนี้ คือ

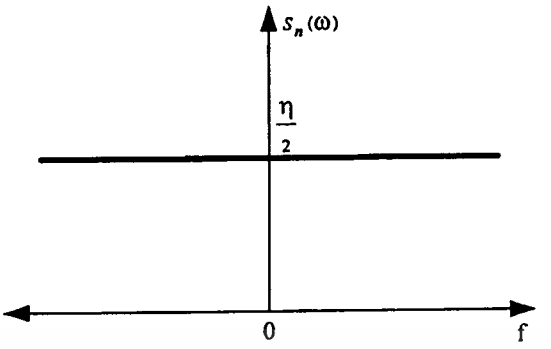
$$NF|_{dB} = 10\log(NF) \tag{3.15}$$

3.3 สัญญาณรบกวนขาว

สัญญาณรบกวน เกิดมาจากสาเหตุต่างๆ กันและมีรูปแบบพีเอสดี ที่แตกต่างกันสัญญาณรบกวนที่ควรสนใจมากที่สุด คือ สัญญาณรบกวนที่มีค่าพีเอสดี เท่ากันที่ทุกความถี่ ซึ่งมีชื่อว่า "สัญญาณรบกวนขาว" (white noise) โดยปรกติเมื่อเรากล่าวถึงสัญญาณรบกวนถ้าไม่ได้มีการกล่าวบอกถึงค่าเฉลี่ยของมัน เราก็จะหมายถึงสัญญาณรบกวนที่มีค่าเฉลี่ยของมันเป็นศูนย์ เพราะปรกติถ้าสัญญาณรบกวนมีค่าเฉลี่ยไม่เป็นศูนย์ เราก็อาจจะคิดเทียบได้ว่า สัญญาณรบกวนนั้น คือ สัญญาณรบกวนที่มีค่าเฉลี่ยเป็นศูนย์ รวมอยู่กับสัญญาณไฟตรงที่มีค่าเท่ากับค่าเฉลี่ยของสัญญาณรบกวนนั้นได้

สัญญาณรบกวนที่มีค่าการแจกแจงของค่ากำลังเท่ากันตลอดทุกความถี่ บนแกนความถี่ข้างเดียวกับ η วัตต์/เฮิรตซ์ จะมีค่าฟังก์ชันพีเอสดี ในรูปแบบของสเปกตรัมสองข้าง คือ

$$S_n(\omega) = \frac{\eta}{2} \qquad \text{ที่ทุกความถี่} \tag{3.16}$$



รูปที่ 3.1 แสดง Power Spectral Density(PSD) ของสัญญาณรบกวนขาว

จากรูปที่ 3.1 เป็นที่ควรสังเกตว่า การกำหนดนิยามสัญญาณรบกวนขาวว่า คือสัญญาณรบกวนที่มี ค่าพีเอสดี คงที่ตลอดทุกความถี่นั้น เป็นการกำหนดโดยความเป็นลักษณะตามอุดมคติ ทั้งนี้เพราะว่าการกำหนดเช่นนี้จะทำให้ค่ากำลังเฉลี่ยของสัญญาณรบกวนชนิดนี้ มีค่ามากอนันต์ กล่าวคือ

$$\overline{n^2(t)} = \frac{1}{2\pi}\int_{\omega\to-\infty}^{\omega\to\infty}\left(\frac{\eta}{2}\right)d\omega \qquad \to\infty \tag{3.17}$$

ซึ่งสัญญาณที่มีค่ากำลังเฉลี่ยมหาศาลเช่นนี้ ย่อมไม่มีในทางปฏิบัติ แต่อย่างไรก็ตามการกำหนดนิยามเช่นนี้ นับได้ว่าเป็นรูปแบบที่ดีสำหรับเมื่อแบนด์วิดท์ของระบบที่กำลังใช้งานนั้นแคบกว่าแบนด์วิดท์ของสัญญาณรบกวนที่มีอยู่มาก ในกรณีเช่นนี้ระบบนั้นก็จะปฏิบัติการอยู่ในช่วงแบนด์วิดท์ที่จำกัดของสัญญาณรบกวนเท่านั้น เพราะฉะนั้นโดยทางปฏิบัติแล้ว สัญญาณรบกวนขาวที่เราสนใจก็จะเป็นสัญญาณรบกวนที่อยู่ในลักษณะ สัญญาณรบกวนขาวที่มีย่านความถี่จำกัด (band-limited white noise) เท่านั้น ซึ่งเมื่อเป็นเช่นนี้ก็จะเห็นว่าเพียงพอสำหรับการวิเคราะห์ระบบนั้น กล่าวอีกนัยหนึ่งได้ว่าในทางปฏิบัตินั้น ถึงแม้ระบบที่เรากำลังใช้อยู่จะถูกรบกวนด้วยสัญญาณรบกวนขาวตามอุดมคติ ส่วนประกอบของสัญญาณรบกวนขาว ที่มีความถี่พ้นไกลไปจากแบนด์วิดท์ของระบบนั้นก็จะไม่มีผลกระทบกับระบบที่เรากำลังให้ความสนใจอยู่เลย องค์ประกอบด้านความถี่ของสัญญาณรบกวนขาวที่มีอิทธิพลต่อระบบนั้น จะเป็นองค์ประกอบซึ่งมีความถี่ซึ่งอยู่ในย่านความถี่ที่จำกัดเท่านั้น

สัญญาณแบบที่มีสัญญาณรบกวนแบบแอดดิทีฟไวท์เกาส์เซียน (Additive White Gaussian Noise : AWGN) ซึ่งช่องสัญญาณแบบนี้จะมีคุณสมบัติดังนี้

1. สัญญาณกับสัญญาณรบกวนที่ไม่มีคอร์รีเลชั่นกัน กล่าวคือถ้า $y(t)=x(t)+n(t)$ จะได้ $\overline{y^2}=\overline{x^2}+\overline{n^2}$

2. สัญญาณรบกวนเป็นแบบไวท์เกาส์เซียนซึ่งมีค่าเฉลี่ยของ $n(t)$ เป็นศูนย์ และกำลังของสัญญาณรบกวน $\overline{n^2}=N=\eta B$ โดยที่ B เป็นแบนด์วิดท์ของช่องสัญญาณ

3. กำลังของสัญญาณอินพุทมีขนาดจำกัด $\overline{x^2} \leq S_i$

บทที่ 4
การหาค่าความถี่มูลฐานของสัญญาณเสียงพูด

4.1 กล่าวนำ

ระดับเสียงสูงต่ำในภาษา หรือในภาษาไทยเรียกว่าเสียงวรรณยุกต์นั้น เกิดจากการสั่นสะบัดเป็นจังหวะของเส้นเสียงในการออกเสียงก้อง ซึ่งคุณสมบัติที่สำคัญของเสียงก้องก็คือมีความเป็นคาบ และระดับเสียงจะสูงหรือต่ำนั้นสามารถสังเกตได้จากค่าความถี่ในการเกิดคาบที่เรียกว่าพิทช์นั่นเอง ซึ่งความถี่ในการเกิดพิทช์นี้เรียกว่าความถี่มูลฐาน โดยถ้ามีการสั่นสะบัดของเส้นเสียงอย่างรวดเร็วความถี่มูลฐานจะมีค่ามากเสียงที่เกิดขึ้นจะเป็นเสียงสูง ในทำนองเดียวกันถ้าความถี่มูลฐานมีค่าน้อยระดับเสียงที่เกิดขึ้นก็จะเป็นเสียงต่ำ ในภาษาไทยระดับเสียงสูง-ต่ำที่แตกต่างกันนี้มีผลต่อความหมายของคำในภาษา ดังนั้นพิทช์หรือค่าความถี่มูลฐานนี้จึงเป็นสิ่งสำคัญในการแยกแยะคำในภาษาไทย

4.2 การวิเคราะห์ในโดเมนเวลา

สัญญาณเสียงพูดเป็นสัญญาณที่เปลี่ยนแปลงไปตามเวลา โดยเกิดในลักษณะแบบสุ่ม (random) แต่ก็ขึ้นกับการควบคุมเสียงของผู้พูดด้วยเพราะเสียงที่เปล่งออกมาในระยะเวลาหนึ่งนั้นจะขึ้นอยู่กับรูปทรงของช่องทางเดินเสียง(vocal tract) และลักษณะการสั่นของเส้นเสียง(vocal cord) เสียงพูดจึงเป็นสัญญาณที่มีคาบเวลาชั่วขณะ(quasi-periodic) คือมีความเป็นคาบคงที่ในเวลาอันสั้นและมีการเปลี่ยนแปลงในช่วงระหว่างเวลานั้น ดังนั้นในการวิเคราะห์จึงต้องทำการแบ่งเสียงพูดออกเป็นช่วงๆ(Frame) โดยมีช่วงเวลาอยู่ระหว่าง 10-30 มิลลิวินาที ในช่วงเวลาดังกล่าวถือว่าเสียงจะมีการเปลี่ยนแปลงคุณสมบัติน้อยมาก ดังนั้นในแต่ละเฟรมจึงสมมติให้เสียงเป็นสัญญาณที่มีคุณลักษณะคงที่ ซึ่งทำให้การวิเคราะห์ทำได้ง่ายขึ้น

4.3 ทฤษฎีการประมาณค่าพิทช์โดยใช้ออโตคอร์รีเลชั่นฟังก์ชั่น

การวิเคราะห์โดยใช้ออโตคอร์รีเลชั่น[10]-[11] เป็นวิธีหนึ่งที่เป็นที่ยอมรับในการใช้ตรวจหาคาบพิทช์ โดยออโตคอร์รีเลชั่นฟังก์ชั่นจะทำหน้าที่ในการแสดงยอดกราฟหลัก (prominent peak) ที่เป็นคาบพิทช์ในแต่ละส่วนของเสียง(section) ซึ่งสามารถหาได้จากรายละเอียดของโครงสร้างของสัญญาณนั้นๆ

4.3.1 การจัดแบ่งการวิเคราะห์สัญญาณออกเป็นช่วงสั้นๆ (Short–Time Autocorrelation Analysis)

ถ้ากำหนดให้ discrete time signal แทนด้วย $x(m)$ ออโตคอร์รีเลชั่นฟังก์ชั่น ของ discrete-time deterministic signal โดยทั่วๆไปเขียนได้เป็น

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \tag{4.1}$$

ซึ่งออโตคอร์รีเลชั่นฟังก์ชั่น ของสัญญาณ โดยพื้นฐานก็คือการแปลงสัญญาณ (transfor mation) ดังนั้นการตรวจวัดค่าคาบพิทช์ สามารถทำได้โดย

ถ้าสัญญาณ x(m) มีความเป็นคาบที่แน่นอนด้วยระยะ P นั่นคือ

$$x(m) = x(m+p) \quad ; \text{สำหรับทุก m}$$

ดังนั้นสามารถเขียนได้ว่า

$$\phi(k) = \phi(k+p) \tag{4.2}$$

นั่นคือ ออโตคอร์รีเลชั่นฟังก์ชั่นก็มีความเป็นคาบด้วยระยะคาบเดียวกัน หรือในทางกลับกันก็คือ "ความเป็นคาบในออโตคอร์รีเลชั่นฟังก์ชั่น เป็นตัวบ่งชี้ให้เห็นถึงความเป็นคาบในสัญญาณ"

คุณสมบัติของออโตคอร์รีเลชั่นฟังก์ชั่นที่สำคัญ คือ

1. เป็นฟังก์ชั่นคู่ นั่นคือ $\phi(k) = \phi(-k)$
2. มีค่ามากที่สุดที่ $k=0$ นั่นคือ $|\phi(k)| \leq \phi(0)$; สำหรับทุกค่า k

ถ้าพิจารณาสมการที่ 4.2 ควบคู่ไปกับคุณสมบัติในข้อ 1 และ 2 จะพบความเป็นคาบของสัญญาณ โดยแซมเปิ้ลของออโตคอร์รีเลชั่นจะมีค่ามากที่สุดที่ $0, \pm p, \pm 2p, \ldots$ โดยไม่ต้องคำนึงถึงเวลาเริ่มต้น (time origin) ของสัญญาณ การคำนวณหาคาบของสัญญาณสามารถประมาณได้จากตำแหน่งแรกที่มีค่ามากที่สุดในออโตคอร์รีเลชั่น ฟังก์ชั่น ซึ่งจากคุณสมบัติเหล่านี้ทำให้ออโตคอร์รีเลชั่นฟังก์ชั่นเป็นหลักการพื้นฐานที่น่าสนใจในการใช้ประมาณค่าความเป็นคาบในสัญญาณทุกชนิด

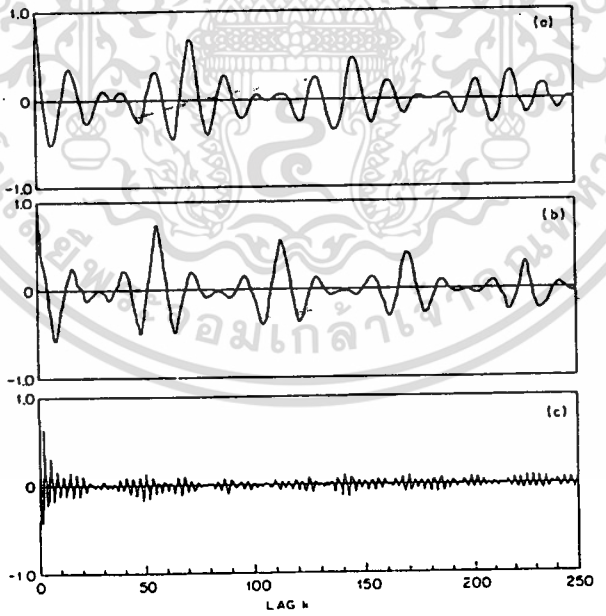
สำหรับสัญญาณที่มีลักษณะเปลี่ยนแปลงอยู่ตลอดเวลา เช่นสัญญาณเสียงพูดจะต้องทำการแบ่งสัญญาณออกเป็นช่วงสั้นๆเพื่อหาสารสนเทศ (information)ที่ต้องการ โดย short-time auto-correlation function สามารถนิยามได้เป็น

$$R(k) = \sum_{m=0}^{N-1-k} x(m)\, x(m+k) \tag{4.3}$$

เมื่อ N คือ จำนวนตัวอย่างสัญญาณ (sample)ต่อเฟรม

k คือ จำนวนจุดที่ใช้ในการคำนวณ ออโตคอร์รีเลชั่น

โดยในการกำหนดช่วงของค่า k จะกำหนดจากค่า k ที่น้อยที่สุดคือช่วงคาบเวลาพิทช์ของเสียงผู้หญิงซึ่งมีค่าเท่ากับ 3 มิลลิวินาที จนถึงค่า k สูงสุดเท่าที่เป็นไปได้คือช่วงคาบเวลาพิทช์ของเสียงผู้ชายซึ่งมีค่าเท่ากับ 20 มิลลิวินาที ดังนั้นถ้าใช้อัตราการชักตัวอย่างที่ 10 kHz ก็จะใช้ k อยู่ในช่วง 30 ถึง 200 และจากการแบ่งสัญญาณออกเป็นเฟรม จำนวนตัวอย่างสัญญาณในเฟรมซึ่งก็คือค่า N จะต้องมีค่ามากกว่า k มากๆ (N>>k) ซึ่งจะกล่าวถึงต่อไป

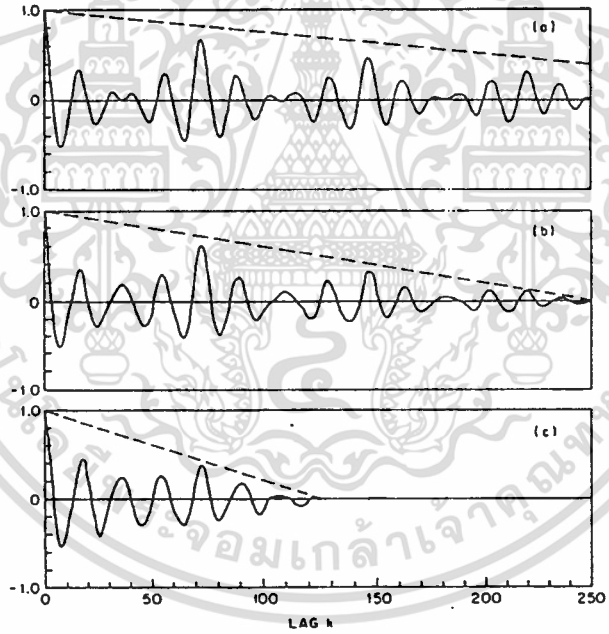


รูปที่ 4.1 ออโตคอร์รีเลชั่น ฟังก์ชั่น (a),(b) คือสัญญาณเสียงก้อง และ (c) คือสัญญาณเสียงไม่ก้อง โดยทั้ง 3 กรณีใช้ N = 401

พิจารณารูปที่ 4.1 แสดงตัวอย่างการคำนวณออโตคอร์รีเลชั่น ฟังก์ชั่นของสัญญาณเสียงพูดที่มีอัตราการชักตัวอย่างด้วยความถี่ 10 kHz โดยใช้สมการคำนวณที่ 4.3 ด้วย N = 401 และ

ค่าการเลื่อนของเวลา (lag) เป็น $0 \le k \le 250$ รูป 4.1(a-b) เป็นส่วนของสัญญาณเสียงที่มีความเป็นคาบ และรูป 4.1(c) คือส่วนของสัญญาณที่ไม่มีความเป็นคาบ จากรูปบน (a) ตำแหน่งสูงสุด (peak) เกิดที่ตำแหน่ง 72 นั่นคือสัญญาณมีคาบที่ระยะ 7.2 msec หรือมีค่าความถี่มูลฐานประมาณ 140 Hz (10 kHz / 72) ในรูป (b) ค่าสูงสุดของออโตคอร์รีเลชั่นเกิดในตำแหน่งที่ 58 แสดงให้เห็นว่ามีค่าเฉลี่ยของคาบในช่วง 5.8 msec ส่วนรูป (c) เป็นส่วนของสัญญาณที่ไม่มีความเป็นคาบ ออโตคอร์รีเลชั่น ฟังก์ชั่นจะประกอบด้วยองค์ประกอบของความถี่สูงคล้ายๆรูปคลื่นของสัญญาณรบกวน (Noise-like waveform)

ในการเลือกค่าของจำนวนตัวอย่าง (N) ที่ใช้ในแต่ละเฟรม รูปคลื่นสัญญาณจะต้องมีความเป็นคาบที่สมบูรณ์ (complete period) อย่างน้อย 2-3 คาบ ซึ่งในความเป็นจริงแล้วความยาวของสัญญาณเสียงพูดมีผลต่อการคำนวณของ $R_n(k)$ เนื่องจากค่าของ $R_n(k)$ จะลดลงเรื่อยๆเมื่อ k มีค่าเพิ่มมากขึ้น ซึ่งมีผลโดยตรงต่อแอมปลิจูดสูงสุด (peak) ของคอร์รีเลชั่นเนื่องจากจะมีค่าลดลงเช่นกัน



รูปที่ 4.2 ออโตคอร์รีเลชั่น ฟังก์ชั่น สำหรับเสียงก้องโดยใช้ค่า N ที่แตกต่างกันคือ (a) N = 401; (b) N = 251 และ (c) N = 125

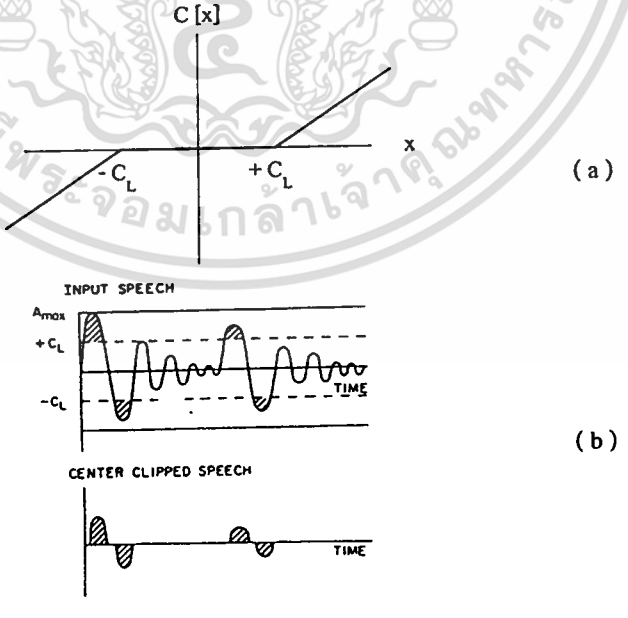
จากรูป 4.2(a) และ 4.2(b) จากการคำนวณพบว่าตำแหน่งคาบจริงอยู่ในตำแหน่งที่ 72 ส่วนในรูป 4.2(c) จะพบว่าค่าสูงสุดของคอร์รีเลชั่นอยู่ในตำแหน่งที่ 15 ทั้งนี้เนื่องจากวินโดว์ที่ใช้มีขนาดสั้นเกินไปเมื่อเทียบกับขนาดของคาบพิทช์ (pitch period) จึงทำให้ตำแหน่งสูงสุดของคอร์รีเลชั่นที่คำนวณได้ผิดพลาดไปจากตำแหน่งจริง

4.3.2 การกำจัดผลของโครงสร้างฟอร์แมนท์ด้วยวิธีเซ็นเตอร์คลิปปิ้ง

จากตัวอย่างรูปที่ 4.2 จะเห็นว่าออโตคอร์รีเลชั่น ฟังก์ชั่นมียอดของกราฟจำนวนมาก ซึ่งยอดของกราฟเหล่านี้เป็นผลมาจากผลตอบสนองทางความถี่ที่เกิดในช่องทางเดินเสียง (vocal tract response) ซึ่งมีผลต่อรูปทรงในแต่ละคาบของสัญญาณเสียงพูด โดยในรูปที่ 4.2(c) ตำแหน่งสูงสุดของออโตคอร์รีเลชั่นผิดไปจากตำแหน่งคาบจริงเนื่องจากวินโดว์มีขนาดสั้นไปเมื่อเทียบกับคาบพิทช์ แต่ในขณะเดียวกันการเปลี่ยนแปลงอย่างรวดเร็วของความถี่ฟอร์แมนท์ก็มีผลให้เกิดปรากฏการณ์ในลักษณะเดียวกันนี้ด้วยเช่นกัน ซึ่งในกรณีนี้ยอดสูงสุดของออโตคอร์รีเลชั่นที่เกิดเนื่องจากผลตอบสนองทางความถี่ในช่องทางเดินเสียงจะมีขนาดใหญ่กว่ายอดกราฟที่เกิดจากความเป็นคาบของแหล่งกำเนิดเสียง (vocal excitation) ซึ่งเหตุการณ์ลักษณะเช่นนี้จะทำให้การเลือกตำแหน่งยอดกราฟที่สูงที่สุดของออโตคอร์รีเลชั่น ฟังก์ชั่น เกิดการผิดพลาดด้วย

ดังนั้นเพื่อที่จะหลีกเลี่ยงปัญหานี้ จึงได้มีการเสนอกรรมวิธีเพื่อที่จะจัดการสัญญาณให้ความเป็นคาบของสัญญาณนี้เด่นชัดขึ้นโดยการขจัดลักษณะของสัญญาณที่จะทำให้เกิดความไขว้เขว (distracting) ออกไป เทคนิคนี้เรียกว่า การทำสเปกตรัมราบเรียบ (spectrum flatteners) ซึ่งมีอยู่หลายวิธี [12] เซ็นเตอร์คลิปปิ้งก็เป็นวิธีหนึ่งที่สะดวกและสามารถคำนวณได้จากสัญญาณโดยตรง ซึ่งถูกเสนอขึ้นโดยใช้การแปลงสัญญาณแบบไม่เป็นเชิงเส้น (nonlinear) (Sondhi. 1968)

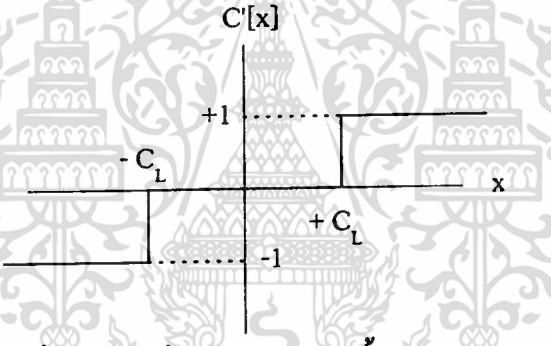
$$y(n) = C[x(m)] \tag{4.4}$$



รูปที่ 4.3 (a) ฟังก์ชั่น เซ็นเตอร์ คลิปปิ้ง (b) ตัวอย่างแสดงการคลิปสัญญาณเสียงพูด

เมื่อ C[x] แสดงได้ดังรูปที่ 4.3(a) วิธีนี้อาศัยหลักการคือสัญญาณเสียงพูดจะถูกนำมาหาค่าแอมปลิจูดสูงสุด A_{max} เพื่อนำมากำหนดระดับในการคลิปสัญญาณ (clipping level : C_L) จากนั้นค่าของสัญญาณที่มีระดับต่ำกว่าระดับคลิปปิ้งจะถูกกำหนดให้มีค่าเป็นศูนย์ ส่วนสัญญาณที่มีระดับสูงกว่าระดับ คลิปปิ้งจะถูกลบออกด้วยระดับคลิปปิ้ง ดังรูปที่ 4.3(b) จะพบว่าสัญญาณยังคงความเป็นคาบของสัญญาณเดิม แต่ส่วนของสัญญาณที่เกิดจากอิทธิพลของโครงสร้างฟอร์แมนท์(อันเนื่องมาจากการตอบสนองทางความถี่ภายในช่องทางเดินเสียง)จะถูกกำจัดออกไป แต่ในการกำหนดระดับการคลิปสัญญาณจะต้องระมัดระวังว่าระดับที่กำหนดจะต้องไม่สูงเกินไปจนทำให้สารสนเทศสูญหาย (Sondhi ใช้ 30% ของค่า A_{max})

จากวิธีดังกล่าวจะเห็นว่า วิธีเซ็นเตอร์คลิปปิ้งเป็นวิธีที่สะดวกในการทำให้สเปคตรัมราบเรียบ ซึ่งต่อมาได้มีการพัฒนาวิธีการนี้บนอุปกรณ์ดิจิตอล (Dobnoeski. 1976) โดยทำการปรับปรุงฟังก์ชั่นของเซ็นเตอร์คลิปปิ้งให้ง่ายต่อการคำนวณ แสดงได้ดังรูปที่ 4.4

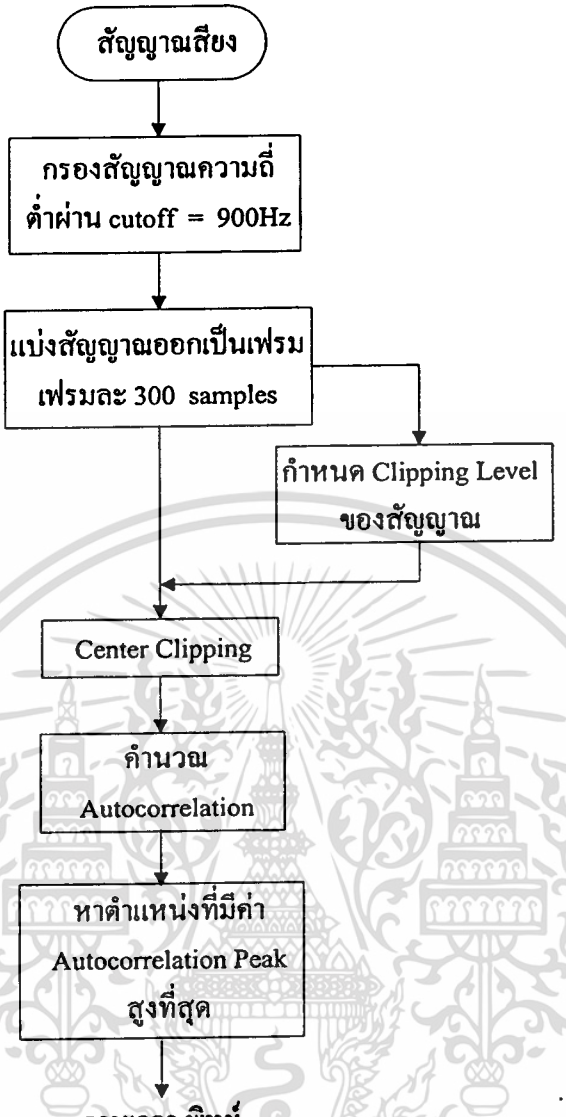


รูปที่ 4.4 ฟังก์ชั่น เซ็นเตอร์คลิปปิ้ง แบบ 3 ระดับ

โดยสัญญาณที่ผ่านการคลิปจะมีค่าเป็น +1 ถ้า $x(m) > C_L$,-1 ถ้า $x(m) < -C_L$ และมีค่าเป็น 0 ถ้า $-C_L \le x(m) \le C_L$ ฟังก์ชั่นนี้เรียกว่า เซ็นเตอร์คลิปปิ้งแบบ 3 ระดับ (3-level center clipping) การกำหนดค่าในลักษณะนี้จะช่วยลดความซับซ้อนในการคำนวณออโตคอร์รีเลชั่นฟังก์ชั่นลง เนื่องจากแต่ละพจน์ในสมการที่ (4.3) อยู่ในรูปของ $x(m)x(m+k)$ และค่า $x(m)$ จะมีค่าได้เพียง 3 ค่า คือ +1, 0, -1 เท่านั้น ดังนั้นผลคูณในสมการที่ (4.3) จึงสามารถมีค่าได้เป็น

$$\begin{aligned} x(m)x(m+k) &= 0 \quad \text{ถ้า} \quad x(m) = 0 \ \text{หรือ} \ x(m+k) = 0 \\ &= +1 \quad \text{ถ้า} \quad x(m) = x(m+k) \\ &= -1 \quad \text{ถ้า} \quad x(m) \neq x(m+k) \end{aligned} \tag{4.5}$$

โดยอัลกอริธึมในการหาค่าพิทช์ที่ถูกพัฒนาขึ้น สามารถแสดงได้ดังรูปที่ 4.5



รูปที่ 4.5 อัลกอริธึม การหาค่าพิทช์

รายละเอียดของขั้นตอนมีดังนี้ คือ

1. นำสัญญาณเสียงพูดที่ได้จากการชักตัวอย่าง มาผ่านตัวกรองความถี่ต่ำผ่านที่มีความถี่คัทออฟประมาณ 900 Hz เพื่อทำการกำจัดอิทธิพลของโครงสร้างความถี่ฟอร์แมนท์ ที่จะเกิดบนออโตคอร์รีเลชั่นฟังก์ชั่น
2. แบ่งสัญญาณออกเป็นเฟรม มีขนาดเฟรมละ 300 ตัวอย่างเพื่อทำการวิเคราะห์ โดยในการเลื่อนเฟรมกำหนดให้มีส่วนของเฟรมซ้อนทับกัน 2 ใน 3 ส่วน
3. ทำการแบ่งข้อมูลในเฟรมออกเป็นส่วนๆ ส่วนละ 100 ตัวอย่าง โดยในส่วนแรกและส่วนที่สามจะถูกนำมาหาค่าแอมปลิจูดสมบูรณ์ที่มีค่าสูงที่สุดของแต่ละส่วน เพื่อนำมากำหนดระดับคลิปปิ้ง โดยเลือกจากค่าแอมปลิจูดที่น้อยกว่าคูณกับเปอร์เซ็นต์ที่กำหนดขึ้น

สามารถคำนวณหาระดับคลิปปิ้งได้จากสมการต่อไปนี้

$$C_L = (S\%) \times min(K_1, K_2) \tag{4.6}$$

โดยที่ C_L = Clipping Level

K_1 = Absolute Amplitude Peak ของ 100 samples แรก ของเฟรม

K_2 = Absolute Amplitude Peak ของ 100 samples ท้าย ของเฟรม

$S\%$ = เปอร์เซ็นต์ที่กำหนดขึ้น (อยู่ภายในช่วง 20-80%)

4. เมื่อกำหนดระดับคลิปปิ้งแล้ว ค่าของสัญญาณอินพุทจะถูกกำหนดใหม่โดยใช้ วิธีเซ็นเตอร์คลิปปิ้งแบบ 3 ระดับ โดยสัญญาณที่มีค่าอยู่ในช่วง $\pm C_L$ จะถูกกำหนดให้มีค่าเป็นไปตามความสัมพันธ์ดังนี้

$$y(m) = \text{sgn}[x(m)] = \begin{cases} 1 & , x(m) \geq C_L \\ 0 & , |x(m)| < C_L \\ -1 & , x(m) \leq -C_L \end{cases} \tag{4.7}$$

เมื่อ $\text{sgn}[x(m)]$ คือ สัญญาณที่ผ่านการคลิป จากนั้นนำค่าที่กำหนดใหม่นี้ไปทำการคำนวณออโตคอร์รีเลชั่นฟังก์ชั่นเพื่อทำการหาคาบพิทช์ของสัญญาณเสียง

5. จากการหาค่าสูงสุดของออโตคอร์รีเลชั่นฟังก์ชั่นจะสามารถจำแนกชนิดของเสียงได้ ว่าเสียงในช่วงนั้นเป็นเสียงก้อง(voice) หรือเสียงไม่ก้อง (unvoiced) โดยนำค่ายอดสูงสุดที่ได้มาเทียบกับระดับที่กำหนด (ประมาณ 30% ของ R(0)) ถ้ายอดนี้มีค่าเกิน ตำแหน่งสูงสุดนั้นให้ถือว่าเป็นคาบพิทช์ของสัญญาณ แต่ถ้ามีค่าต่ำกว่าให้ถือว่าเสียงส่วนนั้นเป็นเสียงไม่ก้อง

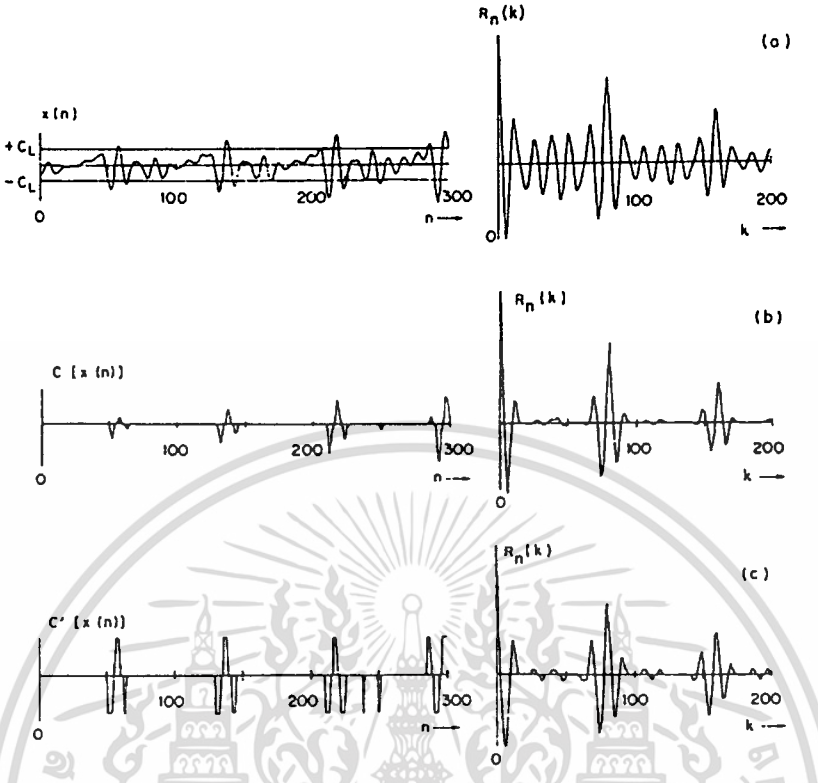
จากค่าคาบเวลาพิทช์ที่ได้นี้สามารถนำมาหาค่าความถี่มูลฐาน F_o ได้จากความสัมพันธ์ คือ

$$F_o = \frac{F_s}{P} \tag{4.8}$$

เมื่อ F_o = ความถี่มูลฐาน (Hz)

F_s = ความถี่ที่ใช้ในการสุ่มสัญญาณ

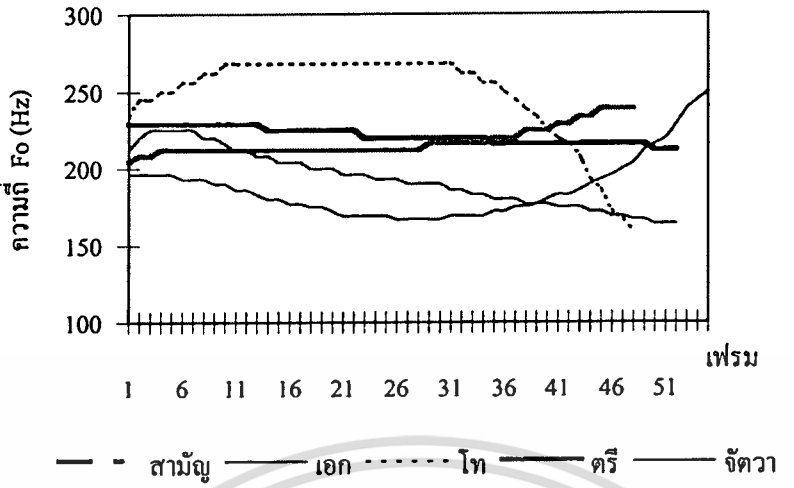
P = คาบเวลา พิทช์



รูปที่ 4.6 ตัวอย่างสัญญาณเสียงและฟังก์ชั่น คอร์รีเลชั่น
(a) ไม่มีการคลิปสัญญาณ
(b) คลิปสัญญาณโดยใช้ เซ็นเตอร์ คลิปปิ้ง
(c) คลิปสัญญาณ โดยใช้ เซ็นเตอร์ คลิปปิ้ง แบบ 3 ระดับ

จากรูปที่ 4.6(a) แสดงผลของการคำนวณออโตคอร์รีเลชั่นจากสัญญาณที่ไม่ผ่านเซ็นเตอร์คลิปปิ้ง จะเห็นว่ารูปกราฟประกอบด้วยจุดยอดจำนวนมากอันเกิดเนื่องจากผลตอบสนองทางความถี่ของช่องทางเดินเสียง ส่วนรูปที่ 4.6(b,c) จะสังเกตเห็นว่าสัญญาณที่ผ่านเซ็นเตอร์คลิปปิ้งจะเหลืออยู่แต่สัญญาณที่มีค่าดาบพิทช์ และผลที่ได้จากการคำนวณออโตคอร์รีเลชั่นจะมีจุดยอดที่จะทำให้เกิดความสับสนเหลืออยู่น้อยมาก

ระดับเสียงของคำในพยางค์หนึ่งๆ จะมีระดับสูงหรือต่ำนั้นสามารถสังเกตได้จากค่าพิทช์หรือค่าความถี่มูลฐาน โดยระดับเสียงของคำในภาษาไทยจะมีระดับเสียงวรรณยุกต์ใดนั้นสามารถสังเกตได้จากแนวทางการเปลี่ยนแปลงของค่าความถี่มูลฐานเมื่อเทียบกับเวลา ดังแสดงในรูปที่ 4.7



รูปที่ 4.7 แสดงค่าความถี่มูลฐานของคำ อา อ่า อ้า อ๊า อ๋า จากผู้ออกเสียงเพศหญิง

จากรูปที่ 4.7 แสดงการเปลี่ยนแปลงค่าความถี่มูลฐานที่ได้จากการคำนวณออโตคอร์รีเลชั่นที่ผ่านกระบวนการเซ็นเตอร์คลิปปิ้ง จะเห็นว่าในแต่ละระดับเสียงวรรณยุกต์จะมีแนวทางการเปลี่ยนแปลงของค่าความถี่มูลฐานที่มีลักษณะเฉพาะตัวแตกต่างกัน

4.4 สรุป

เนื้อหาในบทนี้กล่าวถึงการหาค่าความถี่มูลฐาน ซึ่งเป็นตัวบ่งบอกถึงระดับเสียงสูง-ต่ำของคำในภาษา โดยสัญญาณข้อมูลจะถูกนำมาผ่านการทำสเปกตรัมราบเรียบด้วยวิธี เซ็นเตอร์คลิปปิ้ง แล้วทำการประมาณคาบพิทช์ด้วยวิธีออโตคอร์รีเลชั่น จากนั้นค่าคาบพิทช์จะถูกแปลงให้อยู่ในรูปของค่าความถี่มูลฐาน โดยรูปแบบการเปลี่ยนแปลงของค่าความถี่มูลฐานเมื่อเทียบกับเวลา จะเป็นตัวบ่งบอกถึงระดับเสียงวรรณยุกต์ที่แตกต่างกันของคำหรือพยางค์ในภาษาไทย ซึ่งลำดับของค่าความถี่มูลฐานที่แตกต่างกันนี้จะถูกนำไปเข้าสู่กระบวนการเตรียมข้อมูล เพื่อใช้เป็นข้อมูลฝึกสอนในการสร้างแบบจำลองอ้างอิงการรู้จำระดับเสียงวรรณยุกต์ทั้ง 5 ระดับ

บทที่ 5
การเตรียมข้อมูลเพื่อสร้างแบบจำลอง

5.1 กล่าวนำ

การจดจำเสียงพูด เป็นลักษณะหนึ่งของการจดจำรูปแบบ (Pattern Recognition) คือจะเป็นการเปรียบเทียบระหว่างแบบทดสอบ(Test Pattern) กับแบบอ้างอิง(Reference Pattern) ซึ่งเป็นรูปแบบที่ทราบและเก็บไว้ล่วงหน้า

ขั้นตอนในการจดจำแบ่งเป็น 2 ขั้นตอน ดังนี้

1. ขั้นตอนการเรียนรู้ (Learning)

   จะเป็นการสร้างกลุ่มของแบบอ้างอิงในการจดจำเสียงพูด ในขั้นตอนนี้จะทำการวิเคราะห์เสียงพูดก่อน โดยดึงลักษณะของพารามิเตอร์ที่ต้องการออกมา ซึ่งในวิทยานิพนธ์นี้ก็คือแนวทางการเปลี่ยนแปลงของค่าความถี่มูลฐานของเสียง จากนั้นทำการจัดกลุ่มพารามิเตอร์โดยใช้การ ควอนไตซ์ข้อมูล เพื่อนำไปสร้างแบบจำลองอ้างอิงในการรู้จำต่อไป

2. ขั้นตอนการจดจำ (Recognition)

   จะเป็นการทดสอบการจดจำระหว่างแบบอ้างอิงกับแบบทดสอบ โดยจะทำการเปรียบเทียบพารามิเตอร์ของแบบทดสอบกับแบบอ้างอิงทั้งหมด แบบอ้างอิงที่เลือกคือ แบบอ้างอิงที่มีพารามิเตอร์ใกล้เคียงกับแบบทดสอบที่สุด

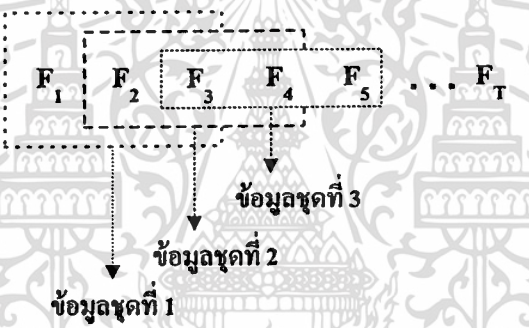
สัญญาณเสียงพูดที่ผ่านขั้นตอนในการหาค่าความถี่มูลฐานแล้ว ก่อนที่จะถูกนำมาเป็นเสียงต้นแบบเพื่อใช้ในกระบวนการสร้างแบบจำลองอ้างอิง หรือใช้เป็นแบบทดสอบ จะต้องนำมาผ่านกระบวนการในการเตรียมข้อมูลเสียก่อน เพื่อที่จะขจัดข้อจำกัดอันเนื่องมาจากความถี่มูลฐานที่แตกต่างกันระหว่างผู้ออกเสียงที่เป็นชายและหญิง โดยมีวัตถุประสงค์เพื่อให้แบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถใช้งานร่วมกันได้ไม่ว่าผู้ออกเสียงจะเป็นชายหรือหญิง ซึ่งกระบวนการเตรียมข้อมูลมี 2 ขั้นตอน คือ ขั้นตอนการปรับปรุงความต่อเนื่องของข้อมูลด้วยวิธีการกรองค่ากลาง และขั้นตอนการควอนไตซ์ทิศทางการเปลี่ยนแปลงของค่าความถี่มูลฐาน โดยรายละเอียดในแต่ละขั้นตอนมีดังนี้

5.2 การปรับปรุงความต่อเนื่องของข้อมูลด้วยวิธีการกรองค่ากลาง (Median Filtering)

สัญญาณเสียงที่ผ่านขั้นตอนในการหาค่าความถี่มูลฐานแล้ว อาจมีความไม่ต่อเนื่องของลำดับความถี่เกิดขึ้น เนื่องจากความไม่ต่อเนื่องของสัญญาณเสียงในช่วงต้นของการออกเสียงพูด และจากการปัดเศษในการคำนวณ ดังนั้น ขั้นตอนแรกของการเตรียมข้อมูลก็คือ นำสัญญาณเสียงที่ผ่านขั้นตอนในการหาค่าความถี่มูลฐานแล้วมาผ่านตัวกรองค่ากลาง เพื่อปรับปรุงให้ข้อมูลมีความต่อเนื่องเพิ่มขึ้น [13] โดยลำดับของความถี่มูลฐานซึ่งเป็นข้อมูลอินพุทจะอยู่ในรูปของข้อมูล 1 มิติ ขนาด [1xT] เมื่อ T คือจำนวนเฟรมของสัญญาณเสียง

ขั้นตอนการทำงานของการกรองค่ากลาง

ทำการจัดเรียงค่าความถี่มูลฐานออกเป็นชุดข้อมูล โดยในแต่ละชุดข้อมูลประกอบด้วยค่าความถี่ 3 ค่า โดยกำหนดให้มีการเลื่อนของชุดข้อมูลแสดงได้ดังรูป 5.1



รูปที่ 5.1 การจัดแบ่งความถี่มูลฐานออกเป็นชุดข้อมูล

เมื่อ F_1 = ค่าความถี่มูลฐานของเฟรมที่ 1

F_2 = ค่าความถี่มูลฐานของเฟรมที่ 2

F_T = ค่าความถี่มูลฐานของเฟรมสุดท้าย

จากนั้นนำค่าความถี่ทั้ง 3 ค่า ในแต่ละชุดข้อมูลมาจัดเรียงใหม่ตามความสัมพันธ์

$$a \leq b \leq c \tag{5.1}$$

โดยที่

a = ความถี่ F_0 ที่มีค่าน้อยที่สุดของแต่ละชุดข้อมูล

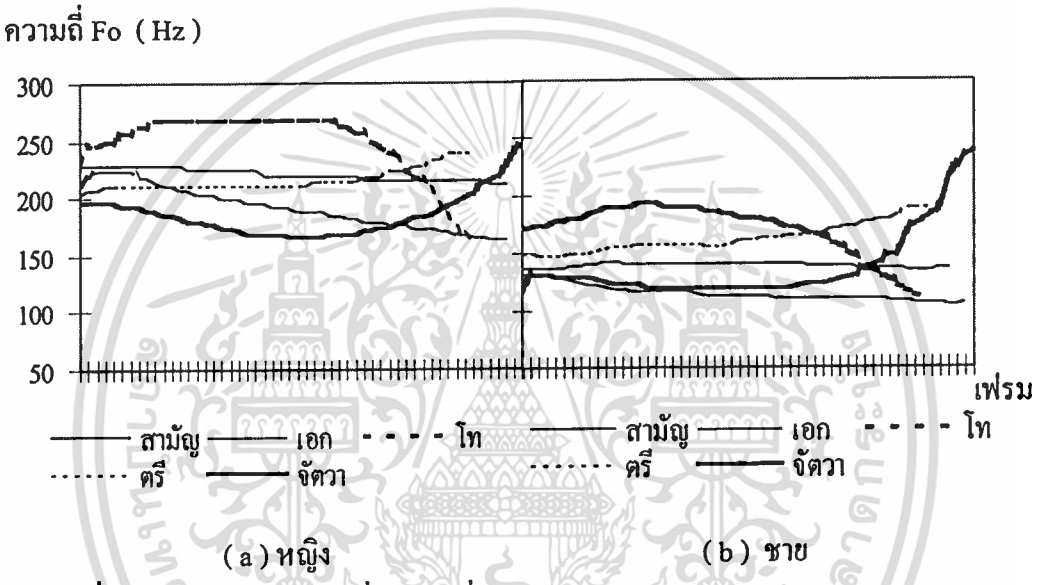
b = ความถี่ F_0 ที่มีค่าอยู่ระหว่างกลาง

c = ความถี่ F_0 ที่มีค่ามากที่สุดของแต่ละชุดข้อมูล

จากนั้นนำความถี่ค่ากลาง (b) ที่ได้จากชุดข้อมูลแต่ละชุดมาจัดเรียงตามลำดับ ก็จะได้ความถี่มูลฐานชุดใหม่ที่ผ่านกระบวนการกรองค่ากลางแล้ว

5.3 การควอนไตซ์ทิศทางการเปลี่ยนแปลงของค่าความถี่มูลฐาน

ขั้นตอนสุดท้ายของการเตรียมข้อมูลก็คือ การควอนไตซ์ทิศทางการเปลี่ยนแปลงของค่าความถี่มูลฐาน จากข้อเท็จจริงที่ว่าระดับความถี่มูลฐานของเสียงชายและหญิงมีความแตกต่างกัน ซึ่งโดยเฉลี่ยแล้ว ในผู้ชายความถี่มูลฐานจะมีค่าอยู่ในช่วง 80-160 Hz และ 160-400 Hz ในผู้ออกเสียงที่เป็นหญิง [14] ดังตัวอย่างในรูป 5.2



รูปที่ 5.2 แสดงระดับความถี่มูลฐานที่แตกต่างกันระหว่าง (a) หญิง และ (b) ชาย

จากรูปจะสังเกตเห็นว่าลักษณะการเปลี่ยนแปลงของค่าความถี่มูลฐานในแต่ละระดับเสียงวรรณยุกต์จะมีรูปแบบการเปลี่ยนแปลงที่มีลักษณะเฉพาะโดยไม่ขึ้นกับผู้ออกเสียงว่าเป็นเพศใด

ดังนั้นในวิทยานิพนธ์นี้จึงได้ดึงเอาลักษณะเด่นนี้มาใช้ในการสร้างแบบจำลองการรู้จำระดับเสียงวรรณยุกต์ โดยทำการจัดกลุ่มค่าความถี่มูลฐานออกเป็น 3 ระดับตามแนวทางการเปลี่ยนแปลงของความถี่ (ΔF) ที่เพิ่มขึ้นหรือลดลงเมื่อเวลาเปลี่ยนไป

โดย

$$\Delta F_t = F_{t+1} - F_t \qquad (5.2)$$

เมื่อ $t = 1, 2, \ldots, (T-1)$ โดย T คือ จำนวนเฟรม

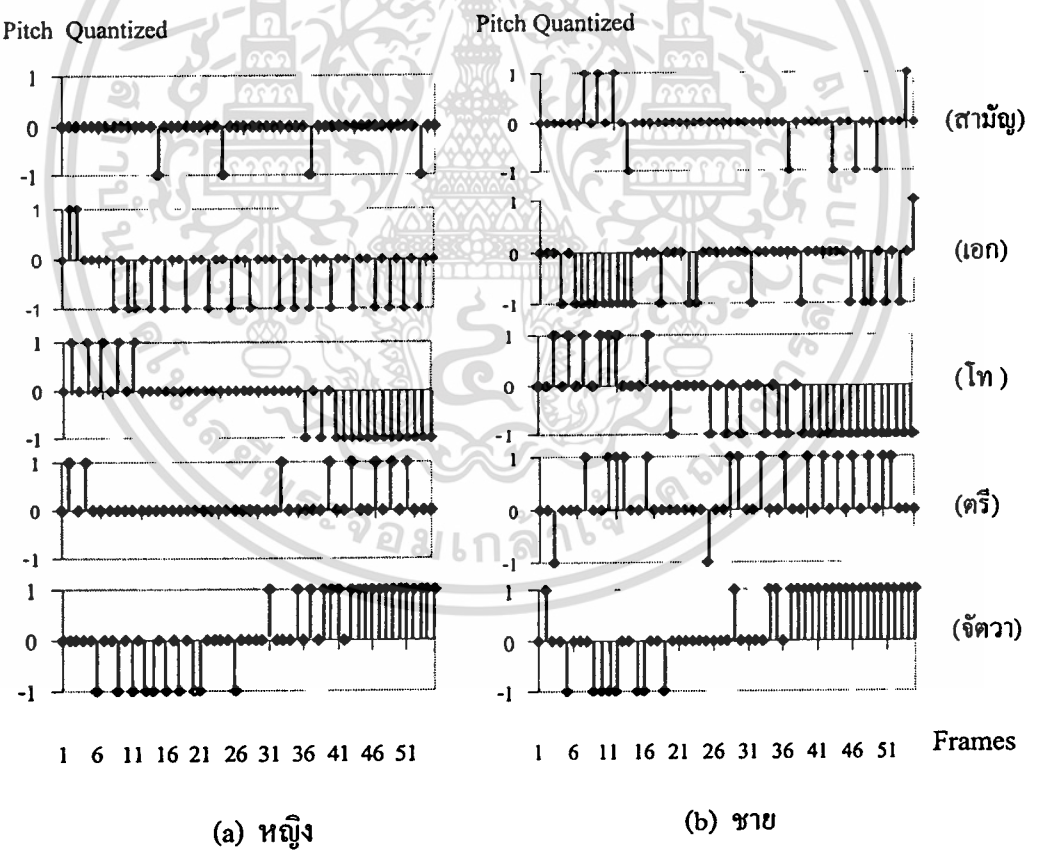
F_t = ความถี่ F_0 ที่เวลา t

F_{t+1} = ความถี่ F_0 ที่เวลา t+1

จากนั้นทำการควอนไตซ์ ΔF โดยแบ่งออกเป็น 3 ระดับตามทิศทางการเปลี่ยนแปลงของความถี่มูลฐาน โดยกำหนดให้

$$V_t = \begin{cases} 1 \; ; \; \Delta F_t > 0 \\ 0 \; ; \; \Delta F_t = 0 \\ -1 \; ; \; \Delta F_t < 0 \end{cases} \qquad (5.3)$$

จากนั้นค่า $V_t = \{-1, 0, 1\}$ จะถูกนำไปใช้เป็นข้อมูลฝึกสอน (Trainning) เพื่อใช้ในการสร้างแบบจำลองอ้างอิงของเสียงวรรณยุกต์ต่อไป ซึ่งจะเห็นว่าการควอนไตซ์ความถี่ออกเป็น 3 ระดับนี้ นอกจากจะขจัดข้อจำกัดของความถี่มูลฐานที่แตกต่างกันระหว่าง ชาย, หญิงแล้ว ยังช่วยลดเนื้อที่ของหน่วยความจำในการจัดเก็บข้อมูล และทำให้การคำนวณทำได้เร็วขึ้นเมื่อเทียบกับการใช้ช่วงความถี่มูลฐานทั้งหมดมาสร้างแบบจำลอง



รูปที่ 5.3 แสดงการจัดแบ่งค่าความถี่มูลฐานออกเป็น 3 ระดับตามทิศทางการเปลี่ยนแปลงต่อเวลาที่เพิ่มขึ้น จากผู้ออกเสียงที่เป็น (a) หญิง และ (b) ชาย

เมื่อนำค่าความถี่มูลฐานในรูปที่ 5.2 มาทำการจัดระดับค่าการเปลี่ยนแปลงความถี่ออกเป็น 3 ระดับจะแสดงได้ดังรูปที่ 5.3 และเมื่อพิจารณาแนวทางการเปลี่ยนแปลงในระดับเสียงวรรณยุกต์ทั้ง 5 ระดับ จะพบว่า

1. เสียงสามัญ ตลอดทั้งเสียง ความถี่มูลฐานมีการเปลี่ยนแปลงลดลงเล็กน้อย
2. เสียงเอก ความถี่มูลฐานของเสียงจะมีค่าลดลงอย่างต่อเนื่อง
3. เสียงโท ความถี่มูลฐานมีค่าเพิ่มขึ้นในช่วงแรก และลดลงอย่างต่อเนื่องในช่วงท้ายของเสียง
4. เสียงตรี ความถี่มูลฐานมีแนวโน้มเพิ่มขึ้น
5. เสียงจัตวา ความถี่มูลฐานมีค่าลดลงในช่วงแรก และเพิ่มขึ้นอย่างต่อเนื่องในช่วงท้ายของเสียง

5.4 สรุป

สัญญาณเสียงพูดที่ผ่านขั้นตอนในการหาค่าความถี่มูลฐานแล้ว ก่อนที่จะถูกนำมาเป็นเสียงต้นแบบเพื่อใช้ในกระบวนการสร้างแบบจำลองอ้างอิง หรือใช้เป็นแบบทดสอบ จะต้องนำมาผ่านกระบวนการในการเตรียมข้อมูลเสียก่อน โดยขั้นแรก ลำดับข้อมูลจะต้องนำมาผ่านการกรองค่ากลาง เพื่อปรับปรุงให้ข้อมูลมีความต่อเนื่องเพิ่มขึ้น จากนั้นจะทำการควอนไตซ์ข้อมูลออกเป็น 3 ระดับตามทิศทางการเปลี่ยนแปลงที่เพิ่มขึ้น คงที่ หรือลดลงของค่าความถี่มูลฐาน เพื่อที่จะขจัดข้อจำกัดอันเนื่องมาจากความถี่มูลฐานที่แตกต่างกัน ทำให้แบบจำลองอ้างอิงที่ถูกสร้างขึ้นนี้สามารถใช้ร่วมกันได้กับผู้ออกเสียงที่เป็นทั้งชายและหญิง อีกทั้งยังเป็นการลดเนื้อที่หน่วยความจำในการจัดเก็บข้อมูล และลดเวลาที่ใช้ในการคำนวณ โดยข้อมูลเอาท์พุทที่ได้จากการควอนไตซ์ออกเป็น 3 ระดับนี้ จะถูกใช้เป็นข้อมูลฝึกสอน หรือข้อมูลทดสอบ ของการสร้างแบบจำลองการรู้จำระดับเสียงวรรณยุกต์ทั้ง 5 ระดับด้วย ฮิดเดน มาร์คอฟ โมเดล

บทที่ 6
การสร้างแบบจำลองการรู้จำด้วยวิธี Hidden Markov Model

6.1 กล่าวนำ

แบบจำลองมาร์คอฟเป็นแบบจำลองทางสถิติซึ่งพัฒนามาเพื่อแบ่งกลุ่มของอนุกรมทางเวลา หรือสัญญาณที่ไม่คงที่ นั่นคือใช้สำหรับจัดกลุ่มของสัญญาณที่ไม่รู้จัก (Unknown signal) ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่งของสัญญาณ ซึ่งแบบจำลองมาร์คอฟได้ถูกนำมาประยุกต์ใช้ในการรู้จำเสียงพูด [15] และเป็นวิธีการที่วิทยานิพนธ์นี้เลือกใช้

แบบจำลองมาร์คอฟ แบ่งออกเป็น 2 ประเภท คือ แบบต่อเนื่อง(Continuous) และแบบไม่ต่อเนื่อง(Discrete-time) ในวิทยานิพนธ์นี้ได้เลือกใช้แบบไม่ต่อเนื่อง เพราะคุณลักษณะของข้อมูลที่ผ่านการควอนไตซ์ ซึ่งใช้เป็นข้อมูลอินพุท มีลักษณะเป็นชนิดไม่ต่อเนื่อง โดยเนื้อหาในบทนี้จะกล่าวถึงทฤษฎีที่ใช้ในการสร้างแบบจำลองการรู้จำจากเสียงต้นแบบ และขั้นตอนในการทดสอบการรู้จำ

6.2 ส่วนประกอบของแบบจำลองมาร์คอฟ

พารามิเตอร์สำคัญที่เกี่ยวข้องในการสร้างแบบจำลองอ้างอิง ที่ต้องรู้จักได้แก่

1. T คือ ความยาวของลำดับข้อมูลที่ได้จากการควอนไตซ์ค่าความถี่มูลฐาน ซึ่งจะใช้เป็นข้อมูลอินพุทในส่วนของ HMM โดยต่อไปจะเรียกแทนว่า "ลำดับของค่าปรากฏ"(Observation sequence) ซึ่งมีขนาดความยาวของลำดับ เท่ากับจำนวนเฟรมทั้งหมดในเสียงแต่ละเสียง
2. N คือ จำนวนสเตทในแบบจำลอง ถ้ากำหนดให้เซ็ตของสเตทเป็น $\{ 1, 2, \ldots, N \}$ จะสามารถแทนสเตทที่เปลี่ยนไปตามเวลา t ด้วย เซ็ตของ $Q = \{ q_1, q_2, \ldots, q_N \}$
3. M คือจำนวนของค่าปรากฏที่สามารถเป็นไปได้ต่อหนึ่งสเตท แทนสัญลักษณ์ ด้วย $V = \{ v_1, v_2, \ldots v_M \}$ ซึ่งจากการจัดระดับของการเปลี่ยนแปลงของความถี่ (ΔF_t) ออกเป็น 3 ระดับ จะได้เซ็ตของค่าปรากฏที่สามารถเป็นไปได้ในแต่ละสเตทมีค่าเป็น $V = \{-1, 0, 1\}$
4. ค่าความน่าจะเป็นในการย้ายสเตท : $A = \{ a_{ij} \}$

   โดย a_{ij} แทนการย้ายสเตทจาก i ไป j

   เมื่อ

$$a_{ij} = P[q_t = j \mid q_{t-1} = i] \quad ; 1 \quad i, j \quad N \tag{6.1}$$

5. การกระจายความน่าจะเป็น ของค่าปรากฏที่สามารถเป็นไปได้ภายในสเตท : B = $\{b_j(k)\}$

โดยที่ $b_j(k) = P[v_k \text{ ที่เวลา } t \mid q_j \text{ ที่เวลา } t] \quad ; 1 \quad k \quad M$ (6.2)

เป็นนิยามการกระจายสัญลักษณ์ในสเตท j เมื่อ $j = 1, 2, \ldots, N$

6. ค่าความน่าจะเป็นของการเป็นสเตทเริ่มต้น : $\pi = \{ \pi_i \}$

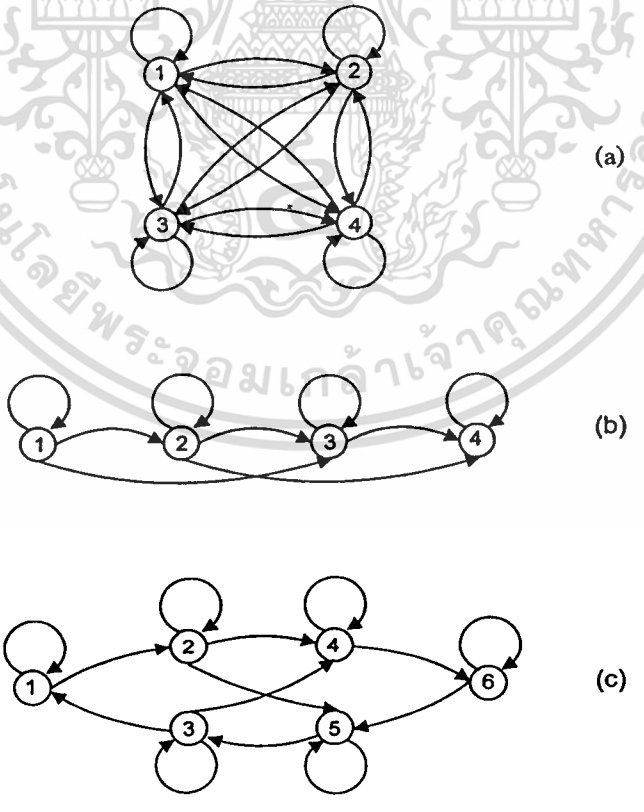
เมื่อ $\pi_i = P[q_i \text{ ที่เวลา } t=1] \quad ; 1 \quad i \quad N$ (6.3)

จะเห็นว่า Hidden Markov Model ต้องการพารามิเตอร์ของแบบจำลองคือ N, M และ กลุ่มของความน่าจะเป็น A,B,π ดังนั้นในการแสดงเซ็ตของพารามิเตอร์ที่สมบูรณ์ของแบบจำลองอ้างอิง จะแทนด้วยสัญลักษณ์

$$\lambda = (A,B,\pi) \quad (6.4)$$

6.3 ชนิดของ HMM

แบ่งชนิดตามการย้ายสเตทของเมตริกซ์ A



รูปที่ 6.1 แบบจำลองชนิดต่างๆของ HMM

1. HMM แบบ Ergodic Model หรือ Fully Connected Model

การย้ายสเตทสามารถย้ายไปยังทุกๆสเตทของแบบจำลอง ดังรูปที่ 6.1(a) เป็นตัวอย่างของแบบจำลองที่มี N = 4 ซึ่งจากรูปนี้มีค่าของเมตริกซ์ A เป็น

$$A = \begin{matrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{matrix}$$

2. HMM แบบ Left-Right Model หรือ Bakis Model

การย้ายสเตทจะย้ายจากซ้ายไปขวาซึ่งจะมีคุณสมบัติของสัมประสิทธิ์ในการย้ายสเตทดังนี้

$$a_{ij} = 0, \qquad j < i$$

คือจะไม่มีการย้ายสเตทไปยังสเตทที่ต่ำกว่าสเตทปัจจุบัน และนอกจากนี้ก็ยังมีความน่าจะเป็นของสเตทเริ่มต้นดังนี้

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases}$$

คือลำดับของสเตทจะต้องเริ่มที่สเตทที่ 1 เสมอ และ Left-Right Model นี้มักมีกฎบังคับการย้ายสเตท เพื่อไม่ให้มีการเปลี่ยนแปลงดัชนีของสเตทมากนัก กล่าวคือ

$$a_{ij} = 0, \qquad j > i + \Delta i$$

ดังรูปที่ 6.1(b) ค่าของ $\Delta i = 2$ คือจะไม่มีการย้ายข้ามสเตทไปเกิน 2 สเตท และมีเมตริกซ์ในการย้ายสเตทเป็น

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

จะเห็นว่าสเตทสุดท้าย สัมประสิทธิ์การย้ายสเตทจะเป็น

$$a_{NN} = 1$$

$$a_{Ni} = 0, \qquad i < N$$

แบบจำลองแบบนี้จะเหมาะกับสัญญาณที่มีลักษณะเปลี่ยนแปลงตามเวลาอย่างต่อเนื่อง เช่น เสียงพูด

3. HMM แบบ parallel Left-Right Model

เป็นแบบจำลองที่มีความยืดหยุ่นมากกว่าแบบที่ 2 แสดงได้ดังรูปที่ 6.1 (c)

6.4 ปัญหาพื้นฐานของแบบจำลอง มาร์คอฟ

ปัญหาของ HMM มี 3 ข้อ ซึ่งต้องใช้อัลกอริธึมวิธีต่างๆในการคำนวณเพื่อแก้ปัญหานี้

ปัญหาที่ 1 เมื่อมีลำดับของค่าปรากฏ $O = \{O_1\ O_2\ O_3 \ldots O_T\}$ และมีแบบจำลอง $\lambda = (A,B,\pi)$ จะคำนวณหาค่าความน่าจะเป็น $P(O|\lambda)$ ของลำดับค่าปรากฏนั้นได้อย่างไร

ปัญหาที่ 2 เมื่อมีลำดับของค่าปรากฏ $O = \{O_1\ O_2\ O_3 \ldots O_T\}$ และแบบจำลอง $\lambda = (A,B,\pi)$ จะคำนวณหาลำดับสเตท $q = \{q_1\ q_2\ q_3 \ldots q_T\}$ ที่เหมาะสมกับลำดับค่าปรากฏนั้นได้อย่างไร

ปัญหาที่ 3 เราจะปรับพารามิเตอร์ของแบบจำลอง $\lambda = (A,B,\pi)$ เพื่อให้ได้ค่า $P(O|\lambda)$ สูงสุดได้อย่างไร

การคำนวณเพื่อแก้ปัญหาของ HMM

การแก้ปัญหาที่ 1 เป็นการคำนวณหาว่าแบบจำลอง λ ใดๆ มีโอกาสจะให้ค่าลำดับเป็นไปตามลำดับของค่าปรากฏนั้น ด้วยค่าของความน่าจะเป็นมาก-น้อยเท่าใด

การแก้ปัญหาสามารถทำได้โดยระบุสเตทให้กับลำดับของค่าปรากฏซึ่งยาว T (โดยที่ค่าปรากฏหนึ่งตัวมีความเป็นไปได้ที่จะอยู่ในสเตทได้ N สเตท) ซึ่งสามารถเป็นไปได้ถึง N^T แบบ ให้สเตทต่างๆแทนด้วย

$$q = q_1\ q_2\ q_3 \cdots q_T \tag{6.5}$$

เมื่อ q_1 เป็นสเตทเริ่มต้นที่เวลา $t = 1$ ความน่าจะเป็นของลำดับของค่าปรากฏ O ที่กำหนดคือ

$$P(O|q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda) \tag{6.6a}$$

ความน่าจะเป็นในการเกิดค่าปรากฏคือ

$$P(O|q,\lambda) = b_{q_1}O_1 \quad b_{q_2}O_2 \quad \ldots \quad b_{q_T}O_T \tag{6.6b}$$

และ ความน่าจะเป็นในการย้ายข้ามสเตท q จะเป็น

$$P(q|\lambda) = \pi_{q_1} \; a_{q_1 q_2} \; a_{q_2 q_3} \; \cdots \; a_{q_{T-1} q_T} \tag{6.7}$$

ดังนั้นเมื่อนำความน่าจะเป็นของการเกิดค่าปรากฏ O และค่าความน่าจะเป็นในการย้ายสเตท q มารวมกัน ซึ่งนั่นก็คือความน่าจะเป็นที่ O และ q จะเกิดขึ้นพร้อมกัน จะได้

$$P(O, q|\lambda) = P(O|q, \lambda) P(q|\lambda) \tag{6.8}$$

$$= (b_{q_1} O_1 \; b_{q_2} O_2 \; \cdots \; b_{q_T} O_T)(\pi_{q_1} \; a_{q_1 q_2} \; a_{q_2 q_3} \; \cdots \; a_{q_{T-1} q_T})$$

โดยที่ความน่าจะเป็นของ O ได้มาจากผลรวมของความน่าจะเป็นที่ O และ q เกิดขึ้นพร้อมกันโดยคิดจากทุกสเตท q ที่จะเป็นไปได้ ดังนี้

$$P(O|\lambda) = \sum_{\text{all } q} P(O|q, \lambda) P(q|\lambda) \tag{6.9}$$

$$= \sum_{q_1 q_2 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_T O_T \tag{6.10}$$

ที่เวลาเริ่มต้น (t = 1) เราจะอยู่ที่สเตท q_1 ด้วยค่าความน่าจะเป็น π_{q_1} และแทนค่าความน่าจะเป็นในการเกิดค่าปรากฏ O_1 ที่สเตทนี้ด้วย $b_{q_1} O_1$

ที่เวลา เพิ่มขึ้นจาก $t \longrightarrow t+1$ (t=2) เราแทนการย้ายสเตทจากสเตท q_1 ไปยัง q_2 ด้วยค่าความน่าจะเป็น $a_{q_1 q_2}$ และแทนค่าความน่าจะเป็นในการเกิดค่าปรากฏเป็น O_2 ด้วยค่าความน่าจะเป็น $b_{q_2} O_2$

จนกระทั่ง ที่เวลา T เราแทนการย้ายสเตทจากสเตท q_{T-1} ไปยัง q_T ด้วยค่าความน่าจะเป็น $a_{q_{T-1} q_T}$ และแทนค่าความน่าจะเป็นในการเกิดค่าปรากฏเป็น O_T ด้วยค่าความน่าจะเป็น $b_{q_T}(O_T)$

จะเห็นว่าสมการนี้มีการคำนวณที่ยุ่งยากเนื่องจากมีการคูณกันเป็นจำนวนมากในรูปของลำดับ $2T*N^T$ ดังนั้นจึงมีการคิดหาวิธีมาช่วย ซึ่งแบ่งออกเป็น

1. กระบวนการไปข้างหน้า (Forward Procedure) ; $\alpha_t(i)$ = Forward variable

นิยาม

$$\alpha_t(i) = P(O_1 O_2 \ldots O_T, q_t = i|\lambda) \tag{6.11}$$

คือ ความน่าจะเป็นของการเกิดลำดับของค่าปรากฏ $O_1 O_2 \ldots O_T$ และอยู่ที่สเตท q_i ณ เวลา t โดยมีแบบจำลองเป็น λ เราสามารถหา $\alpha_t(i)$ ได้ดังนี้

1. <u>การเริ่มต้น (Initialization)</u>

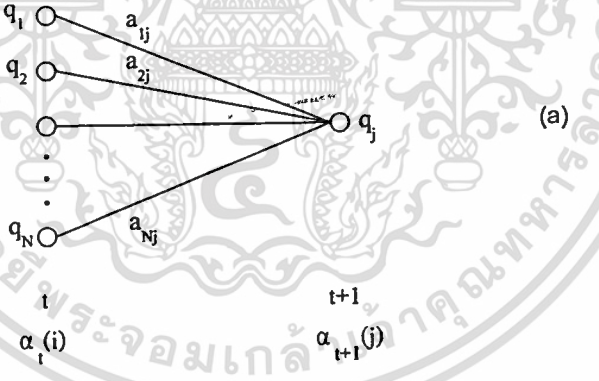
$$\alpha_t(i) = \pi_i b_i O_1 ; \quad 1 \quad i \quad N \tag{6.12}$$

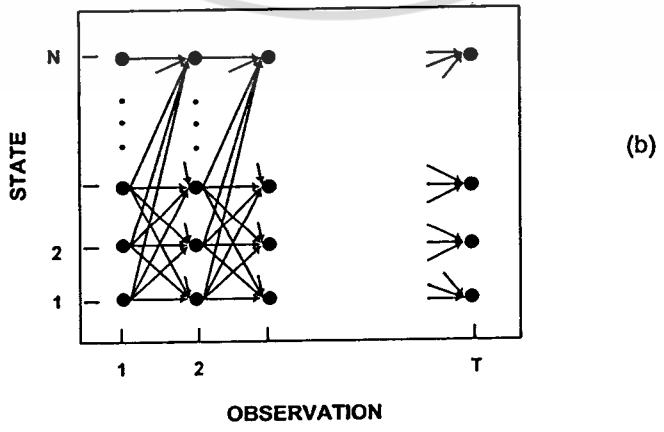
เริ่มด้วยการกำหนดความน่าจะเป็นไปข้างหน้าซึ่งเป็นความน่าจะเป็นร่วมของสเตท i และมีเหตุการณ์เริ่มต้นเป็น O_1

2. <u>การเหนี่ยวนำ (Induction)</u>

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1}) \quad ; \quad \begin{array}{ccc} 1 & t & T\text{-}1 \\ 1 & j & N \end{array} \tag{6.13}$$

หมายความว่า สเตท j ที่เวลา t+1 สามารถมาได้จากสเตทก่อนหน้านี้ซึ่งเป็นไปได้ถึง N สเตท (สเตท i ณ เวลา t โดยที่ 1 i N) ดังรูปที่ 6.2 (a)





รูปที่ 6.2 กระบวนการไปข้างหน้า

จากรูปที่ 6.2 (b) แสดงให้เห็นว่าการคำนวณค่าความน่าจะเป็นแบบไปข้างหน้า (Forward probability) มีโครงสร้างการคำนวณคล้ายๆลักษณะของโครงผลึก และเนื่องจากมีจำนวนสเตท เพียง N สเตท (แทนด้วยจำนวนโหนดในแต่ละช่วงเวลา t ใดๆในโครงผลึก) จำนวนลำดับสเตทจะถูกจัดเรียงลงในโหนดเหล่านี้ โดยในเวลา t = 1 จะทำการคำนวณค่าของ α_t (i) ในทุกๆสเตท ,1 i N และที่เวลา t = 2, 3, ..., T จะทำการคำนวณค่าของ α_t (j) ในทุกๆสเตท, $1 \le j \le N$ โดยในแต่ละค่าจะทำการคำนวณมาจาก α_{t-1} (i) จำนวน N ค่าก่อนหน้านี้

3. <u>การสิ้นสุด (Termination)</u>

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i) \quad ; \quad 1 \quad i \quad N \tag{6.14}$$

เราสามารถหา $P(O|\lambda)$ ได้จากผลรวมของ α_t (i) จากทุกๆสเตท

2. กระบวนการย้อนกลับ (Backward Procedure) ; β_t (i) = Backward variable

นิยาม

$$\beta_t(i) = P(O_{t+1} O_{t+2} \cdots O_T | i_t = q_i, \lambda) \tag{6.15}$$

คือ ความน่าจะเป็นของลำดับค่าปรากฏส่วนหลังจากเวลา t+1 ไปจนจบโดยกำหนดว่าต้องอยู่ที่สเตท i ที่เวลา t และมีแบบจำลองเป็น λ เราจะคำนวณหา β_t (i) ได้ดังนี้

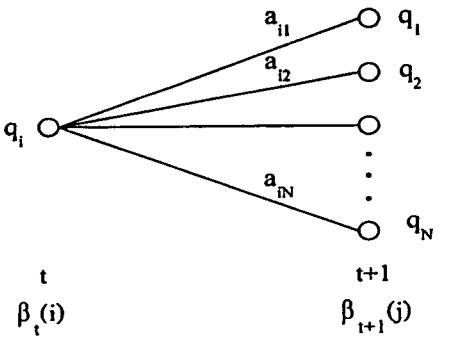
1. <u>การเริ่มต้น (Initialization)</u>

$$\beta_t(i) = 1 \quad ; \quad 1 \quad i \quad N \tag{6.16}$$

2. <u>การเหนี่ยวนำ (Induction)</u>

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{6.17}$$

เมื่อ t = T-1, T-2 , ..., 1 , 1 i N



รูปที่ 6.3 กระบวนการย้อนกลับ

จากรูปที่ 6.3 เพื่อที่จะให้ค่าปรากฏอยู่ที่สเตท i ณ เวลา t โดยคาดคะเนจากลำดับค่าปรากฏจากเวลา t+1 ซึ่งเราจะต้องพิจารณาจากสเตท j ที่เป็นไปได้ทั้งหมด โดยจะขึ้นอยู่กับค่า a_{ij} และ $b_j(O_{t+1})$

การแก้ปัญหาที่ 2 ใช้ วิเทอร์บีอัลกอริธึม [Viterbi Algorithm] เพื่อที่จะหาลำดับสเตทที่ดีที่สุด, $q = (q_1\ q_2\ q_3 \dots q_T)$ ให้กับลำดับของค่าปรากฏ $O = \{O_1 O_2 O_3 \dots O_T\}$ ที่มีอยู่

โดยนิยามให้

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P[q_1\ q_2 \dots q_{t-1}, q_t = i, O_1\ O_2 \dots O_t \mid \lambda] \qquad (6.18)$$

เมื่อ $\delta_t(i)$ คือ ความน่าจะเป็นสูงสุด (highest probability) ของเส้นทาง (path) ซึ่งจะหาได้จากค่าความน่าจะเป็นสูงสุด เมื่อเทียบกับสเตททุกสเตทในการให้ค่าปรากฏเป็นไปตามค่าปรากฏที่กำหนดให้ ที่ขณะเวลา t ใดๆ และจากการอาศัยคุณสมบัติของการเหนี่ยวนำจะได้

$$\delta_{t+1}(j) = [\max_i \delta_t(i)\, a_{ij}]\ b_j(O_{t+1}) \qquad (6.19)$$

โดยกำหนดให้ $\psi_t(j)$ เป็นอาเรย์ที่เก็บตำแหน่งของสเตท ที่ให้ค่าความน่าจะเป็นสูงที่สุดที่คำนวณได้ในแต่ละเวลา t และแต่ละลำดับ j ซึ่งจะสามารถหาลำดับสเตทที่ดีที่สุดได้โดยใช้กระบวนการต่อไปนี้

1. <u>การเริ่มต้น (Initialization)</u>

$$\delta_1(i) = \pi_i b_i(O_1) \quad ; 1 \quad i \quad N \qquad (6.20a)$$

$$\psi_1(i) = 0 \qquad (6.20b)$$

2. <u>การย้อนกลับ (Recursion)</u>

$$\delta_t(j) = [\max_{1 \le i \le N} \delta_{t-1}(i)\, a_{ij}]\; b_j(O_t) \quad ; \begin{matrix} 2 & t & T \\ 1 & j & N \end{matrix} \qquad (6.21a)$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i)\, a_{ij}] \quad ; \begin{matrix} 2 & t & T \\ 1 & j & N \end{matrix} \qquad (6.21b)$$

3. <u>การสิ้นสุด (Termination)</u>

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \qquad (6.22a)$$

$$q_T = \arg \max_{1 \le i \le N} [\delta_T(i)] \qquad (6.22b)$$

4. <u>เส้นทางเดินย้อนกลับ (Backtracking)</u>

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \quad ; t = T-1, T-2, \ldots, 1 \qquad (6.23)$$

การแก้ปัญหาที่ 3 จากที่กล่าวมาแล้วข้างต้นว่าแบบจำลองของเสียงจะแทนด้วยค่าพารามิเตอร์ $\lambda = (A,B,\pi)$ ดังนั้นเมื่อมีลำดับของค่าปรากฏจำนวนหนึ่ง เพื่อที่จะนำมาสร้างแบบจำลองอ้างอิง จะต้องทำการคำนวณหาค่าพารามิเตอร์ A, B, π ของแบบจำลองซึ่งจะอยู่ในรูปของค่าความน่าจะเป็น โดยวิธีที่เลือกใช้ก็คือ วิธีของ บาม-เวลส์ [Baum-Welch method] หรือเรียกอีกชื่อหนึ่งว่า EM (Expectation-Maximization method) โดยมี

นิยาม 1. คือ

$$\gamma_t(i) = P(q_t = i | O, \lambda) \qquad (6.24)$$

เมื่อ $\gamma_t(i)$ คือ ค่าความน่าจะเป็นที่จะอยู่ที่สเตท i ที่ขณะเวลา t โดยให้ลำดับของค่าปรากฏด้วยโมเดล λ โดยที่กำหนดลำดับของค่าปรากฏให้ สามารถแสดงค่า $\gamma_t(i)$ ได้ดังนี้

$$\gamma_t(i) = P(q_t = i|O,\lambda)$$

$$= \frac{P(O, q_t = i|\lambda)}{P(O|\lambda)}$$

$$= \frac{P(O, q_t = i|\lambda)}{\sum_{i=1}^{N} P(O, q_t = i|\lambda)} \quad (6.25)$$

เนื่องจาก $P(O, q_t = i|\lambda)$ มีค่าเท่ากับ $\alpha_t(i)\beta_t(i)$ ดังนั้นสามารถเขียน $\gamma_t(i)$ ได้เป็น

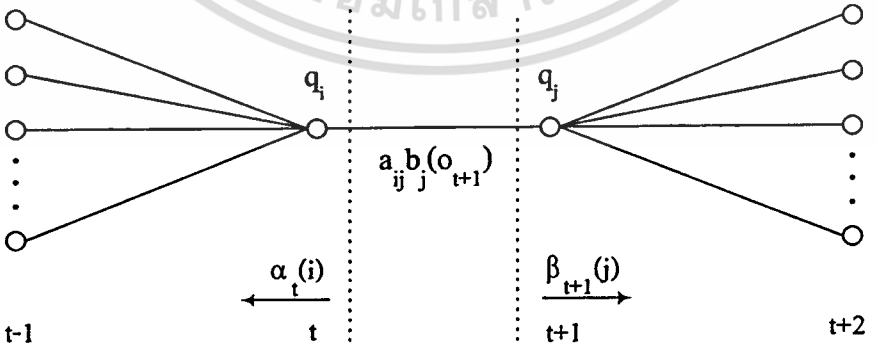
$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \quad (6.26)$$

โดย $\alpha_t(i)$ เริ่มจาก $O_1, O_2, \ldots, O_t$ จนถึงสเตท i ที่เวลา t

โดย $\beta_t(i)$ เริ่มจาก $O_{t+1}, O_{t+2}, \ldots, O_T$ จนถึงสเตท $q_t = i$ ที่เวลา t

นิยาม 2. $\varepsilon_t(i,j) = P(q_t = i, q_{t+1} = j|O, \lambda)$ (6.27)

เมื่อ $\varepsilon_t(i,j)$ คือความน่าจะเป็นที่จะอยู่ที่สเตท i ที่เวลา t และสเตท j ที่เวลา t+1 เมื่อกำหนดแบบจำลองและลำดับค่าปรากฏให้



รูปที่ 6.4 ลำดับการคำนวณการเกิดค่าปรากฏร่วมซึ่งจะอยู่ที่สเตท i ที่เวลา t และอยู่ที่ สเตท j ที่เวลา t+1

จากรูปที่ 6.4 แสดงลำดับการคำนวณการเกิดค่าปรากฏร่วม ซึ่งระบบจะอยู่ในสเตท i ที่เวลา t และอยู่ที่ สเตท j ที่เวลา t+1 โดย $\alpha_t(i)$ เริ่มจากเวลา t = 1 ที่ค่าปรากฏแรก จนถึงสเตท q_i ที่เวลา t และ $a_{ij}b_jO_{t+1}$ เป็นการเปลี่ยนสเตทที่เวลา t ไปเป็น q_j ที่เวลา t+1 และให้ค่าปรากฏเป็น O_{t+1}

ซึ่งจากนิยามของตัวแปรไปข้างหน้า $\alpha_t(i)$ และตัวแปรย้อนกลับ $\beta_t(i)$ สามารถนำมาสัมพันธ์กับ $\varepsilon_t(i,j)$ ได้เป็น

$$\varepsilon_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, O|\lambda)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j)} \qquad (6.28)$$

จากที่ได้นิยาม $\gamma_t(i)$ แล้ว นำมาสัมพันธ์กับ $\varepsilon_t(i,j)$ ได้เป็น

$$\gamma_t(i) = \sum_{j=1}^{N} \varepsilon_t(i,j) \qquad (6.29)$$

เมื่อ $\sum_{t=1}^{T-1} \gamma_t(i)$ = จำนวนของการย้ายสเตทจากสเตท i ในลำดับค่าปรากฏ O (6.30a)

$\sum_{t=1}^{T-1} \varepsilon_t(i,j)$ = จำนวนของการย้ายสเตทจากสเตท i ไป j ในลำดับค่าปรากฏ O (6.30b)

ดังนั้น สามารถคำนวณหาค่าของพารามิเตอร์ได้ดังนี้

π_i = จำนวนครั้งในการอยู่ที่สเตท i ที่เวลา t = 1

$$\pi_i = \gamma_1(i) \quad ; 1 \quad i \quad N \qquad (6.31a)$$

$$a_{ij} = \frac{\text{จำนวนครั้งที่คาดไว้ของการย้ายสเตทจาก i ไป j}}{\text{จำนวนครั้งที่คาดว่าจะย้ายจากสเตท i}}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad (6.31b)$$

$$b_j(k) = \frac{\text{จำนวนครั้งที่คาดว่าจะอยู่ในสเตท j และเกิดค่าปรากฏเป็น } V_K}{\text{จำนวนครั้งที่คาดว่าจะอยู่ที่สเตท j}}$$

$$b_j(k) = \frac{\sum_{t=1, O_t = V_K}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad (6.31c)$$

จากกระบวนการข้างต้นถ้าให้ $\lambda = (A,B,\pi)$ เป็นแบบจำลองปัจจุบัน และใช้ λ นี้คำนวณในด้านขวาของสมการที่(6.31a-c)และให้แบบจำลองที่ได้จากการคำนวณซ้ำเป็น $\lambda = (A,B,\pi)$ เป็นแบบจำลองที่ได้จากด้านซ้ายของสมการที่(6.31a-c) ซึ่งจะได้จุดวิกฤตของฟังก์ชั่นความน่าจะเป็นในกรณีที่ $\lambda = \lambda$ หรือถ้า λ มีความน่าจะเป็นมากกว่าแบบจำลอง λ [$P(O|\lambda) > P(O|\lambda)$] นั่นคือจะได้แบบจำลอง λ ใหม่ ที่น่าจะทำให้เกิดลำดับของค่าปรากฏ O ที่ดีกว่า

6.5 การปรับปรุงค่าพารามิเตอร์ของ HMM

6.5.1 การสเกลลิง (Scaling)

พิจารณาคำจำกัดความของ $\alpha_t(i)$ ในสมการที่ 6.11 จะเห็นว่า $\alpha_t(i)$ ประกอบไปด้วยผลรวมเทอมขนาดใหญ่ที่อยู่ในรูป

$$\prod_{s=1}^{t-1} a_{q_s q_{s+1}} \prod_{s=1}^{t} b_{q_s}(O_s)$$

เนื่องจากค่า a และ b เป็นค่าความน่าจะเป็น ซึ่งโดยทั่วไปแล้วมีค่าน้อยกว่า 1 ด้วยเหตุนี้เมื่อ t มากขึ้นค่าแต่ละเทอมของ $\alpha_t(i)$ จะเข้าสู่ศูนย์ ทำให้ช่วงไดนามิก (Dynamic Range) ของการคำนวณ $\alpha_t(i)$ มีค่าสูงเกินขอบเขตการทำงานของเครื่องคำนวณทำให้ค่าที่ได้ไม่ถูกต้อง ซึ่งเราสามารถแก้ไขปัญหานี้ได้โดยใช้กระบวนการสเกลลิง (Scaling Procedure)

การสเกลลิ่งทำได้โดยการคูณ $\alpha_t(i)$ ด้วยสัมประสิทธิ์การสเกลลิง ซึ่งไม่ขึ้นกับ i (นั่นคือขึ้นอยู่กับค่าของเวลา t เท่านั้น) เพื่อให้ $\alpha_t(i)$ ที่ผ่านการสเกลลิงแล้วมีค่าอยู่ภายในช่วง Dynamic Range ของเครื่องคำนวณในทุกๆค่าเวลาภายใต้ 1 t T และในทำนองเดียวกันจะต้องทำการคำนวณค่าสัมประสิทธิ์การสเกลลิงของค่า $\beta_t(i)$ ด้วย ซึ่งในขั้นตอนสุดท้ายของการคำนวณค่าสัมประสิทธิ์ของการสเกลลิงจะตัดกันหมดไป

เพื่อให้เข้าใจการทำงานของกระบวนการสเกลลิงดีขึ้น เราจะพิจารณาสมการของการย้ายสเตท(a_{ij}) ที่อยู่ในเทอมของตัวแปรไปข้างหน้า และตัวแปรย้อนกลับ

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j)} \qquad (6.32)$$

พิจารณาสัญลักษณ์ในการคำนวณ $\alpha_t(i)$ เมื่อกำหนดให้

$\alpha_t(i)$ แทน α ที่ยังไม่ผ่านการสเกล

$\hat{\alpha}_t(i)$ แทน α ที่สเกลแล้ว

$\hat{\bar{\alpha}}_t(i)$ แทน α แทนเวอร์ชั่นของ α ก่อนการสเกล

เมื่อเวลาเริ่มต้น t = 1

คำนวณ $\alpha_t(i)$ ตามสมการที่ 6.12 และกำหนดให้ $\hat{\bar{\alpha}}_1(i) = \alpha_1(i)$

เมื่อ $c_1 = \frac{1}{\sum_{i=1}^{N} \alpha_1(i)}$

และ $\hat{\alpha}_1(i) = c_1 \alpha_1(i)$

เมื่อเวลา 2 t T

เริ่มแรกทำการคำนวนหา $\hat{\bar{\alpha}}_t(i)$ ตามสมการการเหนี่ยวนำ สมการที่ 6.13 โดยใช้เทอมของค่าที่ผ่านการสเกลแล้ว $\hat{\alpha}_t(i)$ จะได้ดังนี้

$$\hat{\bar{\alpha}}_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji}\, b_i(O_t) \qquad (6.33a)$$

เมื่อกำหนดค่าสัมประสิทธิ์การสเกลลิง ; c_t เป็น

$$c_t = \frac{1}{\sum_{i=1}^{N} \hat{\alpha}_t(i)} \tag{6.33b}$$

เมื่อให้

$$\hat{\alpha}_t(i) = c_t \hat{\alpha}_t(i) \tag{6.33c}$$

จากสมการที่ 6.33 a-c สามารถเขียนสมการได้เป็น

$$\hat{\alpha}_t(i) = \frac{\sum_{j=1}^{N} \hat{\alpha}_{t-1}(j)\, a_{ji}\, b_i(O_t)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{\alpha}_{t-1}(j)\, a_{ji}\, b_i(O_t)} \tag{6.34}$$

และโดยการเหนี่ยวนำสามารถเขียน $\hat{\alpha}_{t-1}(j)$ ได้เป็น

$$\hat{\alpha}_{t-1}(j) = \prod_{T=1}^{t-1} c_T \quad \alpha_{t-1}(j) \tag{6.35a}$$

ดังนั้นสามารถเขียน $\hat{\alpha}_t(i)$ ได้เป็น

$$\hat{\alpha}_t(i) = \frac{\sum_{j=1}^{N} \alpha_{t-1}(j) \prod_{T=1}^{t-1} c_T\, a_{ji}\, b_i(O_t)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{t-1}(j) \prod_{T=1}^{t-1} c_T\, a_{ji}\, b_i(O_t)} = \frac{\alpha_t(i)}{\sum_{i=1}^{N} \alpha_t(i)} \tag{6.35b}$$

นั่นคือการสเกลลิงจะทำได้โดยนำ $\alpha_t(i)$ แต่ละค่า มาหารด้วยผลรวมของ $\alpha_t(i)$ ทุกสเตท จากนั้นทำการคำนวณลักษณะเดียวกันนี้กับเทอมของตัวแปรย้อนกลับ $\beta_t(i)$ โดยใช้สเกลเฟคเตอร์เดียวกัน ในรูปของ

$$\hat{\beta}_t(i) = c_t \beta_t(i) \tag{6.36}$$

พิจารณาสมการ 6.32 ในเทอมของตัวแปรที่ผ่านการสเกล จะได้เป็น

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i)\, a_{ij}\, b_j(O_{t+1})\hat{\beta}_{t+1}(j)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \hat{\alpha}_t(i)\, a_{ij}\, b_j(O_{t+1})\hat{\beta}_{t+1}(j)} \tag{6.37}$$

โดยแต่ละ $\hat{\alpha}_t(i)$, $\hat{\beta}_{t+1}(j)$ สามารถเขียนให้อยู่ในรูปของ

$$\hat{\alpha}_t(i) = \sum_{s=1}^{T} c_s\, \alpha_t(i) = C_t \alpha_t(i) \tag{6.38}$$

$$\hat{\beta}_{t+1}(j) = \sum_{s=t+1}^{T} c_s\, \beta_{t+1}(j) = D_{t+1}\beta_{t+1}(j) \tag{6.39}$$

ดังนั้นสมการ 6.37 สามารถเขียนได้เป็น

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} C_t \alpha_t(i)\, a_{ij}\, b_j(O_{t+1}) D_{t+1}\beta_{t+1}(j)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} C_t \alpha_t(i)\, a_{ij}\, b_j(O_{t+1}) D_{t+1}\beta_{t+1}(j)} \tag{6.40}$$

โดยเทอม C_tD_{t+1} สามารถเขียนให้อยู่ในรูปของ

$$C_tD_{t+1} = \sum_{s=1}^{t} c_s \sum_{s=t+1}^{T} c_s = \sum_{s=1}^{T} c_s = C_T \tag{6.41}$$

จะเห็นว่าเทอม C_tD_{t+1} เป็นค่าที่ไม่ขึ้นกับเวลา ดังนั้นสามารถตัดออกจากทั้งเศษและส่วนของสมการ 6.40 ได้ ซึ่งจะทำให้ได้สูตรของการคำนวณซ้ำ กระบวนการสเกลลิงดังกล่าวนี้จะถูกนำไปใช้กับสัมประสิทธิ์ π และ β การสเกลลิงนี้จะทำให้การคำนวณค่า $P(O|\lambda)$ เปลี่ยนไป เราจะไม่สามารถหาได้จากการรวมเทอมของ $\hat{\alpha}_T(i)$ เนื่องจากเป็นค่าที่ถูกสเกลแล้ว แต่เราสามารถคำนวณได้จากคุณสมบัติ

$$\sum_{t=1}^{T} c_t \sum_{i=1}^{N} \alpha_T(i) = C_T \sum_{i=1}^{N} \alpha_T(i) = 1 \qquad (6.42)$$

ดังนั้นจะได้

$$\sum_{t=1}^{T} c_t \; P(O|\lambda) = 1 \qquad (6.43)$$

หรือ

$$P(O|\lambda) = \frac{1}{\prod_{t=1}^{T} c_t} \qquad (6.44)$$

หรือ

$$\log [\, P(O|\lambda) \,] = \sum_{t=1}^{T} \log c_t \qquad (6.45)$$

นั่นคือ การคำนวณค่า P จะอยู่ในรูป log ของ Pเพื่อไม่ให้เกินช่วงไดนามิก (Dynamic Range) ของเครื่องคำนวณ

6.5.2 ลำดับของค่าปรากฏหลายลำดับ (Multiple Observation Sequences)

ในการสร้างแบบจำลองด้วย Left-Right Model จำเป็นจะต้องใช้จำนวนลำดับของเหตุการณ์หลายๆลำดับเข้ามาเทรนเพื่อให้การประมาณค่าพารามิเตอร์ของแบบจำลองที่ได้มีความน่าเชื่อถือที่สุด ถ้ากำหนดให้ k แทน เซ็ตของลำดับค่าปรากฏ ดังนี้

$$O = [\, O^{(1)}, O^{(2)}, \ldots, O^{(k)} \,] \qquad (6.46)$$

เมื่อ $O^{(k)} = (O_1^{(k)} \; O_2^{(k)} \ldots O_{T_k}^{(k)})$ คือ ลำดับค่าปรากฏอันดับที่ k โดยสมมติให้แต่ละอันดับของค่าปรากฏเป็นอิสระต่อกัน โดยมีจุดประสงค์ เพื่อที่จะปรับค่าพารามิเตอร์ของแบบจำลอง λ ให้มีค่ามากที่สุด

$$P(O|\lambda) = \prod_{k=1}^{K} P(O^{(k)}|\lambda) \qquad (6.47)$$

$$= \prod_{k=1}^{K} P_k \qquad (6.48)$$

ดังนั้นจะได้สมการของการคำนวณซ้ำที่ใช้ในการปรับค่า $\bar{a}_{ij}$ และ $\bar{b}_j(l)$ เป็น

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i)\, a_{ij}\, b_j(O_{t+1}^{(k)})\beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i)\beta_t^k(i)} \tag{6.49}$$

และ

$$\bar{b}_j(l) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1, O_t = v_l}^{T_k - 1} \alpha_t^k(i)\, \beta_t^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i)\beta_t^k(i)} \tag{6.50}$$

ส่วนค่า π_i ไม่ต้องมีการคำนวณซ้ำเนื่องจาก $\pi_1 = 1, \pi_i = 0, i \quad 1$

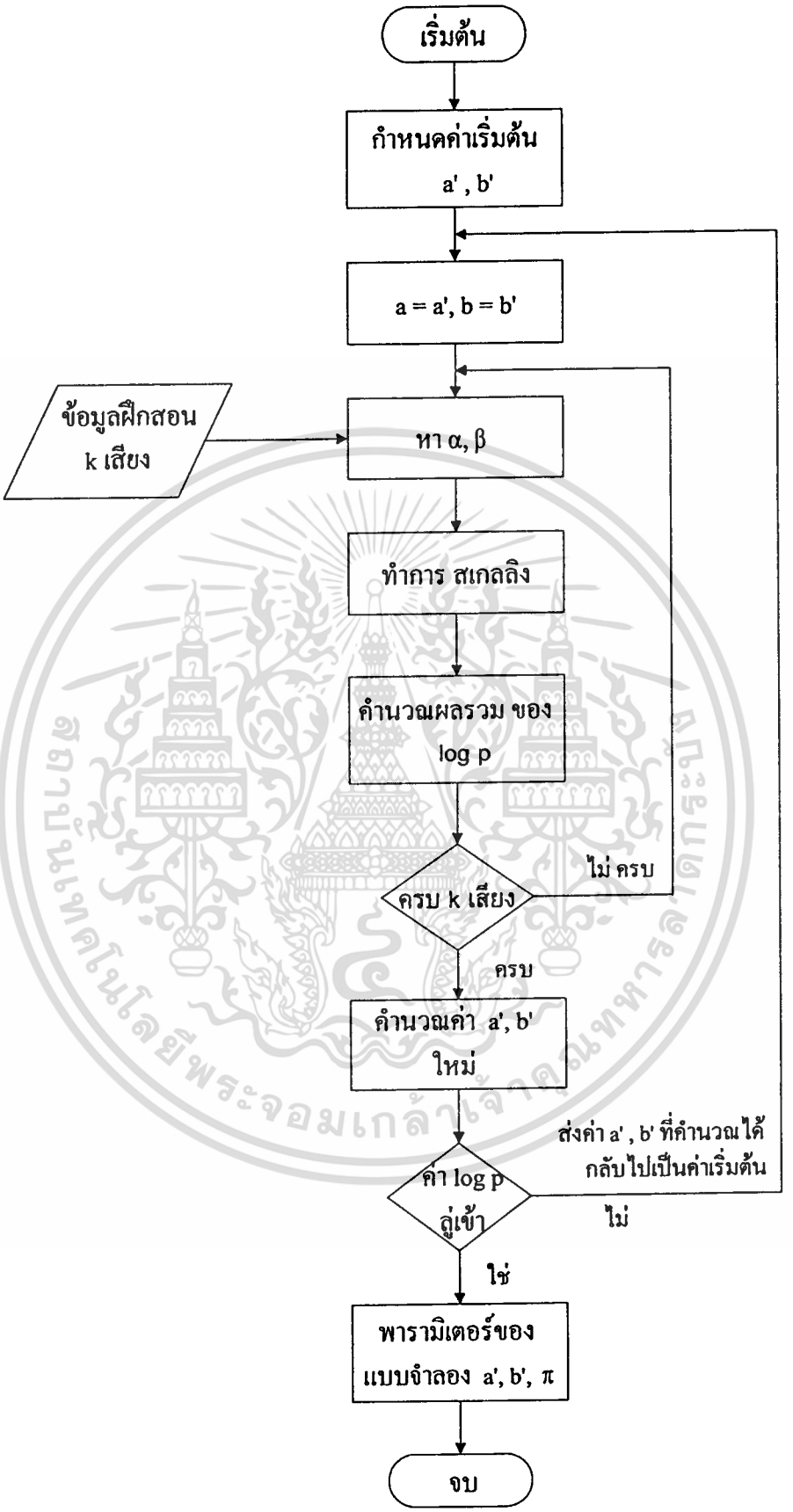
จากสมการของการสเกลลิงสมการที่ 6.49-6.50 เราสามารถเขียนสมการที่อยู่ในเทอมของตัวแปรที่สเกลแล้วได้เป็น

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \hat{\alpha}_t^k(i)\, a_{ij}\, b_j(O_{t+1}^{(k)})\hat{\beta}_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \hat{\alpha}_t^k(i)\hat{\beta}_t^k(i)} \tag{6.51}$$

$$\bar{b}_j(l) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1, O_t = v_l}^{T_k - 1} \hat{\alpha}_t^k(i)\hat{\beta}_t^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \hat{\alpha}_t^k(i)\hat{\beta}_t^k(i)} \tag{6.52}$$

6.6 การสร้างแบบจำลองอ้างอิง

จากหัวข้อ 6.4 ได้กล่าวถึงการแก้ปัญหาทั้ง 3 ข้อของ HMM ซึ่งจะถูกนำมาใช้ในการคำนวณหาพารามิเตอร์ของแบบจำลองอ้างอิงในการรู้จำ โดยขั้นตอนในการคำนวณหาพารามิเตอร์ของแบบจำลองอ้างอิง สามารถแสดงได้ดัง รูปที่ 6.5

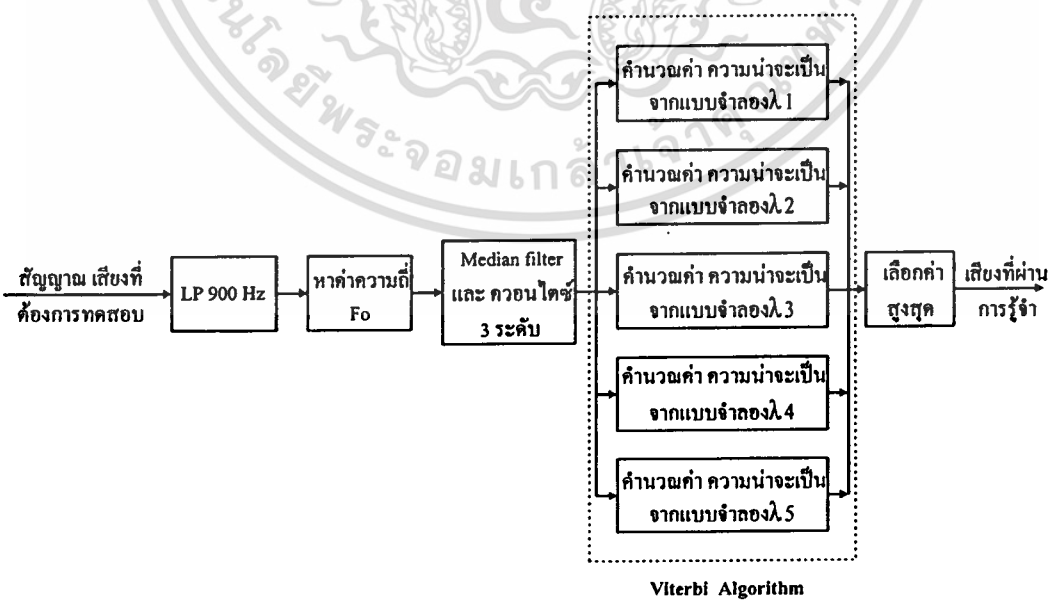


รูปที่ 6.5 โฟลว์ชาร์ต การคำนวณหาค่าพารามิเตอร์ของแบบจำลองอ้างอิง

เมื่อต้องการสร้างแบบจำลองอ้างอิงเสียงวรรณยุกต์ทั้ง 5 ระดับเสียง สิ่งที่จะต้องมีก็คือ กลุ่มเสียงต้นแบบ หรือลำดับของค่าปรากฏทั้ง 5 กลุ่ม เพื่อใช้เป็นข้อมูลฝึกสอน (Training data) จากรูปที่ 6.5 แสดงขั้นตอนของการคำนวณการสร้างแบบจำลองอ้างอิง โดยในขั้นแรกจะต้องทำการกำหนดค่า A, B เริ่มต้น จากนั้น ทำการคำนวณหาค่า α, β โดยใช้การแก้ปัญหาที่ 1 ของ HMM แล้วทำการสเกลลิง เพื่อไม่ให้ค่าที่ได้จากการคำนวณมีค่าเกินช่วงไดนามิกของเครื่องคำนวณ (Dynamic Range) จากนั้นนำค่าสัมประสิทธิ์ของการสเกลลิงมาคำนวณหาค่า ความน่าจะเป็น $P(O|\lambda)$ ซึ่งจะอยู่ในรูปของค่า log P และเนื่องจากในการสร้างแบบจำลองอ้างอิง จำเป็นจะต้องใช้ข้อมูลฝึกสอนจำนวนมาก เพื่อให้แบบจำลองอ้างอิงที่สร้างขึ้น คลอบคลุมความแปรปรวนของลักษณะเสียงให้ได้มากที่สุด ดังนั้นจึงจะต้องมีการคำนวณซ้ำเกิดขึ้น ตามจำนวนของเสียงที่นำมาฝึกสอน จากนั้นทำการหาค่าผลรวมของค่าความน่าจะเป็น (ผลรวมของ log P จากจำนวนเสียงทั้งหมด) เพื่อมาใช้ในการคำนวณหาค่าพารามิเตอร์ของแบบจำลอง $\lambda=(A',B',\pi)$ โดยใช้การแก้ปัญหาที่ 3 ของ HMM จากนั้นทำการคำนวณซ้ำจนกว่าค่าผลรวมของ log P ที่ได้ในแต่ละรอบมีค่าลู่เข้า หรือไม่เปลี่ยนแปลง พารามิเตอร์ของแบบจำลอง $\lambda'=(A',B',\pi)$ ค่าสุดท้าย จะเป็นแบบจำลองที่น่าจะทำให้เกิดลำดับของค่าปรากฏ O ที่ดีกว่า โดยรายละเอียดของขั้นตอนต่างๆได้กล่าวมาแล้วในหัวข้อก่อนหน้านี้

6.7 แบบจำลอง ฮิดเดนมาร์คอฟ ในการรู้จำเสียงวรรณยุกต์ภาษาไทย

การรู้จำระดับเสียงวรรณยุกต์ทั้ง 5 ระดับ ด้วย HMM แสดงได้ดังรูปที่ 6.6



รูปที่ 6.6 บล็อกไดอะแกรม ของการรู้จำระดับเสียงวรรณยุกต์ด้วยแบบจำลองมาร์คอฟ

รูปที่ 6.6 แสดงขั้นตอนในการทดสอบการรู้จำระดับเสียงวรรณยุกต์ทั้ง 5 ระดับ โดยเสียงที่ต้องการทดสอบจะถูกนำมาผ่านขั้นตอนในการหาค่าความถี่มูลฐาน แล้วควอนไตซ์เป็น 3 ระดับ ดังกล่าวมาแล้วในบทที่ 4-5 ซึ่งข้อมูลที่ได้จากการควอนไตซ์ จะถูกนำมาเทียบกับแบบจำลองอ้างอิงเสียงวรรณยุกต์ทั้ง 5 แบบ (λ_1-λ_5) โดยแบบจำลองอ้างอิงใดที่ให้ค่าความน่าจะเป็น(ในการเกิดเหตุการณ์)สูงสุด จะถือว่าคำศัพท์ที่นำมาทดสอบ จะมีระดับเสียงเดียวกันกับแบบจำลองนั้นนั่นเอง โดยขั้นตอนการคำนวณหาค่าความน่าจะเป็นจะใช้การแก้ปัญหาที่ 2 หรือวิทเทอร์บีอัลกอริธึม ดังได้กล่าวถึงรายละเอียดมาแล้วข้างต้น

บทที่ 7

การทดลองและผลการทดลอง

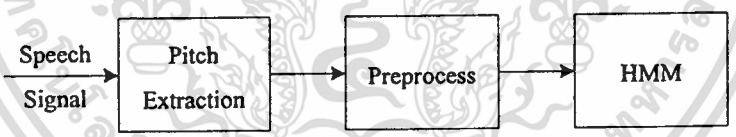
ในบทนี้จะกล่าวถึงวิธีการทดลองและผลการทดลองที่ได้ในขั้นตอนต่างๆของการคำนวณหาค่าพิทช์และสร้างแบบจำลองการรู้จำระดับเสียงวรรณยุกต์ของคำพยางค์เดียวในภาษาไทยในสภาวะที่มีเสียงรบกวน โดยการทดลองจะแบ่งออกเป็น 2 ส่วนใหญ่ๆ ดังนี้

1. ขั้นตอนในการวิเคราะห์และพัฒนาอัลกอริธึมในการสร้างแบบจำลองการรู้จำเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีเสียงรบกวน
2. ขั้นตอนในการทดสอบแบบจำลองอ้างอิงที่สร้างขึ้น

7.1 ขั้นตอนในการวิเคราะห์และพัฒนาอัลกอริธึมในการสร้างแบบจำลองการรู้จำเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีเสียงรบกวน

แบ่งออกเป็น 3 ขั้นตอน ดังแสดงในรูปที่ 7.1

1. การหาค่าพิทช์
2. การเตรียมข้อมูล เพื่อใช้เป็นข้อมูลฝึกสอนในการสร้างแบบจำลองอ้างอิง
3. การสร้างแบบจำลองอ้างอิงด้วย ฮิดเดนมาร์คอฟโมเดล

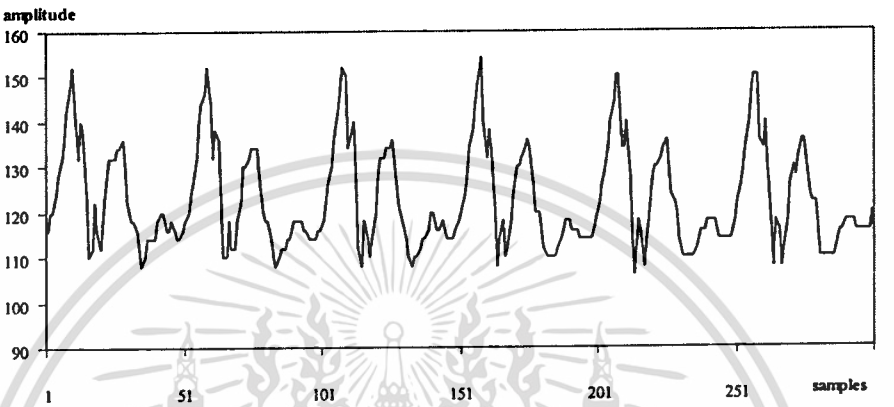


รูปที่ 7.1 ขั้นตอนในการวิเคราะห์

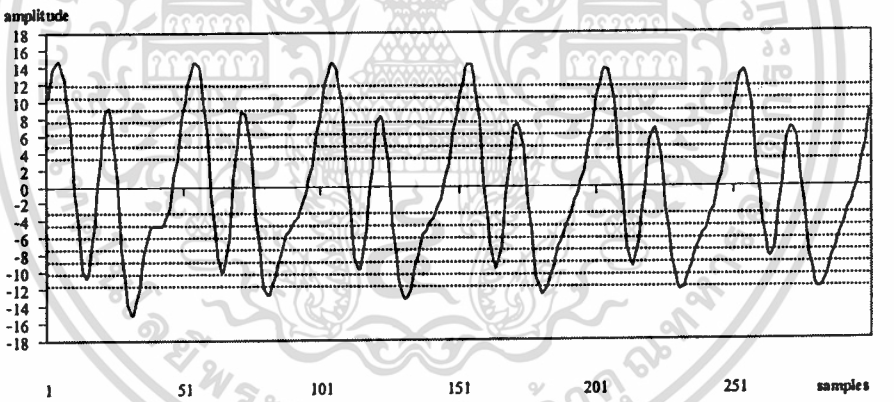
สัญญาณเสียงพูดที่ใช้ในวิทยานิพนธ์นี้ เป็นคำพยางค์เดียวที่ได้จากการเก็บตัวอย่างข้อมูลเสียง ซึ่งข้อมูลจะถูกเก็บอยู่ในรูปของไฟล์ ".wav" โดยข้อมูล 1 ตัวอย่างของเสียงจะถูกแทนด้วยข้อมูล 8 บิท และใช้ความถี่ในการชักตัวอย่างเท่ากับ 11.025 kHz และไฟล์ข้อมูล ".wav" นี้จะถูกนำไปใช้ข้อมูลอินพุทสำหรับการคำนวณของโปรแกรมที่เขียนขึ้น โดยในวิทยานิพนธ์นี้เลือกใช้โปรแกรม Microsoft Visual C++ 6.0 ในการทดลองและพัฒนาวิธีการที่ใช้ในการรู้จำเสียงพูดในสภาวะที่มีเสียงรบกวน

7.1.1 การหาค่าพิทช์

การหาค่าพิทช์ทำได้โดยใช้วิธีออโตคอร์รีเลชั่นฟังก์ชั่นโดยใช้เทคนิคการคลิปยอดของสัญญาณ (Autocorrelation Method using Center Clipping : AUTOC) ซึ่งมีขั้นตอนในการวิเคราะห์ดังกล่าวมาแล้วในบทที่4 หัวข้อที่ 4.3 โดยรูปที่ 7.2 – 7.3 แสดงตัวอย่างสัญญาณที่ผ่านขั้นตอนต่างๆตามลำดับ



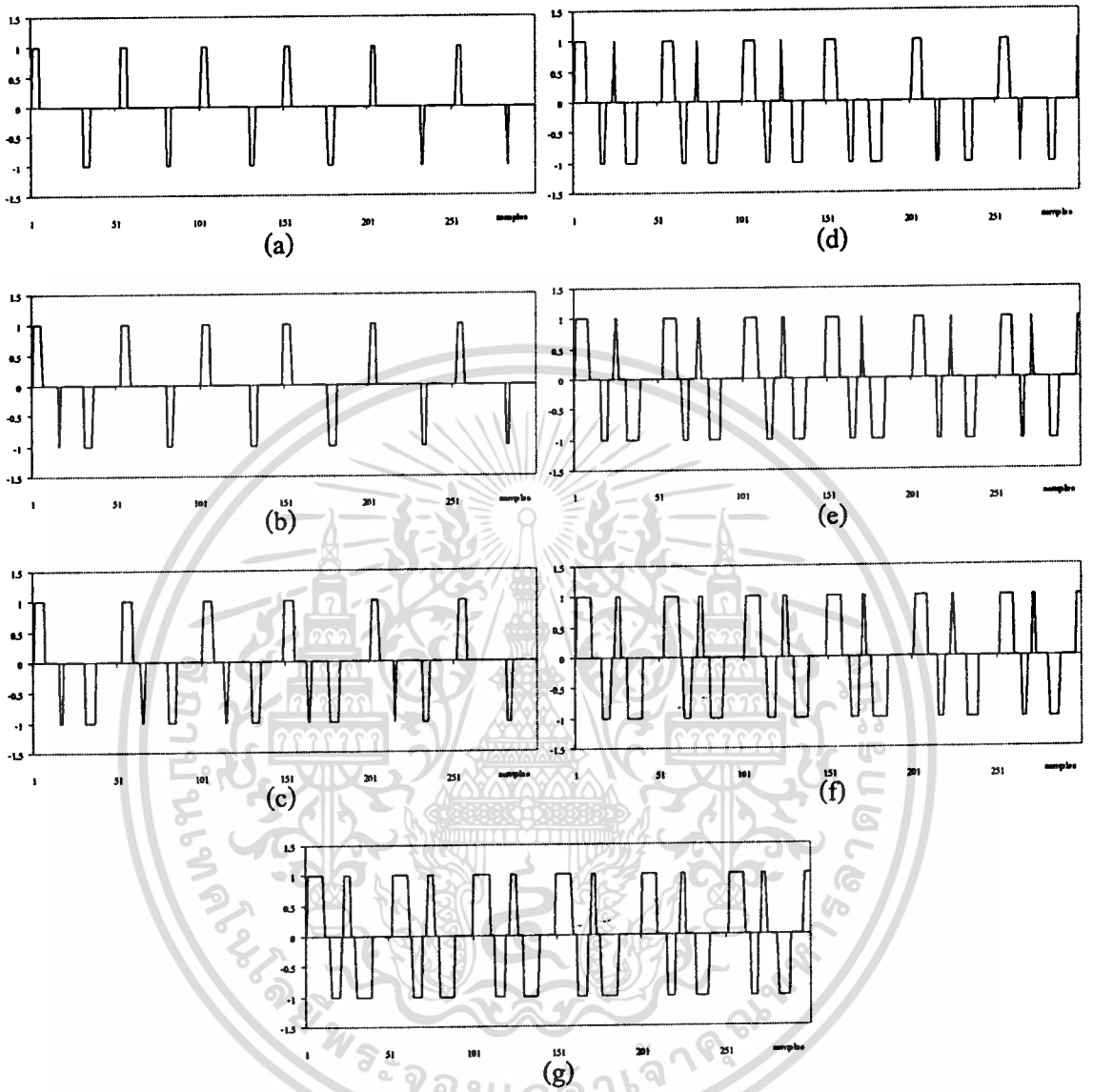
รูปที่ 7.2 แสดงสัญญาณเสียงพูด



รูปที่ 7.3 แสดงสัญญาณเสียงพูดที่ผ่านตัวกรองความถี่ต่ำที่ 900 Hz และมีการกำหนด Clipping Level

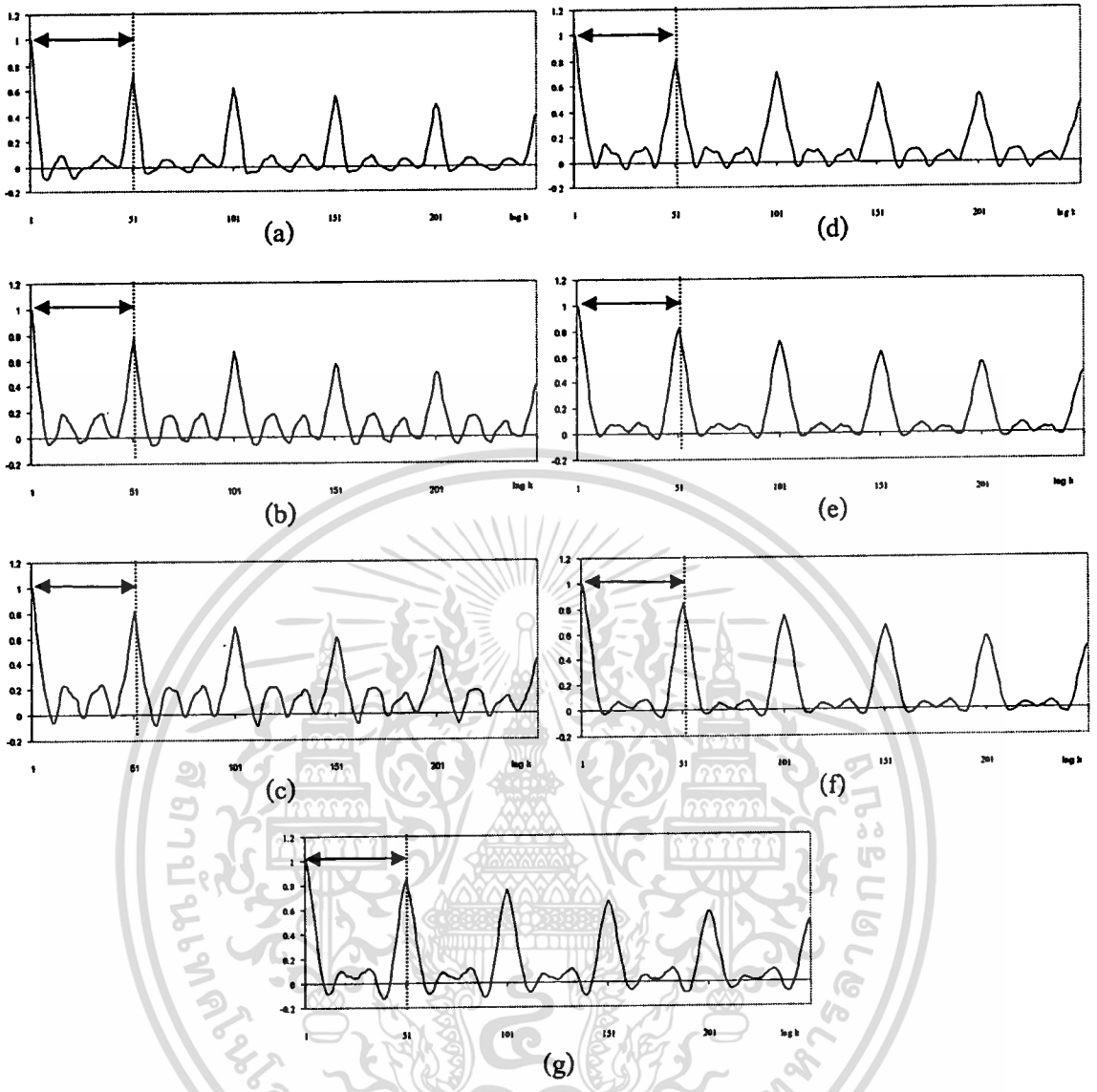
จากรูปที่ 7.2 เป็นสัญญาณเสียงพูดที่เป็นอินพุทที่ใช้ในการวิเคราะห์จะเห็นว่าสัญญาณเสียงพูดจะประกอบด้วยองค์ประกอบของความถี่จำนวนมาก อันเป็นผลมาจากการตอบสนองทางความถี่ภายในช่องทางเดินเสียง ความถี่เหล่านี้อาจมีผลทำให้การกำหนดตำแหน่งพิทช์ในการคำนวณออโตคอร์รีเลชั่นคลาดเคลื่อนไปจากตำแหน่งจริงได้ ดังนั้นเพื่อเป็นการกำจัดผลของความถี่เหล่านี้ออกไป จึงนำสัญญาณเสียงมาผ่านการกรองความถี่ต่ำผ่าน 900 Hz ซึ่งจะได้สัญญาณที่มีความราบเรียบมากขึ้นดังรูปที่ 7.3 โดยเส้นประในรูปแสดงระดับในการคลิปยอดของสัญญาณ ซึ่ง

ในวิทยานิพนธ์นี้ได้ทำการคลิปยอดของสัญญาณในระดับ 80%, 70%, 60%, 50%, 40%, 30% และ 20% ตามลำดับ



รูปที่ 7.4 แสดงสัญญาณที่ถูกคลิปยอดของสัญญาณในระดับต่างๆ (a) 80% (b) 70% (c) 60% (d) 50% (e) 40% (f) 30% (g) 20%

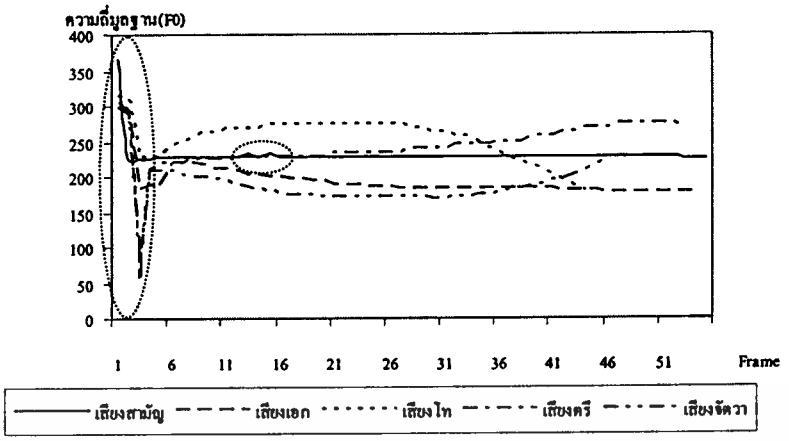
รูปที่ 7.4 แสดงสัญญาณที่ผ่านการคลิปยอดของสัญญาณในระดับต่างๆ ที่กำหนดขึ้น ซึ่งจะสังเกตได้ว่าจุดยอดที่มีจำนวนมากถูกกำจัดไปทำให้ในการคำนวณหาออโตคอร์รีเลชั่นนั้นลดความซับซ้อนลง จากนั้นนำสัญญาณที่ผ่านการคลิปยอดของสัญญาณมาทำการคำนวณออโตคอร์รีเลชั่นเพื่อหาพิทช์ จะได้สัญญาณที่มีลักษณะดังรูปที่ 7.5 โดยระยะห่างระหว่าง R(0) กับจุดยอดที่สูงที่สุดถัดไปก็คือคาบพิทช์ ซึ่งจากรูปได้ตำแหน่งที่ 52 และสามารถหาค่าความถี่มูลฐานได้เท่ากับ 252 Hz (11025/52=252 Hz)



รูปที่ 7.5 แสดงสัญญาณที่ผ่านการคำนวณ Normalized ออโตคอร์รีเลชั่น ที่ถูกคลิปยอดของสัญญาณในระดับต่างๆ (a) 80% (b) 70% (c) 60% (d) 50% (e) 40% (f) 30% (g) 20%

จากรูปที่ 7.5 จะสังเกตว่าสัญญาณที่ถูกคลิปยอดของสัญญาณในระดับที่ 40% และ 30% เมื่อมาคำนวณออโตคอร์รีเลชั่นแล้วจะเห็นว่าสเปกตรัมของสัญญาณมีระดับต่ำเมื่อเทียบกับสัญญาณที่ถูกคลิปยอดในระดับอื่น

เมื่อเสร็จสิ้นกระบวนการออโตคอร์รีเลชั่นจะได้ว่าสัญญาณข้อมูลเสียง 1 เฟรม จะถูกแทนด้วยค่าความถี่มูลฐาน 1 ค่า นั่นคือข้อมูลเสียง 1 เสียงจะถูกแทนด้วยลำดับของความถี่มูลฐานที่มีขนาดเป็น 1xN เมื่อ N คือจำนวนเฟรมทั้งหมดของเสียง แสดงได้ดังรูปที่ 7.6



รูปที่ 7.6 แสดงลักษณะการเปลี่ยนแปลงความถี่มูลฐานในเสียงวรรณยุกต์ภาษาไทยทั้ง 5 เสียง

จากรูปที่ 7.6 เป็นการออกเสียงคำว่า เอ เอ่ เอ้ เอ๊ เอ๋ ในผู้ออกเสียงเพศหญิง ซึ่งจะสังเกตุได้ว่าในระดับเสียงวรรณยุกต์แต่ละระดับจะมีทิศทางการเปลี่ยนแปลงของความถี่มูลฐานที่แตกต่างกัน โดยกราฟที่อยู่ในวงกลมเส้นประจะถูกปรับแต่งให้เรียบขึ้น ซึ่งจะกล่าวถึงในขั้นตอนต่อไป

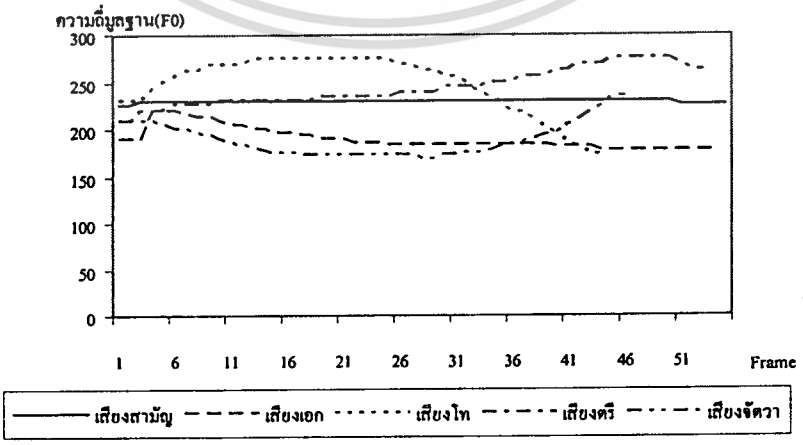
7.1.2 การเตรียมข้อมูล (Pre-process)

เป็นส่วนของการจัดการข้อมูลซึ่งมีอยู่ 2 ขั้นตอน ดังกล่าวไว้ในบทที่ 5 คือ

1. การปรับปรุงความต่อเนื่องของข้อมูลด้วยวิธีการกรองค่ากลาง (Median Filtering)
2. การควอนไตซ์ทิศทางการเปลี่ยนแปลงของค่าความถี่มูลฐาน

โดยแต่ละขั้นตอนสามารถอธิบายได้ดังนี้ คือ

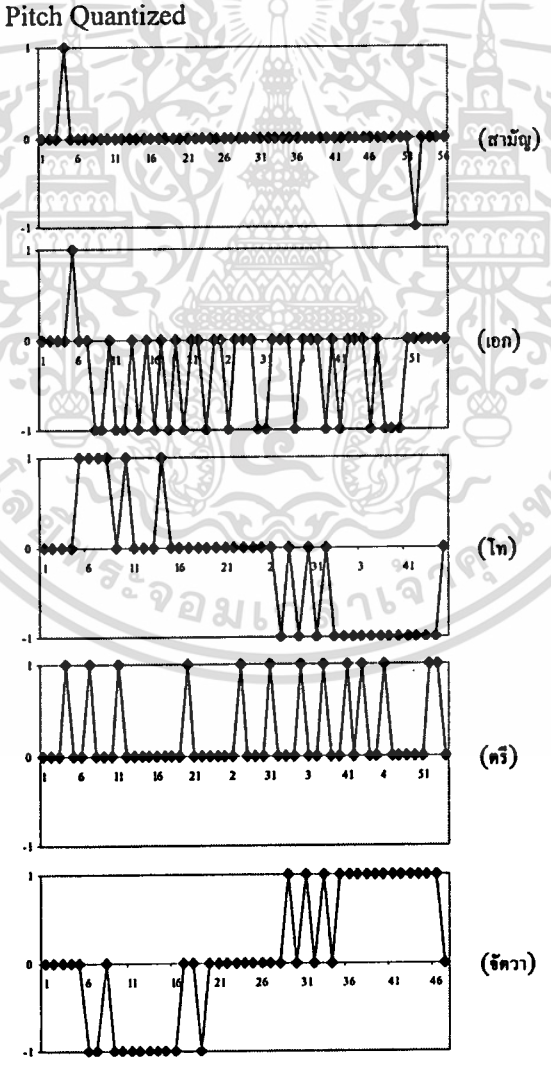
ขั้นตอนที่ 1 การกรองค่ากลาง : เป็นส่วนของการปรับปรุงข้อมูลให้มีความต่อเนื่องมากขึ้น ซึ่งจากรูปที่ 7.6 เมื่อนำมาผ่านขั้นตอนนี้จะได้กราฟที่มีลักษณะดังรูปที่ 7.7



รูปที่ 7.7 ตัวอย่างข้อมูลที่นำมาผ่านตัวกรองค่ากลาง

จากรูปจะพบว่าข้อมูลในช่วงต้นเสียงจะถูกปรับให้มีความต่อเนื่องมากขึ้นเมื่อเทียบกับรูปที่ 7.6 จากนั้นข้อมูลที่ได้นี้จะถูกนำไปทำการควอนไตซ์ค่าการเปลี่ยนแปลงของความถี่ในขั้นตอนต่อไป

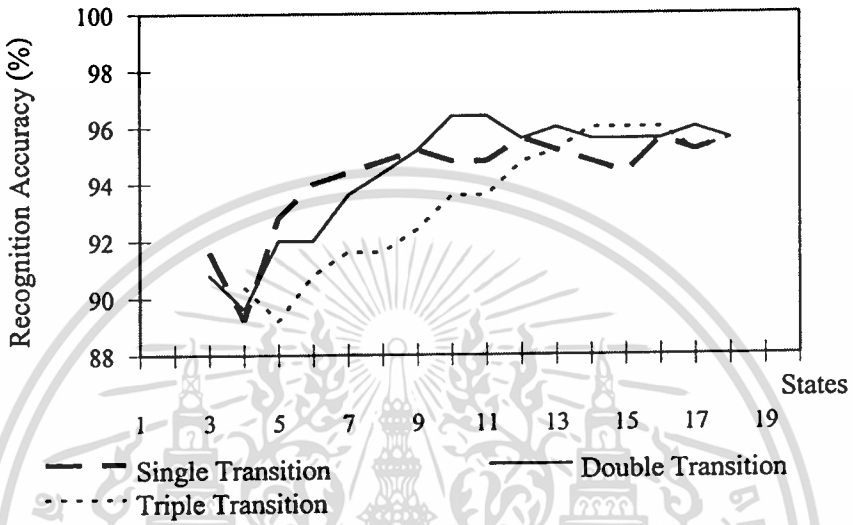
ขั้นตอนที่ 2 การควอนไตซ์ : เป็นการเตรียมข้อมูลเพื่อใช้เป็นข้อมูลฝึกสอนในกระบวนการสร้างแบบจำลองด้วย ฮิดเดนมาร์คอฟโมเดล ซึ่งข้อมูลที่ผ่านการกรองค่ากลางจะถูกนำมาทำการจัดระดับออกเป็น 3 ระดับตามทิศทางการเปลี่ยนแปลงของความถี่มูลฐานต่อเวลา โดยแทนค่าเป็น 1 เมื่อความถี่เพิ่มขึ้นเป็น 0 เมื่อความถี่คงที่ และ –1 เมื่อความถี่ลดลง ตัวอย่างสัญญาณข้อมูลในรูปที่ 7.7 เมื่อนำมาผ่านการควอนไตซ์ออกเป็น 3 ระดับสามารถแสดงได้ดังรูปที่ 7.8 เมื่อผ่านขั้นตอนนี้แล้วข้อมูลจะอยู่ในรูปของ “ลำดับการเปลี่ยนแปลงของความถี่” ซึ่งมีสมาชิกของลำดับเป็น {-1, 0, 1}



รูปที่ 7.8 การควอนไตซ์ข้อมูลออกเป็น 3 ระดับตามทิศทางการเปลี่ยนแปลงของความถี่มูลฐาน

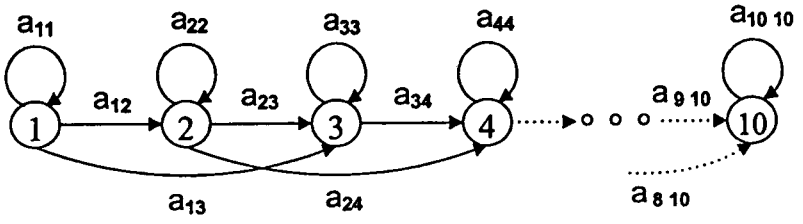
7.1.3 การสร้างแบบจำลองการรู้จำด้วย HMM

แบบจำลองที่ใช้ในวิทยานิพนธ์นี้เป็นแบบ Left-Right Model เพราะเป็นแบบจำลองที่เหมาะสำหรับการรู้จำรูปแบบของคำ ประเภทคำศัพท์เดี่ยว (Isolated word) เนื่องจากสามารถนำเวลาเข้ามาเกี่ยวข้องกับสเตทของแบบจำลองได้โดยตรง และอาจตีความหมายทางกายภาพของสเตทเป็นเสียงที่แตกต่างกันของคำได้ [15]



รูปที่ 7.9 เปอร์เซ็นต์ความถูกต้องเมื่อมีการเปลี่ยนแปลงสเตท และการย้ายข้ามสเตทของ HMM

จากรูปที่ 7.9 [6] แสดงให้เห็นว่าแบบจำลองที่มีจำนวนสเตทเพิ่มมากขึ้นจะให้ผลการรู้จำที่แม่นยำขึ้น และมีค่าค่อนข้างคงที่ตั้งแต่สเตทที่ 12 ขึ้นไป โดยแบบจำลองที่สร้างจาก HMM 4 สเตทให้ผลการรู้จำแม่นยำน้อยที่สุด ทั้งนี้เนื่องจากจำนวนสเตทมีผลโดยตรงต่อค่าพารามิเตอร์ B (ค่าความน่าจะเป็นของค่าปรากฏที่สามารถเป็นไปได้ภายในสเตท {-1, 0, 1}) ซึ่งจะเป็นตัวกำหนดรายละเอียดของข้อมูล โดยอยู่ในรูปของเมตริกซ์ขนาดเท่ากับ "จำนวนสเตทxระดับการควอนไตซ์" ดังนั้นถ้าสเตทมีจำนวนน้อยเมตริกซ์ B จะมีขนาดเล็ก ซึ่งนั่นหมายถึงรายละเอียดของข้อมูลในแบบจำลองจะน้อยตามลงไปด้วยเป็นผลให้การรู้จำมีความแม่นยำลดลง ในขณะเดียวกันแบบจำลอง HMM ที่สร้างจากสเตทจำนวนมาก จากรูปจะสังเกตุได้ว่าตั้งแต่สเตทที่ 12 ถึง 18 ให้ค่าการรู้จำสูงและค่อนข้างคงที่ ทั้งนี้เป็นผลเนื่องมาจากข้อมูลอินพุท คือค่าการควอนไตซ์ความถี่มีระดับเพียงแค่ 3 ระดับซึ่งให้ค่ารายละเอียดของข้อมูลน้อยเกินไปเมื่อเทียบกับจำนวนสเตท ทำให้ไม่ว่าจะเพิ่มจำนวนสเตทขึ้นเท่าใดก็ไม่ทำให้ผลการรู้จำแม่นยำเพิ่มขึ้น แบบจำลองที่ให้ผลการรู้จำดีที่สุดอยู่ที่ 10 สเตท โดยมีรูปแบบของการย้ายข้ามสเตทได้สูงสุด 2 สเตท แสดงได้ดังรูปที่ 7.10 โดยให้ผลการแม่นยำสูงสุดถึง 96.4 %



รูปที่ 7.10 แบบจำลอง HMM ที่ 10 สเตทมีรูปแบบของการย้ายข้ามสเตทได้สูงสุด 2 สเตท

การสร้างแบบจำลองอ้างอิงที่ใช้ในการรู้จำระดับเสียง โดยใช้เสียงต้นแบบจำนวนทั้งสิ้น 1000 เสียง จาก 5 แบบจำลอง (200x5) แบ่งออกเป็นแบบจำลองระดับเสียงสามัญ เสียงเอก เสียงโท เสียงตรี และเสียงจัตวา ระดับเสียงละ 200 เสียง โดยรูปแบบของคำที่ใช้เป็นคำพยางค์เดียว ดังแสดงไว้ในตารางที่ 7.1 ซึ่งเสียงต้นแบบที่ใช้ ได้จากผู้ออกเสียงจำนวนทั้งสิ้น 10 คน แบ่งเป็นชาย 5 คน และหญิง 5 คน เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี ออกเสียงคำในตารางที่ 7.1 คนละ 1 ครั้ง โดยในการสร้างแบบจำลองนั้นจะสร้างแบบจำลองวรรณยุกต์ภาษาไทย 5 ระดับนั้น จะทำจากข้อมูลในการหาค่าพิทช์โดยวิธีออโตคอร์รีเลชั่นฟังก์ชั่นใช้เทคนิคการคลิปยอดของสัญญาณ ในระดับที่ 80%, 70%, 60%, 50%, 40%, 30% และ 20%

เนื่องจากหน่วยเสียงวรรณยุกต์เป็นหน่วยเสียงซ้อนที่วางตัวอยู่เหนือหน่วยเสียงก้อง ดังนั้นรูปแบบของคำที่ใช้ในวิทยานิพนธ์นี้จึงพยายามใช้คำที่มี หน่วยเสียงพยัญชนะต้น หน่วยเสียงสระ และพยัญชนะสะกดต่างๆกัน [16] เพื่อให้ได้แบบจำลองอ้างอิงที่ครอบคลุมความหลากหลายมากที่สุด

ตารางที่ 7.1 กลุ่มคำที่ใช้ในการสร้างแบบจำลองและทดสอบ

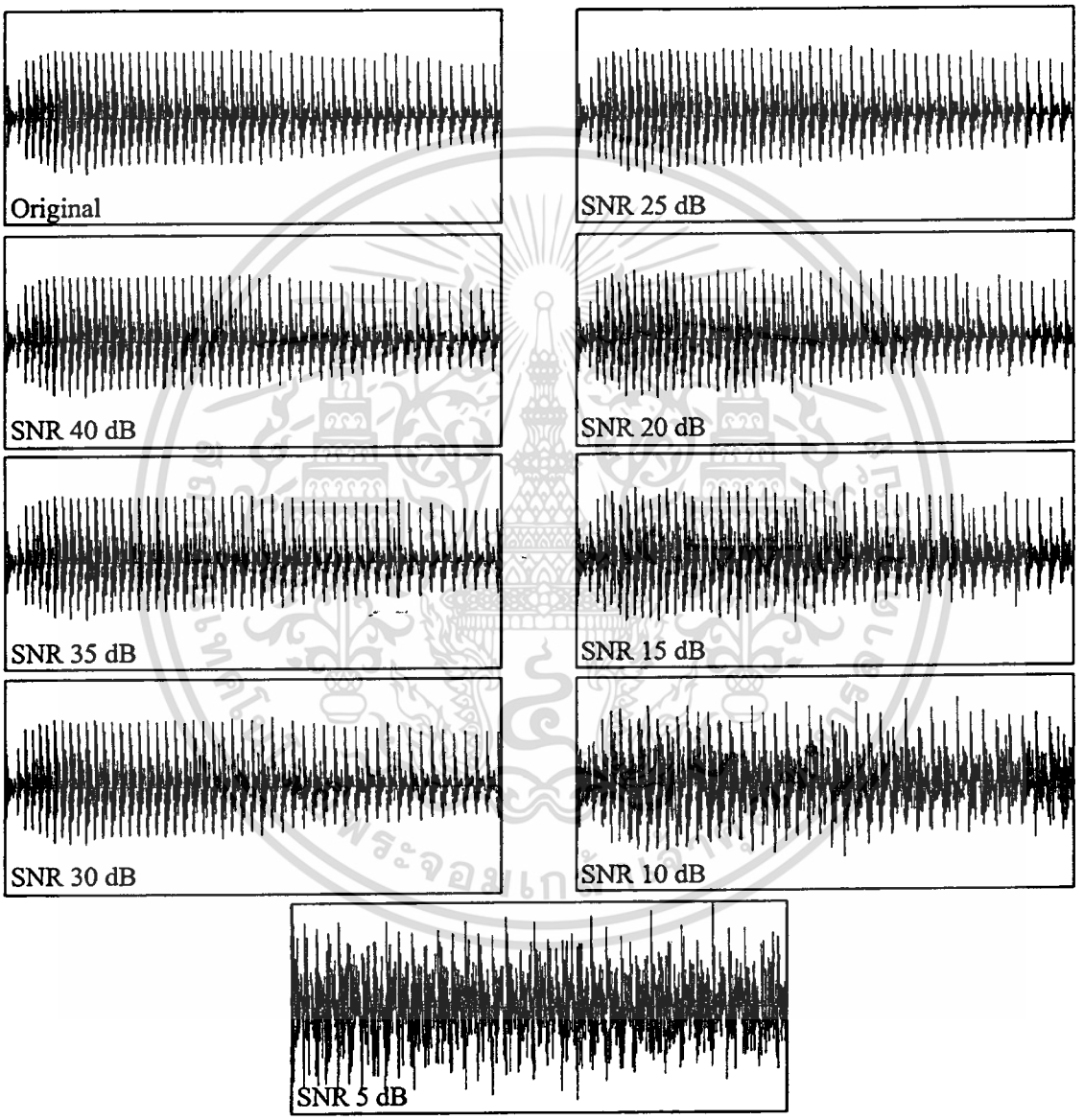
คำที่	ระดับเสียงวรรณยุกต์				
	สามัญ	เอก	โท	ตรี	จัตวา
1	กิน	บอก	ป้า	น้ำ	อ๋อ
2	แก	ไก่	ลูก	ค้า	แจ๋ว
3	เกิน	ตี่	ย่า	ล้อ	ติ๋ม
4	ลอ	อ่าน	น่า	น้อง	โอ๋
5	ปี	เต่า	แก้	มด	ปู๋
6	กาน	แต่	โก้	น้า	หุง
7	ออม	เตะ	ปู้	ผู้	ขาย
8	เกา	หนึ่ง	ยุ่ง	พื้น	หา
9	ดัง	ปู่	เก้า	ล้าน	หมอง
10	ปู	ต่อ	บี้	โย้	หมา
11	กาง	ดิบ	เอ้	นี้	หงอ
12	ลือ	แกะ	อ้อ	ไม้	เก๋
13	ไอ	กก	เก้อ	นก	หนี
14	ตา	ปาก	หนี้	วัด	หมู
15	เปีย	กีบ	เลข	นึก	ไข
16	ตึง	แตก	ลอก	นัด	หวาย
17	เป็น	ปัก	ทาก	มิด	เขียว
18	อำ	เอก	เชื่อ	แม้น	สอง
19	ตอ	ป่า	ชื่อ	ซื้อ	เก๋า
20	ปน	ป่าย	ใบ้	เท้า	ยั๋ง

ผลที่ได้จากการทดลองในส่วนนี้ คือ แบบจำลองอ้างอิงจำนวน 5 แบบจำลอง ได้แก่

1. แบบจำลองอ้างอิง เสียงสามัญ แทนด้วยพารามิเตอร์ $\lambda_1 = (A_1, B_1, \pi)$
2. แบบจำลองอ้างอิง เสียงเอก แทนด้วยพารามิเตอร์ $\lambda_2 = (A_2, B_2, \pi)$
3. แบบจำลองอ้างอิง เสียงโท แทนด้วยพารามิเตอร์ $\lambda_3 = (A_3, B_3, \pi)$
4. แบบจำลองอ้างอิง เสียงตรี แทนด้วยพารามิเตอร์ $\lambda_4 = (A_4, B_4, \pi)$
5. แบบจำลองอ้างอิง เสียงจัตวา แทนด้วยพารามิเตอร์ $\lambda_5 = (A_5, B_5, \pi)$

7.2 ขั้นตอนในการทดสอบแบบจำลองอ้างอิงที่สร้างขึ้น

ทำการทดสอบแบบจำลองที่สร้างขึ้น โดยทดสอบกับเสียงต้นแบบที่ใช้สร้างแบบจำลองอ้างอิงเสียงวรรณยุกต์ λ_1 - λ_5 และทำการสอดแทรกสัญญาณรบกวนขาวที่ระดับ Signal to Noise Ratio : SNR ที่ 40dB, 35dB, 30dB, 25dB, 20dB, 15 dB, 10 dB และ 5 dB ตามลำดับในเสียงต้นแบบ ดังแสดงในรูปที่ 7.11



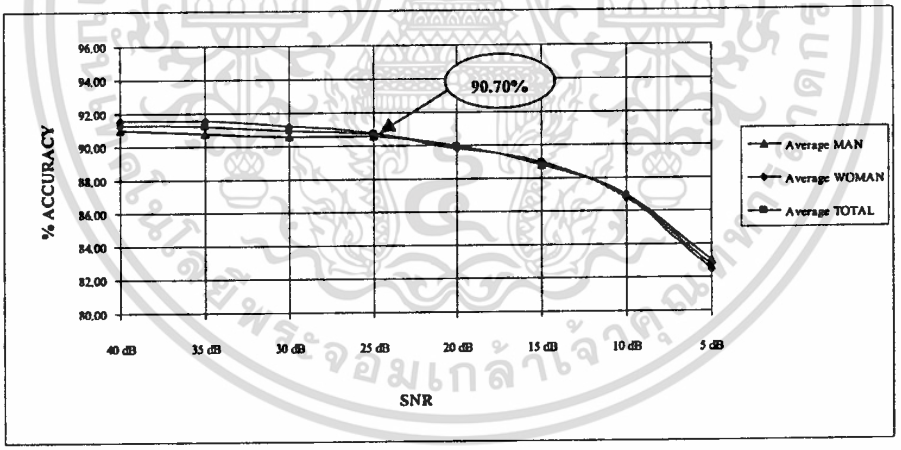
รูปที่ 7.11 แสดงตัวอย่างเสียงที่ใช้ทดสอบกับแบบจำลองอ้างอิงที่สร้างขึ้น

โดยนำมาทดสอบกับกลุ่มคำต้นแบบและใช้เสียงต้นแบบเดิมแต่มีการสอดแทรกสัญญาณรบกวนขาว ซึ่งแต่ละระดับเสียงใช้จำนวนเสียงในการทดสอบเท่ากันคือ 200 เสียง จากผู้ออกเสียง 10 คน และจากการทดสอบ ดังแสดงไว้ในรูปที่ 7.12-7.19

ผู้ออกเสียง	การรู้จำเสียงถูกต้อง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	91.00	91.00	91.00	91.00	91.00	89.00	87.00	83.00
M2	90.00	91.00	91.00	91.00	90.00	90.00	88.00	84.00
M3	91.00	90.00	90.00	90.00	89.00	89.00	87.00	82.00
M4	92.00	91.00	91.00	91.00	90.00	88.00	87.00	84.00
M5	91.00	91.00	90.00	90.00	90.00	88.00	86.00	82.00
W1	92.00	92.00	92.00	91.00	89.00	89.00	87.00	83.00
W2	92.00	93.00	92.00	92.00	91.00	90.00	88.00	84.00
W3	92.00	92.00	91.00	91.00	90.00	89.00	87.00	82.00
W4	92.00	91.00	91.00	90.00	90.00	88.00	86.00	81.00
W5	90.00	90.00	90.00	90.00	89.00	89.00	86.00	82.00
Average	91.30	91.20	90.90	90.70	89.90	88.90	86.90	82.70

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

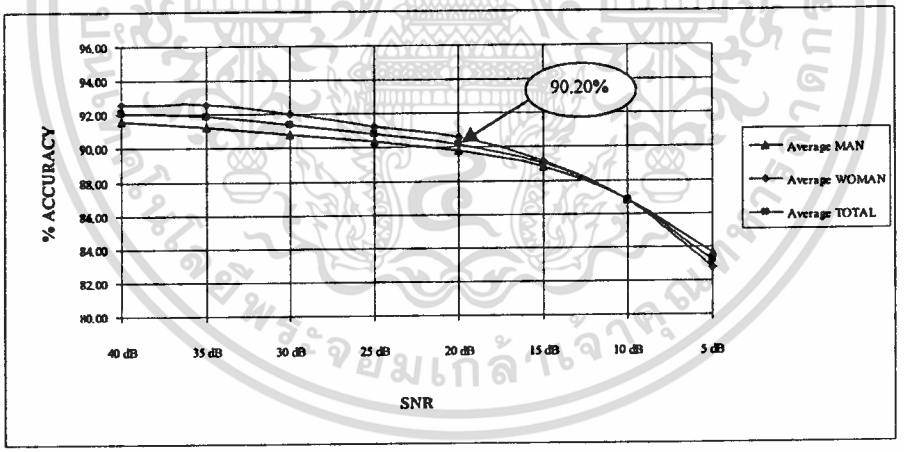
รูปที่ 7.12 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 80 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 80 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณโดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 25 dB โดยมีความแม่นยำเฉลี่ย 90.70 เปอร์เซ็นต์

ผู้ออกเสียง	ความถูกต้องในการรู้จำเสียง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	92.00	92.00	92.00	91.00	91.00	89.00	87.00	85.00
M2	92.00	92.00	91.00	91.00	90.00	90.00	88.00	84.00
M3	91.00	90.00	90.00	90.00	89.00	89.00	87.00	83.00
M4	92.00	91.00	91.00	90.00	90.00	88.00	86.00	84.00
M5	91.00	91.00	90.00	90.00	89.00	88.00	86.00	82.00
W1	93.00	93.00	93.00	92.00	91.00	90:00	87.00	84.00
W2	93.00	93.00	92.00	92.00	91.00	90.00	88.00	84.00
W3	92.00	92.00	92.00	91.00	91.00	89.00	86.00	82.00
W4	93.00	93.00	92.00	91.00	90.00	88.00	86.00	82.00
W5	92.00	92.00	91.00	90.00	90.00	89.00	87.00	82.00
Average	92.10	91.90	91.40	90.80	90.20	89.00	86.80	83.20

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

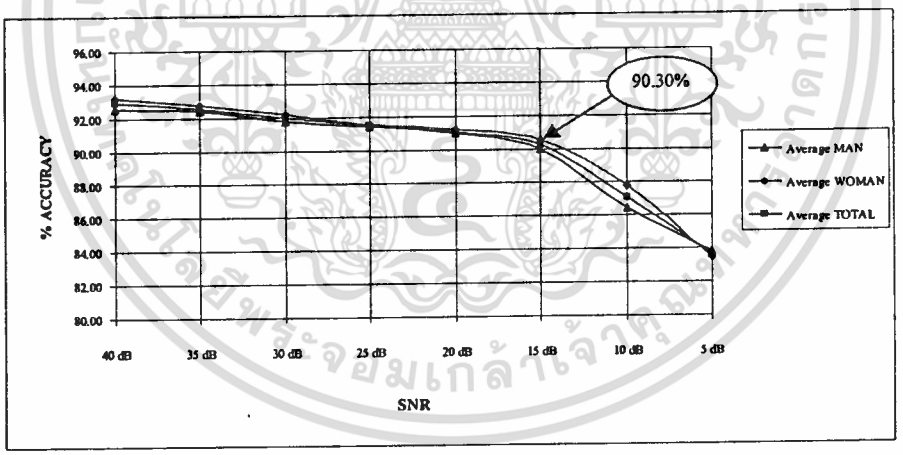
รูปที่ 7.13 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 70 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 70 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณ โดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 20 dB โดยมีความแม่นยำเฉลี่ย 90.20 เปอร์เซ็นต์

ผู้ออกเสียง	การรู้จำเสียงถูกต้อง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	93.00	93.00	92.00	92.00	91.00	90.00	87.00	85.00
M2	92.00	92.00	92.00	91.00	91.00	90.00	86.00	83.00
M3	93.00	93.00	92.00	92.00	92.00	91.00	88.00	84.00
M4	92.00	92.00	91.00	91.00	90.00	90.00	86.00	84.00
M5	93.00	92.00	92.00	91.00	91.00	89.00	85.00	83.00
W1	94.00	94.00	93.00	92.00	92.00	91.00	88.00	84.00
W2	94.00	93.00	93.00	93.00	92.00	92.00	89.00	85.00
W3	93.00	92.00	92.00	91.00	91.00	90.00	88.00	83.00
W4	93.00	93.00	92.00	91.00	91.00	91.00	87.00	83.00
W5	92.00	92.00	91.00	91.00	90.00	89.00	87.00	82.00
Average	92.90	92.60	92.00	91.50	91.10	90.30	87.10	83.60

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

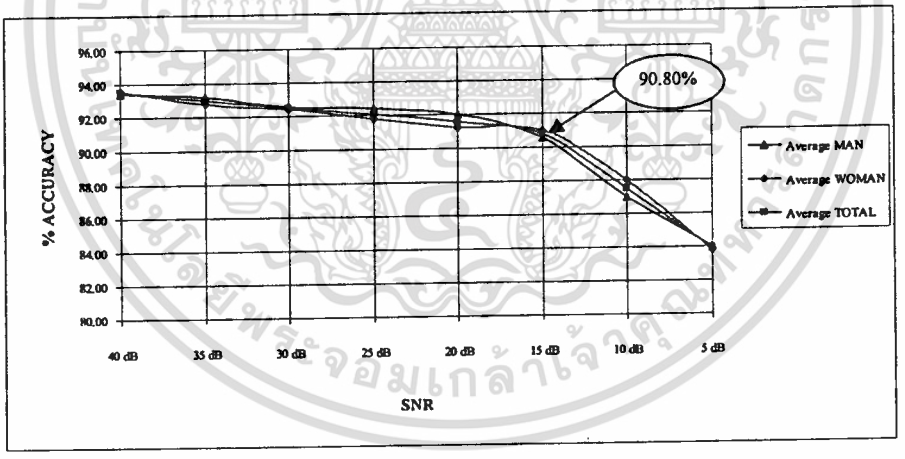
รูปที่ 7.14 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 60 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 60 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณโดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 15 dB โดยมีความแม่นยำเฉลี่ย 90.30 เปอร์เซ็นต์

ผู้ออกเสียง	การจำเสียงถูกต้อง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	94.00	94.00	93.00	93.00	93.00	91.00	88.00	85.00
M2	94.00	94.00	94.00	93.00	93.00	92.00	87.00	84.00
M3	93.00	93.00	92.00	92.00	92.00	90.00	88.00	84.00
M4	93.00	92.00	92.00	92.00	91.00	90.00	86.00	84.00
M5	93.00	93.00	92.00	92.00	91.00	90.00	86.00	83.00
W1	95.00	94.00	94.00	93.00	92.00	92.00	89.00	85.00
W2	94.00	94.00	93.00	93.00	92.00	92.00	89.00	85.00
W3	93.00	92.00	92.00	92.00	91.00	90.00	87.00	83.00
W4	93.00	93.00	92.00	91.00	91.00	91.00	88.00	84.00
W5	93.00	91.00	91.00	90.00	90.00	90.00	87.00	82.00
Average	93.50	93.00	92.50	92.10	91.60	90.80	87.50	83.90

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

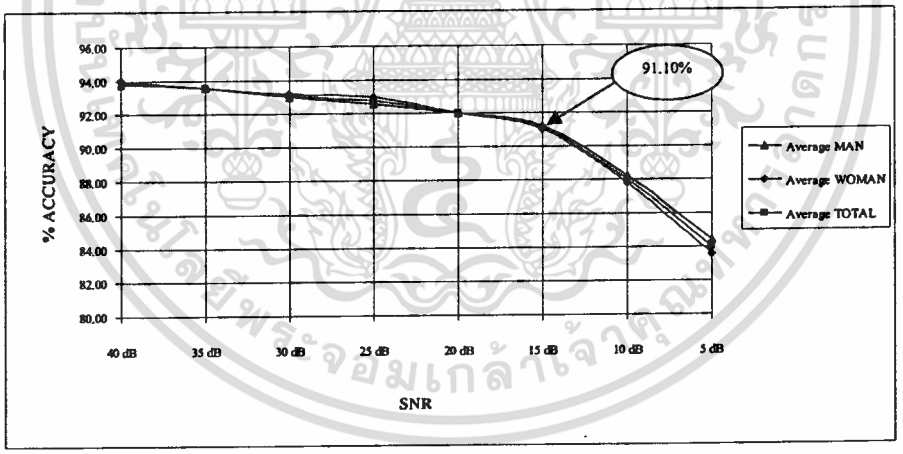
รูปที่ 7.15 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 50 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 50 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณ โดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 15 dB โดยมีความแม่นยำเฉลี่ย 90.80 เปอร์เซ็นต์

ผู้ออกเสียง	อัตราการรู้จำเสียงถูกต้อง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	95.00	95.00	94.00	93.00	93.00	92.00	90.00	86.00
M2	94.00	94.00	94.00	93.00	93.00	92.00	89.00	85.00
M3	94.00	94.00	93.00	93.00	92.00	91.00	88.00	84.00
M4	93.00	92.00	92.00	92.00	91.00	91.00	88.00	84.00
M5	93.00	93.00	92.00	92.00	91.00	90.00	86.00	83.00
W1	95.00	95.00	95.00	94.00	93.00	91.00	89.00	85.00
W2	94.00	94.00	93.00	93.00	92.00	92.00	89.00	85.00
W3	94.00	93.00	93.00	93.00	92.00	91.00	87.00	83.00
W4	94.00	94.00	93.00	93.00	92.00	91.00	87.00	83.00
W5	93.00	92.00	92.00	92.00	91.00	90.00	87.00	82.00
Average	93.90	93.60	93.10	92.80	92.00	91.10	88.00	84.00

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

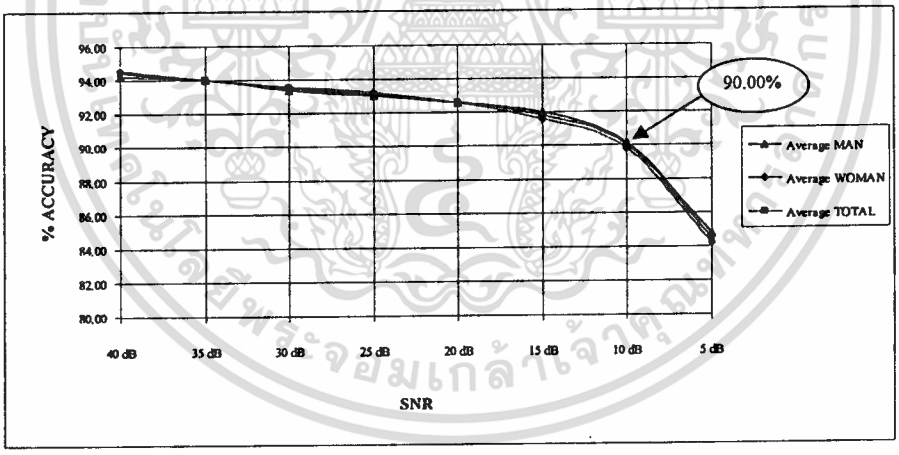
รูปที่ 7.16 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 40 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 40 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณ โดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 15 dB โดยมีความแม่นยำเฉลี่ย 91.10 เปอร์เซ็นต์

ผู้ออกเสียง	ความถูกต้อง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	95.00	95.00	94.00	94.00	93.00	93.00	92.00	87.00
M2	95.00	95.00	94.00	93.00	93.00	92.00	90.00	86.00
M3	94.00	94.00	94.00	93.00	93.00	92.00	91.00	85.00
M4	94.00	93.00	93.00	93.00	93.00	92.00	90.00	84.00
M5	93.00	93.00	92.00	92.00	91.00	91.00	88.00	82.00
W1	96.00	95.00	95.00	94.00	94.00	92.00	90.00	85.00
W2	95.00	95.00	94.00	94.00	93.00	93.00	89.00	84.00
W3	94.00	93.00	93.00	93.00	92.00	91.00	90.00	84.00
W4	94.00	94.00	93.00	93.00	92.00	92.00	91.00	85.00
W5	94.00	93.00	93.00	92.00	92.00	90.00	89.00	83.00
Average	94.40	94.00	93.50	93.10	92.60	91.80	90.00	84.50

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

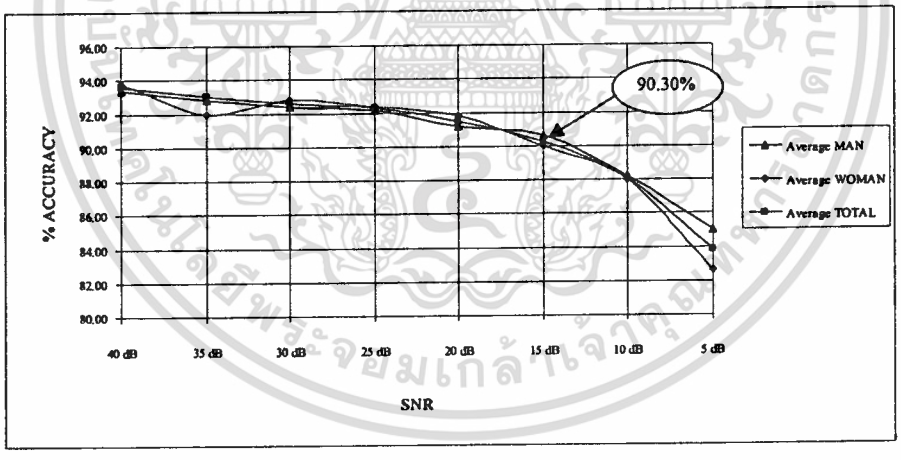
รูปที่ 7.17 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 30 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 30 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณโดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 10 dB โดยมีความแม่นยำเฉลี่ย 90.00 เปอร์เซ็นต์

ผู้ออกเสียง	การรู้จำเสียงถูกต้อง (%)							
	SNR 40 dB	SNR 35 dB	SNR 30 dB	SNR 25 dB	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB
M1	94.00	94.00	93.00	93.00	92.00	92.00	90.00	87.00
M2	94.00	93.00	93.00	93.00	92.00	92.00	90.00	86.00
M3	93.00	92.00	92.00	91.00	90.00	89.00	86.00	84.00
M4	94.00	93.00	93.00	93.00	92.00	90.00	87.00	84.00
M5	92.00	92.00	91.00	91.00	90.00	90.00	88.00	84.00
W1	95.00	94.00	94.00	93.00	93.00	90.00	89.00	83.00
W2	94.00	94.00	93.00	93.00	93.00	91.00	89.00	82.00
W3	94.00	93.00	93.00	93.00	92.00	90.00	88.00	83.00
W4	93.00	93.00	92.00	92.00	91.00	89.00	86.00	83.00
W5	93.00	92.00	92.00	91.00	90.00	90.00	88.00	82.00
Average	93.60	93.00	92.60	92.30	91.50	90.30	88.10	83.80

หมายเหตุ M1- M5 แทนผู้ออกเสียงต้นแบบที่เป็นชาย คนที่ 1-5, W1- W5 แทนผู้ออกเสียงต้นแบบที่เป็นหญิง คนที่ 1-5 เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี

(a) ตารางแสดงผลการรู้จำเสียง



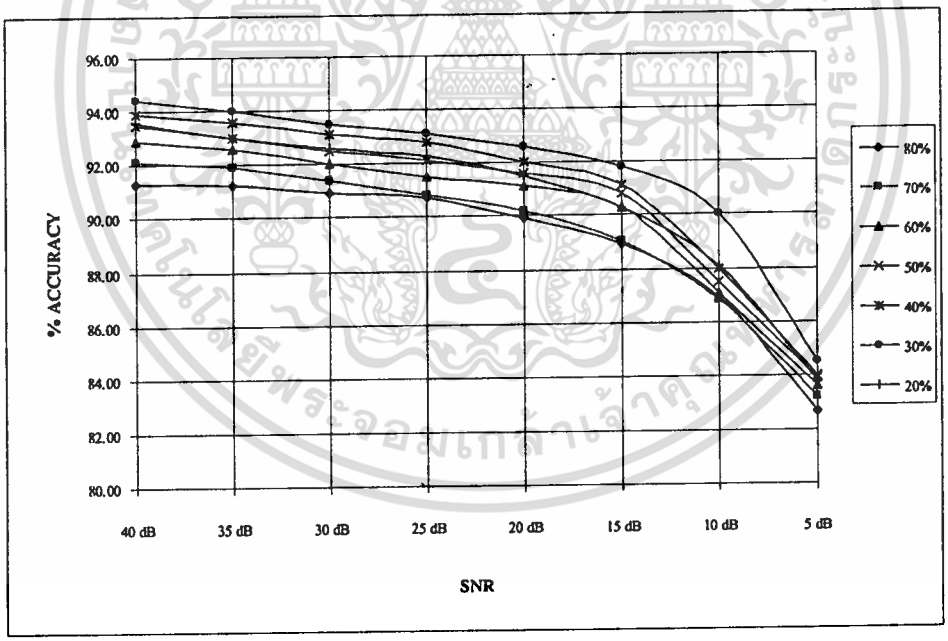
(b) กราฟแสดงค่าเฉลี่ยผลการรู้จำเสียง

รูปที่ 7.18 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยกำหนด Clipping Level ที่ 20 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณเสียง

จากผลการทดสอบแบบจำลองโดยกำหนดการคลิปยอดของสัญญาณที่ 20 เปอร์เซ็นต์ของแอมปลิจูดสูงสุดของสัญญาณโดยให้ผลว่าแบบจำลองอ้างอิงที่สร้างขึ้นนี้สามารถทนทานต่อสัญญาณรบกวนโดยเฉลี่ยที่ระดับสัญญาณต่อสัญญาณรบกวน 15 dB โดยมีความแม่นยำเฉลี่ย 90.30 เปอร์เซ็นต์

ตารางที่ 7.2 ผลการทดสอบแบบจำลองอ้างอิง โดยใช้กลุ่มคำและเสียงต้นแบบ จำนวน 1,000 เสียง โดยแสดงความสัมพันธ์ระหว่าง Clipping Level ของสัญญาณกับเสียงที่มีการสอดแทรกสัญญาณรบกวนขาวในระดับ Signal to Noise Ratio : SNR ต่างๆ

	การรู้จำเสียงถูกต้อง (%)							
SNR / Clipping Level	40 dB	35 dB	30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
80 %	91.30	91.20	90.90	90.70	89.90	88.90	86.90	82.70
70 %	92.10	91.90	91.40	90.80	90.40	89.00	86.80	83.20
60 %	92.90	92.60	92.00	91.50	91.10	89.80	87.10	83.60
50 %	93.50	93.00	92.50	92.10	91.60	90.30	87.50	83.90
40 %	93.90	93.60	93.10	92.80	92.00	90.70	88.00	84.00
30 %	94.40	94.00	93.50	93.10	92.60	91.80	89.00	84.50
20 %	93.60	93.00	92.60	92.30	91.50	90.40	88.10	83.80



รูปที่ 7.19 แสดงกราฟความสัมพันธ์ระหว่าง Clipping Level ของสัญญาณกับเสียงที่มีการสอดแทรกสัญญาณรบกวนขาวในระดับ Signal to Noise Ratio : SNR ต่างๆ

จากตารางที่ 7.2 จะพบว่าผลของการคลิปยอดของสัญญาณที่ระดับ 30 เปอร์เซ็นต์ให้ผลการรู้จำที่แม่นยำถึง 90.00 เปอร์เซ็นต์โดยสามารถทนทานต่อสัญญาณรบกวนในระดับสัญญาณต่อสัญญาณรบกวนได้ถึง 10 dB

บทที่ 8
สรุปผลและข้อเสนอแนะ

วิทยานิพนธ์นี้เป็นการเสนอแนวทางในการสร้างระบบการรู้จำเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีเสียงรบกวน โดยทำการวิเคราะห์ในโดเมนของเวลาในการหาค่าพิทช์ (Pitch Extraction)ใช้วิธีออโตโครีเลชั่นฟังก์ชั่นโดยใช้เทคนิคการคลิปยอดของสัญญาณ (Autocorrelation Method using Center Clipping : AUTOC) โดยทำการคลิปยอดสัญญาณเสียงเพื่อที่จะหาค่าการคลิปที่เหมาะสมและลดผลกระทบของสัญญาณรบกวนที่มีต่อระบบการรู้จำเสียงวรรณยุกต์ภาษาไทย สัญญาณรบกวนที่ใช้ในงานวิจัยนี้ได้ใช้สัญญาณรบกวนเกาส์เซียนขาว (White Gaussian Noise) ซึ่งจะสอดแทรกเข้าไปในสัญญาณเสียงที่จะทำการทดสอบ

8.1 การทดลอง

การทดลองแบ่งออกเป็น 2 ส่วนคือ

1. การวิเคราะห์และพัฒนาอัลกอริธึมในการสร้างแบบจำลองการรู้จำเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีเสียงรบกวน

การหาค่าพิทช์ (Pitch Extraction)ของสัญญาณเสียง จะหาได้โดยใช้วิธีออโตโครีเลชั่นฟังก์ชั่นโดยใช้เทคนิคการคลิปยอดของสัญญาณ (Autocorrelation Method using Center Clipping : AUTOC) โดยจะทำการกำหนดการคลิปยอดของสัญญาณออกเป็น 80%, 70%, 60%, 50%, 40%, 30% และ 20% และนำไปสร้างแบบจำลองของการคลิปแต่ละระดับ ซึ่งค่าพิทช์ที่ได้จะถูกเปลี่ยนให้อยู่ในรูปของค่าความถี่มูลฐาน มีลักษณะรูปแบบการเปลี่ยนแปลงค่าความถี่มูลฐาน (หรือเรียกว่าเส้นทางเดินเสียงวรรณยุกต์) ที่แตกต่างกัน จะเป็นตัวบ่งบอกถึงระดับเสียงของวรรณยุกต์ที่แตกต่างกันในแต่ละคำหรือพยางค์ จากนั้นจึงใช้วิธีการควอนไตซ์โดยทำการจัดกลุ่มค่าความถี่มูลฐานออกเป็น 3 ระดับตามแนวทางการเปลี่ยนแปลงของความถี่มูลฐานที่เพิ่มขึ้น คงที่ หรือลดลงเมื่อเวลาเปลี่ยนไป เพื่อนำไปเป็นข้อมูลฝึกสอนให้กับการสร้างแบบจำลองการรู้จำหน่วยเสียงวรรณยุกต์ภาษาไทยทั้ง 5 ระดับด้วยวิธี Hidden Markov Modeling (HMM) โดยเลือก HMM ขนาด 10 สเตทและมีการย้ายข้ามสเตทได้สูงสุดไม่เกิน 2 สเตท เป็นรูปแบบที่เหมาะสมและให้ความแม่นยำมากที่สุด

ในการสร้างแบบจำลองได้เลือกคำต้นแบบที่มี หน่วยเสียงพยัญชนะต้น ความสั้น-ยาวของหน่วยเสียงสระ และหน่วยเสียงพยัญชนะสะกด ที่แตกต่างกัน เพื่อให้ได้แบบจำลองอ้างอิงที่ครอบคลุมและมีความหลากหลายมากที่สุด โดยใช้คำจำนวนทั้งหมด 100 คำ จากผู้ออกเสียง

10 คน เป็นผู้ชาย 5 คน และผู้หญิง 5 คน เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี ออกเสียงเพื่อใช้เป็นข้อมูลฝึกสอนในการสร้างแบบจำลองอ้างอิงจำนวนทั้งสิ้น 1,000 เสียง

2. การทดสอบแบบจำลองอ้างอิงที่สร้างขึ้น

ใช้ผู้ออกเสียงต้นแบบจำนวน 10 คนเป็นผู้ชาย 5 คนและผู้หญิง 5 คน เป็นคนไทยภาคกลาง อายุระหว่าง 24 – 30 ปี ซึ่งจะทำการสอดแทรกสัญญาณรบกวนขาวในระดับต่างๆที่กำหนด จากนั้นทำการทดสอบกับแบบจำลองอ้างอิงที่สร้างขึ้นโดยสร้างจากข้อมูลที่ได้จากการหาค่าพิทช์ซึ่งกำหนดค่าการคลิปยอดของสัญญาณที่ระดับ 80%, 70%, 60%, 50%, 40%, 30%และ 20% ตามลำดับ ซึ่งให้ผลว่าการคลิปยอดของสัญญาณที่ระดับ 30% ให้ผลการรู้จำที่ดีและสามารถทนทานต่อสัญญาณรบกวนได้ถึง 10 dB ซึ่งให้ความแม่นยำถึง 90.00 เปอร์เซ็นต์

8.2 ปัญหาที่พบในการทดลองและข้อเสนอแนะ

จากวิธีการที่ได้นำเสนอในวิทยานิพนธ์นี้ ก็เป็นวิธีหนึ่งที่สามารถทำให้ระบบการรู้จำเสียงสามารถทนทานต่อสัญญาณรบกวนได้ โดยตามลักษณะการคลิปยอดของสัญญาณแล้วถ้าเรากำหนดระดับสูงเกินไปก็อาจจะทำให้ข้อมูลที่ต้องการสูญหายไปหรือในกรณีที่มีสัญญาณรบกวนมากสัญญาณที่ได้จากการคลิปยอดสัญญาณมาอาจจะเป็นแค่สัญญาณรบกวนเพียงอย่างเดียว ทำให้เมื่อนำมาสร้างแบบจำลองและทำการทดสอบผลที่ออกมามีประสิทธิภาพที่ด้อยลง ในทางกลับกันเมื่อเรากำหนดค่าการคลิปยอดของสัญญาณต่ำเกินไปผลที่ได้ก็คือผลอันเนื่องมาจากการตอบสนองทางความถี่ภายในช่องทางเดินเสียงก็จะไม่ถูกกำจัดออกไป ทำให้เวลาในการคำนวณออโตคอร์รีเลชั่นเพื่อหาค่าคาบพิทช์เกิดความคลาดเคลื่อนไปจากตำแหน่งจริงได้

เอกสารอ้างอิง

[1] ชวดี อิศรปรีดา และคณะ. "การรู้จำเสียงพูด." ปริญญานิพนธ์วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมโทรคมนาคม, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2538.

[2] กรุณา แก้วสมศรี และคณะ. "การต่อหมายเลขโทรศัพท์โดยใช้เสียง." ปริญญานิพนธ์วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมโทรคมนาคม, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2539.

[3] ธันวา ศรีประโมง. "การวิเคราะห์เสียงพูดภาษาไทยในแกนความถี่ฮาร์โมนิค." วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2537.

[4] ณัฐกร ทับทอง. "การรู้จำคำพูดภาษาไทย โดยใช้ลักษณะบ่งความต่างของหน่วยเสียง." วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย. 2538.

[5] ทัศเวท วีระวัฒน์. "การรู้จำเสียงคำไทยเฉพาะบุคคล." วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2541.

[6] จิตรลดา จารุมิศร์. "การออกแบบจำลองในการรู้จำเสียงวรรณยุกต์สำหรับภาษาไทยโดยใช้เทคนิคการควอนไตซ์พิทช์ และ Hidden Makov Modelling." วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถานบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2542.

[7] วิไลวรรณ ขนิษฐานันท์. **ภาษาและภาษาศาสตร์. พิมพ์ครั้งที่ 5.** กรุงเทพฯ :สำนักพิมพ์มหาวิทยาลัยธรรมศาสตร์. 2533.

[8] ราตรี ธันวารชร. **การศึกษาภาษาไทยตามแนวภาษาศาสตร์ เล่ม๑ เสียงและระบบเสียงในภาษาไทย.** กรุงเทพฯ : คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์. 2537.

[9] ชาคริต อนันทราวัน. **หลักภาษาไทย.** กรุงเทพฯ : โอเดียนสโตร์. 2524.

[10] อภิชาต ตั้งทางธรรม. "การเปลี่ยนความเร็วของเสียงพูด." **การประชุมทางวิศวกรรมไฟฟ้าครั้งที่ 17,** 2537. หน้า 329-332.

[11] L.R. Rabiner, R.W. Schafer. **Digital Processing of Speech Signals.** New Jersey : Prentice-Hall, Inc. 1978.

[12] L.R. Rabiner, M.J. Cheng et. Al. "A Comparative Performance Study of Several Pitch Detection Algorithms." **IEEE Trans. Acoust., Speech, Signal Processing,** vol.ASSP-24, no.5, pp. 399-418, Oct. 1976.

[13] A. Rosenfeld, A.C. Kak. **Digital picture Processing.** Orlando Florida : Academic Press Inc. 1982.

[14] W. Thomas. **Voice and Speech Processing.** New York : McGraw-Hill, Inc. 1987.

[15] L.R. Rabiner, B.H. Juang. **Fundamentals of Speech Recognition.** New Jersey : Prentice Hall, Inc. 1993.

[16] เรืองเดช ปันเขื่อนขัติย์. **แบบทดสอบระบบเสียงวรรณยุกต์ภาษาถิ่น.** สถาบันวิจัยภาษาและวัฒนธรรมเพื่อพัฒนาชนบท มหาวิทยาลัยมหิดล. 2532.

ภาคผนวก ก.
ผลงานวิจัยที่ได้รับการตีพิมพ์

1. กฤษกร สุนทรมนทกานติ , ไกรสิน ส่งวัฒนา "การเปรียบเทียบระหว่างวิธีการ Pitch Extraction สองวิธีสำหรับการรู้จำเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีสัญญาณรบกวนขาว", วิศวกรรมสาร ลาดกระบัง, ปีที่ 21 ฉบับที่ 4 เดือนธันวาคม 2547

การเปรียบเทียบระหว่างวิธีการ Pitch Extraction สองวิธี สำหรับการรู้จำเสียงวรรณยุกต์ภาษาไทย ในสภาวะที่มีสัญญาณรบกวนขาว

Comparison of two Pitch Extraction Method for Recognition of Thai Tone in the Presence of White Noise

กฤษกร สุนทรมนทกานติ ไกรสิน ส่งวัฒนา

ภาควิชาวิศวกรรมโทรคมนาคม คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

บทความนี้เปรียบเทียบวิธีการ Pitch Extraction ที่ใช้ในการรู้จำหน่วยเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีสัญญาณรบกวนขาว โดยจะเปรียบเทียบวิธี Pitch Extraction 2 วิธีคือ 1) Autocorrelation Method using Center Clipping (AUTOC) กับ 2) Average Magnitude Difference Function (AMDF) ค่าคาบเวลาพิทช์ของสัญญาณเสียงจาก 2 วิธีนี้ จะถูกนำมาทำการควอนไตซ์เบี่ยงเบนออกเป็น 3 ระดับของความถี่มูลฐานของคาบเวลาพิทช์นั้นๆ เพื่อเป็นข้อมูลฝึกสอนให้กับการสร้างแบบจำลองการรู้จำหน่วยเสียงวรรณยุกต์ภาษาไทยทั้ง 5 ระดับด้วยวิธี Hidden Markov Modeling (HMM) จากนั้นทำการทดลองความทนทานการรู้จำหน่วยเสียงวรรณยุกต์ภาษาไทยในสภาวะที่มีสัญญาณรบกวนขาว โดยการสุ่มป้อนข้อมูลเสียงวรรณยุกต์ในภาษาไทยทั้ง 5 ระดับที่ได้จากเพศชาย 5 คน และเพศหญิง 5 คน ในระดับอัตราสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio : SNR) ที่ต่างกัน โดยใช้สัญญาณรบกวนขาว พบว่าผลการรู้จำหน่วยเสียงวรรณยุกต์ภาษาไทยโดยการ Pitch Extraction วิธี Autocorrelation Method using Center Clipping (AUTOC) มีความทนทานต่อสัญญาณรบกวนดีกว่าวิธี Average Magnitude Difference Function (AMDF) มีความแม่นยำในการรู้จำเฉลี่ยมากกว่า 90 เปอร์เซ็นต์ ในระดับอัตราสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio : SNR) 20 dB.

Abstract

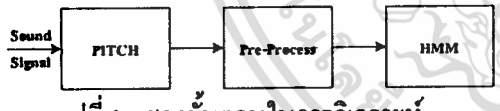
This paper presents pitch extraction method for recognition of Thai tone in the presence of white noise. We have compared 2 methods of pitch extraction : Autocorrelation Method using Center Clipping (AUTOC) and Average Magnitude Difference Function (AMDF). The pitch estimation by 2 solutions will be used to find the pitch differences and the sequence of pitch differences are then grouped into three quantized levels as training data to build model for recognition of 5 Thai tone levels using Hidden Markov Model (HMM). The speech from 5-male and 5-female. Thai subjects was collected and noise were added at different signal to noise ratio(SNR) levels. The testing results show recognition of Thai tone by pitch extraction used Autocorrelation Method using Center Clipping (AUTOC) method is more robust to noise than Average Magnitude Difference Function (AMDF) method. The average accuracy of recognition is more than 90 percent at signal to noise ratio (SNR) of 20 dB.

1. บทนำ

ในปัจจุบันการพัฒนาระบบการรู้จำเสียงพูด ได้พัฒนากันมาอย่างต่อเนื่องซึ่งสามารถนำมาใช้งานได้จริง โดยในการใช้งานจริงนั้นสิ่งที่เราไม่สามารถมองผ่านไปได้คือสภาวะแวดล้อมซึ่งมีผลต่อโดยตรงกับระบบการรู้จำเสียงพูดอาจทำให้ประสิทธิภาพของการรู้จำเสียงพูดด้อยลง บทความนี้ได้เปรียบเทียบแนวทางในการหาค่าคาบเวลาพิทช์ในสภาวะที่มีสัญญาณรบกวนขาว โดยใช้หลักการออโตโครีเลชั่น คือวิธี Autocorrelation Method using Center Clipping (AUTOC) กับวิธี Average Magnitude Difference Function (AMDF) ค่าคาบเวลาพิทช์ที่ได้มาคือค่าความถี่มูลฐาน (F_0) จะนำมาผ่าน median filter เพื่อจะปรับปรุงให้มีความต่อเนื่องของข้อมูล จากนั้นนำมาทำการควอนไตซ์เบี่ยงเบนออกเป็น 3 ระดับของความถี่มูลฐาน (F_0) และค่าการเปลี่ยนแปลงนี้จะนำไปใช้เป็นข้อมูลสำหรับ Training ของ HMM เพื่อสร้างแบบจำลองของหน่วยเสียงวรรณยุกต์ภาษาไทยทั้ง 5 เสียง [1]

2. ขั้นตอนการวิเคราะห์เสียงวรรณยุกต์

ในการวิเคราะห์และสร้างแบบจำลองเสียงวรรณยุกต์ แบ่งออกได้เป็น 3 ขั้นตอนดังแสดงในรูปที่ 1

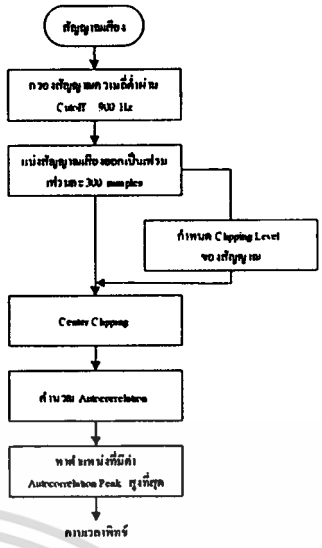


รูปที่ 1 แสดงขั้นตอนในการวิเคราะห์

2.1 การหาค่าพิทช์

2.1.1 Autocorrelation Method using Center Clipping (AUTOC) [1-2] มีขั้นตอนวิเคราะห์ดังรูปที่ 2

นำสัญญาณเสียงที่ได้จากการ sampling ที่ความถี่ 11.025 KHz มาผ่านตัวกรองสัญญาณความถี่ต่ำ 900 Hz เพื่อกำจัดความถี่ฮาร์โมนิกส์ที่ไม่ต้องการออกไป นำสัญญาณที่ผ่านตัวกรองสัญญาณความถี่ต่ำมาทำการแบ่งสัญญาณออกเป็นช่วงๆหรือเฟรม(frame) เพื่อที่จะคำนวณหาพารามิเตอร์ที่อยู่ในช่วงเวลานั้นๆออกมา โดยในแต่ละเฟรมกำหนดให้มีตัวอย่างสัญญาณ 300 samples การวิเคราะห์จะทำทีละเฟรม โดยกำหนดให้มีช่วงของการเลื่อนเฟรมครั้งละ 100 samples นั่นคือ แต่ละเฟรมจะมีช่วงของการซ้อนทับกัน 2 ใน 3 เฟรม จากนั้นเพื่อเป็นการกำจัดสัญญาณที่มีขนาดแอมปลิจูดต่ำออกไป จึงทำการ clip สัญญาณที่อยู่ในช่วง 65 เปอร์เซ็นต์ของ Absolute Amplitude Peak [2] โดยค่าเปอร์เซ็นต์ของการ clip สัญญาณที่กำหนดนี้จะกำจัดผลอันเนื่องจากการตอบสนองทางความถี่ของช่องทางเดินเสียงที่เกิดขึ้นทำให้สัญญาณที่ได้เหลืออยู่แต่สัญญาณที่มีค่าคาบพิทช์ ซึ่งจะสามารถคำนวณหา Clipping Level ได้จากสมการต่อไปนี้



รูปที่ 2 ขั้นตอนการวิเคราะห์เสียงวรรณยุกต์วิธี Autocorrelation Method using Center Clipping

$$C_L = (65\%) \times \min(K_1, K_2) \qquad (1)$$

โดยที่ C_L = Clipping Level

K_1 = Absolute Amplitude Peak ของ 100 samples แรก ของเฟรม

K_2 = Absolute Amplitude Peak ของ 100 samples ท้าย ของเฟรม

เมื่อได้ Clipping Level ของสัญญาณแล้ว สัญญาณที่มีค่าอยู่ในช่วงจะถูกกำหนด $\pm C_L$ ตามความสัมพันธ์ดังต่อไปนี้

$$y(n) = \mathrm{sgn}[x(n)] = \begin{cases} 1 & , x(n) \geq C_L \\ 0 & , |x(n)| < C_L \\ -1 & , x(n) \leq -C_L \end{cases} \qquad (2)$$

โดยที่ $\mathrm{sgn}[x(n)]$ คือ สัญญาณที่ผ่านการ clip จากนั้นนำสัญญาณที่ผ่านการ clip คือค่า y(n) มาคำนวณออโตโครีเลชั่น ตามสมการต่อไปนี้

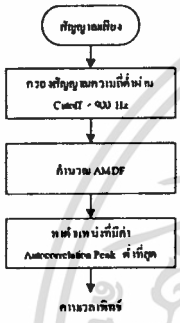
$$R(k) = \sum_{n=0}^{N-1-k} y(n) \times y(n+k) \qquad (3)$$

เมื่อ k = การเลื่อนไปของเวลา ในที่นี้ให้มีค่าเป็น 250

N = จำนวนตัวอย่างสัญญาณ(sample)ต่อเฟรม

จากคุณสมบัติของออโตโครีเลชั่นฟังก์ชั่น ถ้าสัญญาณมีความเป็นคาบที่ระยะ P ค่าที่มากที่สุดของ R(k) จะเกิดที่ตำแหน่ง $k = 0, \pm P, \pm 2P, \ldots$ จากนั้นทำการหาตำแหน่งที่มี Autocorrelation Peak สูงที่สุดเมื่อเทียบกับ R(0) ซึ่งระยะที่ได้ก็คือ คาบเวลาพิทช์

2.1.2 Average Magnitude Difference Function (AMDF) [2-4] มีขั้นตอนวิเคราะห์ดังรูปที่ 3



รูปที่ 3 ขั้นตอนการวิเคราะห์เสียงวรรณยุกต์วิธี Average Magnitude Difference Function

นำสัญญาณเสียงที่ได้จากการ sampling ที่ความถี่ 11.025 KHz มาผ่านตัวกรองสัญญาณความถี่ต่ำ 900 Hz เพื่อกำจัดความถี่ฮาร์โมนิกส์ที่ไม่ต้องการออกไป นำสัญญาณที่ผ่านตัวกรองสัญญาณความถี่ต่ำมาคำนวณออโตโครีเลชั่น ตามสมการต่อไปนี้

$$R(k) = \sum_{n=0}^{N-1-k} \left| y(n) - y(n+k) \right| \qquad (4)$$

เมื่อ k = การเลื่อนไปของเวลา ในที่นี้ให้มีค่าเป็น 250

N = จำนวนตัวอย่างสัญญาณ(sample)ต่อเฟรม

จากคุณสมบัติของออโตโครีเลชั่นฟังก์ชั่น ถ้าสัญญาณมีความเป็นคาบที่ระยะ P ค่าที่น้อยที่สุดของ R(k) จะเกิดที่ตำแหน่ง $k = 0, \pm P, \pm 2P, \ldots$ จากนั้นทำการหาตำแหน่งที่มี Autocorrelation Peak ต่ำที่สุดเมื่อเทียบกับ R(0) ซึ่งระยะที่ได้ก็คือ คาบเวลาพิทช์

จากค่าคาบเวลาพิทช์ที่ได้จาก 2 วิธี สามารถนำมาหาค่าความถี่มูลฐาน (F_0)ได้จากความสัมพันธ์ คือ

$$F_0 = \frac{F_s}{P} \qquad (5)$$

เมื่อ F_0 = ความถี่มูลฐาน (Hz)

F_s = ความถี่ที่ใช้ในการสุ่มสัญญาณ ในบทความนี้ใช้ 11.025 KHz

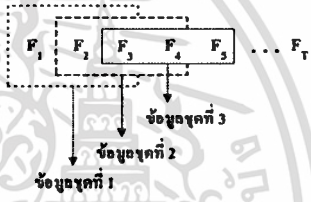
P = คาบเวลาพิทช์

2.2 การ Pre-process ข้อมูล

ขั้นตอนนี้ประกอบด้วยขั้นตอนย่อย 2 ขั้นตอน [1] คือ

2.2.1 median filter

นำค่าความถี่มูลฐาน (F_0) ที่คำนวณได้จากการหาค่าพิทช์มาทำการปรับปรุงให้มีความต่อเนื่องของข้อมูล โดยนำ median filter มาใช้ ทำการจัดเรียงค่าความถี่มูลฐานที่คำนวณได้ออกเป็นชุดข้อมูล โดยในแต่ละชุดข้อมูลประกอบด้วยค่าความถี่ 3 ค่า การเลื่อนของชุดข้อมูลแสดงได้ดังรูปที่ 4



รูปที่ 4 การจัดแบ่งความถี่มูลฐานออกเป็นชุดข้อมูล

จากนั้นนำค่าความถี่ 3 ค่า ในแต่ละชุดข้อมูลมาจัดเรียงใหม่ตามความสัมพันธ์

$$a \leq b \leq c \qquad (6)$$

โดยที่

a = ความถี่ F_0 ที่มีค่าน้อยที่สุดของแต่ละชุดข้อมูล

b = ความถี่ F_0 ที่มีค่าอยู่ระหว่างกลาง

c = ความถี่ F_0 ที่มีค่ามากที่สุดของแต่ละชุดข้อมูล

นำความถี่ค่ากลาง (b) ที่ได้จากชุดข้อมูลแต่ละชุดมาจัดเรียงตามลำดับ ก็จะได้ความถี่มูลฐาน (F_0) ชุดใหม่ที่ผ่านกระบวนการ median filter

2.2.2 ทำการ Quantized ทิศทางการเปลี่ยนแปลงของค่าความถี่มูลฐาน (F_0)

ทำการจัดกลุ่มค่าความถี่มูลฐานออกเป็น 3 ระดับตามแนวทางการเปลี่ยนแปลงของความถี่ (ΔF) ที่เพิ่มขึ้นหรือลดลงเมื่อเวลาเปลี่ยนไป

$$\Delta F_t = F_{t+1} - F_t \qquad (7)$$

เมื่อ t = 1, 2 ,...,(T-1) ; T = จำนวนเฟรม

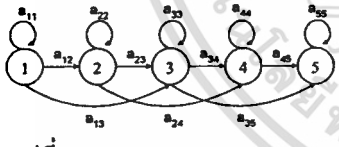
F_t = ความถี่ F_0 ที่เวลา t

$$V_t = \begin{cases} 1 & ; \Delta F_t > 0 \\ 0 & ; \Delta F_t = 0 \\ -1 & ; \Delta F_t < 0 \end{cases} \qquad (8)$$

จากนั้นค่า V_t จะถูกนำไปใช้เป็นข้อมูล Training เพื่อสร้างแบบจำลองของเสียงวรรณยุกต์ภาษาไทยต่อไป

2.3 การสร้างแบบจำลองเสียงวรรณยุกต์โดยใช้เทคนิค HMM

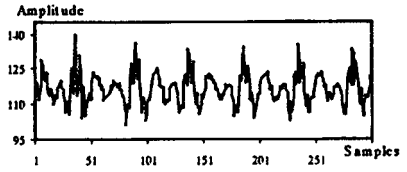
ด้วยเทคนิค HMM นี้ การสร้างแบบจำลองเสียงพูดสามารถทำได้โดยใช้จำนวนสเตทที่แตกต่างกัน จำนวนหนึ่งคำศัพท์แต่ละคำที่ใช้ในการรู้จำจะถูกสร้างใหม่ โดยเปลี่ยนจากรูปของการเปลี่ยนแปลงความถี่ให้อยู่ในรูปของการจัดเรียงตัวของสเตท โดยมีการย้ายจากสเตทเริ่มต้นไปยังสเตทถัดไปตามการเลื่อนไปของ discrete time ด้วยเซตของความน่าจะเป็นที่เกี่ยวข้องกับสเตทนั้นๆจนกระทั่งได้ output ที่มีลักษณะเป็นการเรียงตัวกันของสเตทที่ใช้แทนรูปแบบของคำนั้นๆ ในบทความนี้เลือกใช้ HMM แบบ Left-Right Model ที่ประกอบไปด้วยสเตททั้งหมด 5 สเตท [1] โดยมีรูปแบบในการย้ายสเตทที่สามารถเป็นไปได้ดังรูปที่ 5



รูปที่ 5 Left-Right Model 5 state

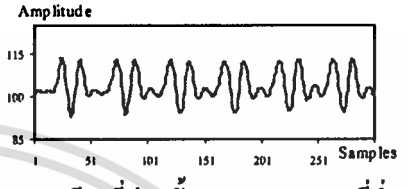
3. ผลการทดลอง

ทำการวิเคราะห์เสียงคำพยางค์เดียวจำนวน 25 คำ จากคำว่า อา อี อู เอ และ โอ ทั้ง 5 ระดับวรรณยุกต์ของเพศหญิง 5 คนและเพศชาย 5 คน แล้วนำลักษณะของการเปลี่ยนแปลงความถี่ของวรรณยุกต์ทั้ง 5 เสียงมาสร้างแบบจำลอง รูปที่ 6 แสดงลักษณะสัญญาณเสียงที่ได้จากการ sampling เสียง "เอ" ของผู้หญิง โดยในบทความนี้ใช้ความถี่ sampling ที่ 11.025 KHz จะเห็นว่าสัญญาณที่ได้จากการ sampling จะประกอบด้วยความถี่จำนวนมาก



รูปที่ 6 สัญญาณเสียงที่ได้จากการ sampling ที่ 11.025 KHz

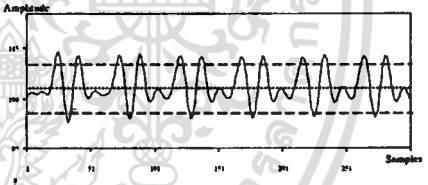
เมื่อนำสัญญาณที่เสียงที่ได้จากการ sampling มาผ่านขั้นตอนกรองความถี่ต่ำ 900 Hz จะได้สัญญาณที่มีลักษณะเรียบขึ้นดังรูปที่ 7



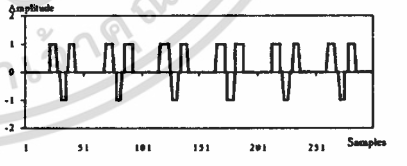
รูปที่ 7 สัญญาณเสียงที่ผ่านขั้นตอนกรองความถี่ต่ำ 900 Hz

3.1 Autocorrelation Method using Center Clipping

รูปที่ 8 แสดงเส้นประ Clipping Level ของสัญญาณ ซึ่งในบทความนี้ใช้ Clipping Level ของสัญญาณเป็น 65%ของ Absolute Amplitude Peak ในแต่ละเฟรม ดังกล่าวมาแล้วข้างต้น สัญญาณที่ได้จากการ Clip จะเป็นดังรูปที่ 9

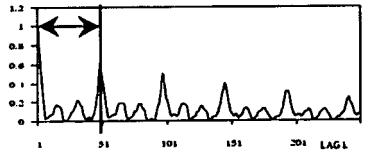


รูปที่ 8 แสดง Clipping Level 65 % ของสัญญาณที่ผ่านขั้นตอนกรองความถี่ต่ำ 900 Hz



รูปที่ 9 สัญญาณเสียงที่ผ่านขั้นตอนการ Clip

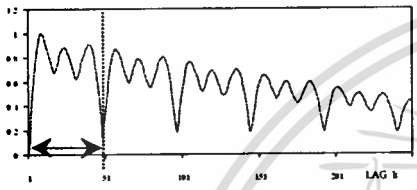
นำสัญญาณเสียงที่ผ่านขั้นตอนการ Clipping Level ไปทำการคำนวณออโตโครีเลชั่น ตามสมการที่ (3) จะได้สัญญาณที่มีลักษณะดังรูปที่ 10 ระยะห่างระหว่าง R(0) กับจุดยอดที่สูงที่สุดถัดไปก็คือคาบพิทช์ ซึ่งสามารถนำมาหาค่าความถี่มูลฐานได้ตามสมการที่ (5) จากรูปที่ 10 ค่าความถี่มูลฐานเท่ากับ 225 Hz (11025 / 49 = 225)



รูปที่ 10 สัญญาณที่ผ่านการคำนวณ Normalized Autocorrelation Method using Center Clipping

3.2 Average Magnitude Difference Function

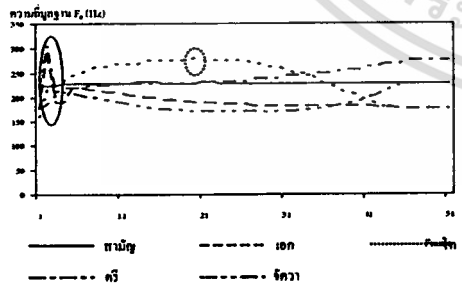
นำสัญญาณเสียงที่ผ่านขั้นตอนกรองความถี่ต่ำ 900 Hz ไปคำนวณออโตโครีเลชั่น ตามสมการที่ (4) จะได้สัญญาณที่มีลักษณะดังรูปที่ 11



รูปที่ 11 สัญญาณที่ผ่านการคำนวณ Normalized Average Magnitude Difference Function

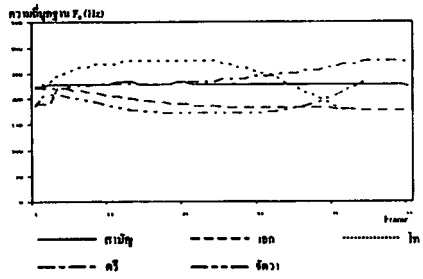
ระยะห่างระหว่าง R(0) กับจุดยอดที่ต่ำสุดถัดไปก็คือค่าคาบเวลาพิทช์ ซึ่งสามารถนำมาหาค่าความถี่มูลฐาน (F_0) ตามสมการ (5) จากรูปที่ 11 ได้ค่าความถี่มูลฐานเท่ากับ 225 Hz (11025 / 49 = 225)

จากวิธีการหาค่าคาบเวลาพิทช์ทั้ง 2 วิธี ลักษณะของการเปลี่ยนแปลงความถี่มูลฐาน (F_0) เทียบกับเวลาของวรรณยุกต์ภาษาไทยทั้ง 5 เสียง จะมีรูปแบบการเปลี่ยนแปลงที่แตกต่างกันไปดังรูปที่ 12



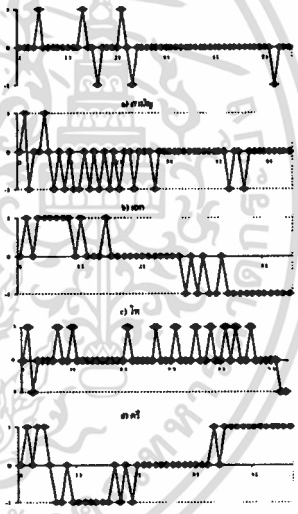
รูปที่ 12 ลักษณะของการเปลี่ยนแปลงของความถี่มูลฐาน (F_0)ในวรรณยุกต์ภาษาไทยทั้ง 5 เสียงของผู้พูดเพศหญิง

จากรูปที่ 12 จะเห็นว่าในวงกลมเส้นประ ค่าความถี่มูลฐาน (F_0) ไม่มีความเรียบ ดังนั้นจึงนำค่าความถี่มูลฐาน (F_0) มาผ่าน median filter ในรูปที่ 13 จะสังเกตเห็นว่าค่าความถี่มูลฐาน (F_0) จะถูกปรับให้มีความเรียบขึ้น



รูปที่ 13 ความถี่มูลฐาน (F_0)ในวรรณยุกต์ภาษาไทยทั้ง 5 เสียงที่ผ่าน median filter

นำความถี่มูลฐาน (F_0) ในวรรณยุกต์ภาษาไทยทั้ง 5 เสียงมาทำการควอนไตซ์ ตามทิศทางการเปลี่ยนแปลงของความถี่มูลฐาน (F_0) ออกเป็น 3 ระดับ {-1,0,1} เพื่อจะนำไปใช้เป็นข้อมูล Training เพื่อสร้างแบบจำลองของเสียงวรรณยุกต์ภาษาไทยทั้ง 5 เสียง ดังรูปที่ 14



รูปที่ 14 เสียงวรรณยุกต์ภาษาไทยทั้ง 5 เสียงที่ผ่านการควอนไตซ์

นำเสียงวรรณยุกต์ภาษาไทยทั้ง 5 เสียงที่ผ่านขั้นตอนการควอนไตซ์ เข้าสู่ขั้นตอน HMM เพื่อสร้างเป็นแบบจำลองของหน่วยเสียงวรรณยุกต์ภาษาไทยทั้ง 5 เสียง โดยทำการทดสอบการรู้จำออกเป็น 3 กรณีคือ กรณีที่ 1 และกรณีที่ 2 ใช้เสียงต้นแบบเป็นผู้หญิง 5 คนและผู้ชาย 5 คน ออกเสียง อา อี อู เอ โอ ทั้ง 5 ระดับ คนละ 1 ครั้ง และกรณีที่ 3 ใช้เสียงต้นแบบเป็นผู้หญิง 3 คนและผู้ชาย 2 คน ออกเสียง อา อี อู เอ โอ ทั้ง 5 ระดับ คนละ 1 ครั้ง จากนั้นทำการทดสอบการรู้จำเทียบกับเสียงต้นแบบจำนวน 125 เสียงทั้ง 3 กรณีและทำการสอดแทรกสัญญาณรบกวนขาว

ที่ระดับสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio : SNR) ที่ระดับ 80dB 60dB 40dB 20dB 15dB และ 10dB ดังแสดงในตารางต่อไปนี้

ตารางที่ 1 กรณีที่ 1 เสียงผู้หญิง 5 คน

SNR	จำนวนเสียงที่ถูกต้อง		ความถูกต้อง(%)	
	AUTOC	AMDF	AUTOC	AMDF
Original	121	120	96.80	96.00
80 dB	119	119	95.20	95.20
60 dB	119	118	95.20	94.40
40 dB	118	116	94.40	92.80
20 dB	116	104	92.80	83.20
15 dB	115	98	92.00	78.40
10 dB	107	95	85.60	76.00

ตารางที่ 2 กรณีที่ 2 เสียงผู้ชาย 5 คน

SNR	จำนวนเสียงที่ถูกต้อง		ความถูกต้อง (%)	
	AUTOC	AMDF	AUTOC	AMDF
Original	119	118	95.20	94.40
80 dB	118	118	94.40	94.40
60 dB	118	116	94.40	92.80
40 dB	117	114	93.60	91.20
20 dB	115	104	92.00	83.20
15 dB	112	99	89.60	79.20
10 dB	106	95	84.80	76.00

ตารางที่ 3 กรณีที่ 3 เสียงผู้หญิง 3 คนและผู้ชาย 2 คน

SNR	จำนวนเสียงที่ถูกต้อง		ความถูกต้อง (%)	
	AUTOC	AMDF	AUTOC	AMDF
Original	121	121	96.80	96.80
80 dB	120	121	96.00	96.80
60 dB	120	119	96.00	95.20
40 dB	119	117	95.20	93.60
20 dB	117	108	93.60	86.40
15 dB	115	102	92.00	81.60
10 dB	110	98	88.00	78.40

กรณีที่ 1 เสียงผู้หญิง 5 คน วิธี AUTOC สามารถให้ผลการรู้จำแม่นยำได้ถึง 92 % ที่ระดับ SNR 15 dB. แต่วิธี AMDF สามารถให้ผลในการรู้จำได้เพียง 78.40 %

กรณีที่ 2 เสียงผู้ชาย 5 คน วิธี AUTOC สามารถให้ผลการรู้จำแม่นยำได้ถึง 92 % ที่ระดับ SNR 20 dB. แต่วิธี AMDF สามารถให้ผลในการรู้จำได้เพียง 83.20 %

กรณีที่ 3 เสียงผู้หญิง 3 คนและผู้ชาย 2 คน วิธี AUTOC สามารถให้ผลการรู้จำแม่นยำได้ถึง 92 % ที่ระดับ SNR 15 dB. แต่วิธี AMDF สามารถให้ผลในการรู้จำได้เพียง 81.60 %

4. สรุป

จากการทดสอบการรู้จำพบว่าวิธี Pitch Extraction แบบ Autocorrelation Method using Center Clipping เมื่อนำมาหาค่าคาบเวลาพิทช์ในสภาวะที่มีสัญญาณรบกวนกับแบบจำลองเสียงวรรณยุกต์ภาษาไทย สามารถทนทานต่อสัญญาณรบกวนได้ถึง 20 dB. โดยมีความแม่นยำในการรู้จำเฉลี่ยมากกว่า 90 % ในทุกกรณีของกลุ่มทดสอบแต่แบบ Average Magnitude Difference Function สามารถทนทานต่อสัญญาณรบกวนได้เพียง 40 dB. มีความแม่นยำในการรู้จำเฉลี่ยมากกว่า 90 % และจากผลการทดสอบจะพบอีกว่าในกรณีที่ 1 และกรณีที่ 3 วิธีแบบ Autocorrelation Method using Center Clipping สามารถทนทานต่อสัญญาณรบกวนได้ถึง 15 dB. มีความแม่นยำในการรู้จำเฉลี่ยมากกว่า 90 % ซึ่งจะเห็นว่าวิธีเซ็นเตอร์คลิปปิ้งเป็นวิธีที่ทำให้สเปคตรัมราบเรียบสามารถลดสัญญาณรบกวนได้ทำให้ผลการรู้จำมีประสิทธิภาพที่ดี

5. เอกสารอ้างอิง

[1] K.Songwattana, I.Arungsrisangchai, C.Charumit, "Tone Recognition Model for Thai language using Pitch Quantization and Hidden Makov Modeling Techniques," National Conference on Computer Technology and Computer Engineering , Kasetsart University , Thailand, 1998.

[2] L.R. Rabiner, R.W. Schafer,"Digital Processing of Speech Signals,"New Jersey : Prentice-Hall, 1978.

[3] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," IEEE Trans.Acoust., Speech, Signal Processing, Vol.ASSP-22, pp.353-362, Oct.1974

[4] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms,"IEEE Trans. Acoust., Speech, and Signal Proc., Vol.ASSP-24, No. 5, pp.399-418, October 1976.