บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทนี้จะนำเสนอทฤษฎีพื้นฐานและแนวคิดที่เกี่ยวข้องกับการพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์และการพัฒนาระบบรู้จำเสียงพูด โดยเริ่มจากทฤษฎีทางภาษาศาสตร์ ซึ่งเป็นทฤษฎีที่ศึกษาปรากฏการณ์ของเสียงพูดในด้านต่างๆดังนี้

- สรีรสัทศาสตร์ (Articulatory Phonetics)
- สวนสัทศาสตร์ (Acoustic Phonetics)

จากนั้นจะนำเสนอทฤษฎีการวิเคราะห์เสียงพูดในโดเมนความถี่ด้วยการแปลงฟูริเยร์แบบวิยุต (Discrete Fourier Transform) และสเปกโตรแกรม (Spectrogram) ถัดจากนั้นจะนำเสนอการสกัดลักษณะสำคัญ แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model) การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ และ ซัพพอร์ตเวกเตอร์แมชชีน - เอสวีเอ็ม (Support Vector Machine - SVM) ตามลำดับ รวมทั้งเสนองานวิจัยต่างๆที่เกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์

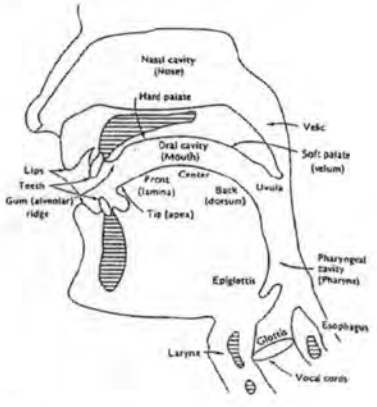
ทฤษฎีที่เกี่ยวข้อง

1. สรีรสัทศาสตร์

สรีรสัทศาสตร์เป็นการศึกษาเสียงพูดจากอวัยวะและการเคลื่อนไหวของอวัยวะที่ทำให้เกิดเสียงพูด โดยในหัวข้อนี้จะนำเสนอเกี่ยวกับอวัยวะที่ทำให้เกิดเสียง และเสียงในภาษาไทยซึ่งได้แก่ เสียงพยัญชนะ เสียงสระ และเสียงวรรณยุกต์ ตามลำดับ

1.1 อวัยวะที่ทำให้เกิดเสียง (The Organs of Speech) [5]

ตำแหน่งของอวัยวะที่ทำให้เกิดเสียงของมนุษย์สามารถแสดงได้ดังรูปที่ 2.1

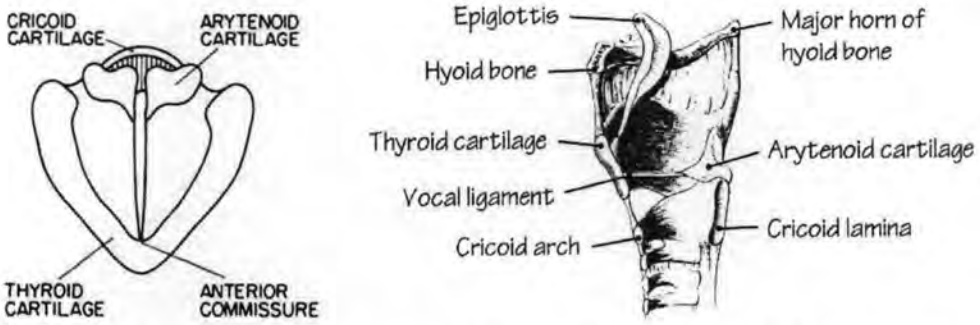


รูปที่ 2.1 อวัยวะภายในของระบบการพูดของมนุษย์ [6]

อวัยวะที่ใช้ในการออกเสียงทั้งหมดมีดังนี้

1. **ริมฝีปาก** (Lips) เป็นอวัยวะส่วนที่สามารถเคลื่อนไหวได้และทำให้เสียงแตกต่างกันได้มาก เราอาจจะบังคับริมฝีปากให้ปิดสนิท ให้เปิดเล็กน้อย ให้เปิดกว้างขึ้น ให้ยื่นออกมา ให้ห่อกลมหรือทำเป็นรูปรีก็ได้ ลักษณะต่างๆ ของริมฝีปากล้วนมีผลต่อการออกเสียงทั้งสิ้น เสียงพยัญชนะที่เกิดจากการกักที่ริมฝีปากเรียกว่าเสียงโอฐชะ (Bilabial Sound)
2. **ฟัน** (Teeth) เป็นอวัยวะที่เป็นฐานหรือตำแหน่งที่เกิดของเสียงหลายชนิด เช่น เมื่อฟันบนกดลงบนริมฝีปากล่าง ลมที่ผ่านออกมาโดยแรงจะลอดช่องที่พอจะผ่านได้ออกมา ทำให้เกิดเป็นเสียงชนิดที่เรียกว่า เสียงเสียดแทรกที่เกิดระหว่างฟันกับริมฝีปาก ถ้าฟันบนกดกับฟันล่าง ลมที่ผ่านออกมาโดยแรงจะทำให้ได้เสียงเสียดแทรกที่เกิดที่ฟัน เป็นต้น นอกจากนี้เนื่องจากปลายลิ้นอยู่ใกล้กับฟัน ปลายลิ้นจึงมักจะทำอาการต่างๆ บริเวณฟันและหลังฟันบ่อยๆ ทำให้เกิดเสียงทันตชะ (Dental Sound)
3. **ปุ่มเหงือก** (Alveolus, Gum Ridge, Tooth Ridge) เป็นส่วนที่นูนออกมาตรงบริเวณหลังฟันด้านบน ถ้าเอาลิ้นแตะดูจะรู้สึกว่ามีลักษณะเป็นคลื่น ลิ้นอาจแตะหรือวางอยู่ใกล้บริเวณปุ่มเหงือก ซึ่งทำให้เกิดเสียงมุทธชะ (Alveolar Sound)
4. **เพดานแข็ง หรือเพดานปาก** (Palate, Hard Palate) หมายถึงส่วนโค้งของเพดานปากส่วนที่เป็นกระดูกแข็ง ซึ่งอยู่ถัดจากปุ่มเหงือกเข้ามา ถ้าลิ้นแตะหรือวางใกล้เพดานแข็งจะทำให้เกิดเสียงตาลุชะ (Palatal Sound)
5. **เพดานอ่อน** (Velum, Soft Palate) คือ ส่วนของเพดานที่อยู่ต่อเพดานแข็งเข้าไปข้างในเป็นกระดูกอ่อนที่ขยับขึ้นลงได้เล็กน้อย เวลาหายใจเพดานอ่อนและลิ้นไก่ซึ่งอยู่ปลายเพดานอ่อนจะลดระดับลงมาเปิดช่องให้ลมออกทางจมูก ฉะนั้นเวลาที่ไม่พูด เพดานอ่อนและลิ้นไก่จะลดระดับลงมา เวลาพูดส่วนใหญ่เพดานอ่อนและลิ้นไก่จะถูกยกขึ้นไปจดกับผนังคอ จะมีแต่เวลาออกเสียงนาสิกเท่านั้นที่เพดานอ่อนจะลดระดับลงมาเพื่อให้ลมออกไปทางจมูกได้ ถ้าลิ้นแตะหรือวางใกล้เพดานอ่อนจะทำให้เกิดเสียงที่เกิดที่เพดานอ่อน (Velar Sound)
6. **ลิ้นไก่** (Uvula) เป็นก้อนเนื้อเล็กๆ อยู่ต่อจากปลายเพดานอ่อนเข้าไปข้างใน และห้อยอยู่ตรงกลางปาก สามารถสั่นรัวได้ เวลาอ้าปากมักจะเห็น ลิ้นไก่ใช้ออกเสียงในบางภาษา เช่น ภาษาเยอรมัน ฝรั่งเศส นอร์เวย์ อาหรับ และอิสราเอล เป็นต้น
7. **ช่องจมูก** (Nasal Cavity) หมายถึง โพรงในช่องจมูกซึ่งอยู่เหนือลิ้นไก่ขึ้นไป เป็นช่องที่ลมซึ่งผ่านเส้นเสียงขึ้นมาจะผ่านออกไปทางจมูกได้เมื่อเวลาหายใจและเวลาออกเสียงนาสิก ในเวลาเปล่งเสียงอื่นๆลิ้นไก่จะถูกยกขึ้นไปปิดช่องจมูกเพื่อให้ลมออกมาทางช่องปาก

8. **ลิ้น** (Tongue) เป็นส่วนที่เคลื่อนไหวได้มากที่สุดในการออกเสียงพูด ส่วนที่เคลื่อนไหวของลิ้นแต่ละส่วนมีผลต่อการออกเสียง เราจึงแบ่งลิ้นออกเป็น 3 ส่วนด้วยกันตามหน้าที่ที่มีในการออกเสียงคือ
   1. ปลายลิ้น (Tip of the Tongue) หรือ ลิ้นส่วนปลายสุด หมายถึงส่วนปลายของลิ้นซึ่งสามารถจะยกขึ้นไปแตะอวัยวะส่วนต่างๆ ในปากตอนบนได้โดยง่าย
   2. หน้าลิ้น (Blade of the Tongue) หรือ ลิ้นส่วนหน้า หมายถึงลิ้นส่วนที่อยู่ตรงข้ามกับเพดานแข็ง ในขณะที่วางลิ้นราบกับปากตอนไม่ได้พูด
   3. หลังลิ้น (Back of the Tongue) หรือ ลิ้นส่วนหลัง หมายถึงส่วนของลิ้นที่อยู่ตรงข้ามกับเพดานอ่อน ในขณะที่วางลิ้นราบกับปากตอนไม่ได้พูด
9. **แผ่นเนื้อปากหลอดลม** (Epiglottis) หรือ ลิ้นปิดกล่องเสียง เป็นก้อนเนื้อเล็กๆ คล้ายลิ้นไก่อยู่ต่อโคนลิ้นลงไปในคอ มีหน้าที่ปิดเปิดช่องหลอดลม เพื่อป้องกันมิให้อาหารตกลงไปในหลอดลม ในเวลาที่กลืนอาหาร แผ่นเนื้อปากหลอดลมปิดลงให้อาหารผ่านไปลงหลอดอาหาร แต่ในเวลาที่พูด แผ่นเนื้อนี้จะเปิดออกเพื่อให้ลมจากหลอดลมออกมา
10. **โพรงคอ** (Pharynx) เป็นโพรงซึ่งอยู่ถัดปากลงไปจากช่องปากจนถึงเส้นเสียงหรือสายเสียง
11. **เส้นเสียง หรือสายเสียง** (Vocal Cords) เป็นอวัยวะสำคัญที่ทำให้เกิดเสียง เส้นเสียงประกอบด้วยเส้นเอ็นและกล้ามเนื้อเป็นแผ่น 2 แผ่น มีความยาวประมาณ 1.2-1.7 เซนติเมตร กว้างประมาณ 0.2-0.3 เซนติเมตร ปิดขวางอยู่ตรงปากของช่องหลอดลม โดยจะวางตัวจากด้านหลังมายังด้านหน้าอยู่ตรงกลางของกล่องเสียง เส้นเสียงทั้งสองสามารถที่จะดึงออกให้ห่างจากกันหรือดึงเข้ามาให้ชิดกันก็ได้ เส้นเสียงเป็นส่วนสำคัญที่ทำให้เกิดเสียงพูด โดยจะเปิดให้ลมผ่านในเวลาหายใจตามปกติ แต่จะอยู่ชิดกันเมื่อมีการเปล่งเสียง
12. **กล่องเสียง** (Larynx) ตั้งอยู่ตอนบนของหลอดลมตรงตำแหน่งที่เรียกว่าลูกกระเดือก (Adam's Apple) กล่องเสียงประกอบด้วยกระดูกอ่อนหลายส่วนด้วยกัน ส่วนที่อยู่ด้านหน้า คือ กระดูกอ่อนไทรอยด์ (Thyroid Cartilage) ปลายด้านหนึ่งของเส้นเสียงทั้งสองจะเชื่อมอยู่กับกระดูกอ่อนไทรอยด์นี้และอยู่ชิดกัน ส่วนปลายอีกด้านหนึ่งของเส้นเสียงทั้งสอง จะเชื่อมอยู่กับกระดูกอ่อนอาริตินอยด์ (Arytenoids Cartilages) ซึ่งเป็นกระดูกอ่อนอีกสองชิ้น กระดูกอ่อนอาริตินอยด์และกล้ามเนื้อในกล่องเสียงจะทำให้เส้นเสียงทั้งสองอยู่ชิดติดกันหรือห่างจากกันได้ เมื่อเส้นเสียงอยู่ห่างจากกันจะเกิดเป็นช่องสามเหลี่ยม ซึ่งเป็นทางให้ลมผ่านเข้าไปถึงปอด หรือผ่านออกมาจากปอดได้ ดังรูปที่ 2.2



รูปที่ 2.2 กล่องเสียง (ซ้าย : กล่องเสียงด้านหน้า, ขวา : กล่องเสียงด้านหลัง) [7]

13. **ช่องระหว่างเส้นเสียง** (Glottis) จะเปิดอยู่ระหว่างที่หายใจเข้าออกตามปกติ แต่จะปิดลงเมื่อมีการเปล่งเสียง ก่อให้เกิดการสั่น และเป็นเสียงดังขึ้น
14. **ช่องปาก** (Oral Cavity) ทำหน้าที่เป็นช่องกำทอน (Resonant Chamber) ซึ่งสามารถเปลี่ยนให้มีรูปร่างต่างๆ กัน ตามท่าทางของอวัยวะภายในช่องปาก โดยอวัยวะภายในช่องปากอาจสามารถแบ่งได้ดังนี้
    1. อวัยวะส่วนกระทำอาการ (Articulator) หมายถึงอวัยวะส่วนที่เคลื่อนไหวเพื่อผลักหรือกักลมในที่ต่างๆ อวัยวะส่วนกระทำอาการที่สำคัญคือลิ้น ซึ่งเคลื่อนไหวได้มากที่สุด อวัยวะส่วนกระทำอาการอาจเรียกว่า "กรณ์"
    2. อวัยวะส่วนเกิดอาการ (Point of Articulation) หมายถึง ตำแหน่งที่อวัยวะส่วนกระทำอาการเคลื่อนไหวไป เพื่อผลักหรือกักลมไว้ อาจเรียกอวัยวะส่วนนี้ว่า "ฐาน" ที่เกิดของหน่วยเสียงต่างๆ ฐานภายในช่องปากที่สำคัญได้แก่ ริมฝีปาก ฟัน ปุ่มเหงือก เพดานแข็ง และเพดานอ่อน
15. **หลอดลม** (Trachea) เป็นทางเดินอากาศจากปอดถึงกล่องเสียง

1.2 เสียงพยัญชนะในภาษาไทย (Thai Consonants) [5]

1.2.1 เสียงพยัญชนะ

เสียงพยัญชนะ (Consonant) หมายถึงเสียงของลมที่ผ่านปอดขึ้นมายังกล่องเสียงแล้วปะทะกับอวัยวะต่างๆ ในช่องปาก ทำให้ลมเพียงส่วนหนึ่งหรือทั้งหมดพบกับอุปสรรคที่อยู่เหนือช่องของเส้นเสียง โดยอุปสรรคเหล่านี้เกิดจากการทำงานประสานกันของอวัยวะในช่องปาก เสียงพยัญชนะที่เกิดขึ้นมาจึงมีหลายแบบแตกต่างกัน ซึ่งเสียงที่แตกต่างกันมักจะทำให้ความหมายของเสียงในภาษาแตกต่างกันไปด้วย คุณสมบัติที่ทำให้เสียงพยัญชนะแตกต่างกันมีดังนี้

1. **คุณสมบัติความก้องของเสียง** ความก้องของเสียงเป็นคุณสมบัติที่ใช้ในการแบ่งแยกเสียงพยัญชนะออกได้เป็นสองชนิด คือ

1. เสียงพยัญชนะโฆษะ (Voiced) หรือเสียงก้อง เป็นเสียงพยัญชนะที่เส้นเสียงสั่นสะเทือนขณะที่เปล่งเสียง
2. เสียงพยัญชนะอโฆษะ (Voiceless) หรือเสียงไม่ก้อง เป็นเสียงพยัญชนะที่เส้นเสียงไม่สั่นสะเทือนขณะที่เปล่งเสียง

2. **ลักษณะของลมที่ผ่านเส้นเสียง** เสียงพยัญชนะสามารถแบ่งตามลักษณะลมที่ผ่านเส้นเสียงออกมาได้ดังนี้
   1. เสียงพยัญชนะหยุด (Stop) อาจแบ่งออกเป็น 2 ลักษณะย่อยๆ ได้แก่ เสียงพยัญชนะระเบิด (Plosive Stop) และเสียงพยัญชนะกัก (Unreleased Stop) เสียงพยัญชนะระเบิดเกิดจากการที่ลมซึ่งเปล่งออกมาถูกกักเอาไว้ ณ ที่ใดที่หนึ่งในช่องปาก แล้วช่องที่กักนั้นเปิดให้ลมพุ่งออกมา เสียงพยัญชนะระเบิดแบ่งออกได้อีกเป็นเสียงพยัญชนะระเบิดมีลม (Aspirated Plosive) หรือธนิต ซึ่งจะมีลมหายใจพ่นออกมาหลังเปล่งเสียงและเสียงพยัญชนะระเบิดไม่มีลม (Unaspirated Plosive) หรือสิถิล ซึ่งไม่มีลมหายใจพ่นออกมา ส่วนเสียงพยัญชนะกักเกิดจากลมซึ่งเปล่งออกมาถูกกักไว้ ณ ที่ใดที่หนึ่งในช่องปาก โดยเสียงพยัญชนะกักนี้มักจะเป็นเสียงตัวสะกดท้ายพยางค์
   2. เสียงพยัญชนะเสียดแทรก (Fricative) เป็นเสียงพยัญชนะที่เมื่อออกเสียงแล้วลมที่ผ่านขึ้นมาถูกบังคับให้ต้องบีบตัวผ่านช่องแคบๆ ที่ใดที่หนึ่งในช่องปาก ซึ่งเสียงเสียดแทรกนี้เราจะทำค้างไว้นานเท่าใดก็ได้ ตราบเท่าที่ลมหายใจจะอำนวย
   3. เสียงพยัญชนะนาสิก (Nasal) เป็นเสียงพยัญชนะที่มีลมผ่านออกมาทางจมูก ซึ่งเกิดจากการที่ลมมากักอยู่ในช่องปาก แล้วเพดานอ่อนและลิ้นไก่ลดระดับลง
   4. เสียงพยัญชนะข้างลิ้น (Lateral) เป็นเสียงที่เกิดจากการนำลิ้นปิดบริเวณปุ่มเหงือกและเพดานแข็งส่วนกลางไว้ แล้วปล่อยให้ลมผ่านออกมาทางข้างลิ้น
   5. เสียงพยัญชนะรัว (Trill) เกิดจากการที่อวัยวะส่วนใดส่วนหนึ่งในช่องปากกระทบกับอวัยวะอีกส่วนหนึ่งในขณะที่ลมถูกพ่นผ่านอวัยวะนั้นออกมาอย่างรุนแรง
   6. เสียงพยัญชนะกึ่งสระ (Semi-vowel, Approximant) หรืออรรธสระ เป็นเสียงเลื่อน (Glide) ที่เกิดขึ้นระหว่างเสียงสระสองเสียง ในการเปล่งเสียงพยัญชนะกึ่งสระอวัยวะที่ใช้ในการออกเสียงจะอยู่ในตำแหน่งของการออกเสียงสระใดสระหนึ่งก่อน แล้วจึงเปล่งเสียงออกมาขณะที่เปลี่ยนตำแหน่งอวัยวะไปสู่การออกเสียงของอีกสระหนึ่ง

3. **ฐานที่เกิดของเสียง** ไม่ว่าลมที่ใช้ในการออกเสียงพยัญชนะนั้นจะมาถูกกัก ถูกดัดแปลงจนเกิดการกัก หรือการเสียดแทรก จำเป็นต้องมีตำแหน่งที่เกิดอยู่ด้วยเสมอในช่องปาก

1.2.2 เสียงพยัญชนะภาษาไทย

พยัญชนะในภาษาไทยมีทั้งหมด 44 รูป 21 หน่วยเสียง แบ่งเป็น 2 กลุ่มใหญ่ๆ คือ พยัญชนะกัก (Stop Consonants) 11 หน่วยเสียง และพยัญชนะไม่กัก (Non-stop Consonants) 10 หน่วยเสียง ดังแสดงในตารางที่ 2.1 ทั้งนี้หน่วยเสียงพยัญชนะทั้ง 21 หน่วยเสียง สามารถที่จะอยู่ในต้นพยางค์ได้ทุกหน่วยเสียง แต่จะมีหน่วยเสียงพยัญชนะที่ปรากฏท้ายพยางค์ได้เพียง 9 หน่วยเสียงเท่านั้น คือ เสียงพยัญชนะกัก 4 หน่วยเสียง (/p/, /t/, /k/, /ʔ/) เสียงพยัญชนะนาสิก 3 หน่วยเสียง (/m/, /n/, /ŋ/) และเสียงพยัญชนะกึ่งสระ 2 หน่วยเสียง (/w/, /j/) ส่วนพยัญชนะต้นควบกล้ำในภาษาไทยแท้เป็นได้ 11 หน่วยเสียง คือ /pr/, /pʰr/, /pl/, /pʰl/, /tr/, /kr/, /kʰr/, /kl/, /kʰl/, /kw/, /kʰw/ ส่วนพยัญชนะต้นควบกล้ำในภาษาไทยทับศัพท์อังกฤษมีได้ 6 หน่วยเสียง คือ /br/, /bl/, /dr/, /fr/, /fl/, /tr/ ส่วนคำไทยที่ยืมมาจากภาษาสันสกฤตก็ควบ /tr/ ได้เช่นกัน

ตารางที่ 2.1 เสียงพยัญชนะภาษาไทย

1 (*) ปรากฏท้ายพยางค์ได้

2 (/.../) ปรากฏเฉพาะในคำไทยทับศัพท์ภาษาอังกฤษ

3 [/.../] ปรากฏในคำไทยทับศัพท์ภาษาอังกฤษ หรือคำไทยที่ยืมมาจากภาษาสันสกฤต

หน่วยเสียง[1]	หน่วยเสียงควบกล้ำ	ลักษณะของลม	การพ่นลม	ความก้อง	ฐานที่เกิด	รูปพยัญชนะ
/p/ (*)	/pr/, /pl/	กัก	ไม่พ่นลม	ไม่ก้อง	ริมฝีปาก	ป
/pʰ/	/pʰr/,/pʰl/	กัก	พ่นลม	ไม่ก้อง	ริมฝีปาก	ผ พ ภ
/b/	/br/,/bl/	กัก	ไม่พ่นลม	ก้อง	ริมฝีปาก	บ
/t/ (*)	/tr/	กัก	ไม่พ่นลม	ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ฏ ต
/tʰ/	/tʰr/	กัก	พ่นลม	ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ฐ ฑ ฒ ถ ท ธ
/d/	/dr/	กัก	ไม่พ่นลม	ก้อง	ฟัน หรือ ปุ่มเหงือก	ฎ ด
/c/		กัก	ไม่พ่นลม	ไม่ก้อง	เพดานแข็ง	จ
/cʰ/		กัก	พ่นลม	ไม่ก้อง	เพดานแข็ง	ฉ ช ฌ
/k/ (*)	/kr/, /kl/, /kw/	กัก	ไม่พ่นลม	ไม่ก้อง	เพดานอ่อน	ก
/kʰ/	/kʰr/, /kʰl/, /kʰw/	กัก	พ่นลม	ไม่ก้อง	เพดานอ่อน	ข ฃ ค ฅ ฆ
/ʔ/ (*)		กัก	ไม่พ่นลม	ไม่ก้อง	เส้นเสียง	อ
/m/ (*)		นาสิก		ก้อง	ริมฝีปาก	ม
/n/ (*)		นาสิก		ก้อง	ฟัน หรือ ปุ่มเหงือก	ณ น
/ŋ/ (*)		นาสิก		ก้อง	เพดานอ่อน	ง
/f/	/fr/, /fl/	เสียดแทรก		ไม่ก้อง	ริมฝีปาก	ฝ ฟ
/s/		เสียดแทรก		ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ซ ศ ษ ส
/h/		เสียดแทรก		ไม่ก้อง	เส้นเสียง	ห ฮ
/r/		รัว		ก้อง	ฟัน หรือ ปุ่มเหงือก	ร
/l/		ข้างลิ้น		ก้อง	ฟัน หรือ ปุ่มเหงือก	ล ฬ
/w/ (*)		กึ่งสระ		ก้อง	ริมฝีปาก–เพดานอ่อน	ว
/j/ (*)		กึ่งสระ		ก้อง	เพดานแข็ง	ญ ย

[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

1.3 เสียงสระในภาษาไทย (Thai Vowels) [5]

1.3.1 เสียงสระ

เสียงสระ (Vowel) เป็นเสียงซึ่งถูกเปล่งออกมาทางช่องปากหรือช่องจมูกโดยไม่มีอวัยวะส่วนใดในปากมาเป็นอุปสรรคปิดกั้นทางลมไว้เลย เสียงสระเกิดจากการที่ลมผ่านเส้นเสียงในตำแหน่งที่เส้นเสียงทั้งสองอยู่ชิดกันมากจนเกือบปิดสนิท ทำให้ลมต้องดันตัวออกมาอย่างรุนแรงจนเส้นเสียงเกิดการสั่นสะเทือน และส่งผลทำให้เกิดเสียงดังที่เป็นเสียงก้อง โดยคุณสมบัติที่ทำให้เสียงสระมีความแตกต่างกันมีดังนี้

1. **ส่วนของลิ้นที่ใช้ในการเปล่งเสียง** (Place of Articulation) ในขณะที่ออกเสียงสระต่างๆ ลิ้นหลายส่วนที่ใช้ในการออกเสียงสระ ไม่ว่าจะเป็นลิ้นส่วนหน้า ลิ้นส่วนกลาง หรือลิ้นส่วนหลัง โดยลิ้นส่วนนั้นๆ จะยกขึ้นใกล้เพดานปากในขณะที่ออกเสียงสระหนึ่งๆ ก่อให้เกิดเสียงสระที่แตกต่างกัน โดยถ้าลิ้นส่วนหน้ายกขึ้นให้จุดสูงสุดอยู่ใกล้เพดานแข็ง เราก็จะเรียกเสียงสระนั้นว่าเสียงสระส่วนเพดานแข็ง หรือสระหน้า (Front Vowel) เช่น สระอี สระเอ สระแอ เป็นต้น แต่ถ้าการออกสียงสระใดใช้ลิ้นส่วนหลัง โดยทำการยกลิ้นส่วนหลังขึ้นให้จุดสูงสุดอยู่ใกล้เพดานอ่อนเราก็จะเรียกเสียงสระนั้นว่าเป็นเสียงสระส่วนเพดานอ่อน หรือสระหลัง (back vowel) เช่นสระอู สระโอ สระออ เป็นต้น ส่วนถ้าในการออกเสียงสระใดลิ้นส่วนกลางถูกยกขึ้นไปยังส่วนกลางของเพดานปาก เราก็จะเรียกเสียงสระนั้นว่า สระกลาง (Central Vowel) เช่นสระอือ สระเออ สระอา เป็นต้น
2. **ระยะห่างระหว่างลิ้นและเพดานปากหรือความสูงของลิ้น** (Degree of Stricture) ระยะห่างระหว่างลิ้นและเพดานปากเป็นลักษณะที่สำคัญอย่างหนึ่งในการแบ่งชนิดของเสียงสระ โดยระยะห่างนี้จะเป็นตัวกำหนดว่าเสียงสระที่เปล่งออกมาเป็นสระเปิดหรือสระปิด ถ้าหากลิ้นอยู่ห่างจากเพดานปากมาก หรือลิ้นอยู่ในระดับต่ำ ทำให้ช่องโพรงปากกว้างลมก็จะผ่านออกมาได้มาก เสียงสระที่ได้จะเป็นสระเปิด (Open Vowel) เช่น สระอา ในทางตรงกันข้าม ถ้าตำแหน่งของลิ้นอยู่ใกล้กับเพดานปากมาก หรือลิ้นอยู่ในระดับสูง ช่องโพรงในปากก็จะแคบ ทำให้ลมผ่านออกมาได้น้อย เสียงสระที่ได้จะเป็นสระปิด (Close Vowel) เช่น สระอี สระอู เป็นต้น แต่ถ้าระยะห่างระหว่างลิ้นกับเพดานปากอยู่ในระหว่างสระเปิดและสระปิด เช่นเสียงสระที่เปิดกว้างกว่าสระปิดเล็กน้อย เราก็จะเรียกว่าเป็นสระกลางปิดหรือสระกึ่งปิด (Close-mid, Half-close Vowel) เช่น สระเอ สระโอ เป็นต้น แต่ถ้าเปิดกว้างขึ้นอีก จะเรียกว่าเป็นสระกลางเปิดหรือสระกึ่งเปิด (Open-mid, Half-open Vowel) เช่นสระ แอ สระออ เป็นต้น

3. **การห่อริมฝีปาก** (Labialization) หมายถึงการที่ริมฝีปากทั้งสองเคลื่อนไหวโดยยื่นตัวไปข้างหน้า แล้วห่อกลมมากน้อยเพียงใด ถ้าริมฝีปากยื่นออกไปข้างหน้าแล้วห่อกลมมากเสียงสระที่ได้จะเรียกว่าสระกลม (Rounded Vowel) เช่น สระอู สระโอ สระออ เป็นต้น แต่ถ้าริมปากทั้งสองฉีกออกหรือไม่ห่อกลมขณะเปล่งเสียง สระที่ได้ก็จะเป็นสระไม่กลม (Unrounded Vowel) เช่น สระอี สระเอ สระแอ สระอา เป็นต้น
4. **ลักษณะนาสิก** (Nasalization) เป็นลักษณะในการออกเสียงสระที่ทำให้เกิดเสียงสระขึ้นจมูกหรือสระนาสิก (Nasal Vowel) ขึ้น ซึ่งจะทำให้เสียงแตกต่างจากสระโอฐชะ (Oral Vowel) กล่าวคือ ในการเปล่งเสียงสระโอฐชะนั้น เพดานอ่อนจะยกขึ้นปิดโพรงจมูกอากาศจึงไม่สามารถผ่านออกไปทางช่องจมูกได้ แต่ออกมาทางปากทั้งหมด สำหรับสระนาสิกนั้นเพดานอ่อนจะลดต่ำลง และปล่อยให้อากาศผ่านออกทางช่องจมูกด้วยในเวลาเดียวกัน โดยในการเขียนสัทอักษรสากลจะใช้เครื่องหมาย ~ กำกับอยู่เหนือสระที่ออกเสียงแบบสระนาสิก เช่นในภาษาฝรั่งเศสจะมีหน่วยเสียงนาสิกอยู่ 4 หน่วยเสียงด้วยกัน แต่สำหรับภาษาไทยปกติแล้วไม่มีการออกเสียงสระนาสิก แต่ในบางครั้งก็อาจได้รับอิทธิพลจากการเปล่งเสียงพยัญชนะนาสิกที่อยู่ใกล้เคียง เช่น คำว่า นั้น เป็นต้น
5. **ความยาวในการออกเสียง** (Duration) ความสั้นยาวของการออกเสียงนั้นมีความสำคัญมากในภาษาไทย เพราะหน่วยเสียงที่ใช้ความยาวในการออกเสียงต่างกันจะทำให้ความหมายของพยางค์แตกต่างกันได้ เช่นคำว่า ขุด และคำว่า ขูด โดยสระจะถูกแบ่งออกเป็นสองประเภทตามความสั้นยาวของสระ คือ สระเสียงสั้น (รัสสระ) และสระเสียงยาว (ทีฆสระ) โดยในการเขียนสัทอักษรสากลจะใช้สัญลักษณ์ : (Length Mark) เพื่อแสดงว่าเสียงสระนั้นถูกยืดออกไป

1.3.2 เสียงสระภาษาไทย

สระในภาษาไทยตามไวยากรณ์ดั้งเดิม [8] มีทั้งหมด 21 รูป 32 หน่วยเสียง แบ่งเป็น 3 กลุ่ม คือ

1. สระเดี่ยว (Monophthongs) เป็นสระเสียงแท้ ซึ่งการออกเสียงสระตั้งแต่เริ่มต้นจนสิ้นสุดไม่มีการเปลี่ยนรูปร่างของลิ้นและช่องปาก สระเดี่ยวในภาษาไทยมีทั้งสิ้น 18 หน่วยเสียงเป็นสระเสียงสั้น 9 หน่วยเสียง และสระเสียงยาว 9 หน่วยเสียง
2. สระประสม (Diphthongs) เป็นสระที่เกิดจากการออกเสียงผสมกันของสระแท้ โดยลิ้นและช่องปากจะเปลี่ยนจากรูปร่างการออกเสียงของสระหนึ่งไปยังอีกสระหนึ่งอย่างค่อนข้างกลมกลืนและรวดเร็ว สระประสมในภาษาไทยมีทั้งสิ้น 6 หน่วยเสียง เป็นสระเสียงสั้น 3 หน่วยเสียง และสระเสียงยาว 3 หน่วยเสียง

3. สระเกิน (Vowel Letter) ในภาษาไทยมีรูปสระที่เกิดจากการรวมของเสียงสระกับตัวสะกดหรือคำควบเข้าไว้ด้วยกัน ซึ่งมีทั้งหมด 8 หน่วยเสียง เสียงสระในภาษาไทยสามารถสรุปได้ดังตารางที่ 2.2

ตารางที่ 2.2 ตารางเสียงสระภาษาไทย

หน่วยเสียง[1]	ส่วนของลิ้นที่ใช้เปล่งเสียง	ความสูงของลิ้น	การห่อริมฝีปาก	ความยาวเสียง	รูปสระ
สระเดี่ยว					
/i/	หน้า	ปิด	ไม่ห่อ	สั้น	อิ
/iː/	หน้า	ปิด	ไม่ห่อ	ยาว	อี
/e/	หน้า	กึ่งปิด	ไม่ห่อ	สั้น	เอะ
/eː/	หน้า	กึ่งปิด	ไม่ห่อ	ยาว	เอ
/æ/	หน้า	กึ่งเปิด	ไม่ห่อ	สั้น	แอะ
/æː/	หน้า	กึ่งเปิด	ไม่ห่อ	ยาว	แอ
/ɨ/	หลัง ค่อนมาทางกลาง	ปิด	ไม่ห่อ	สั้น	อึ
/ɨː/	หลัง ค่อนมาทางกลาง	ปิด	ไม่ห่อ	ยาว	อือ
/ɜ/	หลัง ค่อนมาทางกลาง	กึ่งปิด	ไม่ห่อ	สั้น	เออะ
/ɜː/	หลัง ค่อนมาทางกลาง	กึ่งปิด	ไม่ห่อ	ยาว	เออ
/a/	กลาง	เปิด	ไม่ห่อ	สั้น	อะ
/aː/	กลาง	เปิด	ไม่ห่อ	ยาว	อา
/ɯ/	หลัง	ปิด	ห่อ	สั้น	อุ
/uː/	หลัง	ปิด	ห่อ	ยาว	อู
/o/	หลัง	กึ่งปิด	ห่อ	สั้น	โอะ
/oː/	หลัง	กึ่งปิด	ห่อ	ยาว	โอ
/ɔ/	หลัง	กึ่งเปิด	ห่อ	สั้น	เอาะ
/ɔː/	หลัง	กึ่งเปิด	ห่อ	ยาว	ออ
สระประสม	ส่วนประกอบ				
/ia/	/i/ + /a/			สั้น	เอียะ
/iːa/	/iː/ + /a/			ยาว	เอีย
/ɨa/	/ɨ/ + /a/			สั้น	เอือะ
/ɨːa/	/ɨː/ + /a/			ยาว	เอือ
/ua/	/u/ + /a/			สั้น	อัวะ
/uːa/	/uː/ + /a/			ยาว	อัว
สระเกิน	ส่วนประกอบ				
/am/	/a/ + /m/			สั้น	อำ
/aj/	/a/ + /j/			สั้น	ไอ ใอ
/aw/	/a/ + /w/			สั้น	เอา
/ri/, /rɨ/	/r/ + /i/, /r/ + /ɨ/			สั้น	ฤ
/riː/, /rɨː/	/r/ + /iː/, /r/ + /ɨː/			ยาว	ฤา
/li/, /lɨ/	/l/ + /i/, /l/ + /ɨ/			สั้น	ฦ
/liː/, /lɨː/	/l/ + /iː/, /l/ + /ɨː/			ยาว	ฦา

[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

1.4 เสียงวรรณยุกต์ในภาษาไทย (Thai Tones) [5]

เสียงวรรณยุกต์นั้นคือเสียงสูงต่ำในภาษา ซึ่งเกิดจากการสั่นสะเทือนของเส้นเสียงในอัตราความถี่ที่ต่างกันไป ดังนั้นเสียงวรรณยุกต์จะปรากฏอยู่ในส่วนของเสียงสระ เพราะเสียงสระเป็นเสียงที่เกิดจากการสั่นของเส้นสียง นอกจากนี้ยังอาจมีเสียงวรรณยุกต์ปรากฏอยู่บ้างในบางส่วนของเสียงพยัญชนะ แต่จะต้องเป็นส่วนของเสียงพยัญชนะที่เป็นเสียงก้องหรือพยัญชนะนาสิกเท่านั้น เพราะเสียงพยัญชนะไม่ก้องนั้นไม่ได้เกิดจากการสั่นของเส้นเสียง จึงไม่สามารถมีเสียงวรรณยุกต์อยู่ด้วยได้

สำหรับในภาษาไทย วรรณยุกต์นั้นถือได้ว่าเป็นหน่วยเสียงที่สำคัญ เพราะสามารถใช้แยกแยะความแตกต่างทางความหมายของคำในภาษาไทยได้ ตรงกันข้ามกับบางภาษา เช่น ภาษาอังกฤษ ซึ่งไม่จัดว่าเสียงวรรณยุกต์เป็นหน่วยเสียงในภาษา เพราะไม่ว่าเราจะพูดภาษาอังกฤษด้วยเสียงสูงต่ำอย่างไร ก็ไม่ทำให้ความหมายของคำเปลี่ยนไป แต่ก็อาจจะต้องมีการใช้เสียงวรรณยุกต์ประกอบบ้าง ทั้งนี้เพื่อใช้เน้นให้ความหมายโดยรวมของประโยคเปลี่ยนไป ภาษาไทยจึงจัดได้ว่าเป็นภาษามีวรรณยุกต์ (Tonal Language) เสียงวรรณยุกต์ภาษาไทยสามารถแบ่งออกเป็น 2 ชนิดใหญ่ๆ คือ

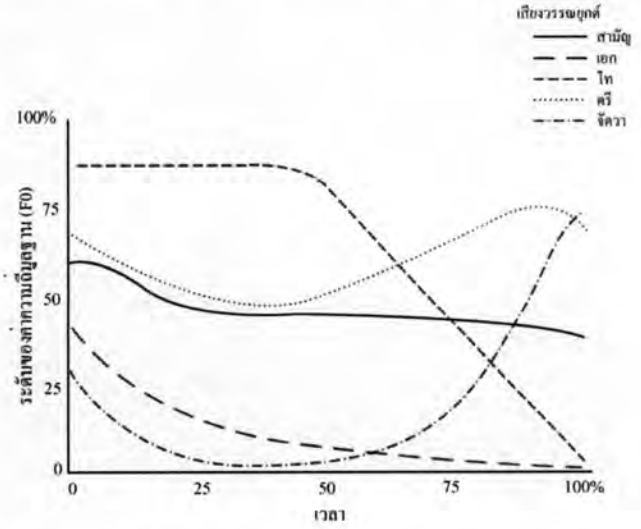
1. เสียงวรรณยุกต์ระดับ (Level Tone) เป็นเสียงวรรณยุกต์ที่มีระดับความถี่ค่อนข้างคงที่ตลอดพยางค์ ถึงแม้ว่าในการออกเสียงพูดโดยปกตินั้น เสียงต้นพยางค์มักจะไม่ได้มีความถี่และความดังเท่ากันกับเสียงท้ายพยางค์ โดยเสียงต้นพยางค์มักมีมีระดับความถี่สูงกว่าและดังกว่าเสียงท้ายพยางค์ แต่ในทางสัทศาสตร์แล้ว ความถี่ที่ต่างกันหรือการเปลี่ยนแปลงของระดับเสียงนี้ถือว่าเล็กน้อยมาก เมื่อเทียบกับการเปลี่ยนระดับความถี่ของเสียงในพยางค์อีกจำพวกหนึ่งซึ่งจะได้กล่าวต่อไป สำหรับเสียงวรรณยุกต์ระดับในภาษาไทยนั้น มีอยู่ด้วยกัน 3 เสียงดังนี้คือ
    1. เสียงวรรณยุกต์สามัญ (Mid Tone) เสียงวรรณยุกต์นี้มีระดับความถี่ปานกลางประมาณ 120 เฮิรตซ์ และคงที่อยู่ที่ระดับนั้นจนกระทั่งปลายพยางค์ จึงจะลดต่ำลงมาจนเกือบถึงประมาณ 110 เฮิรตซ์ เสียงวรรณยุกต์สามัญนี้จะไม่ปรากฏในพยางค์ที่มีพยัญชนะกักเป็นพยัญชนะท้าย หรือที่เรียกกันว่าคำตาย
    2. เสียงวรรณยุกต์เอก (Low Tone) เสียงวรรณยุกต์นี้มีระดับความถี่ต้นเสียงปานกลางประมาณ 120 เฮิรตซ์ แล้วลดต่ำลงมาเหลือประมาณ 100 เฮิรตซ์อย่างรวดเร็ว และคงที่อยู่ในระดับนี้ สำหรับเสียงวรรณยุกต์เอกจะปรากฏกับพยางค์ได้ทุกรูปแบบทั้งคำเป็นและคำตาย
    3. เสียงวรรณยุกต์ตรี (High Tone) เสียงวรรณยุกต์นี้มีระดับความถี่ค่อนข้างสูง โดยจะค่อยๆ สูงขึ้นทีละน้อยจากต้นพยางค์ซึ่งมีความถี่ประมาณ 125 เฮิรตซ์ ไปจนถึง

ประมาณ 135 – 140 เฮิรตซ์เมื่อสิ้นพยางค์ หรืออาจจะลดต่ำลงตอนปลายพยางค์มาอยู่ที่ประมาณ 130 เฮิรตซ์ก็ได้ ขึ้นอยู่กับว่าพยางค์นั้นๆ จบลงด้วยเสียงประเภทใด ถ้าพยางค์นั้นคำเป็น ระดับของเสียงตอนปลายของพยางค์จะไม่ลดต่ำลงมา แต่ถ้าพยางค์นั้นเป็นคำตาย ระดับเสียงตอนปลายจะลดต่ำลงอย่างรวดเร็ว

2. เสียงวรรณยุกต์เปลี่ยนระดับ (Contour Tone) เป็นเสียงวรรณยุกต์ที่มีระดับความถี่ของการออกเสียงเปลี่ยนแปลงมากในช่วงพยางค์หนึ่งๆ เช่น ต้นพยางค์ออกเสียงให้มีระดับสูงแล้วลดระดับเสียงลงอย่างรวดเร็วไปสู่ระดับต่ำที่ท้ายพยางค์ หรือต้นพยางค์ออกเสียงให้มีระดับต่ำ แล้วเพิ่มระดับเสียงอย่างรวดเร็วไปเป็นระดับสูงที่ท้ายพยางค์ นอกจากนี้ ยังอาจจะเกิดจากการเปลี่ยนระดับเสียงจากสูงแล้วไปต่ำแล้วไปสูงอีก หรือเปลี่ยนจากต่ำแล้วไปสูงแล้วไปต่ำอีกก็ได้ สำหรับในภาษาไทยนั้นมีเสียวรรณยุกต์เปลี่ยนระดับอยู่ 2 เสียงดังนี้

   1. เสียงวรรณยุกต์โท (Falling Tone) ระดับเสียงจะเริ่มต้นที่ระดับความถี่ประมาณ 140 เฮิรตซ์ แต่เมื่อถึงประมาณ 1 ใน 4 ของความยาวช่วงพยางค์ ระดับความถี่จะเริ่มลดลงเรื่อยๆ จนต่ำกว่า 100 เฮิรตซ์ที่ปลายพยางค์ หรืออาจจะมีการเปลี่ยนระดับความถี่สูงขึ้นจากต้นพยางค์เล็กน้อยก่อนที่จะลดระดับเสียงลงอย่างรวดเร็วก็ได้ เสียงวรรณยุกต์โทนี้จะไม่ปรากฏในคำตาย ยกเว้นในคำเลียนเสียงธรรมชาติ หรือคำลงท้ายประโยคบางคำ เช่น "พลัก" หรือ "ละ" เป็นต้น
   2. เสียงวรรณยุกต์จัตวา (Rising Tone) ระดับเสียงจะเริ่มที่ระดับความถี่ประมาณ 110 เฮิรตซ์ แล้วมักจะลดลงเล็กน้อยก่อนจะเพิ่มความถี่ขึ้นอย่างรวดเร็วจนสูงถึงประมาณ 140 เฮิรตซ์ที่ท้ายพยางค์ เสียงวรรณยุกต์จัตวานี้จะไม่ปรากฏที่คำตาย

การเปลี่ยนแปลงความถี่ของเสียงในวรรณยุกต์ภาษาไทยสามารถแสดงได้ดังรูปที่ 2.3



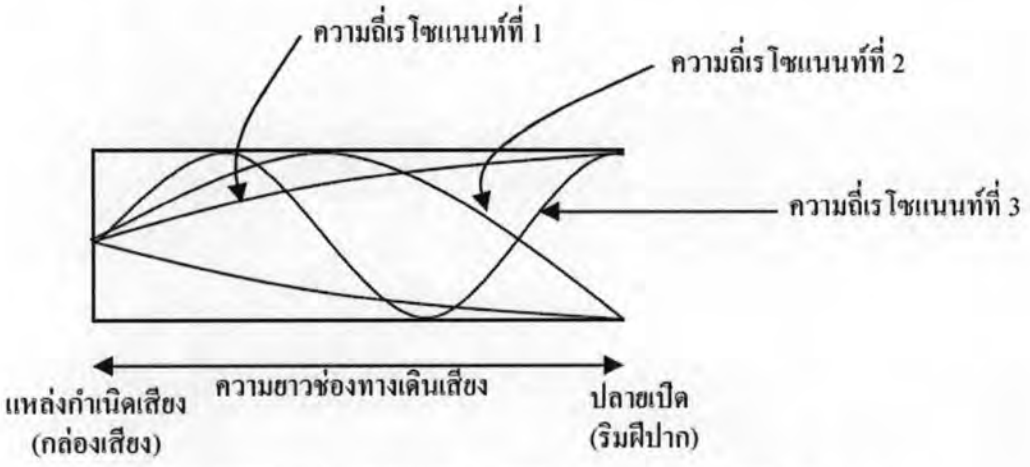
รูปที่ 2.3 การเปลี่ยนแปลงความถี่ของเสียงในววรณยุกต์ภาษาไทย

2. สวนสัทศาสตร์

2.1 กระบวนการสร้างเสียงพูด (Speech Production)

ระบบเสียงพูดสามารถพิจารณาได้ว่าประกอบไปด้วยลำดับของท่อ และช่องที่ต่อออกมาจากปอดไปยังปากและจมูก ท่อและช่องนี้จะมีความยาวโดยประมาณ 7 นิ้ว เส้นเสียงจะอยู่ในตำแหน่งปลายที่ตรงข้ามกับขั้วปอด ทำหน้าที่ควบคุมการไหลของลมให้ผ่านปอดเข้าสู่ช่องทางเดินเสียง (Vocal tract) ภายใต้การควบคุมของกล้ามเนื้อ ส่วนประกอบของช่องทางเดินเสียงที่มีลักษณะเป็นท่อจะสามารถเปลี่ยนรูปร่างได้ในอัตราถึง 10 ครั้งต่อวินาที ส่วนเส้นเสียงนั้นจะสามารถเปิด-ปิดด้วยอัตราเร็วประมาณ 100 - 300 ครั้งต่อวินาที การเปลี่ยนแปลงรูปร่างของช่องทางเดินเสียงและรูปร่างและตำแหน่งของสื่อกลางที่ทำให้เกิดเสียงดังกล่าวนี้รวมเรียกว่ากระบวนการทำให้เกิดเสียง

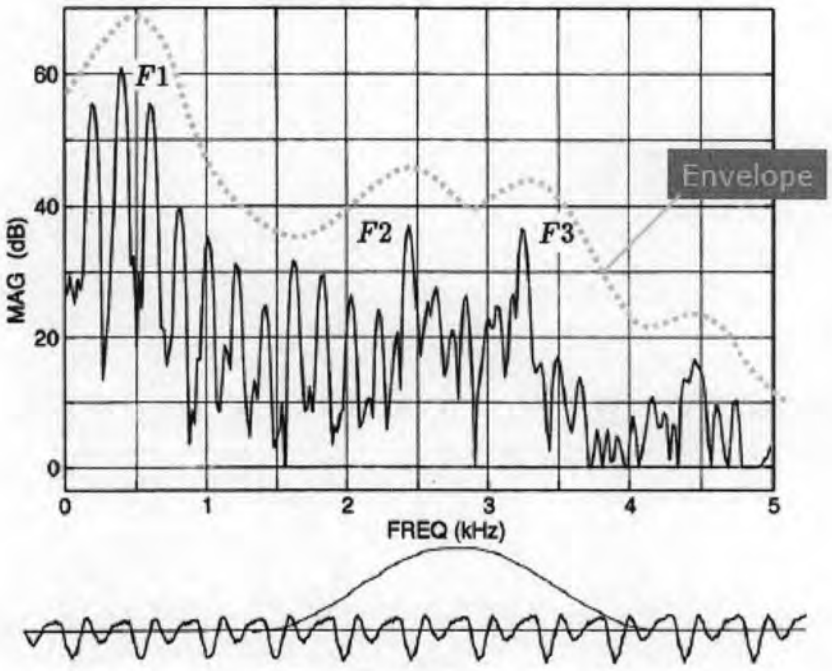
รูปแบบจำลองอย่างง่ายของช่องทางเดินเสียง ก็อาจมองได้เป็นลักษณะของท่อทรงกระบอกที่มีต้นกำเนิดเสียงอยู่ที่ปลายข้างหนึ่ง (ส่วนของกล่องเสียง) และปลายอีกข้างหนึ่งจะเปิด (ส่วนของปาก) ดังรูปที่ 2.4 ดังนั้นมันจะเกิดเรโซแนนท์ภายในท่อได้ที่ความยาวเท่ากับ 4L, 4L/3, 4L/5, ... เฮิร์ต โดยที่ L คือความยาวท่อ ถ้าคิดเป็นความถี่ที่เกิดเรโซแนนท์จะได้ความถี่ที่ c/4L, 3c/4L, 5c/4L, ... เฮิร์ต โดยที่ c คือ ค่าความเร็วของเสียงในอากาศ และถ้าจะคำนวณหาความถี่ในการเรโซแนนท์ของช่องทางเดินเสียงของคน ซึ่งปกติช่องทางเดินเสียงของคนเราจะมีความยาวประมาณ 7 นิ้ว หรือ 17 เซนติเมตร และ c มีค่าเท่ากับ 340 เมตรต่อวินาที ดังนั้นจึงมีเรโซแนนท์ที่ความถี่ประมาณ 500 เฮิรต์, 1500 เฮิรต์, 2500 เฮิรต์, ... เป็นต้น



รูปที่ 2.4 การเกิดเรโซแนนท์ภายในแบบจำลองของช่องทางเดินเสียง

ความถี่เรโซแนนท์ต่างๆ จะเปลี่ยนไปเมื่อพื้นที่หน้าตัดของท่อในบริเวณต่างๆ เปลี่ยนแปลงไป การเปลี่ยนแปลงพื้นที่หน้าตัดของท่อตามเวลานี้ เป็นการจำลองรูปร่างที่เปลี่ยนไปของช่องทางเสียง ซึ่งมีสาเหตุจากการเคลื่อนไหวของอวัยวะส่วนการกระทำอาการ

เสียงพูดของมนุษย์เกิดจากต้นกำเนิดเสียง ที่อาจเป็นสัญญาณกึ่งคาบ (Quasi-periodic Signal) เช่น ต้นกำเนิดที่เกิดจากการสั่นของเส้นเสียง หรือสัญญาณที่มีลักษณะเหมือนสัญญาณรบกวน (Noise) เมื่อต้นกำเนิดเป็นสัญญาณกึ่งคาบ ช่องทางเสียงที่ถูกกระตุ้นด้วยต้นกำเนิดเสียง จะสร้างคลื่นเสียงที่มีองค์ประกอบทางความถี่สูงเด่นขึ้นมาในบริเวณของความถี่เรโซแนนท์ ในขณะนั้นความถี่เรโซแนนท์ในกรณีนี้ถูกเรียกว่า ความถี่ฟอร์แมนต์ (Formant Frequency)



รูปที่ 2.5 สเปกตรัมของพลังงานเสียง [7]

ฟอร์แมนท์ที่มีความถี่ต่ำที่สุดจะเรียกว่าฟอร์แมนท์ที่หนึ่ง ซึ่งจะมีค่าประมาณ 200 – 1200 เฮิรต์ ทั้งนี้ขึ้นอยู่กับขนาดของช่องทางเดินเสียงด้วย ส่วนฟอร์แมนท์ที่สองที่อยู่ถัดไปก็จะมีค่าประมาณ 500 – 2500 เฮิรต์ และฟอร์แมนท์ที่สามจะมีค่าประมาณ 1500 – 3500 เฮิรต์ เป็นต้น ในทำนองเดียวกัน เมื่อต้นกำเนิดเสียงเป็นลักษณะสัญญาณรบกวน เสียงที่ถูกสร้างจากช่องทางเสียงจะมีลักษณะเป็นสัญญาณรบกวนเช่นเดียวกับต้นกำเนิด แต่ขนาดของสเปกตรัมที่ความถี่เรโซแนนท์จะถูกขยายขึ้นอย่างมาก รูปร่างของสเปกตรัมของเสียงนี้สามารถนำมาวิเคราะห์เพื่อประมาณเหตุการณ์ที่เกิดขึ้นในช่องทางเสียงของผู้พูดได้

2.2 สัทลักษณ์ (Phonetic Features)

สัทลักษณ์โดยทั่วไปที่ใช้ในการวิเคราะห์และอธิบายคุณสมบัติของแต่ละประเภทหน่วยเสียงนั้นแบ่งออกเป็น 3 ประเภทคือคุณสมบัติความก้องของเสียงตามลักษณะของแหล่งกำเนิดเสียง (Source Characteristics), ลักษณะการออกเสียง (Manner of Articulation) และ ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง (Place of Articulation) Chomsky และ Halle [9] ได้ศึกษาและให้นิยามของ สัทลักษณ์โดยกำหนดให้แต่ละสัทลักษณ์มีค่าเป็น + หรือ – เท่านั้น

2.2.1 คุณสมบัติความก้องของเสียง

เสียงพูดของมนุษย์เมื่ออากาศถูกดันออกมาจากปอดด้วยแรงดันที่มากจะทำให้เส้นเสียงสั่นทำให้สัญญาณที่เกิดจากแหล่งกำเนิดเสียงเป็นลักษณะเป็นคาบจะแสดงได้ด้วยค่าคุณสมบัติความเป็นเสียงก้อง (Voiced) แทนด้วยสัญลักษณ์ '+' หรือ [+voiced] ส่วนสัญญาณเสียงที่เกิดจากโดยไม่มีการสั่นของเส้นเสียงและมีลักษณะไม่เป็นคาบจะแสดงได้ด้วยค่าคุณสมบัติความเป็นเสียงไม่ก้อง (Unvoiced) แทนด้วยสัญลักษณ์ '–'หรือ [–voiced] ทั้งสัญญาณเสียงที่ก้องและไม่ก้องอาจปรากฏอยู่ในเสียงที่มนุษย์ได้ยินเนื่องจากมีการเปลี่ยนแปลงของสัญญาณเสียงในช่องทางเดินเสียง

2.2.2 ลักษณะการออกเสียง

ลักษณะการออกเสียงพิจารณาจากลักษณะของช่องทางเดินเสียงว่ามีการเปิด/ปิดอย่างไร มีการกักเสียงไว้มากน้อยแค่ไหน และการออกเสียงนั้นอากาศไหลผ่านบริเวณช่องปากหรือผ่านไปในช่องโพรงจมูกบ้างหรือไม่ เสียงที่ผ่านช่องปากไปโดยไม่ได้ถูกกักเอาไว้เพียงพอต่อการสร้างเสียงรบกวนหรือกักการไหลของอากาศจะเรียกว่าเสียงที่มีเสียงสั่น (Sonorant) ซึ่งประกอบไปด้วยเสียงสระ (Vowels) เสียงกึ่งสระ (Semi-vowels) และเสียงนาสิก (Nasals) คุณสมบัติความเป็นเสียงที่มีเสียงสั่นแสดงได้ด้วยสัญลักษณ์ [+sonorant] ส่วนเสียงเสียงที่ไม่มีคุณสมบัติความเป็นความเป็นเสียงที่มีเสียงสั่น ได้แก่ เสียงพยัญชนะกัก (Stop Consonants) และเสียงเสียดแทรก (Fricatives) จะแสดงได้ด้วยสัญลักษณ์ [–sonorant] เสียงที่มีและไม่มีคุณสมบัติเป็นความเป็นเสียงที่มีเสียงสั่นสามารถแยกต่อไปได้อีก ตารางที่ 2.3 แสดงความสัมพันธ์ระหว่างการแบ่งประเภทของหน่วยเสียงกลุ่มใหญ่ๆ (เสียงสระ, เสียงพยัญชนะกัก, เสียงเสียดแทรก และ เสียงนาสิกและเสียงกึ่งสระ) กับคุณสมบัติของลักษณะการออกเสียงในแต่ละกลุ่ม คุณสมบัติความเป็นเสียงที่เป็นเสียงพยางค์ (Syllabic) คือเสียงที่เกิดจากการช่องทางเดินเสียงเปิดอยู่แทนด้วยสัญลักษณ์ [+syllabic] ซึ่งประกอบด้วยพวกเสียงสระ ส่วนเสียงที่เกิดขึ้นโดยที่ช่องทางเดินเสียงปิดอยู่จะแทนด้วยสัญลักษณ์

[–syllabic] ซึ่งประกอบไปด้วยเสียงกึ่งสระและเสียงนาสิก ส่วนเสียงที่มีคุณสมบัติความเป็นเสียงที่มีการกักเสียงเอาไว้โดยที่ช่องทางเดินเสียงอยู่ในสภาพปิดอยู่แต่ยังปิดไม่สมบูรณ์ (Continuant) ตัวอย่างเช่น เสียง /ส/ ที่ใช้ฟันในการกักไม่ให้อากาศไหลผ่านช่องปากได้อย่างเต็มที่ทำให้เกิดเป็นเสียงเสียดแทรก แสดงได้ด้วยสัญลักษณ์ [+continuant] ส่วนเสียงพยัญชนะกักแสดงได้ด้วยสัญลักษณ์ [–continuant]

ตารางที่ 2.3 ตารางแสดงความสัมพันธ์ระหว่างประเภทของหน่วยเสียง เปรียบเทียบกับสัทลักษณ์ของลักษณะการออกเสียง

สัทลักษณ์	[-continuant]	[+continuant]
[-sonorant]	เสียงพยัญชนะกัก	เสียงเสียดแทรก
[+sonorant, -syllabic]	เสียงนาสิกและเสียงกึ่งสระ	
[+sonorant, +syllabic]	เสียงสระ	

2.2.3 ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง

ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียงในกรณีของเสียงพยัญชนะกักและเสียงพยัญชนะเสียดแทรกจะพิจารณาจากตำแหน่งที่มีใช้อวัยวะเช่นฟัน ลิ้น หรือริมฝีปากในการกั้นไม่ให้อากาศไหลผ่านช่องปากออกไปได้โดยตรง ในกรณีของเสียงสระจะพิจารณาจากตำแหน่งของลิ้นในขณะที่อากาศเดินทางผ่านช่องปากออกไป ดังจำแนกด้วยคุณสมบัติดังตารางที่ 2.4

ตารางที่ 2.4 ตารางแสดงความสัมพันธ์ระหว่างสัทลักษณ์ของตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียงเปรียบเทียบกับเสียงพยัญชนะกักในภาษาไทย

สัทลักษณ์	อวัยวะที่มีความสัมพันธ์กับการออกเสียง	/บ/ /พ/	/ด/ /ท/	/ค/ /ก/
Velar	เกิดการกักเสียงระหว่างตัวลิ้นกับเพดานอ่อน	–	–	+
Alveolar	เกิดการกักเสียงปลายลิ้นกับเพดานส่วนหน้า	–	+	–
Labial	เกิดการกัดเสียงบริเวณริมฝีปาก	+	–	–

3. การรู้จำเสียงพูด

การรู้จำเสียงพูด (Speech Recognition) เป็นกระบวนการสกัดลำดับของคำ (Sequence of Words) ที่อยู่ในสัญญาณเสียงออกมา โดยในเชิงทฤษฎีสารสนเทศ (Information Theory) เราพิจารณาเสียงพูดที่ได้ยินว่าเป็นสัญญาณที่ถูกรบกวนผ่านทางช่องสัญญาณรบกวน (Noisy Channel) และการรู้จำเสียงพูดคือการถอดรหัส (Decode) สัญญาณนั้น

3.1 ปัญหาการรู้จำเสียงพูด

ปัญหาการรู้จำเสียงพูดหรือการถอดรหัสของสัญญาณเสียงพูด สามารถมองได้ว่าเป็นการหาลำดับของคำที่ดีที่สุดสำหรับลำดับของข้อมูลทางเสียงที่สังเกตได้ (Observation Sequence) โดยกำหนดให้ สัญลักษณ์ของลำดับข้อมูลที่สังเกตได้เป็น $O = o_1 o_2 o_3 ... o_t$ สัญลักษณ์ของลำดับของคำเป็น $W = w_1 w_2 w_3 ... w_n$ และ L แทนภาษาที่พิจารณา

ให้ W^* เป็นลำดับของคำที่ดีที่สุด เพราะฉะนั้นเราจะได้ว่า

$$W^* = \arg\max_{W \in L} P(W \mid O)$$

ซึ่งในทางปฏิบัติแล้ว การหาค่า $P(W \mid O)$ โดยตรงทำได้ยาก จึงเขียนสูตรข้างต้นใหม่โดยใช้กฎของเบย์ได้ดังนี้

$$\begin{aligned} W^* &= \arg\max_{W \in L} \frac{P(O \mid W) P(W)}{P(O)} \\ &= \arg\max_{W \in L} P(O \mid W) P(W) \end{aligned}$$

ในที่นี้ $P(W)$ คือความน่าจะเป็นที่ลำดับของคำ W จะเกิดขึ้นในภาษา จึงเรียก $P(W \mid O)$ ว่าน่าจะเป็นก่อนหน้า (Prior) หรือแบบจำลองทางภาษา (Language Model) ส่วน $P(O \mid W)$ เป็นความน่าจะเป็นที่ลำดับของข้อมูลที่สังเกตได้เป็น O เมื่อลำดับของคำที่เป็นที่มาของ O คือ W และเรียก $P(O \mid W)$ ว่าความเป็นไปได้ (Likelihood) หรือแบบจำลองทางเสียง (Acoustic Model)

3.2 ประเภทของระบบรู้จำเสียงพูด

เราสามารถแบ่งระบบรู้จำเสียงพูดตามเกณฑ์ต่างๆ ซึ่งเป็นส่งผลต่อความยากง่ายในการพัฒนา และประสิทธิภาพของระบบรู้จำเสียงพูด ได้ดังต่อไปนี้

1. **จำนวนคำศัพท์** (Vocabulary Size)

ระบบที่รองรับจำนวนคำศัพท์น้อย เช่น รู้จำเสียงพูดของตัวเลข ศูนย์ ถึง เก้า สามารถพัฒนาได้ง่ายกว่าและให้ความผิดพลาดน้อยกว่าระบบที่ต้องรองรับจำนวนคำศัพท์มาก เช่น รู้จำเสียงพูดของทุกคำในภาษา ซึ่งอาจมีมากถึง 20,000 คำ

2. **ความขึ้นต่อผู้พูด** (Speaker Dependency)
    1. การรู้จำเสียงพูดแบบคำโดดเดี่ยว (Isolated Word Recognition) จะพิจารณาหนึ่งคำต่อการเปล่งเสียงหนึ่งครั้ง โดยระบบจะพยายามตัดสินใจให้รู้จำคำที่ใกล้เคียงที่สุด
    2. การรู้จำเสียงพูดแบบเสียงพูดต่อเนื่อง (Continuous Speech Recognition) จะพิจารณาเป็นวลี หรือประโยค ซึ่งการรู้จำแบบนี้ระบบจะแบ่งการทำงานออกเป็น 2 ขั้นตอน คือขั้นตอนของการแบ่งเสียงที่เข้ามาออกเป็นคำย่อยๆ และขั้นตอนของการรู้จำประโยคโดยอาศัยบริบท เพื่อเข้าใจความหมายของประโยค นอกจากนี้ต้องพิจารณาการเชื่อมเสียงระหว่างคำต่างๆ ในประโยค โดยใช้ข้อมูลทางไวยกรณ์ของภาษาที่ซับซ้อน
3. **ลักษณะการพูด** (Speaking Style)
    1. ขึ้นกับผู้พูด (Speaker-Dependent) จะทำให้แบบจำลองที่เราใช้อ้างอิงต้องเปลี่ยนแปลงตามผู้พูด
    2. ไม่ขึ้นกับผู้พูด (Speaker-Independent) ระบบจะสามารถรู้จำคำได้ไม่ว่าอินพุทของเสียงที่เข้ามาในระบบจะเป็นของผู้พูดคนไหน แต่ระบบแบบนี้ระบบของเราจะต้องมีความซับซ้อนอย่างมาก เพราะระบบของเราจะต้องสามารถรองรับสถานการณ์ทุกรูปแบบของเสียงที่จะเป็นอินพุทให้ได้
4. **ระดับของหน่วยเสียง** (Sound Unit)
    1. หน่วยเสียงระดับคำ (Word-Based) เป็นการรู้จำคำพูดโดยใช้หน่วยของ "คำ" เป็นหน่วยย่อยที่สุดในการรู้จำคำพูด โดยจะใช้โมเดลแต่ละโมเดลเพื่อรู้จำคำแต่ละคำ และไม่มีการใช้โมเดลร่วมกันระหว่างคำ ซึ่งจะทำให้มีการใช้พื้นที่หน่วยความจำสูงมาก และจะประสบปัญหาในการพิจารณาการรู้จำคำในประโยคที่เราต้องพิจารณาการเชื่อมเสียงระหว่างคำเป็นบริบท
    2. หน่วยเสียงระดับเล็กกว่าคำ (Sub-Word Based) เป็นการรู้จำคำพูดโดยใช้หน่วยเสียงที่เป็นหน่วยที่เล็กกว่า "คำ" ในการรู้จำคำพูด โดยจะใช้โมเดลแต่ละโมเดลเพื่อรู้จำแต่ละหน่วยที่เล็กกว่าคำ ซึ่งเราสามารถใช้โมเดลร่วมกันหลายๆ โมเดลได้เพื่อรู้จำคำ โดยแต่ละหน่วยที่เล็กกว่าคำ จะต้องพึ่งบริบทในการพิจารณาการรู้จำคำศัพท์ เพื่อพิจารณาการเชื่อมเสียง นอกจากนี้วิธีนี้ยังใช้พื้นที่ในหน่วยความจำไม่สูงนัก

4. การแปลงฟูริเยร์แบบวิยุต

ในกระบวนการการวิเคราะห์เสียงพูดในโดเมนความถี่นั้น เราจะใช้การแปลงฟูริเยร์แบบวิยุตกับเสียงพูดที่ได้รับเข้ามาให้อยู่ในรูปของเวกเตอร์คุณสมบัติของเสียงเพื่อนำไปวิเคราะห์เปรียบเทียบกับเสียงตัวอย่าง โดยเสียงตัวอย่างที่ถูกส่งเข้ามาในระบบนี้จะเป็นสัญญาณที่อยู่ในโดเมนเวลาซึ่งเป็นฟังก์ชันคาบ และเมื่อได้รับการแปลงฟูริเยร์แบบวิยุตแล้วจะทำให้สัญญาณถูกเปลี่ยนไปอยู่ในโดเมนความถี่ แต่เนื่องจากระบบคอมพิวเตอร์จะจัดเก็บสัญญาณดังกล่าวขึ้นอยู่กับอัตราการสุ่ม (Sampling Rate) ดังนั้นเราจึงสามารถทำการวิเคราะห์สัญญาณในโดเมนความถี่ดังกล่าว ให้อยู่ในรูปแบบฟังก์ชันคาบอีกครั้งหนึ่ง

การแปลงฟูริเยร์แบบต่อเนื่อง (Continuous Fourier Transform) ได้นิยามไว้ ดังนี้

$$f(v) = \mathbf{F}[f(t)](v) = \int_{-\infty}^{+\infty} f(t)e^{-2\pi ivt}\,dt$$

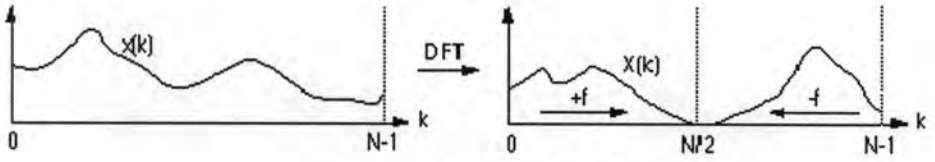
จากนิยามของการแปลงฟูริเยร์แบบต่อเนื่อง เราจะทำการพิจารณาในกรณีที่เป็นฟังก์ชันไม่ต่อเนื่อง $f(t) \to f(t_k)$ โดยให้ $f_k \equiv f(t_k)$ และ $t_k \equiv k\Delta$ เมื่อให้ $k = 0,1,2,...,N-1$ จะได้การแปลงฟูริเยร์แบบวิยุตดังนี้ $F_n = \mathbf{F}_k[\{f_k\}_{k=0}^{N-1}](n)$

$$F_n = \sum_{k=0}^{N-1} f_k e^{-2\pi ink/N}$$

การแปลงฟูริเยร์แบบวิยุต จะทำให้เห็นถึงช่วงคาบและความสัมพันธ์ของแต่ละช่วงสัญญาณที่เข้ามาอย่างชัดเจน ซึ่งการแปลงฟูริเยร์แบบวิยุตของลำดับจำนวนจริงจะเท่ากับการแปลงฟูริเยร์แบบวิยุตของลำดับจำนวนเชิงซ้อนที่มีจำนวนพจน์เท่ากัน โดยสรุปแล้ว ถ้า f_k เป็นจำนวนจริงแล้ว F_{N-n} และ F_n จะมีความสัมพันธ์ดังนี้

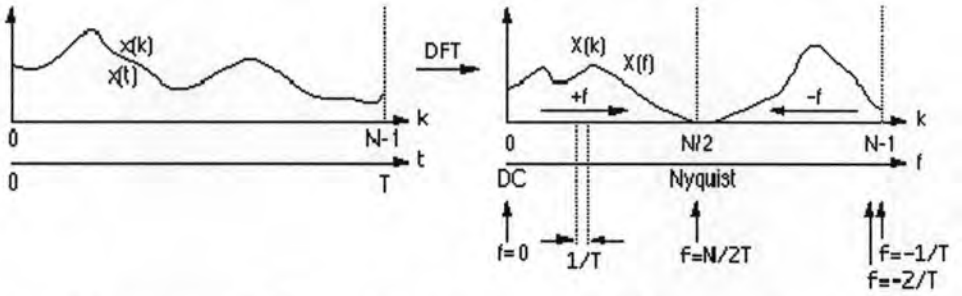
$$F_{N-n} = \overline{F}_n$$

สำหรับ $n = 0,1,2,...,N-1$ และ $\overline{z}$ แทนจำนวนเชิงซ้อนผัน (Complex Conjugate) ซึ่งหมายความว่าส่วน F_0 จะเป็นจำนวนจริงเสมอ และจากความสัมพันธ์ดังกล่าว ฟังก์ชันคาบ (Periodic Function) จะประกอบด้วยจุดสูงสุดของสัญญาณ 2 จุดด้วยกัน ซึ่งหมายความว่าช่วงคาบของสัญญาณที่เข้ามาเป็นสองส่วน คือช่วงความถี่บวก (Positive Frequency) และช่วงความถี่ลบ (Negative Frequency) ซึ่งเป็นช่วงความถี่ที่เป็นจำนวนเชิงซ้อน



รูปที่ 2.6 การแปลงฟูริเยร์แบบวิยุต

กราฟในรูปที่ 2.6 แสดงความสัมพันธ์ระหว่างหน่วยของอนุกรมและหน่วยของความถี่ (กราฟดังกล่าวเป็นเพียงการจำลองภาพขึ้นเท่านั้น) ซึ่งในปกติแล้วการแปลงฟูริเยร์แบบวิยุตจะมีข้อมูลเป็นอนุกรมของจำนวนเชิงซ้อน ดังตัวอย่าง ถ้าอนุกรมแทนลำดับของเวลาที่มีความยาว T แล้ว จะสามารถแสดงค่าความถี่ในรูปที่ 2.7 ได้ดังนี้



รูปที่ 2.7 การแปลงฟูริเยร์แบบวิยุต (เมื่ออนุกรมแทนลำดับของเวลามีความยาว T)

การแปลงฟูริเยร์แบบเร็ว (Fast Fourier Transform - FFT) เป็นอัลกอริทึมของการแปลงฟูริเยร์แบบวิยุต ซึ่งสามารถลดเวลาในการคำนวณได้สำหรับกลุ่มตัวอย่างสัญญาณจำนวน N จุดจาก $2N^2$ ให้เหลือเพียง $2N\log N$ เท่านั้น อัลกอริทึมของการแปลงฟูริเยร์แบบเร็วสามารถนั้นแยกออกเป็นสองคลาส คือ การลดจำนวนของหน่วยเวลา และการลดจำนวนของหน่วยความถี่

ในการลดจำนวนของหน่วยเวลา โดยการใช้การแปลงฟูริเยร์แบบเร็วด้วยอัลกอริทึมของคูเลย์-เตอกีย์ (Cooley-Turkey) จะทำการจัดเรียงข้อมูลที่เข้ามาให้อยู่ในรูปของบิตที่เรียงลำดับถอยหลัง (Bit-reversed Order) แล้วทำการคำนวณผลลัพธ์ออกมาด้วยวิธีนี้จะเป็นการแบ่งการแปลงฟูริเยร์ขนาดความยาว N ให้ออกเป็นการแปลงฟูริเยร์สองส่วนที่มีความยาวเป็น $N/2$

$$\sum_{n=0}^{N-1} a_n e^{-2\pi ink/N} = \sum_{n=0}^{N/2-1} a_{2n} e^{-2\pi i(2n)k/N} + \sum_{n=0}^{N/2-1} a_{2n+1} e^{-2\pi i(2n+1)k/N}$$

$$= \sum_{n=0}^{N/2-1} a_n^{even} e^{-2\pi ink/(N/2)} + \sum_{n=0}^{N/2-1} a_n^{odd} e^{-2\pi ink/(N/2)}$$

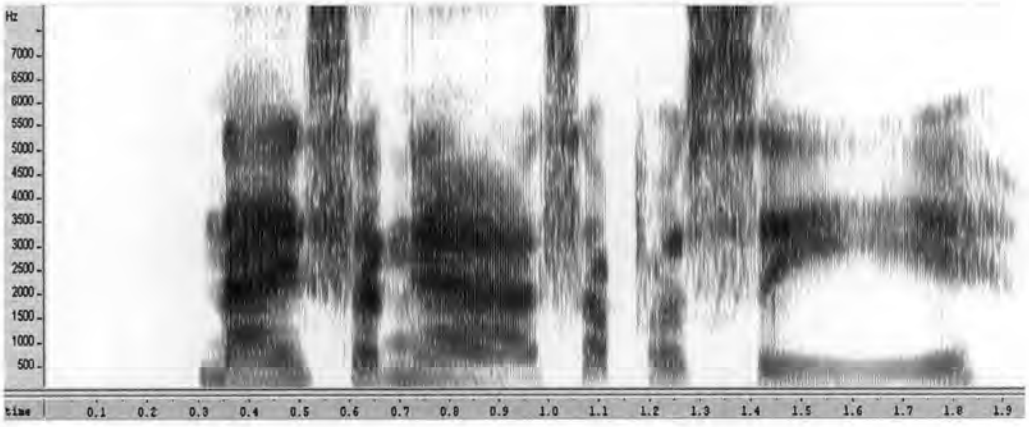
5. สเปกโตรแกรมของเสียงพูด

สัญญาณเสียงมีลักษณะรูปร่างเป็นคลื่นที่สั่นแกว่งไปมา เราไม่สามารถจะอ่านหน่วยเสียงในรูปแบบของคลื่นในโดเมนเวลา แต่ถ้าเราวิเคราะห์รูปแบบคลื่นในโดเมนความถี่ เราจะได้สเปกโตรแกรมซึ่งสามารถจะนำมาถอดรหัสได้

ช่วงความถี่ที่มนุษย์สามารถได้ยินจะอยู่ระหว่าง 20 - 20000 เฮิร์ต (20 กิโลเฮิร์ต) มนุษย์ไม่สามารถได้ยินความสั่นสะเทือนซึ่งเกิดขึ้นที่ความถี่ต่ำกว่า 20 ครั้งต่อวินาที และไม่สามารถรับรู้ความถี่ที่สูงกว่า 20 กิโลเฮิร์ต เสียงคำพูดจะประกอบด้วยพลังงานที่ระดับความถี่ต่างๆในช่วงที่มนุษย์สามารถได้ยินได้ โดยที่เสียงทั้งหมดจะอยู่ที่ระดับต่ำกว่า 8,000 เฮิร์ต

เรานำเทคนิคทางคณิตศาสตร์ที่เรียกว่าการวิเคราะห์ฟูริเยร์มาใช้กับรูปแบบคลื่นคำพูด เพื่อที่จะหาว่า ความถี่ใดที่เกิดขึ้นในเวลาต่างๆ ในสัญญาณคำพูด ผลจากการทำการวิเคราะห์ฟูริเยร์ก็คือ สเปกตรัม (Spectrum) หลังจากเราคำนวณสเปกตรัมสำหรับช่วงเวลาสั้นๆ (5-20 มิลลิวินาที) ของคำพูด เราจะคำนวณสเปกตรัมของช่วงต่อไปเรื่อยๆจนสิ้นสุดรูปแบบของคลื่นโดยทั่วๆไป สเปกตรัมที่อยู่ติดกันจะเปลี่ยนแปลงอย่างช้าๆและราบรื่น สะท้อนถึง (ชี้ให้เห็นถึง) การเคลื่อนไหวอย่างช้าๆของเสียงเทียบกับช่วงเวลาที่เราวิเคราะห์

การวิเคราะห์ดังกล่าวเรียกว่า การวิเคราะห์ฟูริเยร์ในช่วงเวลาสั้น (Short-time Fourier analysis) ซึ่งชุดของสเปกตรัมที่ได้มาจากช่วงเวลาต่างๆ นั้น สามารถนำไปแสดงผลได้ในรูปแบบของสเปกโตรแกรม

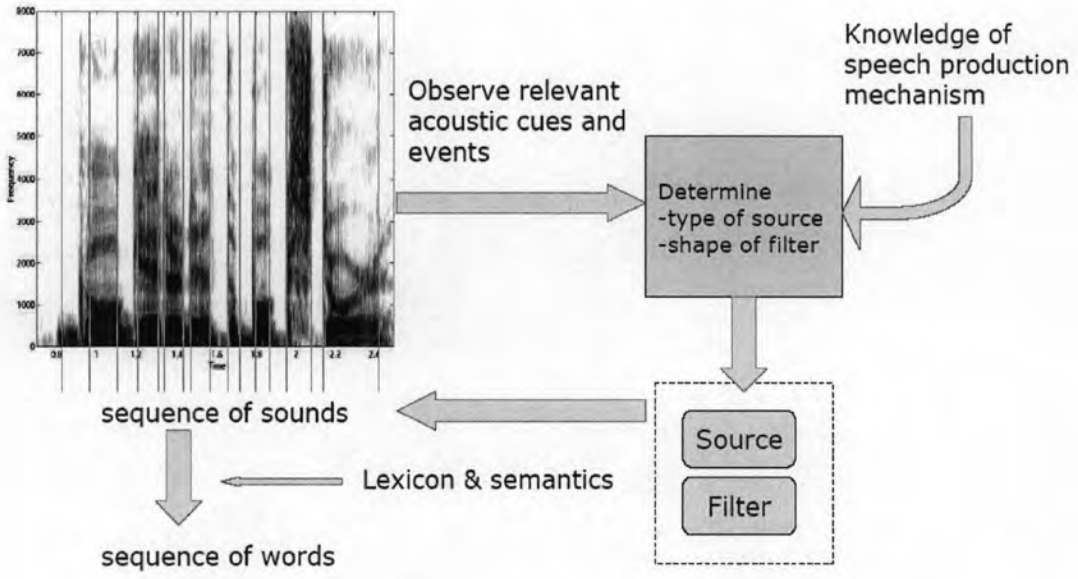


รูปที่ 2.8 สเปกโตรแกรมของเสียงพูดคำว่า "นายสง่าสรรพศรี"

สเปกโตรแกรมตัวอย่างในรูปที่ 2.8 เกิดขึ้นโดยการแสดงสเปกตรัมทั้งหมดที่คำนวณจากรูปแบบคลื่นของคำพูด แกนตั้งของสเปกตรัมแสดงความถี่โดยที่ระดับฐานจะเท่ากับ ศูนย์เฮิร์ต เส้นที่มองเห็นได้ในสเปกโตรแกรมแต่ละเส้นแสดงระดับ 1000 เฮิร์ต ตามแกนความถี่ ดังนั้น สเปกโตรแกรมจะประกอบด้วยความถี่ทั้งหมด 8000 เฮิร์ต สเปกตรัมที่คำนวณจากการแปลงฟูริเยร์ทั้งหมดจะถูกแสดงขนานกับแกนตั้ง แกนนอนแสดงถึงเวลา การเลื่อนไปทางขวาตามแกน x แสดงถึง

สเปกตรัมตามเวลาที่เพิ่มขึ้น สเปกโตรแกรมจะถูกคำนวณค่าพลังงานของเสียงและเก็บไว้ในอาเรย์ขนาดสองมิติ สำหรับสเปกโตรแกรม S ใดๆ ความแรงของสัญญาณความถี่ f ที่เวลา t ในสัญญาณเสียงพูดจะแสดงโดยความเข้มหรือสีในกราฟที่จุด $S(t,f)$

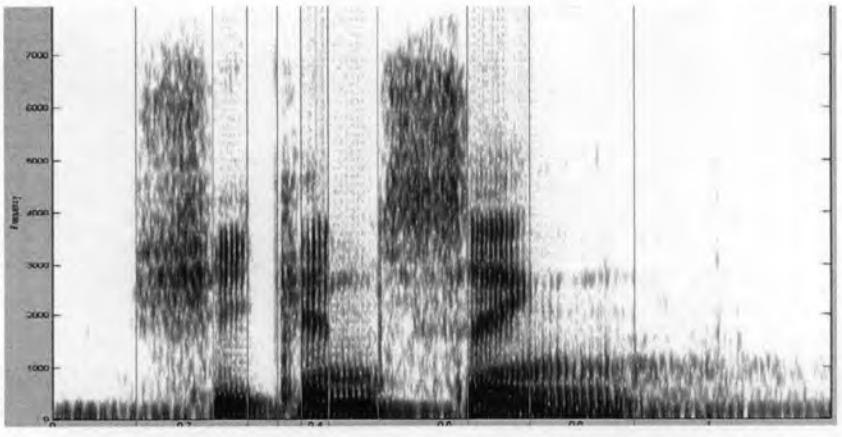
การอ่านสเปกโตรแกรมจะใช้พื้นฐานความรู้ทางด้านการออกเสียงคำพูดเพื่อแบ่งแยกลำดับของคำพูดจากสัญญาณเสียงที่ส่งเข้ามา โดยวิเคราะห์จากสเปกโตรแกรม



รูปที่ 2.9 การแบ่งแยกเสียงจากสัญญาณเสียงที่ได้รับเข้ามา

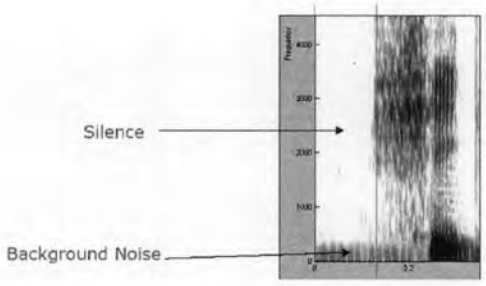
จากรูปที่ 2.9 จะเห็นได้ว่าเมื่อเราได้รับชุดลำดับของเสียงเข้ามา เราจะทำการแบ่งแยกชุดเสียงโดยใช้หลักเกณฑ์ต่างๆ เช่น ประเภทของแหล่งกำเนิด, รูปทรงของตัวกรองเสียง และจะใช้ความยาวกับไวยกรณ์เพื่อแปลความหมายจากชุดลำดับของเสียง ให้เกิดเป็นชุดลำดับของคำ โดยขั้นตอนการแปลความหมายของสเปกโตรแกรมมีดังนี้

1. ทำการกำกับขอบเขตหน่วยเสียงซึ่งสามารถพิจารณาได้จากสเปกโตรแกรม



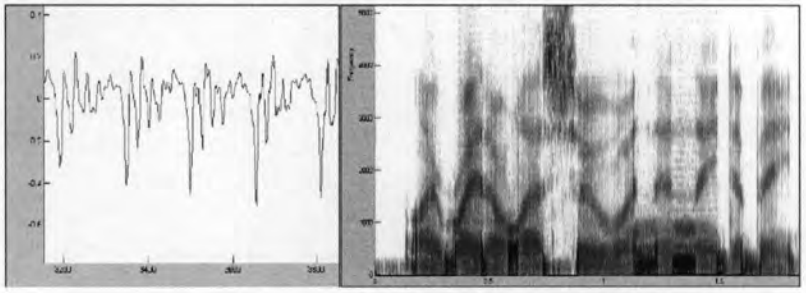
รูปที่ 2.10 การกำกับขอบเขตหน่วยเสียง

2. ทำการแบ่งประเภทกลุ่มของเสียง (Classes of Sound) ซึ่งประกอบด้วย
   1. เสียงเงียบ เป็นเสียงที่ไม่มีพลังงานใดๆ



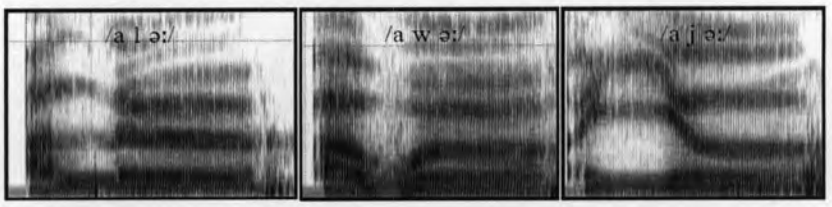
รูปที่ 2.11 สเปกโตรแกรมแสดงความแตกต่างระหว่างเสียงเงียบกับเสียงประเภทอื่นๆ

   2. เสียงสระ (Vowels) จะมีพลังงานสูง และลมมาก ซึ่งมีรูปแบบที่ชัดเจน



รูปที่ 2.12 สเปกโตรแกรมแสดงสัญญาณเสียงสระ

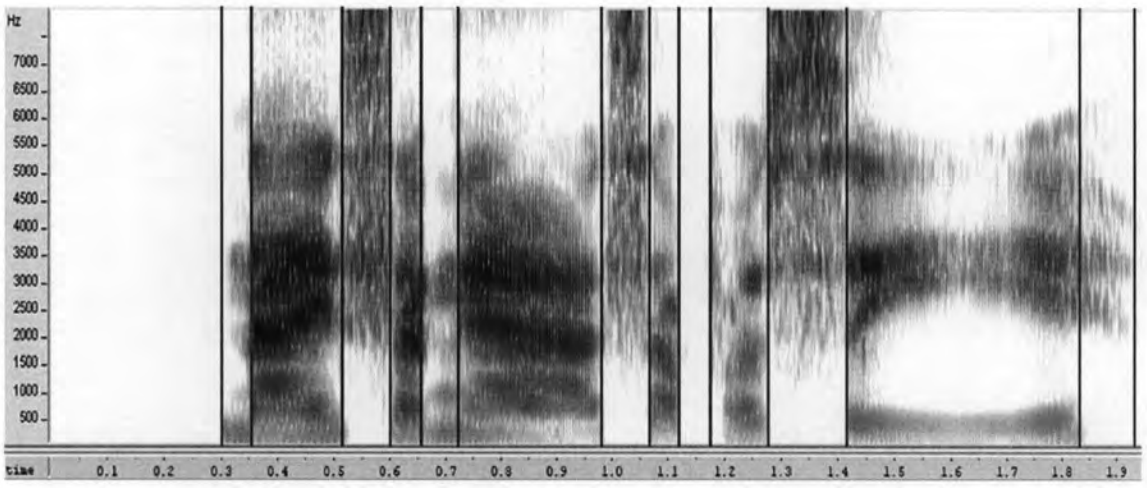
3. เสียงพยัญชนะ (Consonants)
   1. เสียงพยัญชนะกัก (Stop Consonant) ไม่มีพลังงานในช่วงปิด และมีเสียงระเบิดในช่วงเปิดปาก
   2. เสียงเสียดแทรก (Fricatives) มีรูปของเสียงรบกวน (noise) ในช่วงต้น
   3. เสียงกึ่งเสียดแทรก (Affricates) เป็นเสียงกักและต่อด้วยเสียงเสียดแทรก
   4. เสียงนาสิก (Nasal Consonants) พลังงานช่วงความถี่กลางและสูงมีน้อย ความถี่ F1 เพิ่มขึ้น
   5. เสียงกึ่งสระ (Semi-Vowel) มีการเคลื่อนตำแหน่งของความถี่ฟอร์แมนท์เร็วกว่าเสียงสระแต่ไม่มากเท่าเสียงพยัญชนะ



รูปที่ 2.13 สเปกโตรแกรมแสดงสัญญาณเสียงกึ่งสระ

6. ขอบเขตของหน่วยเสียง

ขอบเขตของหน่วยเสียงคือตำแหน่งบอกเวลาเริ่มต้นและสิ้นสุดของหน่วยเสียง ดังแสดงได้ด้วยสเปกโตรแกรมในรูปที่ 2.14 โดยเส้นตรงแต่ละเส้นแทนขอบเขตของหน่วยเสียง



รูปที่ 2.14 สเปกโตรแกรมแสดงการกำกับขอบเขตของหน่วยเสียง

การหาขอบเขตของหน่วยเสียงสามารถทำได้โดยอาศัยการอ่านสเปกโตรแกรม วิเคราะห์รูปแบบสเปกโตรแกรมของหน่วยเสียงแต่ละประเภท เช่นขอบเขตของเสียงสระจะเป็นส่วนที่มีแถบเข้มมากเนื่องจากเป็นเสียงที่มีพลังงานมาก เป็นต้น

ข้อมูลเสียงที่มีการกำกับขอบเขตของหน่วยเสียงไว้แล้วนี้ สามารถนำไปใช้ประโยชน์ด้านการสร้างระบบรู้จำเสียงพูด หรือการสร้างคลังข้อมูลเสียงสำหรับระบบสังเคราะห์เสียงได้

7. การสกัดลักษณะสำคัญ

การสกัดค่าลักษณะสำคัญ คือการวิเคราะห์หาค่าที่จะใช้แทนสัญญาณเสียง เพื่อนำไปใช้ในขั้นตอนการรู้จำ แบ่งได้เป็น 3 กลุ่มหลัก กลุ่มแรกเป็นค่าลักษณะสำคัญระดับสูง (High level feature) ได้แก่ สำเนียงการพูด รูปแบบในการพูด และความเร็วในการพูด เป็นต้น ในกลุ่มที่สอง จะใช้ค่าลักษณะสำคัญทางฉันทลักษณ์ (Prosodic feature) เช่น ค่าความถี่มูลฐาน (Fundamental frequency) ความถี่ฟอร์แมนท์ (Formant frequency) และระดับพลังงาน (Energy profile) เป็นต้น ถึงแม้ว่าค่าลักษณะสำคัญแบบนี้จะมีประสิทธิภาพสูงในการรู้จำ แต่ยากในการสกัดจากสัญญาณ กลุ่มสุดท้ายเรียกว่าค่าลักษณะสำคัญแบบเอนเวโลปเชิงสเปกตรัม (Spectral envelop feature) [10] เป็นกลุ่มที่นิยมใช้กันมาก เนื่องจากค่าลักษณะสำคัญส่วนใหญ่สำหรับการรู้จำเสียงจะรวมอยู่ในข้อมูลเชิงสเปกตรัมนี้ อีกทั้งยังง่ายและสะดวกในการคำนวณหาค่าด้วย ตัวอย่างค่าลักษณะสำคัญแบบนี้ได้แก่ สัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear prediction coefficients: LPC), สัมประสิทธิ์เซปสตรัม (Cepstral coefficient) และพัฒนาการอีกมากมายจากเซปสตรัมปกติ [11] อาทิเช่น สัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel frequency cepstral coefficients: MFCC) เซปสตรัมแบบหักลบค่าเฉลี่ย (Cepstral mean substraction: CMS) และเซปสตรัมแบบผ่านตัวกรองภายหลัง (Post filtered cepsturm: PFL) เป็นต้น นอกจากนั้น ยังมีการคำนวณค่าการเปลี่ยนแปลง (Derivative หรือ Delta) ของสัมประสิทธิ์เหล่านี้มาใช้เป็นค่าลักษณะสำคัญเพิ่มเติมได้ด้วย

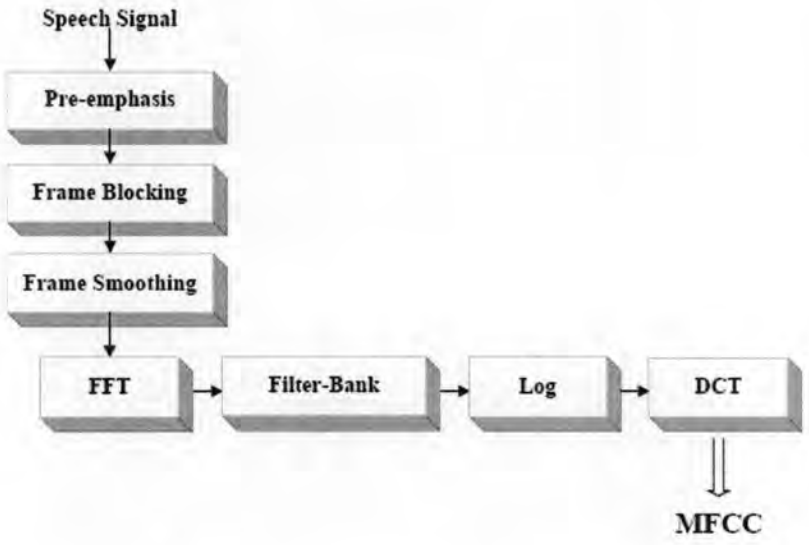
สำหรับการคำนวณค่าลักษณะสำคัญแบบเอนเวโลปของสเปกตรัมจะมีขั้นตอนดังนี้ [12]

1. การเน้นสัญญาณขั้นต้น (Preemphasis) เป็นขั้นตอนในการบีบอัดสัญญาณเสียงโดยนำสัญญาณเสียงผ่านตัวกรองลำดับหนึ่ง (First-order filter) ซึ่งจะเพิ่มอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to noise ratio)
2. การแบ่งเป็นส่วนย่อย (Frame) เป็นขั้นตอนในการแบ่งสัญญาณเสียงเป็นส่วนย่อยขนาดความยาวประมาณ 10 – 40 มิลลิวินาที ซึ่งทำให้สัญญาณเสียงมีคุณสมบัติเปลี่ยนแปลงตามเวลาน้อยมาก หรือไม่มีเลย เพื่อให้สามารถสร้างแบบจำลองการกระจายของหน่วยสัญญาณเสียงย่อยทางสถิติได้
3. การลดขอบด้วยฟังก์ชันหน้าต่างสำหรับปรับสัญญาณให้ราบเรียบ (Smoothing window)
4. การสกัดค่าลักษณะสำคัญ (Feature extraction) ในส่วนนี้ จะทำการคำนวณค่าลักษณะสำคัญของสัญญาณเสียงในแต่ละส่วนย่อย ผลลัพธ์อยู่ในรูปแบบของเวกเตอร์ของค่าลักษณะสำคัญ (Feature vector) สำหรับแต่ละส่วนย่อย

7.1 สัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel frequency cepstral coefficients - MFCC)

ค่าสัมประสิทธิ์เซปสตรัมเป็นค่าลักษณะสำคัญที่นิยมมากทั้งในระบบรู้จำผู้พูดและเสียงพูด โดยพื้นฐานแล้วเซปสตรัมสามารถคำนวณได้จาก การแปลงโคซายน์แบบไม่ต่อเนื่อง (Discrete cosine transformation) ของค่าลอการิธึม (Logarithm) ของสเปกตรัม ของสัญญาณเสียงแต่ละส่วนย่อย สเปกตรัมของสัญญาณเสียงสามารถหาได้โดยการแปลงฟูริเยร์แบบวิยุต หรือการแปลงฟูริเยร์แบบเร็ว ขั้นตอนดังกล่าวตั้งอยู่บนพื้นฐานแนวคิดที่ว่า สเปกตรัมของสัญญาณเสียงกำเนิดจากส่วนประกอบ 2 ส่วนคือ เอนเวโลปของสเปกตรัม (Spectral envelop) และโครงสร้างรายละเอียดของสเปกตรัม (Spectral fine structure) ทั้ง 2 ส่วนสามารถแยกกันได้ด้วยการใส่ลอการิธึม สัมประสิทธิ์เซปสตรัมเป็นการแทนสัญญาณในส่วนเอนเวโลปของสเปกตรัมเท่านั้น

พัฒนาการหนึ่งของเซปสตรัม คือการผ่านสเปกตรัมของสัญญาณเสียงเข้าไปในกลุ่มของตัวกรอง (Filter bank) ซึ่งกระจายอยู่บนสเกลความถี่แบบไม่สม่ำเสมอ เช่น การกระจายตามสเกลเมล (Mel scale) [10] ซึ่งออกแบบมาให้เหมาะสมกับการรับฟังของหู เป็นต้น ค่าพลังงานของสเปกตรัมของเสียงที่ได้จากตัวกรองแต่ละตัวจะถูกนำมาใช้คำนวณค่าสัมประสิทธิ์เซปสตรัมแทนค่าสเปกตรัมปกติ ค่าสัมประสิทธิ์เซปสตัมที่ได้จากการกระทำเช่นนี้จึงได้ชื่อว่า MFCC



รูปที่ 2.15 ขั้นตอนการคำนวณค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล

8. แบบจำลองฮิดเดนมาร์คอฟ

ระบบแบบจำลองฮิดเดนมาร์คอฟ เป็นขั้นตอนวิธีการจำแนกรูปแบบที่ดีที่สุดวิธีการหนึ่งที่มีอยู่ในขณะนี้ [13] ซึ่งอาศัยวิธีการทางสถิติ เหตุผลที่ระบบแบบจำลองฮิดเดนมาร์คอฟ เป็นที่นิยมมีด้วยกันสองประการคือ

ประการแรก แบบจำลองนี้อาศัยโครงสร้างทางคณิตศาสตร์ และสามารถเปลี่ยนแปลงทฤษฎีพื้นฐานเพื่อประยุกต์ใช้งานได้อย่างกว้างขวาง

ประการที่สอง แบบจำลองนี้สามารถทำงานได้เป็นอย่างดีเมื่อประยุกต์ใช้อย่างเหมาะสม โดยเราจะพิจารณาเนื้อหาของแบบจำลองฮิดเดนมาร์คอฟ โดยแบ่งออกเป็นส่วนย่อยๆ ดังนี้

8.1 องค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟ

แบบจำลองฮิดเดนมาร์คอฟ ประกอบไปด้วยพารามิเตอร์ต่างๆดังนี้

1. จำนวนสถานะ (State) ที่อยู่ภายในแบบจำลอง (Model) ใช้สัญลักษณ์แทนด้วย " N " แต่ละสถานะแสดงได้ด้วย $S=\{S_1,S_2,...,S_N\}$ โดยมีสถานะที่เวลา t แสดงได้ด้วย q_t มีค่าแปรเปลี่ยนได้ ตามที่กำหนดจนกว่าจะได้ผลลัพธ์การรู้จำที่น่าพอใจ
2. จำนวนสัญลักษณ์ของค่าสังเกตต่อสถานะ ใช้สัญลักษณ์แทนด้วย " M " แต่ละสัญลักษณ์แสดงได้ด้วย $V=\{v_1,v_2,...,v_M\}$
3. การแจกแจงของความน่าจะเป็นในการเปลี่ยนสถานะ (State Transition Probability Distribution) ใช้สัญลักษณ์แทนด้วยเมตริก (Matrix) " A " โดย $A=\{a_{ij}\}$ เมื่อ $a_{ij}=P[q_{t+1}=S_j \mid q_t=S_i],\ 1\le i,j<N$ ในกรณีเฉพาะที่สถานะใดๆ สามารถเข้าถึงสถานะอื่นๆได้ภายในขั้นตอนเดียว จะกำหนดให้ ส่วนในกรณีอื่นนอกเหนือจากนี้จะกำหนดให้ สำหรับ (i,j) เพียงคู่เดียวหรือมากกว่า
4. การแจกแจงของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต (Observation Symbol Probability Distribution) ใช้สัญลักษณ์แทนด้วยเมตริก (Matrix) " B " โดย $B=\{b_j(k)\}$ ในสถานะที่ j เมื่อ

$$b_j(k)=P[v_k \text{ at } t \mid q_t=S_j],\ 1\le j\le N, 1\le k\le M$$

5. การแจกแจงสถานะเริ่มต้น (Initial State Distribution) ซึ่งก็คือความน่าจะเป็นของแบบจำลองที่จะเริ่มต้นแบบจำลองด้วยสถานะ i ใดๆ ใช้สัญลักษณ์แทนด้วย " π " โดย $\pi=\pi_i$ เมื่อ

$$\pi_i=P[q_1=S_i],\ 1\le i\le N$$

โดยการกำหนดค่าที่เหมาะสมให้กับองค์ประกอบ N, M, A, B, π ของแบบจำลองฮิดเดนมาร์คอฟซึ่งใช้ในการกำหนดลำดับค่าสังเกต เมื่อแต่ละค่าสังเกต O_t เป็นสัญลักษณ์ที่ได้จาก V และ T เป็นจำนวนค่าสังเกตทั้งหมดที่มีในลำดับซึ่งมีขั้นตอนวิธีการดังนี้

ตารางที่ 2.5 ขั้นตอนการกำหนดค่าองค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟ

ขั้นตอนที่	รายละเอียดการดำเนินการ
1	เลือกสถานะเริ่มต้น $q_1 = S_i$ ที่สัมพันธ์กับการกระจายของสภาวะเริ่มต้น π
2	กำหนดให้ $t = 1$
3	เคลื่อนย้ายไปยังสถานะใหม่ $q_{t+1} = S_j$ ที่สัมพันธ์กับการกระจายความน่าจะเป็นในการเปลี่ยนแปลงสถานะสำหรับสถานะ S_i เช่น a_{ij}
4	กำหนดให้ $t = t+1$ แล้วกลับไปทำซ้ำขั้นตอนที่ 3 ใหม่ถ้า $t < T$ นอกเหนือจากนี้ให้ยุติกระบวนการ

ขั้นตอนดังกล่าวนี้เป็นได้ทั้งการกำเนิดค่าสังเกต และเป็นแบบจำลองเพื่อบอกถึงความเหมาะสมในการกำเนิดลำดับค่าสังเกตด้วยแบบจำลองฮิดเดนมาร์คอฟ ดังนั้นการกำหนดคุณสมบัติเฉพาะของแบบจำลองฮิดเดนมาร์คอฟ จึงต้องการคุณสมบัติเฉพาะของพารามิเตอร์ของแบบจำลอง (นั่นคือ N และ M) คุณสมบัติเฉพาะของสัญลักษณ์ของค่าสังเกต และ คุณสมบัติเฉพาะของการวัดค่าความน่าจะเป็นอันได้แก่ A, B, π โดยทั้งหมดนี้สามารถเขียนให้อยู่ในรูปแบบ แบบย่อเพื่อบ่งบอกถึงชุดพารามิเตอร์ที่สมบูรณ์ของแบบจำลองได้ดังนี้

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

แบบจำลองฮิดเดนมาร์คอฟก็เปรียบเสมือนเครื่องจักรสถานะ (State Machine) แบบหนึ่งซึ่งนำไปใช้อธิบายรูปแบบการเกิดของเวกเตอร์คุณสมบัติของเสียง (Feature Vector) โดยการที่จะรู้ว่าเวกเตอร์คุณสมบัติของเสียงนั้นเกิดได้อย่างไรสามารถทำได้ด้วยการย้อนรอยสถานะเพื่อหาเส้นทางการรู้จำของเสียงว่าเวกเตอร์คุณสมบัติของเสียงนั้นได้มาจากหน่วยเสียงใดบ้างตามสถานะที่เสียงนั้นๆผ่าน

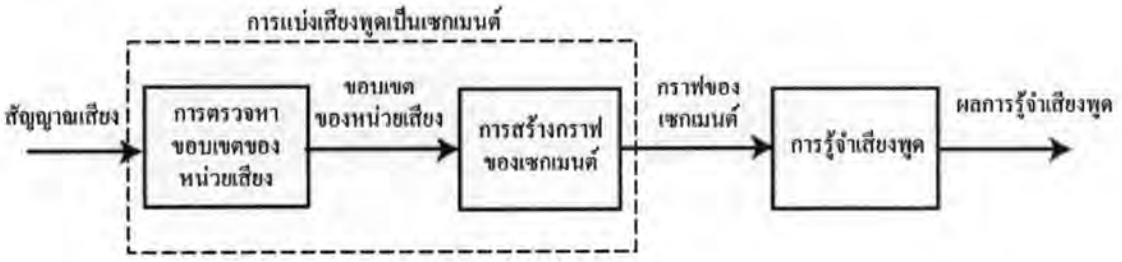
9. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

กว่าสองทศวรรษที่แบบจำลองฮิดเดนมาร์คอฟได้รับความนิยมมากในการนำมาใช้กับการพัฒนาเครื่องรู้จำเสียงพูด สาเหตุหนึ่งมาจากการที่แบบจำลองฮิดเดนมาร์คอฟมีรากฐานทางคณิตศาสตร์ที่ออกแบบมาเป็นอย่างดี และสามารถนำไปใช้แก้ปัญหาการรู้จำเสียงพูดได้เป็นอย่างมีประสิทธิภาพ แต่แบบจำลองเสียงพูดในเครื่องรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟมีคุณสมบัติสำคัญที่ทำให้เกิดข้อจำกัดสองประการ [14] คือ

คุณสมบัติประการแรก คือการพิจารณาลำดับของเวกเตอร์ลักษณะสำคัญที่คำนวณมาจากสัญญาณเสียงในแต่ละกรอบเวลาสั้นๆที่มีขนาดตายตัว (โดยทั่วไปมีขนาด 10 มิลลิวินาทีต่อหนึ่งกรอบเวลา) แต่ในกระบวนการสร้างเสียงพูด อวัยวะส่วนกระทำอาการจะเคลื่อนไหวช้า ทำให้เวกเตอร์ลักษณะสำคัญที่คำนวณจากสัญญาณเสียงในกรอบเวลาที่อยู่ติดกันมีความคล้ายคลึงกันและความเกี่ยวพันกันสูงโดยเฉพาะอย่างยิ่งเวกเตอร์ลักษณะสำคัญที่คำนวณมาจากสัญญาณเสียงพูดที่มีหน่วยเสียงเหมือนกัน ซึ่งขัดกับคุณสมบัติหนึ่งของแบบจำลองฮิดเดนมาร์คอฟที่สมมุติให้เวกเตอร์ลักษณะสำคัญแต่ละเวกเตอร์เป็นอิสระต่อกันแบบมีเงื่อนไข

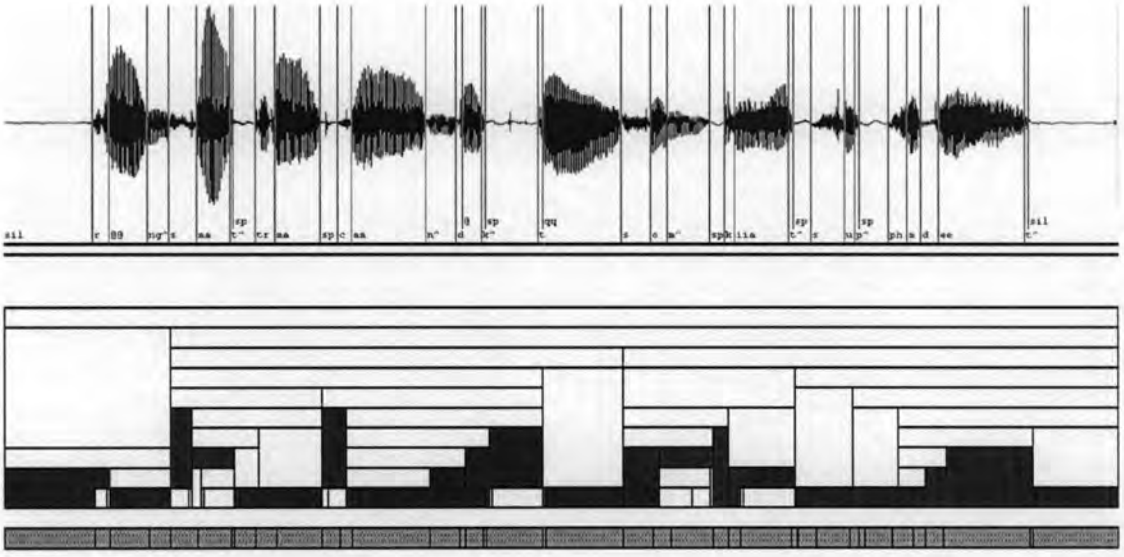
คุณสมบัติประการที่สอง คือการคำนวณหาเวกเตอร์ลักษณะสำคัญจากสัญญาณเสียงในแต่ละกรอบเวลาจะต้องเลือกใช้ลักษณะสำคัญที่เหมือนกันทั้งหมด (เช่น ใช้สัมประสิทธิ์เซปสตรัมบนสเกลเมลเป็นลักษณะสำคัญเพียงแบบเดียว) ทำให้ยากต่อการนำสารสนเทศทางสวนสัทศาสตร์ หรือลักษณะสำคัญอื่นๆ เช่น เสียงวรรณยุกต์ มารวมกันเพื่อเพิ่มประสิทธิภาพการรู้จำเสียงพูดให้ดีขึ้น

เพื่อผ่อนปรนข้อจำกัดเหล่านี้ จึงมีการเสนอวิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ [1] ซึ่งเป็นวิธีการรู้จำเสียงพูดทางสถิติ ซึ่งแตกต่างจากการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟตรงที่การรู้จำเสียงพูดจะพิจารณาข้อมูลเสียงเป็นเซกเมนต์ โดยที่แต่ละเซกเมนต์ไม่จำเป็นต้องมีขนาดเท่าๆกัน แล้วจำแนกว่าแต่ละเซกเมนต์นั้นมีหน่วยเสียงเป็นอะไร ด้วยเหตุนี้สิ่งที่จำเป็นและส่งผลต่อประสิทธิภาพของเครื่องรู้จำเสียงแบบอาศัยเซกเมนต์โดยตรง คือการมีข้อมูลเสียงที่มีการกำกับขอบเขตของหน่วยเสียงไว้ก่อนแล้ว ซึ่งเป็นผลมาจากกระบวนการแบ่งเสียงพูดเป็นเซกเมนต์ที่มีความถูกต้องแม่นยำและทำงานได้อย่างรวดเร็ว



รูปที่ 2.16 แผนภาพส่วนประกอบของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์

ภาพรวมส่วนประกอบของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์แสดงได้ด้วยรูปที่ 2.16 การรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะพิจารณาหน่วยเสียงให้เป็นเซกเมนต์ โดยขั้นตอนการทำงานจะประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนที่ทำงานต่อเนื่องกัน คือ ขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ และขั้นตอนการรู้จำเสียงพูด จุดประสงค์ของขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ ก็เพื่อสันนิษฐานหาขอบเขตของเซกเมนต์แล้วนำมาประกอบกันเป็นกราฟของเซกเมนต์ โดยในขั้นตอนการรู้จำเสียงพูด กราฟนี้จะถูกใช้เป็นข้อมูลอินพุตเพื่อค้นหาลำดับของหน่วยเสียงที่ดีที่สุดออกมา ดังรูปที่ 2.17

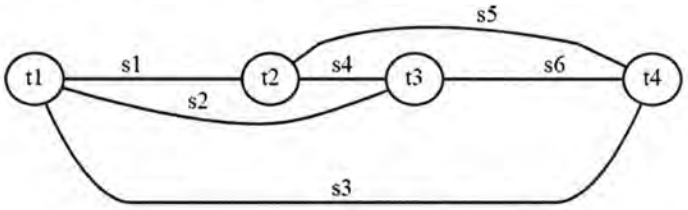


รูปที่ 2.17 สัญญาณเสียงที่กำกับขอบเขตของหน่วยเสียงไว้แล้ว (บน) กราฟของเซกเมนต์ (ล่าง)
(สี่เหลี่ยมสีเข้มแทนเซกเมนต์ที่ให้ระบบรู้จำเสียงพูดเลือกเป็นคำตอบ สี่เหลี่ยมสีขาวแทนเซกเมนต์ของหน่วยเสียงอื่นๆที่เป็นไปได้ และสี่เหลี่ยมสีเทาแทนเซกเมนต์ของเสียงพูดที่ถูกต้อง)

9.1 การแบ่งเสียงพูดเป็นเซกเมนต์

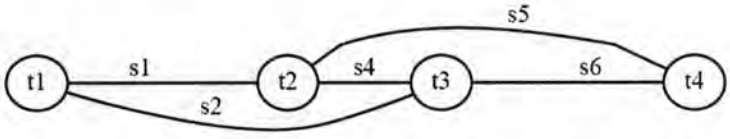
การแบ่งเสียงพูดเป็นเซกเมนต์ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ คือการหาขอบเขตของเซกเมนต์ในที่นี้คือหน่วยเสียง แล้วนำมาประกอบกันสร้างเป็นกราฟของเซกเมนต์ โดยการทำงานจะประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนคือ ขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง และขั้นตอนการสร้างกราฟของเซกเมนต์ โดยในแต่ละขั้นตอนสามารถใช้วิธีการได้หลายวิธี ตัวอย่างเช่น อาจใช้วิธีรู้จำเสียงพูดในระดับหน่วยเสียงออกมาก่อนแล้วจึงพิจารณาเวลาเริ่มต้นหรือสิ้นสุดลงของหน่วยเสียงที่รู้จำออกมาได้ให้เป็นเขตของหน่วยเสียง หรืออาจใช้วิธีดูการเปลี่ยนแปลงสเปกตรัมของสัญญาณเสียงแล้วพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงมากเกินกว่าระดับที่กำหนด

ไว้ให้เป็นขอบเขตของหน่วยเสียง หรืออาจใช้วิธีการจำแนกเสียงพูดออกเป็นประเภทกว้างๆ แล้วพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงประเภทหน่วยเสียงนั้นๆให้เป็นขอบเขตของหน่วยเสียงก็ได้



รูปที่ 2.18 ตัวอย่างกราฟของเซกเมนต์แบบเชื่อมต่อกันหมด

ลำดับของขอบเขตของหน่วยเสียงที่ตรวจหามาได้ จะนำมาประกอบกันสร้างเป็นกราฟของเซกเมนต์ กราฟที่ดีจะต้องมีขนาดเล็กและครอบคลุมหน่วยเสียงที่แท้จริงไว้ได้เป็นจำนวนมาก ในทางทฤษฎีแล้วการรู้จำเสียงพูดแบบอาศัยเซกเมนต์สามารถใช้กราฟของเซกเมนต์ที่ทุกปมของกราฟเชื่อมต่อกันหมดมารู้จำเสียงพูดก็ได้ แต่เนื่องจากจำนวนเซกเมนต์ในกราฟนั้นมีขนาดใหญ่มาก ทำให้ขั้นตอนการรู้จำใช้เวลาในการทำงานมากตามไปด้วย ด้วยเหตุนี้วิธีการแบบอาศัยเซกเมนต์ส่วนใหญ่จึงต้องมีการลดขนาดของกราฟลงก่อนด้วยการตัดเส้นเชื่อมบางเส้นออกไป ตัวอย่างเช่นรูปที่ 2.18 เป็นกราฟของเซกเมนต์ที่ทุกปมเชื่อมต่อกันหมด และรูปที่ 2.19 แสดงกราฟของเซกเมนต์แบบที่ตัดเซกเมนต์ s_3 ออกไป



รูปที่ 2.19 กราฟของเซกเมนต์แบบที่ยอมให้มีการเชื่อมต่อกันบางส่วนเท่านั้น

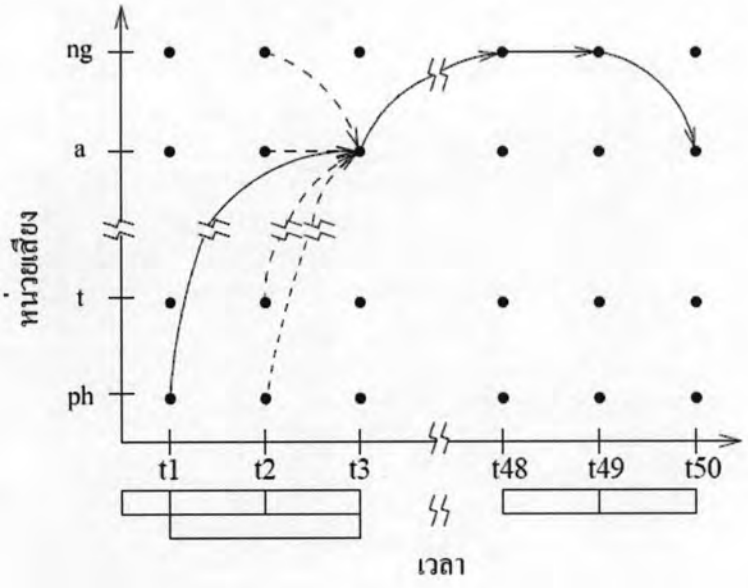
คุณภาพของกราฟที่ได้จากขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์จะเป็นการแลกเปลี่ยนกันในแง่ของขนาดของกราฟและความครอบคลุม ซึ่งจะส่งผลถึงเวลาที่ใช้ในการทำงานและความถูกต้องในการรู้จำเสียงพูดในขั้นตอนการค้นหาและรู้จำเสียงพูดต่อไป

9.2 การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จะเป็นการใช้อัลกอริทึมแบบกำหนดการพลวัต (Dynamic Programming) ค้นหาลำดับของเซกเมนต์ที่ดีที่สุดออกมาจากกราฟของเซกเมนต์ โดยอาศัยค่าลักษณะสำคัญทั้งแบบอาศัยกรอบเวลาและแบบอาศัยเซกเมนต์ ซึ่งค่าลักษณะสำคัญแบบอาศัยกรอบเวลาจะคำนวณมาจากสัญญาณเสียงในแต่ละกรอบเวลาสั้นๆที่มีขนาดตายตัว ตัวอย่างเช่น ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล หรือ ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น

เป็นต้น ส่วนค่าลักษณะสำคัญแบบอาศัยเซกเมนต์ จะคำนวณมาจากสัญญาณเสียงของแต่ละเซกเมนต์ ตัวอย่างเช่น ค่าความยาวของเซกเมนต์ เป็นต้น โดยค่าลักษณะสำคัญเหล่านี้จะนำมาใช้คำนวณเป็นค่าคะแนนสะท้อนความน่าจะเป็นว่าสัญญาณเสียงของแต่ละเซกเมนต์นั้นน่าจะเป็นหน่วยเสียงใด

การค้นหาลำดับของเซกเมนต์ที่ดีที่สุดออกมาจากกราฟของเซกเมนต์จะพิจารณากราฟของเซกเมนต์ควบคู่ไปกับหน่วยเสียงประเภทต่างๆ ดังรูปที่ 2.20 โดยในแนวแกนนอนคือเวลา แสดงลำดับของเซกเมนต์ แกนตั้งแสดงเซตของหน่วยเสียงทั้งหมด และเส้นโค้งเชื่อมต่อระหว่างปมในกราฟ จะเป็นเส้นทางเดินของเหตุการณ์ที่เซกเมนต์นั้นๆมีหน่วยเสียงเป็นอะไร ตัวอย่างเช่นเส้นโค้งที่เชื่อมต่อจาก t_1 ไป t_3 ซึ่งชี้ตรงไปยังปมของหน่วยเสียง a ก็คือเส้นทางการเดินของเหตุการณ์ที่เซกเมนต์ซึ่งมีเวลาเริ่มต้นอยู่ที่ t_1 และสิ้นสุดที่เวลา t_3 เป็นหน่วยเสียง a โดยที่แต่ละเส้นโค้งนั้นจะมีการคำนวณค่าคะแนนกำกับไว้ และการค้นหาลำดับของเซกเมนต์ที่ดีที่สุดจะต้องพิจารณาเส้นทางเดินในกราฟของเซกเมนต์ให้ครบทุกเส้นทางเพื่อเลือกเส้นทางที่มีคะแนนรวมสูงที่สุดเป็นผลการรู้จำเสียงพูด



รูปที่ 2.20 การค้นหาลำดับของเซกเมนต์จากกราฟของเซกเมนต์

10. ซัพพอร์ตเวกเตอร์แมชชีน – เอสวีเอ็ม [15]

ซัพพอร์ตเวกเตอร์แมชชีน [16] เป็นเทคนิคการเรียนรู้เชิงสถิติที่ Vapnik เริ่มคิดค้นตั้งแต่ช่วงปี ค.ศ. 1960 แต่ยังไม่เป็นที่นิยมจนถึงช่วงทศวรรษที่ผ่านมาเริ่มได้รับความสนใจมาก สาเหตุหนึ่งมาจากการที่มีการนำไปใช้แก้ปัญหาต่างๆและพบว่าได้ผลดีมาก เนื้อหาในหัวข้อนี้จะกล่าวถึงแนวคิดของเอสวีเอ็มและเอสวีเอ็มแบบหลายประเภท

10.1 แนวคิดพื้นฐานของซัพพอร์ตเวกเตอร์แมชชีน

เอสวีเอ็มเกิดจากแนวคิดพื้นฐานดังนี้

1. การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด (Structural risk minimization) เป็นแนวคิดที่แสดงขอบเขตของความเสี่ยงของเครื่องเรียนรู้ ซึ่งขึ้นกับความเสี่ยงเชิงทดลองที่มาจากความผิดพลาดในการสอน กับช่วงความเชื่อมั่นที่เป็นฟังก์ชันของมิติวีซี (VC dimension) ที่แสดงถึงว่าฟังก์ชันที่ใช้จำแนกมีลักษณะทั่วไปเพียงใด ในทางปฏิบัติเราไม่สามารถลดความเสี่ยงที่แท้จริงได้ จึงพยายามลดความเสี่ยงจากความผิดพลาดให้ต่ำสุดแทน
2. ระนาบหลายมิติแบ่งแยกดีสุด (Optimal separating hyperplane) ระนาบหลายมิตินี้จะต่างจากของข่ายงานประสาทเทียมตรงที่เป็นระนาบที่แบ่งคอนเวกซ์ฮัล (Convex hull) ของข้อมูลสองกลุ่มออกจากกันด้วยระยะห่างที่กว้างที่สุด
3. ฟังก์ชันเคอร์เนล (Kernel Function) เป็นเทคนิคที่ช่วยขยายความสามารถของเอสวีเอ็มให้จัดการกับปัญหาที่ไม่สามารถแบ่งแยกแบบเชิงเส้นได้ โดยการแมปข้อมูลในปริภูมินำเข้า (Input Space) ไปสู่ปริภูมิคุณลักษณะ (Feature Space) ที่มีอันดับสูงขึ้นไปซึ่ง ณ ปริภูมิอันดับสูงนี้ จะสามารถใช้ระนาบหลายมิติแบบเชิงเส้นในการแยกข้อมูลสองกลุ่มออกจากกันได้

10.2 การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด

ในปัญหาการรู้จำแบบเราต้องการหาฟังก์ชันที่มาประมาณตัวจำแนกประเภท (Classifier) ที่แท้จริง ซึ่งค่าความผิดพลาดคือผลต่างระหว่างค่าที่ได้จากฟังก์ชันประมาณกับค่าที่ได้จากฟังก์ชันที่แท้จริง ซึ่งโดยปกติเราจะต้องการลดความเสี่ยงจากการจำแนกประเภทผิดให้ต่ำที่สุด แต่ในทางปฏิบัติเราไม่รู้ฟังก์ชันที่แท้จริง จึงไม่สามารถคำนวณหาค่าผิดพลาดที่แท้จริงได้

ทางหนึ่งที่ทำได้คือ ในการสอนเราจะพิจารณาค่าผิดพลาดจากกลุ่มตัวอย่างแทน และลดค่าของความเสี่ยงเชิงโครงสร้างซึ่งประกอบด้วย ความเสี่ยงเชิงทดลอง (Empirical risk) กับช่วงความ

เชื่อมั่น (confidence interval) ให้ต่ำสุดแทน ความเสี่ยงเชิงโครงสร้างนี้มีพื้นฐานมาจากข้อเสนอของ Vapnik เรื่องขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่องกับมิติวีซี

10.2.1 ขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง

Vapnik ได้เสนอขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง โดยสมมติให้เรามีข้อมูลอยู่ l ตัวซึ่งแต่ละข้อมูลประกอบด้วยคู่ของ เวกเตอร์ $\mathbf{x}$ กับฉลากแสดงประเภท y โดยที่

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in \mathrm{R}^{\mathrm{N}} \times \{\pm 1\}$$

ตัวอย่างเช่น $\mathbf{x}$ อาจหมายถึงค่าสีของแต่ละจุดภาพของรูปตัวอักษรที่ต้องการรู้จำ ส่วน y แสดงถึงประเภทของรูปภาพว่าเป็นตัวอักษรที่สนใจหรือไม่ ถ้า y เป็น 1 หมายถึงว่าเป็นตัวอักษรที่สนใจ ส่วนถ้า y เป็น -1 ก็แสดงว่าไม่ใช่ตัวอักษรที่สนใจ และหากลองพิจารณาในแง่ของการรู้จำแบบ ภายใต้การแจกแจงความน่าจะเป็น $P(\mathbf{x}, y)$ ใดๆ สมมติว่าเราต้องการเครื่องที่จะเรียนรู้การจำแบบ $\mathbf{x}_i \Rightarrow y_i$ ซึ่งเรานิยามให้เป็นเซตของการรู้จำแบบที่เป็นไปได้ $\mathbf{x} \Rightarrow f(\mathbf{x}, \alpha)$ ซึ่ง α นี้เป็นตัวแปรที่สามารถเปลี่ยนแปลงค่าได้ และเครื่องที่ได้รับการสอนแล้วก็จะมีค่า α ที่แน่นอนของตัวเอง อย่างเช่นในข่ายงานประสาทเทียม ค่า α หมายถึงน้ำหนักและค่าขีดแบ่ง

เรานิยามค่าผิดพลาดสำหรับเครื่องที่ได้รับการสอนนี้เป็น

$$R(\alpha) = \int \frac{1}{2} |y - f(\mathbf{x}, \alpha)| \, dP(\mathbf{x}, y)$$

โดยที่ค่า $R(\alpha)$ เป็นความผิดพลาดที่แท้จริง และค่า $1/2|y - f(\mathbf{x}, \alpha)|$ จะมีค่าเป็น 0 หรือ 1 ส่วน $R_{emp}(\alpha)$ เป็นค่าความผิดพลาดโดยเฉลี่ยในข้อมูลสอนเท่านั้นและสามารถหาค่าได้ดังนี้

$$R_{emp}(\alpha) = \frac{1}{2l} \sum^{l} |y - f(\mathbf{x}, \alpha)|$$

ค่า $R_{emp}(\alpha)$ จะคงที่สำหรับ α และข้อมูลสอน $\{(\mathbf{x}_i, y_i)\}$ ชุดหนึ่งๆ เราเรียก $R_{emp}(\alpha)$ ว่าความเสี่ยงเชิงทดลอง

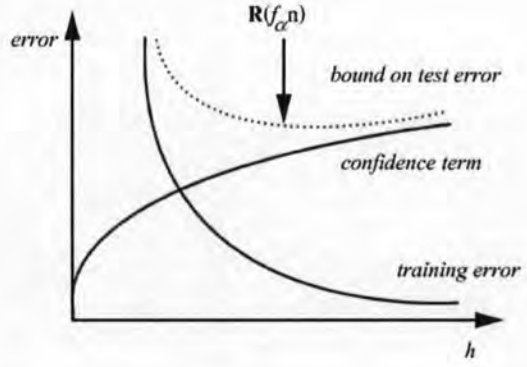
เมื่อเลือก η ตัวหนึ่งโดยที่ $0 \le \eta \le 1$ จะได้ว่า ที่ระดับความเชื่อมั่น $1 - \eta$ ความผิดพลาดที่แท้จริงจะมีขอบเขตดังนี้

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h\left(\log\left(\frac{2l}{h}\right) + 1\right) - \log\left(\frac{\eta}{4}\right)}{l}}$$

เราเรียก h ซึ่งเป็นจำนวนเต็มบวกหรือศูนย์ เรียกว่ามิติวีซี (Vapnik Chervonenkis dimension – VC dimension) และเรียกพจน์ขวามือว่าขอบเขตความเสี่ยง (Risk bound) ซึ่งพบว่าขอบเขตนี้ไม่ขึ้นกับการแจกแจงความน่าจะเป็น $P(\mathbf{x}_i, y)$

โดยทั่วไป เราไม่สามารถคำนวณค่าของพจน์ทางซ้ายมือได้แต่ถ้ารู้ h จะสามารถคำนวณพจน์ทางขวามือได้ ดังนั้นโดยหลักการแล้ว ในการเรียนรู้ เราจึงกำหนดค่า η เป็นค่าคงที่ต่ำๆแล้วเลือกเครื่องที่ลดค่าทางขวามือให้ต่ำที่สุด เราก็จะได้เครื่องที่ให้ขอบเขตความเสี่ยงต่ำที่สุด ซึ่งแนวคิดนี้เป็นแนวคิดที่สำคัญของการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด

เราเรียกพจน์ที่สองทางขวามือของสมการขอบเขตความผิดพลาดแท้จริงข้างต้นนี้ว่าช่วงความเชื่อมั่น พบว่าช่วงความเชื่อมั่นนี้เป็นฟังก์ชันของมิติ VC ที่แทนด้วย h ซึ่งการลดค่า h ให้ต่ำที่สุดจะเป็นการลดความเสี่ยงเชิงโครงสร้างด้วย แต่โดยปกติเมื่อค่า h ลดลงความเสี่ยงเชิงทดลองจะสูงขึ้นตามดังนั้นในการลดความเสี่ยงเชิงโครงสร้างจึงต้องหาจุดที่ผลจากทั้งความเสี่ยงเชิงทดลองและช่วงความเชื่อมั่นต่ำที่สุด โดยเลือกฟังก์ชันในการเรียนรู้เครื่อง $f(\mathbf{x},\alpha)$ ที่มีขอบเขตความเสี่ยงต่ำสุดดังในรูปที่ 2.21

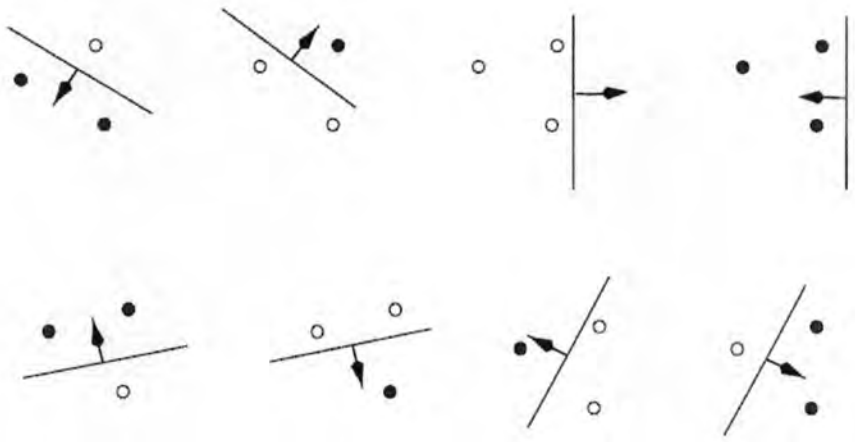


รูปที่ 2.21 ความสัมพันธ์ระหว่าง VC(h) กับค่าผิดพลาด

10.2.2 มิติวีซี

มิติวีซีเป็นคุณลักษณะของเซตของฟังก์ชัน $\{f(\alpha)\}$ ที่แสดงถึงความสามารถของฟังก์ชันในการแบ่งแยกกลุ่มของจุดข้อมูลออกจากกัน โดยทั่วไปฟังก์ชันที่มีมิติวีซีสูงจะมีความสามารถในการแบ่งแยกสูง ในที่นี้จะขออธิบายเฉพาะฟังก์ชันสำหรับกรณีการรู้จำแบบที่มี 2 ประเภทเท่านั้นคือ $f(\mathbf{x},\alpha)\in\{-1,1\}\ \forall\mathbf{x},\alpha$

กำหนดให้มีขอบเขตของจุด l จุด ซึ่งสามารถกำหนดว่าจุดแต่ละจุดอยู่ประเภทใดได้ทั้งหมด 2^l แบบและถ้าในแบบแต่ละแบบมีฟังก์ชันอย่างน้อยหนึ่งฟังก์ชันใน $\{f(\alpha)\}$ ที่สามารถกำหนดประเภทให้กับจุดแต่ละจุดได้อย่างถูกต้อง จะเรียกได้ว่าเซตของจุดเหล่านี้ถูกทำให้แตก (Shattered) โดยเซตของฟังก์ชันนี้ ดังแสดงในรูปที่ 2.22



รูปที่ 2.22 เซตของจุด 3 จุดใน R^2 ถูกทำให้แตกโดยเส้นที่มีทิศทาง

เรานิยามมิติวีซีของฟังก์ชัน $\{f(\alpha)\}$ ให้เป็นจำนวนสูงสุดของจุดในข้อมูลสอนที่สามารถทำให้แตกด้วย $\{f(\alpha)\}$ ได้ โดยมีข้อสังเกตคือ ถ้าให้มิติวีซีมีค่าเป็น h แล้ว จะมีอย่างน้อยหนึ่งเซตของจุดจำนวน h จุดที่ถูกทำให้แตกออกได้ แต่โดยทั่วไปไม่จำเป็นว่าทุกเซตของจุดจำนวน h จุดจะถูกทำให้แตกออกเป็นส่วนได้เสมอไป

ทฤษฎีบทที่ 2.1 พิจารณาเซตที่ประกอบด้วยจุด m จุดใน R^n ถ้าเลือกจุดใดจุดหนึ่งเป็นจุดกำเนิด จะได้ว่าจุด m จุดเหล่านี้จะถูกจำแนกประเภทโดยระนาบหลายมิติแบบมีทิศทาง (Oriented hyperplane) ได้ก็ต่อเมื่อเวกเตอร์บอกตำแหน่งของจุดที่เหลือเป็นอิสระเชิงเส้นต่อกัน

ผลที่ตามมาคือมิติวีซีของเซตของฟังก์ชันระนาบหลายมิติแบบมีทิศทางใน R^n มีค่าเป็น n+1 ซึ่งพิสูจน์ได้ดังนี้ เนื่องจากเราสามารถเลือกจุดจำนวน n+1 จุดแล้วให้จุดหนึ่งจุดเป็นจุดกำเนิด ดังนั้นจุดที่เหลือ n จุดย่อมต้องเป็นอิสระเชิงเส้นต่อกันแน่นอน (เช่น ให้จุดแต่ละจุดอยู่บนแกนแต่ละแกนใน R^n)

มีข้อสังเกตว่าฟังก์ชันที่มีพารามิเตอร์มากไม่จำเป็นต้องมีมิติวีซีมาก และในทางกลับกันฟังก์ชันที่มีพารามิเตอร์เพียงหนึ่งเดียวอาจมีมิติวีซีเป็นอนันต์ได้ แต่แม้ฟังก์ชันจะมีมิติวีซีเป็นอนันต์ก็อาจจะไม่สามารถแยะจุดเพียงไม่กี่จุดได้

ในทางปฏิบัติมิติวีซีที่ต่ำๆย่อมทำให้ขอบเขตของความผิดพลาดแท้จริงต่ำด้วย แต่ก็ไม่ได้หมายความว่าฟังก์ชันที่มีมิติวีซีสูงๆจะใช้งานได้ไม่ดี เช่น เอสวีเอ็มแบบอาร์บีเอฟ (RBF-Radial Basis Function) ซึ่งมีมิติวีซีเป็นอนันต์ก็เป็นฟังก์ชันที่ใช้งานได้ดีฟังก์ชันหนึ่ง อย่างไรก็ตามการเลือกฟังก์ชันเคอร์เนลฟังก์ชันสำหรับเครื่องเรียนรู้ ไม่ได้มีรูปแบบที่แน่นอน หากแต่ต้องการอ้างอิงจากผลการทดลอง เนื่องจากยังไม่มีข้อสรุปทางทฤษฎีว่าเคอร์เนลแบบใดเหมาะสมกับปัญหาแบบใด

10.2.3 วิธีการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด

ขอบเขตบนของความผิดพลาดอันอาจเกิดจากการทำให้มีลักษณะทั่วไปสามารถเขียนให้อยู่ในรูปสมการดังต่อไปนี้

$$R(\alpha) \le R_{emp}(\alpha) + \Phi\left(\frac{l}{h}\right)$$

ซึ่งพจน์แรกคือ ความเสี่ยงเชิงทดลอง ส่วนพจน์หลังคือ ช่วงความเชื่อมั่นซึ่งเป็นฟังก์ชันของมิติวีซี (h) ในการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด เราจะต้องคำนึงถึงพจน์ทั้งสองของสมการข้างต้น โดยมีอยู่สองแนวทางดังนี้

แนวทางแรก ให้คงค่าช่วงความเชื่อมั่นให้คงที่แล้วลดความเสี่ยงเชิงทดลองให้ต่ำสุด แนวทางนี้จะลดพจน์แรกของสมการให้ต่ำสุดโดยการออกแบบเครื่องเรียนรู้ให้ซับซ้อน แต่ถ้าออกแบบเครื่องที่ซับซ้อนเกินไป จะทำให้ช่วงความเชื่อมั่นกว้างขึ้น ซึ่งแม้จะสามารถลดความเสี่ยงเชิงทดลองลงจนหมดสิ้นไปได้ แต่ความผิดพลาดที่เกิดขึ้นในการใช้งานจริงยังอาจสูงอยู่ ปรากฏการณ์ที่เกิดขึ้นนี้เรียกว่าการปรับเหมาะเกินไป อย่างไรก็ตามถ้าเราเลือกเครื่องที่มีความซับซ้อนต่ำ เพื่อให้ช่วงความเชื่อมั่นแคบ เราก็จะประสบปัญหาในการหาฟังก์ชันที่จะนำมาใช้ประมาณปัญหาได้ยาก อันจะทำให้เกิดความผิดพลาดเชิงทดลองสูง เพื่อที่จะลดปัญหาการประมาณที่ไม่ดีและการปรับเหมาะเกินไปเราจะต้องเลือกสถาปัตยกรรมของเครื่องที่เหมาะสม โดยอาศัยความรู้เกี่ยวกับลักษณะของปัญหา แล้วจึงหาฟังก์ชันในเครื่องนี้ที่สามารถลดค่าผิดพลาดในข้อมูลสอนให้ต่ำที่สุดได้

แนวทางที่สอง ให้คงค่าของความเสี่ยงเชิงทดลองให้คงที่ (อาจเป็นศูนย์) แล้วลดช่วงความเชื่อมั่นให้ต่ำสุด แนวทางนี้ทำโดยการกำหนดขอบเขตความเสี่ยงเชิงทดลองสูงสุดที่ยอมรับได้ แล้วเปลี่ยนรูปแบบของฟังก์ชันเพื่อลดช่วงความเชื่อมั่นให้ต่ำสุด แนวทางนี้พบในเอสวีเอ็ม โดยเอสวีเอ็มจะแบ่งกลุ่มของฟังก์ชันออกเป็นเซตย่อย แล้วหาเซตของเคอร์เนลที่ให้ความเสี่ยงเชิงทดลองต่ำสุด จากนั้นจึงหาฟังก์ชันในเซตนั้นที่มีมิติวีซีต่ำที่สุด แล้วบันทึกค่าขอบเขตความเสี่ยงที่ได้ ทำการพิจารณาเช่นเดิมกับเคอร์เนลกลุ่มถัดมาที่ให้ความเสี่ยงเชิงทดลองต่ำสุด ทำวนซ้ำจนได้ฟังก์ชันที่ให้ขอบเขตความผิดพลาดที่แท้จริงต่ำสุด

10.3 ระนาบหลายมิติแย่งแยกดีสุด

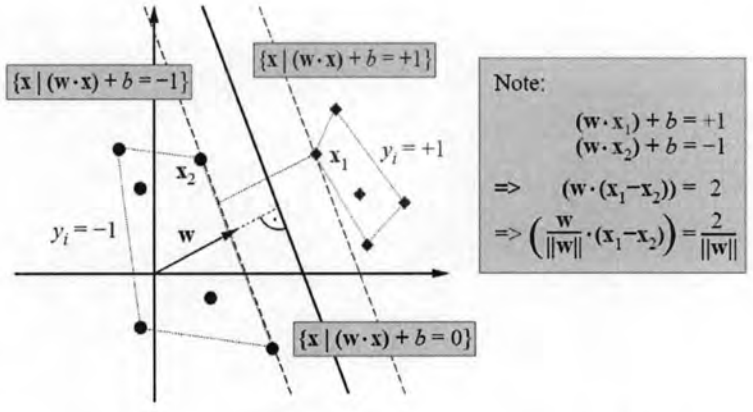
ในการออกแบบขั้นตอนการเรียนรู้ เราจะต้องใช้ฟังก์ชันที่สามารถคำนวณขีดความสามารถได้ ตัวจำแนกเอสวีเอ็มมีพื้นฐานจากฟังก์ชันประเภทระนาบหลายมิติ โดยที่ระนาบหลายมิติทำหน้าที่คล้ายเป็นตัวแยกเขตแดนระหว่างข้อมูลทั้งสองประเภท ดังสมการ

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0, \quad \mathbf{w} \in \mathrm{R}^{\mathrm{N}},\ b \in \mathrm{R}$$

และสอดคล้องกับฟังก์ชันตัดสินใจ

$$f(\mathbf{x}) = sign((\mathbf{w}\cdot\mathbf{x})+b)$$

ระนาบหลายมิติแย่งแยกดีสุดมีนิยามเป็นระนาบที่มีระยะห่างของการแบ่งแยกระหว่างข้อมูลทั้งสองประเภทมากที่สุด เราพบว่าระยะห่างระหว่างกลุ่มข้อมูลทั้งสองประเภทที่เรียกว่าระยะขอบ (Margin) คือ $2/\|\mathbf{w}\|$ ดังแสดงในรูปที่ 2.23 โดยที่ $y(\mathbf{w}\cdot\mathbf{x}+b)\geq 1$ ต้องเป็นจริงด้วย ในรูปนี้จะมีข้อมูลที่สำคัญจริงๆ ซึ่งส่งผลต่อตำแหน่งของระนาบหลายมิติแบ่งแยกดีสุด ซึ่งคือ x_1 (ของตัวอย่างประเภท +1) และ x_2 กับ x_3 (ของตัวอย่างประเภท -1) เราเรียกข้อมูลทั้งสามตัวนี้ว่าซัพพอร์ตเวกเตอร์ (Support Vector) ซึ่งทำหน้าที่สนับสนุนการสร้างระนาบหลายมิติแบ่งแยกดีสุดนี้ ส่วนข้อมูลอื่นๆ แม้ว่าจะถูกตัดออกไป ก็จะไม่ส่งผลกระทบต่อการสร้างระนาบ



รูปที่ 2.23 ระนาบหลายมิติที่ใช้แยกดีสุดจะมีระยะห่างระหว่างข้อมูลทั้งสองกลุ่มเป็น $2/\|\mathbf{w}\|$

ปัญหาของการหาระนาบหลายมิติแบ่งแยกดีสุดนี้มีลากรองเกียนคือ

$$L(w,b,\alpha) = \tfrac{1}{2}\mathbf{w}\cdot\mathbf{w} - \sum\alpha_i((\mathbf{w}\cdot\mathbf{x}_i+b)y_i - 1)$$

ซึ่ง α คือตัวคูณลากรองจ์โดยต้องหาค่าต่ำสุดเมื่อเทียบกับ $\mathbf{w},b$ และหาค่าสูงสุดเมื่อเทียบกับ $\alpha_i \geq 0$ ที่จุดที่ดีที่สุด คำตอบ $\mathbf{w}_0, b_0$ และ α_i^0 จะสอดคล้องกับคุณลักษณะของระนาบหลายมิติที่ดีที่สุดดังนี้

$$\sum\alpha_i^0 y_i = 0,\ \ \alpha_i^0 \geq 0,\ \ i = 1,...,l$$
$$\mathbf{w}_0 = \sum\alpha_i^0 y_i \mathbf{x}_i\ \ \alpha_i^0 \geq 0,\ \ i = 1,...,l$$

โดยที่จริงแล้วเฉพาะสัมประสิทธิ์ α_i^0 ของซัพเพอร์เวกเตอร์เท่านั้นที่ไม่เป็นศูนย์ ดังนั้นในการหาค่าของเวกเตอร์ $\mathbf{w}_0$ เราจึงหาจากผลรวมเชิงเส้นของข้อมูลตัวที่เป็นซัพพอร์ตเวกเตอร์เท่านั้น

$$b_0 = \frac{1}{2}\left[(\mathbf{w}_0\cdot\mathbf{x}^*(1)) + (\mathbf{w}_0\cdot\mathbf{x}^*(-1))\right]$$

โดยที่ $\mathbf{x}^*(1)$ คือซัพพอร์ตเวกเตอร์ใดๆ ที่อยู่ในประเภท +1 และ $\mathbf{x}^*(-1)$ คือซัพพอร์ตเวกเตอร์ใดๆที่อยู่ในประเภท -1 และจะได้ฟังก์ชันการตัดสินใจในรูปแบบ

$$f(\mathbf{x}) = sign(\sum_{SupportVector} \nu_i(\mathbf{w}_i \cdot \mathbf{x}_i) + b_0)$$

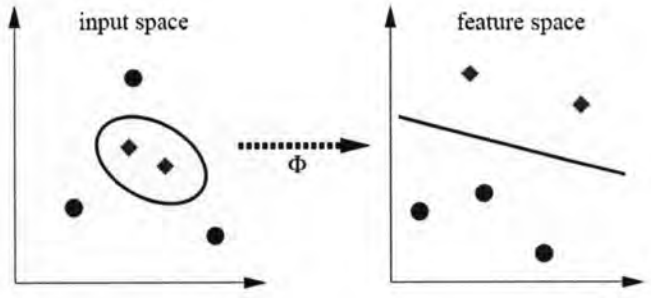
$$\nu_i = y_i\alpha_i^0$$

คำตอบนี้ใช้ได้เฉพาะกรณีที่สามารถแบ่งแยกแบบเชิงเส้นได้เท่านั้น แต่สำหรับกรณีที่แบ่งแยกแบบเชิงเส้นไม่ได้ ต้องมีการปรับข้อจำกัดเล็กน้อย คือ α_i จะมีขอบเขต $0 \le \alpha_i \le C$ โดยเราสามารถแปรค่า C ได้ โดยเป็นการแลกเปลี่ยนระหว่างความถูกต้องกับความเร็วในการสอน

สังเกตได้ว่าทั้งการแก้ปัญหาการหาฟังก์ชันของระนาบหลายมิติและตัวฟังก์ชันการตัดสินใจต่างก็ขึ้นอยู่กับผลคูณเชิงสเกลาร์ระหว่างเวกเตอร์ สิ่งนี้เองที่ทำให้เราสามารถขยายอัลกอริทึมสำหรับกรณีที่ไม่เป็นเชิงเส้นได้ในปริภูมิอันดับสูง

10.4 ปริภูมิระดับสูงและเคอร์เนล

การแมปข้อมูลเข้าไปสู่ปริภูมิคุณลักษณะที่มีระดับสูงขึ้น จะช่วยให้สามารถแยกข้อมูลสองประเภทออกจากกันได้โดยใช้ฟังก์ชันเชิงเส้น โดยมีสมมติฐานว่าข้อมูลที่มีความสัมพันธ์ไม่เป็นเชิงเส้นในปริภูมิอันดับต่ำ เมื่อแมปไปสู่ปริภูมิอันดับสูงจะมีความสัมพันธ์แบบเชิงเส้นได้ อย่างไรก็ตามการแมปไปสู่ปริภูมิอันดับสูงอาจต้องการการคำนวณที่สูงเกินไป ฟังก์ชันเคอร์เนลช่วยหลีกเลี่ยงปัญหาการคำนวณหาฟังก์ชันในการแมปได้ โดยยอมให้คำนวณผลคูณสเกลาร์ของตัวแปรสองตัวในปริภูมิอันดับสูงได้โดยไม่ต้องคำนวณหาฟังก์ชันที่ใช้แมปการแสดงด้วยสมการจะช่วยให้เข้าใจถึงแนวคิดนี้ได้ง่ายขึ้น



รูปที่ 2.24 แนวคิดการแมปแบบไม่เชิงเส้น

แนวคิดเบื้องต้นของเอสวีเอ็ม ดังแสดงในรูปที่ 2.24 เป็นการแมปข้อมูลไปสู่ปริภูมิผลคูณเชิงสเกลาร์อันดับสูงขึ้นไป (เขียนแทนด้วย F) ผ่านการแมปแบบไม่เชิงเส้น

$$\Phi : \mathrm{R}^{\mathrm{N}} \rightarrow \mathrm{F}$$

แล้วจึงใช้ขั้นตอนวิธีเชิงเส้นนี้ใน F ซึ่งสิ่งที่ต้องทำก็เพียงการหาค่าของผลคูณสเกลาร์

$$k(\mathbf{x},\mathbf{y}) := (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$$

ถ้า F มีอันดับสูง ย่อมเท่ากับว่าพจน์ด้านขวามือของสมการข้างต้นจะคำนวณได้ยากมาก อย่างไรก็ตามในบางกรณีจะมีเคอร์เนล k ที่ง่ายต่อการคำนวณตัวอย่างเช่นเคอร์เนลแบบพหุนาม

$$k(\mathbf{x},\mathbf{y}) := (\mathbf{x} \cdot \mathbf{y})^d$$

ซึ่งสามารถแสดงให้เห็นว่าสอดคล้องกับการแมป Φ ไปสู่ปริภูมิที่แสปนโดยผลคูณทั้งหมดของอันดับ d ใน R^{N} ตัวอย่างเช่นในกรณีที่ $d = 2$ และ $x, y \in \mathrm{R}^2$ ตัวอย่างเช่น เรามี

$$\begin{aligned}(\mathbf{x} \cdot \mathbf{y})^2 &= ((x_1, x_2) \cdot (y_1, y_2))^2 \\ &= ((x_1^2, \sqrt{2}x_1x_2, x_2^2) \cdot (y_1^2, \sqrt{2}y_1y_2, y_2^2)) \\ &= (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))\end{aligned}$$

ที่นิยามให้ $\Phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$

นอกเหนือจากการใช้เคอร์เนลแบบพหุนามแล้วเอสวีเอ็มยังสามารถนำมาใช้กับเคอร์เนลแบบอาร์บีเอฟ (RBF) ซึ่งนิยามดังนี้คือ

$$k(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$$

และเคอร์เนลแบบซิกมอยด์ (ที่มีเกน κ และออฟเซต Θ)

$$k(\mathbf{x},\mathbf{y}) = \tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) + \Theta)$$

10.5 การเรียนรู้ซัพพอร์ตเวกเตอร์แมชชีน

เอสวีเอ็มเป็นเทคนิคการเรียนรู้ที่มีขีดความสามารถสูงสำหรับการจำแนกข้อมูลแบบสองประเภท โดยสามารถลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด เอสวีเอ็มใช้ประโยชน์จากการแมปและฟังก์ชันเคอร์เนล ซึ่งระนาบหลายมิติที่ใช้แยกประเภทข้อมูลจะอยู่ในปริภูมิอันดับสูง ถึงจุดนี้จะแทนที่ข้อมูลแต่ละตัวในชุดสอน $\mathbf{x}_i$ ด้วย $\Phi(\mathbf{x}_i)$ และหาระนาบหลายมิติแบ่งแยกดีสุดใน F เนื่องจากเราใช้เคอร์เนล ดังนั้นจึงได้ผลเป็นฟังก์ชันการตัดสินใจในรูปแบบ

$$f(\mathbf{x}) = sign(\sum_{SupportVector} \nu_i \bullet k(\mathbf{x}, \mathbf{x}_i) + b_0)$$

$$\nu_i = y_i \alpha_i^0$$

โดยที่พารามิเตอร์ ν_i สามารถคำนวณได้โดยเป็นคำตอบของปัญหาการโปรแกรมกำลังสอง (Quadratic Programming) โดยลดค่าของฟังก์ชันวัตถุประสงค์ให้ต่ำสุด

$$\frac{1}{2}\sum \alpha_i Q_j \alpha_j - \sum \alpha_i$$

โดยขึ้นกับข้อกำหนดต่อไปนี้

$$0 \le \alpha_i \le C$$
$$\sum y_i \alpha_i = 0$$

Q อยู่ในรูปของ $y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ เป็นเมทริกซ์มิติ N×N ซึ่งขึ้นอยู่กับขนาดของข้อมูลสอน $\mathbf{x}_i$ แบะฉลากของประเภท y_i กับรูปแบบของฟังก์ชันที่จะใช้ ส่วน C เป็นค่าคงที่ที่สามารถกำหนดได้ เมื่อเราลดค่าฟังก์ชันวัตถุประสงค์โดยการแก้ปัญหาการโปรแกรมกำลังสองแล้ว เราจะได้พารามิเตอร์ $\alpha_i, C, \mathbf{w}_0, b_0$ ที่ให้ฟังก์ชันต่ำสุด ซึ่งก็คือการเรียนรู้ของเอสวีเอ็ม

ในปริภูมิอันดับสูงจะได้ระนาบหลายมิติเป็นแบบเชิงเส้น ส่วนในปริภูมิอันดับต่ำระนาบหลายมิติจะสอดคล้องกับฟังก์ชันการตัดสินใจแบบไม่เชิงเส้นซึ่งรูปแบบจะถูกกำหนดโดยเคอร์เนล และโดยการเปลี่ยนเคอร์เนล เราจะได้สถาปัตยกรรมที่ต่างออกไป ดังเช่นตัวจำแนกแบบพหุนาม ตัวจำแนกแบบอาร์บีเอฟ และข่ายงานประสาทเทียมแบบสามระดับ เป็นต้น ในปัญหาต่างกัน เราต้องการเคอร์เนลที่อาจไม่เหมือนกัน แต่โดยปกติแล้วไม่ว่าจะใช้เคอร์เนลชนิดใด ก็มักได้ซัพพอร์ตเวกเตอร์ชุดที่ใกล้เคียงกัน สิ่งนี้สนับสนุนแนวคิดที่ว่าซัพพอร์ตเวกเตอร์เป็นคุณลักษณะเฉพาะสำหรับปัญหาหนึ่งๆ

งานวิจัยที่เกี่ยวข้อง

การศึกษาวิจัยส่วนใหญ่เกี่ยวกับการพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ตั้งแต่อดีตจนถึงปัจจุบันนั้น จะมุ่งเน้นการหาลำดับของหน่วยเสียงที่เรียงต่อเนื่องกันเป็นเส้นตรงเพียงลำดับเดียว (Linear Sequence of Segments) [17 – 21] อย่างไรก็ตามการจะนำลำดับของเซกเมนต์เพียงลำดับเดียวมาใช้กับระบบรู้จำเสียงแบบอาศัยเซกเมนต์นั้น ลำดับของเซกเมนต์จะต้องมีความถูกต้องแม่นยำ โดยแม้จะมีความผิดพลาดเพียงเล็กน้อย ก็อาจทำให้ผลของการรู้จำเสียงพูดผิดพลาดไปมาก ในปัจจุบันยังไม่มีวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่สามารถทำงานได้ถูกต้องสมบูรณ์แบบ งานวิจัยบางส่วนจึงมุ่งเน้นไปที่การแบ่งเสียงพูดเป็นเซกเมนต์ให้ผลลัพธ์ออกมาอยู่ในรูปกราฟของเซกเมนต์แทน [22] ในหัวข้อนี้จะนำเสนองานวิจัยเกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งสามารถแบ่งได้เป็นสามกลุ่มดังนี้

1. การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยการจำแนกเสียงพูดเป็นประเภทกว้าง

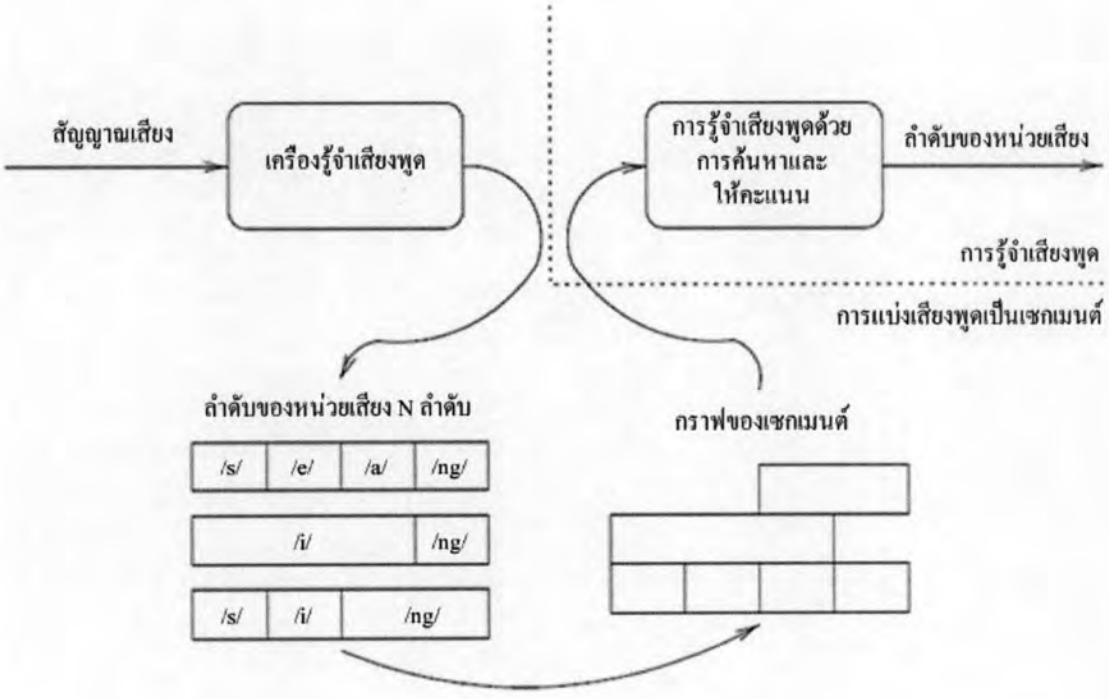
Cole และ Fanty [23] เสนอการแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้การจำแนกหน่วยเสียงในแต่กรอบเวลาสั้นๆ แบ่งหน่วยเสียงออกเป็นประเภทกว้างๆ (Segmentation using Broad-class Classification) ในการค้นหาขอบเขตของหน่วยเสียง รวมทั้งนำโครงข่ายประสาทเทียมมาใช้ในการจำแนกประเภทของเสียงในแต่ละกรอบเวลาออกเป็น 22 ประเภทเพื่อใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับนำไปใช้ในระบบรู้จำเสียงพูดตัวอักษรภาษาอังกฤษที่มีลักษณะของเสียงพูดแบบไม่ต่อเนื่อง โดยสามารถรู้จำอักษรในภาษาอังกฤษได้เปอร์เซ็นต์ความถูกต้องสูงถึง 95%

2. การแบ่งเสียงพูดเป็นเซกเมนต์จากการเปลี่ยนแปลงทางสัญญาณเสียง

การแบ่งเสียงพูดเป็นเซกเมนต์จากการเปลี่ยนแปลงทางสัญญาณเสียง (Acoustic Segmentation) จะใช้วิธีวัดจากปริมาณที่แสดงถึงการเปลี่ยนแปลงหรือความไม่ต่อเนื่องกันของสัญญาณเสียง (Acoustic Discontinuity) โดยอาศัยหลักการที่ว่าสัญญาณเสียงที่บริเวณขอบเขตของหน่วยเสียงจะมีความไม่ต่อเนื่องสูงกว่าบริเวณที่เป็นหน่วยเสียง Wang, Lu และ Zhang [24] เสนอวิธีในการแบ่งเสียงพูดเป็นเซกเมนต์โดยอาศัยการวัดปริมาณสำคัญๆอย่าง เวลาของแต่ละหน่วยเสียง (Duration), อัตราการออกเสียง (Rate of Speech: ROS) และลำดับของการออกเสียง (Phonetic Sequence) มาคำนวณเป็นปริมาณที่ใช้แสดงถึงความไม่ต่อเนื่องของเสียงเพื่อใช้ระบุหาขอบเขตของหน่วยเสียง ผลลัพธ์ที่ได้คือกราฟของเซกเมนต์ ที่ได้จากการเชื่อมกันของขอบเขตทั้งหมด โดยแม้ว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบนี้จำทำงานได้อย่างรวดเร็ว แต่ขนาดของกราฟของเซกเมนต์ก็มีใหญ่มากเกินความจำเป็น

3. การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด [4] จะหาขอบเขตของหน่วยเสียงโดยอาศัยเครื่องรู้จำเสียงพูดในระดับหน่วยเสียงแบบอาศัยกรอบเวลา (Frame-based Phonetic Recognizer) มารู้จำเสียงพูดออกมาเป็นลำดับของหน่วยเสียง N ลำดับที่ดีที่สุด แล้วจึงนำเซกเมนต์ของหน่วยเสียงที่รู้จำได้มาประกอบรวมกันเป็นกราฟของเซกเมนต์ดังแสดงด้วยรูปที่ 2.25 โดยที่ N ซึ่งเป็นตัวแปรที่กำหนดจำนวนลำดับที่ดีที่สุดที่ต้องการรู้จำ จะเป็นตัวกำหนดขนาดของกราฟของเซกเมนต์อีกที เครื่องรู้จำเสียงพูดในที่นี้จะสร้างโดยอาศัยแบบจำลองฮิดเดนมาร์คอฟ หรืออาจใช้การค้นหาด้วยวิธีของไวเทอร์บิแบบไปข้างหน้า (Forward Viterbi) และวิธีของ A^* แบบย้อนกลับ (Backward A^*) ก็ได้ แม้การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีการดังกล่าวจะสามารถทำงานได้ผลดี แต่ก็มีข้อเสียอยู่ที่ต้องใช้การคำนวณนาน ไม่เหมาะกับระบบรู้จำเสียงแบบที่ต้องการความรวดเร็ว



รูปที่ 2.25 การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด