การรู้จำเสียงพูดต่อเนื่องภาษาไทยโดยใช้นิวรอลเน็ตเวิร์ก

นายประเสริฐศักดิ์ ผุงประเสริฐยิ่ง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

THAI CONTINUOUS SPEECH RECOGNITION USING NEURAL NETWORKS

Mr. Prasertsak Pungprasertying

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2007

หัวข้อวิทยานิพนธ์	การรู้จำเสียงพูดต่อเนื่องภาษาไทยโดยใช้นิวรอลเน็ตเวิร์ก
โดย	นายประเสริฐศักดิ์ ผุงประเสริฐยิ่ง
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์)

..... อาจารย์ที่ปรึกษา

(รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... กรรมการ

(อาจารย์ ดร.อติวงศ์ สุชาโต)

..... กรรมการ

(อาจารย์ ดร.สุกรี สินธุภิญโญ)

ประเสริฐศักดิ์ ผุงประเสริฐยิ่ง : การรู้จำเสียงพูดต่อเนื่องภาษาไทยโดยใช้นิวรอลเน็ตเวิร์ก. (THAI CONTINUOUS SPEECH RECOGNITION USING NEURAL NETWORKS) อ. ที่ปรึกษา : รศ.ดร.บุญเสริม กิจศิริกุล, 113 หน้า.

งานวิจัยชิ้นนี้มีจุดมุ่งหมายเพื่อพัฒนาระบบรู้จำเสียงพูดต่อเนื่องอัตโนมัติภาษาไทย โดยใช้นิวรอลเน็ตเวิร์กรู้จำหน่วยเสียงในกรอบการวิเคราะห์ระดับเฟรม แล้วจึงนำผลการรู้จำนี้ประกอบกับแบบจำลองทางภาษาและกระบวนการค้นหา จนได้ลำดับของคำในภาษาออกมาเป็นผลลัพธ์ จากนั้นทำการวิเคราะห์ประสิทธิภาพของระบบโดยใช้ฐานข้อมูลเสียงพูดซื่อไทย และฐานข้อมูลเกี่ยวกับสัตว์ภาษาไทย โดยทดลองปรับค่าพารามิเตอร์ต่างๆ คือ ชุดหน่วยเสียง อันดับของพีแอลพี และจำนวนเฟรมที่ใช้ แล้วแสดงความถูกต้องของการรู้จำ ทั้งในระดับเฟรม และในระดับคำ ทั้งในชุดข้อมูลสำหรับการเรียนรู้ และในชุดข้อมูลสำหรับการทดสอบ

ในชุดข้อมูลสำหรับการทดสอบ ฐานข้อมูลเสียงพูดซื่อไทยมีความถูกต้องสูงสุดระดับเฟรมอยู่ที่ประมาณ 70% และระดับคำอยู่ที่ประมาณ 90% ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยมีความถูกต้องสูงสุดระดับเฟรมอยู่ที่ประมาณ 60% และระดับคำอยู่ที่ประมาณ 40%

ภาควิชา.....วิศวกรรมคอมพิวเตอร์..... ลายมือชื่อนิสิต.....

สาขาวิชา.....วิศวกรรมคอมพิวเตอร์..... ลายมือชื่ออาจารย์ที่ปรึกษา.....

ปีการศึกษา.....2550.....

4670660321 : MAJOR COMPUTER ENGINEERING

KEY WORD : SPEECH RECOGNITION / NEURAL NETWORKS

PRASERTSAK PUNGPRASERTYING : THAI CONTINUOUS SPEECH RECOGNITION USING NEURAL NETWORKS. THESIS ADVISOR : ASSOC. PROF. BOONSERM KIJSIRIKUL, Ph. D., 113 pp.

The purpose of this research is to develop an automatic Thai continuous speech recognition system by applying neural networks to frame-based recognition of phonemes. The recognition results are then combined with the language model and the search process to provide the sequence of words as an outcome. The system performance has been analyzed with Thai First Names Speech Corpus and Thai Animal Speech Corpus. The experiments are performed by adjusting the system parameters which are the phoneme set, the PLP order and the number of frames. We present the recognition accuracy at the frame level and the word level, both in the training set and the test set.

For the test set of the Thai First Names Speech Corpus, the system achieves about 70% and 90% maximum accuracy in the frame level and the word level respectively, while for that of the Thai Animal Speech Corpus, the system provides about 60% and 40% maximum accuracy in the frame level and the word level respectively.

Department.....Computer Engineering..... Student's signature.....

Field of study.....Computer Engineering... Advisor's signature.....

Academic year.....2007.....

กิตติกรรมประกาศ

เหนือสิ่งอื่นใด ผมขอกราบขอบพระคุณ รศ. ดร. บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษา ผู้ริเริ่ม เป็นแรงบันดาลใจ ให้คำชี้แนะ และขัดเกลา จนวิทยานิพนธ์ชิ้นนี้สำเร็จเป็นรูปเป็นร่างขึ้นมา ทั้งยังเปี่ยมด้วยความเมตตากรุณาต่อลูกศิษย์คนนี้อย่างหาที่สุดมิได้ ไม่เฉพาะในยามที่ประสบปัญหาทางด้านการเรียนและการวิจัยเท่านั้น หากแม้ในช่วงชีวิตที่ต้องเผชิญกับความยากลำบากอย่างมากล้น อาจารย์ก็ให้ความช่วยเหลืออย่างเต็มที่ ตลอดระยะเวลาหลายปีที่ผ่านมา

ขอกราบขอบพระคุณ ผศ. ดร. ญาใจ ลิ่มปิยะกรณ์ อ. ดร. สุกรี สินธุภิญโญ อ.ดร. อติวงศ์ สุชาโต และ อ. ดร. ณัฐกร ทับทอง ที่ให้ความรู้ทางด้านการทำเหมืองข้อมูล การเรียนรู้ของเครื่อง และการรู้จำเสียงพูด รวมทั้งคำแนะนำที่เป็นประโยชน์ต่างๆ มากมาย ด้วยความยินดีและความเอาใจใส่อย่างสูงยิ่ง

ขอขอบคุณจุง และเพื่อนสมาชิกในห้องปฏิบัติการทุกคน ที่เคยร่วมเสวนาขยายความคิดเชิงวิชาการ และยังคอยเป็นกำลังใจให้ด้วยดีเสมอมา

สุดท้ายนี้ ลูกขอน้อมรำลึกถึงพระคุณของบิดามารดา ผู้เป็นพรหมในบ้าน และเป็นอาจารย์คนแรกของลูก

สารบัญ

หน้า

สารบัญภาพ

หน้า

สารบัญตาราง

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การรู้จำเสียงพูด โดยทั่วไปหมายถึงกระบวนการแปลงเสียงพูด[1] ไปสู่ลำดับของคำในภาษา[2] อาจกล่าวได้ว่าเสียงพูดคือ*สื่อ*ที่ทำการเข้ารหัส*สาร*คือลำดับของคำในภาษาไว้[3] และการรู้จำเสียงพูดคือการถอดรหัสสารภาษาออกมาจากสื่อเสียงพูดนั่นเอง

จากวิวัฒนาการอันยาวนาน ทำให้มนุษย์เป็นสิ่งมีชีวิตที่สามารถติดต่อสื่อสารกันได้ด้วยเสียงพูด ซึ่งกลไกการสร้างเสียงพูดและการรู้จำเสียงพูดของมนุษย์นั้นเป็นธรรมชาติเสียจนเราไม่ได้ตระหนักว่าแท้จริงแล้วนี่คือปรากฏการณ์ที่สุดแสนจะซับซ้อน โดยเฉพาะในด้านการรู้จำเสียงพูดนั้น แม้เทคโนโลยีทางนี้จะก้าวหน้าขึ้นเพียงใด แต่ความสามารถในการรู้จำเสียงพูดของเครื่องจักรก็ยังห่างไกลจากความสามารถในการรับฟังของมนุษย์อยู่อีกมาก ทั้งนี้เนื่องจากการรู้จำเสียงพูดมีความยุ่งยากอยู่นานัปการ ประการแรกก็คือ การรู้จำเสียงพูดเป็นการรู้จำรูปแบบที่มีข้อมูลเข้าอยู่ในรูปอนุกรมเวลา ทำให้ระบบรู้จำเสียงพูดต้องมีวิธีการที่เหมาะสมสำหรับจัดการกับความเป็นพลวัตของข้อมูล ซึ่งต่างจากการรู้จำรูปแบบทั่วไปที่ไม่มีเรื่องของเวลาเข้ามาเกี่ยวข้อง อีกปัญหาสำคัญที่ระบบรู้จำเสียงพูดต้องเผชิญคือความหลากหลายของลักษณะทางเสียง ซึ่งโดยปกติแล้วเสียงพูดที่เข้าสู่ระบบ แม้จะเป็นการออกเสียงของหน่วยย่อยทางภาษาเดียวกัน แต่ก็มีความแปรผันไปได้มาก ความแปรผันเหล่านี้อาจมีสาเหตุมาจากทั้งความแตกต่างของอวัยวะการออกเสียงและการเคลื่อนไหวของอวัยวะการออกเสียงในผู้พูดแต่ละคน รวมถึงสภาพเสียงแวดล้อม และวิธีการบันทึกเสียง นอกจากนี้ลักษณะทางเสียงของแต่ละหน่วยย่อยทางภาษายังขึ้นอยู่กับหน่วย

---

[1] *เสียงพูด* คือการสะเทือนอย่างต่อเนื่องแบบพลวัตของมวลอากาศ ที่เกิดจากการสั่นของเส้นเสียงและการกรองในช่องทางเดินเสียงของมนุษย์ - เสียงพูดถือเป็นลักษณะทางกายภาพ

[2] *ลำดับของคำในภาษา* เป็นการร้อยเรียงต่อกันของหน่วยย่อยทางภาษา (อาจจะเป็น*หน่วยเสียง*ในภาษาพูด หรือ*ตัวอักษร*ในภาษาเขียน) โดยมีความหมายและไวยากรณ์เป็นตัวกำหนดวิธีการร้อยเรียงนั้น - ลำดับของคำในภาษาถือเป็นลักษณะทางนามธรรม

[3] นอกจากข้อมูลทางภาษาแล้ว ในเสียงพูดยังบรรจุข้อมูลชนิดอื่นอยู่อีกมาก เช่น ข้อมูลที่สามารถใช้ระบุถึงตัวผู้พูด และข้อมูลที่แสดงถึงอารมณ์ของผู้พูด เป็นต้น แต่ในการรู้จำเสียงพูดจะสนใจเฉพาะข้อมูลทางภาษาเท่านั้น

ย่อยทางภาษาข้างเคียง โดยเป็นผลจากการออกเสียงร่วม ซึ่งระบบรู้จำเสียงพูดก็จำเป็นอีกเช่นกันที่จะต้องรับมือกับความแปรผันเหล่านี้

การรู้จำเสียงพูดในระยะแรกเป็นการรู้จำเสียงพูดไม่ต่อเนื่องในจำนวนคำศัพท์ที่น้อย จนต่อมาได้พัฒนาเป็นระบบรู้จำเสียงพูดต่อเนื่องอัตโนมัติที่รู้จำคำศัพท์ได้เป็นจำนวนมากขึ้น สถาปัตยกรรมของระบบรู้จำเสียงพูดก็ได้ถูกปรับปรุงอยู่เสมอ เพื่อให้สามารถรองรับงานการรู้จำเสียงพูดที่ยากขึ้นและใหญ่ขึ้น โดยทั่วไป ระบบการรู้จำเสียงพูดต่อเนื่องอัตโนมัตินั้นจะประกอบด้วยส่วนหลักสองส่วน คือ ส่วนการรู้จำหน่วยย่อยทางภาษาของเสียงพูด (อาจจะเป็นหน่วยเสียง พยางค์ หรือคำ) และส่วนที่ประกอบหน่วยย่อยทางภาษาของเสียงพูดเหล่านั้นออกมาให้เป็นลำดับของคำในเสียงพูด

ในส่วนการรู้จำหน่วยย่อยทางภาษาของเสียงพูดนั้น วิธีการที่ใช้สามารถแบ่งออกเป็นกลุ่มใหญ่ๆ ได้ดังนี้

1. *การจับคู่แผ่นแบบ* เป็นการรู้จำโดยนำเสียงพูดที่เข้ามามาเปรียบเทียบกับแผ่นแบบเสียงพูดที่มีอยู่ เสียงพูดที่เข้ามาถ้าใกล้เคียงกับแผ่นแบบของเสียงใดมากที่สุดก็จะถูกจัดให้เป็นเสียงนั้น โดยการเปรียบเทียบในที่นี้ใช้วิธีวาร์ปเวลาแบบพลวัต ซึ่งสามารถจัดการกับความแปรผันทางเวลาได้ดี แต่การจับคู่แผ่นแบบทำได้แค่เพียงจับคู่ระหว่างหนึ่งแผ่นแบบต่อหนึ่งเสียงพูดที่เข้ามาเท่านั้น ทำให้ยังรองรับกับความแปรผันทางเสียงได้ไม่ดีนัก
2. *การใช้ความรู้ทางสวนศาสตร์-สัทศาสตร์* การรู้จำเสียงพูดชนิดนี้จะให้ผู้เชี่ยวชาญใส่ความรู้ที่เป็นกฎเข้าไปในระบบ และระบบจะใช้กฎเหล่านี้ในการจำแนกเสียงพูด อย่างไรก็ตาม การใส่กฎให้ครอบคลุมทุกความแปรผันในเสียงพูดนั้นเป็นเรื่องยากและต้องอาศัยแรงงานของมนุษย์มาก
3. *การสร้างแบบจำลองเฟ้นสุ่ม* วิธีนี้จะมองว่าแต่ละเสียงพูดที่ออกมาเกิดจากกระบวนการเชิงสุ่มกระบวนการหนึ่ง การรู้จำเสียงพูดจึงจำเป็นต้องสร้างแบบจำลองเฟ้นสุ่มขึ้นเพื่อใช้อธิบายและจำแนกเสียงพูด แบบจำลองเฟ้นสุ่มที่นิยมใช้ในปัจจุบันคือแบบจำลองมาร์คอฟซ่อนตัว ซึ่งสามารถสร้างจากตัวอย่างเสียงพูด และนำมาใช้ในการรู้จำเสียงพูดได้ผลดี การแบ่งหนึ่งหน่วยย่อยทางภาษาของเสียงพูดออกเป็นหลายสถานะในแบบจำลองมาร์คอฟซ่อนตัวทำให้สามารถรองรับกับปัญหาความแปรผันทางเวลาได้ อย่างไรก็ตาม แบบจำลองมาร์คอฟซ่อนตัวไม่ได้ถูกสร้างขึ้นมาเพื่อการแบ่งแยกโดยเฉพาะ และหลายข้อกำหนดในแบบจำลองยังไม่สอดคล้องกับความเป็นจริงเท่าใดนัก

4. *การใช้การเรียนรู้แบบแบ่งแยก* เป็นการรู้จำเสียงพูดโดยสร้างแบบจำลองเพื่อแบ่งแยกความแตกต่างระหว่างเสียงพูดของแต่ละหน่วยย่อยทางภาษาโดยตรง ทำให้ผลที่ได้จากการรู้จำดีขึ้น และประหยัดพารามิเตอร์กว่า แต่การใช้การเรียนรู้แบบแบ่งแยกกับข้อมูลในรูปอนุกรมเวลานั้นยังมีข้อจำกัด และติดขัดอยู่หลายประการ การเรียนรู้แบบแบ่งแยกที่นิยมใช้กันมากคือนิวรอลเน็ตเวิร์ก ซึ่งสามารถสร้างฟังก์ชันการแบ่งแยกได้โดยตรงจากการปรับค่าน้ำหนัก ซึ่งเป็นพารามิเตอร์ในแบบจำลอง

ทั้งการสร้างแบบจำลองเฟ้นสุ่มและการใช้การเรียนรู้แบบแบ่งแยก อาจเรียกรวมกันได้ว่าเป็น*การรู้จำเสียงพูดโดยใช้วิธีทางสถิติ* เนื่องจากเป็นการรู้จำเสียงพูดที่อาศัยหลักการเรียนรู้ของเครื่องสำหรับสร้างแบบจำลองทางสถิติจากข้อมูลตัวอย่างเพื่อนำมาใช้ในการรู้จำ

อีกส่วนหนึ่งในระบบรู้จำเสียงพูด คือส่วนการประกอบหน่วยย่อยทางภาษาของเสียงพูดออกมาให้เป็นลำดับของคำในภาษานั้น โดยทั่วไปจะใช้วิธีการค้นหาแบบวิเทอบี โดยมีพจนานุกรมและแบบจำลองทางภาษาเข้ามาช่วย การค้นหานี้จะรับหน่วยย่อยทางภาษาของเสียงพูดที่รู้จำได้และให้ผลลัพธ์คือลำดับของคำที่ดีที่สุดออกมาเป็นคำตอบ

การรู้จำเสียงพูดภาษาไทยที่ผ่านมาโดยมากเป็นการรู้จำเสียงพูดไม่ต่อเนื่อง การรู้จำเสียงพูดต่อเนื่องอัตโนมัติภาษาไทยส่วนใหญ่จะทำการรู้จำหน่วยย่อยทางภาษาของเสียงพูดโดยใช้แบบจำลองมาร์คอฟซ่อนตัว งานวิจัยนี้จะทดลองสร้างระบบรู้จำเสียงพูดต่อเนื่องอัตโนมัติภาษาไทยโดยใช้นิวรอลเน็ตเวิร์กเป็นหลักในการรู้จำหน่วยย่อยทางภาษาของเสียงพูด ด้วยหวังว่าความสามารถในการแบ่งแยกของนิวรอลเน็ตเวิร์กจะช่วยจัดการกับความแปรผันทางเสียง และทำให้ผลลัพธ์การรู้จำออกมาดีขึ้น รวมทั้งเป็นจุดเริ่มต้นในการใช้การเรียนรู้แบบแบ่งแยกเพื่อรู้จำเสียงพูดต่อเนื่องอัตโนมัติภาษาไทยต่อไป

มนุษย์มีความใฝ่ฝันมาช้านานในการสร้างเครื่องจักรที่มีความฉลาดทัดเทียมตัวมนุษย์เอง ความสามารถในการรู้จำเสียงพูดก็เป็นหนึ่งในคุณสมบัติที่เครื่องจักรอันชาญฉลาดนั้นควรจะมี ซึ่งระบบรู้จำเสียงพูดต่อเนื่องอัตโนมัติจะช่วยเติมเต็มความใฝ่ฝันในส่วนนี้ นอกจากนั้นในหลายๆ งานเมื่อนำการรู้จำเสียงพูดมาใช้เป็นตัวเชื่อมประสานระหว่างมนุษย์กับคอมพิวเตอร์ จะทำให้งานนั้นสะดวก รวดเร็ว และมีประสิทธิภาพขึ้นมาก ตัวอย่างเช่น ในการสอบถามข้อมูลทางโทรศัพท์ หรือการบอกจดในโปรแกรมประมวลผลคำ เป็นต้น ซึ่งผลจากงานวิจัยชิ้นนี้คาดว่าจะมีประโยชน์ในการนำไปประยุกต์ใช้กับงานต่างๆ ที่เหมาะสมสืบไป

1.2 วัตถุประสงค์

เพื่อศึกษาวิจัยการรู้จำเสียงพูดต่อเนื่องอัตโนมัติภาษาไทย โดยใช้นิวรอลเน็ตเวิร์กเป็นหลักในการรู้จำหน่วยย่อยทางภาษาของเสียงพูด

1.3 ขอบเขตของการวิจัย

1. พัฒนาระบบรู้จำเสียงพูดต่อเนื่องอัตโนมัติแบบไม่ขึ้นกับผู้พูด โดยใช้นิวรอลเน็ตเวิร์กเป็นหลักในการรู้จำหน่วยย่อยทางภาษาของเสียงพูด
2. ทำการทดลองกับฐานข้อมูลเสียงพูดชื่อไทย และฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

1.4 ขั้นตอนและวิธีดำเนินงานวิจัย

1. ศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง
2. ออกแบบและพัฒนาสถาปัตยกรรมของระบบรู้จำเสียงพูด
3. ทดสอบประสิทธิภาพของระบบ
4. วิเคราะห์และสรุปผล
5. จัดทำวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยที่มีความถูกต้องในการรู้จำมากขึ้น
2. เป็นแนวทางในการใช้การเรียนรู้แบบแบ่งแยกเพื่อรู้จำเสียงพูดต่อเนื่องอัตโนมัติภาษาไทยต่อไป
3. สามารถปรับปรุงระบบรู้จำเสียงพูดเพื่อนำไปใช้งานกับโปรแกรมประยุกต์อื่นๆ ได้ เช่น ในการสอบถามข้อมูลทางโทรศัพท์

1.6 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

วิทยานิพนธ์นี้แบ่งเนื้อหาออกเป็น 4 บทดังนี้ เริ่มจากบทที่ 1 คือบทนำบทนี้ซึ่งกล่าวถึงที่มาและความสำคัญของปัญหา รวมทั้งวัตถุประสงค์ของงานวิจัยชิ้นนี้ บทที่ 2 เป็นการรวบรวมทฤษฎีพื้นฐานและงานวิจัยที่เกี่ยวข้อง บทที่ 3 กล่าวถึงรายละเอียดของระบบรู้จำเสียงพูดที่ได้พัฒนาขึ้น รวมถึงการทดลองต่างๆ และสุดท้าย บทที่ 4 จะเป็นข้อสรุปและข้อเสนอแนะจากการวิจัย

1.7 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตีพิมพ์เป็นบทความทางวิชาการและนำเสนอในงานประชุมวิชาการ มีรายละเอียดดังนี้

1. Prasertsak Pungprasertying and Boonserm Kijsirikul. 2004. An Automatic Dialing System Using Speech of Thai Names. The 8th National Computer Science and Engineering Conference (NCSEC2004), October, Songkhla, Thailand.
2. Prasertsak Pungprasertying and Boonserm Kijsirikul. 2006. Phoneme Recognition in Thai Speech Using Neural Networks. The 10th National Computer Science and Engineering Conference (NCSEC2006), October, Khonkaen, Thailand.

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การรู้จำเสียงพูดภาษาไทยโดยใช้นิวรอลเน็ตเวิร์กนั้นจำเป็นต้องพึ่งพางานวิจัยที่เกี่ยวข้องมากมาย และอาศัยพื้นฐานทางทฤษฎีหลายประการ โดยในขั้นแรกจะกล่าวถึงการรู้จำเสียงพูดไม่ต่อเนื่อง ซึ่งเป็นการรู้จำรูปแบบประเภทหนึ่ง ที่ข้อมูลเข้าอยู่ในรูปอนุกรมเวลาและมีความแปรผันทางเสียง โดยจะชี้ให้เห็นถึงปัญหาในการรู้จำข้อมูลที่เป็นอนุกรมเวลา และสาเหตุของความแปรผันทางเสียงเหล่านั้น

ในส่วนต่อไปจะบรรยายถึงวิธีการที่ใช้ในการรู้จำเสียงพูดไม่ต่อเนื่อง ซึ่งได้แจกแจงออกเป็น 4 ประเภทหลัก ได้แก่ การจับคู่แผ่นแบบ การใช้ความรู้ทางสวนศาสตร์-สัทศาสตร์ การสร้างแบบจำลองเฟ้นสุ่ม และการใช้การเรียนรู้แบบแบ่งแยก โดยสองประเภทหลังนั้นเป็นการรู้จำเสียงพูดโดยใช้วิธีทางสถิติ ซึ่งให้ประสิทธิภาพสูงในปัจจุบัน ในที่นี้จะเน้นกล่าวถึงแบบจำลองมาร์คอฟซ่อนตัว อันเป็นแบบจำลองเฟ้นสุ่มที่ใช้กันโดยทั่วไปในการรู้จำเสียงพูด และนิวรอลเน็ตเวิร์ก ซึ่งเป็นการเรียนรู้แบบแบ่งแยกที่นิยมอย่างแพร่หลายในการรู้จำรูปแบบต่างๆ

ส่วนการรู้จำเสียงพูดต่อเนื่องอัตโนมัตินั้น จำเป็นต้องใช้แบบจำลองทางเสียง แบบจำลองทางภาษา และกระบวนการค้นหา เข้ามาประกอบกัน ซึ่งแบบจำลองทางเสียงสามารถสร้างได้โดยใช้วิธีเดียวกับการรู้จำเสียงพูดไม่ต่อเนื่อง ส่วนแบบจำลองทางภาษาและกระบวนการค้นหานั้น จะเน้นที่แบบจำลองทางภาษาแบบเอ็นแกรมและกระบวนการค้นหาโดยใช้อัลกอริทึมการผ่านโทเคนตามลำดับ

ในส่วนของงานวิจัยที่เกี่ยวข้อง จะสรุปงานวิจัยที่ใช้นิวรอลเน็ตเวิร์กมาทำการรู้จำเสียงพูดทั้งเสียงพูดไม่ต่อเนื่อง ซึ่งเป็นงานวิจัยในระยะแรก มาสู่การรู้จำเสียงพูดต่อเนื่องอัตโนมัติ ซึ่งจำเป็นต้องใช้นิวรอลเน็ตเวิร์กมาผสมผสานกับแบบจำลองมาร์คอฟซ่อนตัว ซึ่งพบว่าให้ผลการรู้จำที่ดีกว่าการใช้แบบจำลองมาร์คอฟซ่อนตัวเพียงอย่างเดียว

เป็นที่แน่นอนว่าในการรู้จำเสียงพูดจะขาดเสียไม่ได้เลยซึ่งความรู้พื้นฐานเกี่ยวกับที่มาและลักษณะของเสียงพูด แต่เพื่อไม่ให้เป็นการบั่นทอนความกระชับของบท จึงขอยกเนื้อหาในส่วนนี้ไปกล่าวไว้ที่ภาคผนวก ก ว่าด้วยเรื่องธรรมชาติของเสียงพูด อนึ่ง ในบทนี้มีการยกตัวอย่างโดยใช้สัญลักษณ์แทนหน่วยเสียงภาษาไทยไว้พอสมควร ซึ่งส่วนรายละเอียดสามารถติดตามได้จากภาคผนวก ก ในเรื่องสัทศาสตร์ภาษาไทย และภาคผนวก ข อันแสดงหน่วยเสียงที่ใช้

2.1 การรู้จำเสียงพูดไม่ต่อเนื่อง

2.1.1 ลักษณะของปัญหา

ปัญหาการรู้จำเสียงพูดไม่ต่อเนื่องมีลักษณะดังนี้

1. เป็นการรู้จำหน่วยย่อยทางภาษาของเสียงพูดที่มีช่วงเวลาไม่นานนัก เช่น หน่วยเสียง พยางค์ หรือคำ
2. หน่วยย่อยทางภาษาที่ใช้ในการรู้จำมีจำนวนจำกัด
3. เสียงพูดมีลักษณะไม่ต่อเนื่อง โดยอาจจะพูดหน่วยย่อยทางภาษาหนึ่ง แล้วคั่นด้วยเสียงเงียบ ก่อนจะพูดหน่วยย่อยทางภาษาถัดไป หรือเสียงพูดหนึ่งอาจจะบรรจุหน่วยย่อยทางภาษาเพียงหน่วยเดียว
4. การรู้จำทำทีละหน่วยย่อยทางภาษา และไม่มีการประกอบหน่วยย่อยทางภาษาเหล่านั้นเข้าด้วยกัน

2.1.2 มุมมองต่อปัญหา

การรู้จำเสียงพูดไม่ต่อเนื่องอาจมองได้ว่าเป็นปัญหาการรู้จำรูปแบบชนิดหนึ่ง หรือถ้าจะกล่าวให้เฉพาะเจาะจงกว่านั้นก็คือ การรู้จำเสียงพูดไม่ต่อเนื่อง ISR เป็นฟังก์ชัน หรืออัลกอริทึม ซึ่งรับข้อมูลเข้าเป็นความดังของเสียงพูดที่ได้จากการชักตัวอย่าง ณ เวลาต่างๆ คือ $s_1, s_2, ..., s_T$ และให้ผลลัพธ์เป็นหน่วยย่อยทางภาษา u หรือ

$$u = ISR(s_1, s_2, ..., s_T) \tag{2.1}$$

แต่การรู้จำโดยให้ข้อมูลเข้าเป็นอนุกรมเวลาของความดังในเสียงพูดโดยตรงเลยนั้นเป็นเรื่องที่ยากมาก จากการศึกษาเสียงพูดของมนุษย์พบว่าลักษณะสำคัญที่เป็นเครื่องบ่งชี้ความแตกต่างของหน่วยย่อยทางภาษาในเสียงพูดนั้นคือลักษณะเชิงความถี่ และมีค่าคงที่ในช่วงระยะเวลาหนึ่ง ซึ่งถ้าให้ข้อมูลเข้าของระบบเป็นลักษณะสำคัญนี้ในแต่ละช่วงของเสียงพูด จะทำให้ปัญหาการรู้จำเสียงพูดทำได้ง่ายขึ้น และสามารถเขียนระบบรู้จำเสียงพูดใหม่ให้อยู่ในรูปทั่วไปได้ว่า

$$u = ISR(x_1, x_2, ..., x_N) \tag{2.2}$$

เมื่อ $x_n = f(s_{t(n-1)+1}, ..., s_{t(n)})$ โดย f เป็นฟังก์ชันสำหรับหาลักษณะสำคัญของเสียงพูด และ $t(n)$ คือตำแหน่งสุดท้ายของตัวอย่างเสียงพูดในการหาลักษณะสำคัญ x_n

แม้จะผ่านการหาลักษณะสำคัญแล้ว ข้อมูลเข้าของระบบรู้จำเสียงพูดไม่ต่อเนื่องก็ยังมีลักษณะพิเศษที่ทำให้การรู้จำทำได้ยากอยู่สองประการ คือ

2.1.2.1 ข้อมูลเข้าของระบบเป็นข้อมูลอนุกรมเวลา

เมื่อไม่มีเรื่องของเวลาเข้ามาเกี่ยวข้อง การรู้จำรูปแบบ PR สามารถเขียนให้อยู่ในรูปทั่วไปได้ว่า

$$o = PR(x_1, x_2, ..., x_N) \tag{2.3}$$

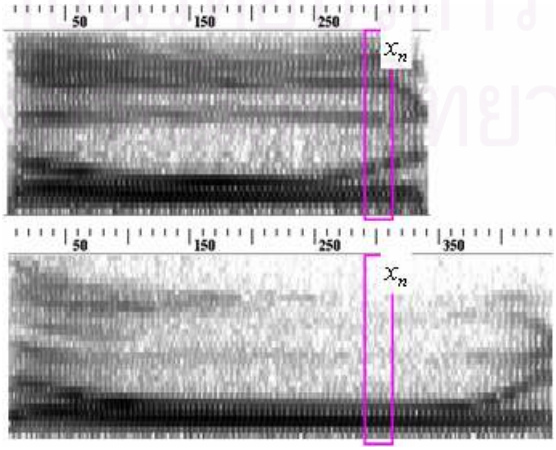
เมื่อ x_n เป็นค่าคุณสมบัติตัวที่ n และ o คือผลลัพธ์ที่ได้จากการรู้จำ

ในปัญหาการรู้จำรูปแบบส่วนมาก พบว่า N หรือจำนวนข้อมูลเข้าใน (2.3) มีค่าคงที่เหมือนกันทั้งหมดทุกชุดข้อมูล ขณะที่ N ในการรู้จำรูปแบบข้อมูลอนุกรมเวลานั้นอาจมีค่าแตกต่างกันไปในแต่ละชุดข้อมูล เช่น N ในการรู้จำเสียงพูด (2.2) นั้นจะมีค่าเปลี่ยนแปลงไปตามระยะเวลาการออกเสียง

นอกจากนี้ ในการรู้จำรูปแบบของบางปัญหา คุณสมบัติ x_n จะเป็นคุณสมบัติเดียวกันเหมือนกันทั้งหมดทุกชุดข้อมูล เช่น ปัญหา *PlayTennis* (Mitchell [1]) คุณสมบัติ x_1 จะเป็นคุณสมบัติ *Outlook* เสมอ ไม่ว่าจะเป็นชุดข้อมูลใด ขณะที่ในการรู้จำเสียงพูด เราไม่อาจตีความคุณสมบัติของ x_n ได้อย่างชัดแจ้ง เช่นรูปที่ 2.1 รูปที่ 2.1 ซึ่งแสดงสเปกโตรแกรมของเสียงพูดคำว่าศูนย์ (/suu4n^/) สองเสียงพูด เมื่อพิจารณา x_n ที่ n เดียวกัน พบว่า x_n ของเสียงพูดแรกแสดงลักษณะของหน่วยเสียง /n^/ ขณะที่ x_n ของเสียงพูดหลังแสดงลักษณะของหน่วยเสียง /u/

ยิ่งไปกว่านั้น ในปัญหาการรู้จำรูปแบบทั่วไป จะถือว่าทุกคุณสมบัติในข้อมูลเป็นอิสระต่อกัน แต่สำหรับข้อมูลอนุกรมเวลาแล้ว $x_1, x_2, ..., x_N$ จะมีความขึ้นต่อกันอยู่

สังเกตว่าถ้าหดแกนเวลาของสเปกโตรแกรมที่แทนเสียงพูดหลังให้สั้นลง หรือยืดแกนเวลาของสเปกโตรแกรมที่แทนเสียงพูดแรกให้ยาวขึ้น จะพบว่าเสียงพูดทั้งสองมีลักษณะคล้ายๆ กัน จึงสรุปได้ว่าการเลื่อนทางเวลาเป็นสาเหตุที่ทำให้ข้อมูลโดยรวมของเสียงพูดทั้งสองแตกต่างกัน โดยอาจเรียกปรากฏการณ์นี้ว่า *ความแปรผันทางเวลา* ของเสียงพูด



รูปที่ 2.1 ตัวอย่างความแปรผันทางเวลา

เพื่อให้เกิดความเข้าใจที่ง่ายขึ้นในเรื่องการรู้จำรูปแบบข้อมูลอนุกรมเวลา เราอาจทำการแบ่งนับข้อมูลเข้า จนได้ c_n เป็นอักขระที่สามารถใช้แทน x_n และการรู้จำเสียงพูดอาจเขียนได้ว่า

$$u = ISR(c_1, c_2, ..., c_N) \tag{2.4}$$

ระบบรู้จำรูปแบบอนุกรมเวลาจำเป็นต้องไม่ยึดติดกับตำแหน่งเวลาที่ข้อมูลค่าต่างๆ ปรากฏ หากแต่ต้องพิจารณาลำดับและความสัมพันธ์ของข้อมูลค่าต่างๆ เป็นสำคัญ เช่น

$$ISR(\text{/s/}, \text{/s/}, \text{/u/}, \text{/u/}, \text{/u/}, \text{/n\^{}/}, \text{/n\^{}/}) = \text{/suu4n\^{}/} \tag{2.5}$$

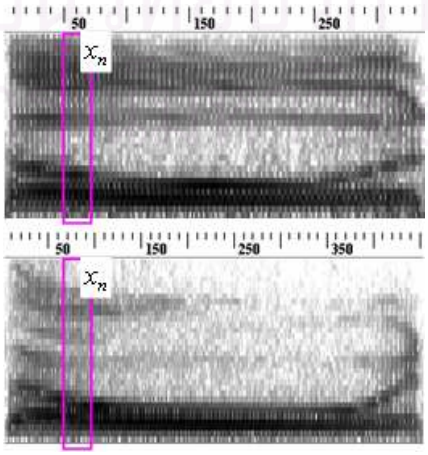
และ $$ISR(\text{/s/}, \text{/s/}, \text{/u/}, \text{/u/}, \text{/u/}, \text{/u/}, \text{/u/}, \text{/n\^{}/}, \text{/n\^{}/}) = \text{/suu4n\^{}/} \tag{2.6}$$

โดย (2.5) อาจเทียบได้กับการรู้จำเสียงพูดในรูปที่ 2.1 ด้านบน และ (2.6) อาจเทียบได้กับการรู้จำเสียงพูดในรูปที่ 2.1 ด้านล่าง สังเกตว่าแม้ผลลัพธ์จะเป็น /suu4n^/ เหมือนกัน แต่ที่ n เท่ากับ 6 ข้อมูลอาจจะมีค่าเป็น /n^/ ดังใน (2.5) หรือ /u/ ดังใน (2.6) ก็ได้ ระบบรู้จำรูปแบบข้อมูลอนุกรมเวลาจึงไม่ควรเน้นหนักในส่วนนี้ แต่ควรจะรู้จำลำดับและความสัมพันธ์ของ /s/, /u/ และ /n^/ มากกว่า โดยสามารถกล่าวให้แรงขึ้นไปอีกได้ว่า $ISR(x)$ ควรจะเท่ากับ /suu4n^/ ในทุกๆ x ที่อยู่ใน $\{\text{/s/}\}^+\{\text{/u/}\}^+\{\text{/n\^{}/}\}^+$ เมื่อ $^+$ คือ ตัวดำเนินการคลีนที่ไม่รวมเซตว่างเป็นผลลัพธ์

เป็นที่น่าสังเกตว่าข้อมูลอนุกรมเวลาที่เข้าสู่ระบบรู้จำเสียงพูดมีลักษณะเช่นเดียวกับสตริงในภาษาฟอร์มัล ทว่าจริงๆ แล้วการรู้จำเสียงพูดซับซ้อนกว่านั้น เนื่องจากข้อมูลเสียงพูดยังมีความแปรผันทางเสียงอยู่ด้วย ซึ่งจะกล่าวถึงในหัวข้อถัดไป

2.1.2.2 **ข้อมูลเข้าของระบบมีความแปรผันทางเสียง**

จากรูปที่ 2.1 ถ้าเราหดแกนเวลาของสเปกโตรแกรมเสียงพูดหลังให้สั้นลงโดยมีความยาวเท่ากับเสียงพูดแรก จะได้สเปกโตรแกรมดังรูปที่ 2.2 เมื่อพิจารณา x_n ของเสียงพูดทั้งสองที่ n เดียวกัน พบว่ายังไม่เหมือนกันนัก แม้ x_n ของเสียงพูดทั้งสองจะแสดงลักษณะของเสียงพยัญชนะต้นเดียวกัน ซึ่งอาจเรียกว่าเป็น *ความแปรผันทางเสียง*



รูปที่ 2.2 ตัวอย่างความแปรผันทางเสียง

เพื่อให้เกิดความเข้าใจที่ง่ายขึ้นในเรื่องความแปรผันทางเสียง เราอาจใช้กรรมวิธี (2.4) มายกตัวอย่างว่าระบบรู้จำเสียงพูดที่ดีจะต้องรองรับความแปรผันทางเสียงได้ เช่น

$$ISR(/s/,/f/,/u/,/u/,/u/,/n\hat{}/,/n\hat{}/) = /suu4n\hat{}/ \quad (2.7)$$

และ $$ISR(/s/,/s/,/u/,/u/,/u/,/n\hat{}/,/n\hat{}/) = /suu4n\hat{}/ \quad (2.8)$$

โดย (2.7) อาจเทียบได้กับการรู้จำเสียงพูดในรูปที่ 2.2 ด้านบน และ (2.8) อาจเทียบได้กับการรู้จำเสียงพูดในรูปที่ 2.2 ด้านล่าง สังเกตว่าแม้ผลลัพธ์จะเป็น /suu4n^/ เหมือนกัน แต่ที่ n เท่ากับ 2 ข้อมูลอาจจะมีค่าเป็น /f/ ดังใน (2.7) หรือ /s/ ดังใน (2.8) ก็ได้ ซึ่งทั้ง (2.7) และ (2.8) เป็นเพียงการสมมติปัญหาความแปรผันทางเสียงให้อยู่ในรูปอย่างง่ายเท่านั้น เห็นได้ว่าความแปรผันทางเสียงนี่เองที่ทำให้เรามองเสียงพูดเป็นสตริงในภาษาฟอร์มัลไม่ได้

สาเหตุสำคัญที่ก่อให้เกิดความแปรผันทางเสียงอาจแยกได้ดังนี้

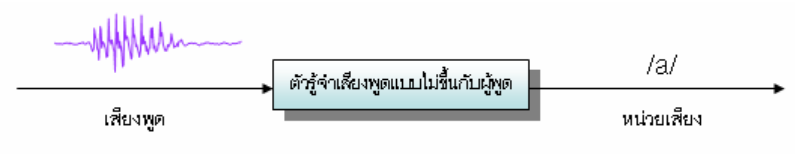
การออกเสียงจากผู้พูดหลายคน

ผู้พูดแต่ละคนจะมีอวัยวะการออกเสียงและการเคลื่อนไหวของอวัยวะการออกเสียงที่ต่างกัน นอกจากนี้ ผู้พูดแต่ละคนยังมีลีลาการพูดที่ไม่เหมือนกัน เช่นในภาษาไทย ผู้พูดบางคนออกเสียงพยัญชนะควบกล้ำได้ชัดเจน ขณะที่ผู้พูดบางคนออกเสียงพยัญชนะควบกล้ำเพียงเล็กน้อยเท่านั้น และในบางครั้ง ภูมิลำเนาก็ส่งผลต่อสำเนียงการพูดได้มาก ซึ่งสิ่งเหล่านี้ทั้งหมดล้วนเป็นปัจจัยที่ทำให้เสียงพูดที่ออกมาจากผู้พูดต่างคนให้ลักษณะทางเสียงต่างกัน แม้จะเป็นการออกเสียงของหน่วยย่อยทางภาษาเดียวกันก็ตาม

ระบบรู้จำเสียงพูดบางระบบถูกสร้างมาให้ขึ้นกับผู้พูด ซึ่งทำให้รู้จำเสียงพูดได้เฉพาะคนหรือเฉพาะกลุ่มเท่านั้น ขณะที่ระบบรู้จำเสียงพูดที่ไม่ขึ้นกับผู้พูดจะรู้จำเสียงพูดได้หมดไม่ว่าใครจะเป็นผู้พูด โดยระบบรู้จำเสียงพูดแบบไม่ขึ้นกับผู้พูดจะพัฒนาได้ยากกว่าระบบรู้จำเสียงพูดแบบขึ้นกับผู้พูด เนื่องจากความแปรผันทางเสียงที่มีสาเหตุมาจากการออกเสียงของผู้พูดหลายคนนั่นเอง

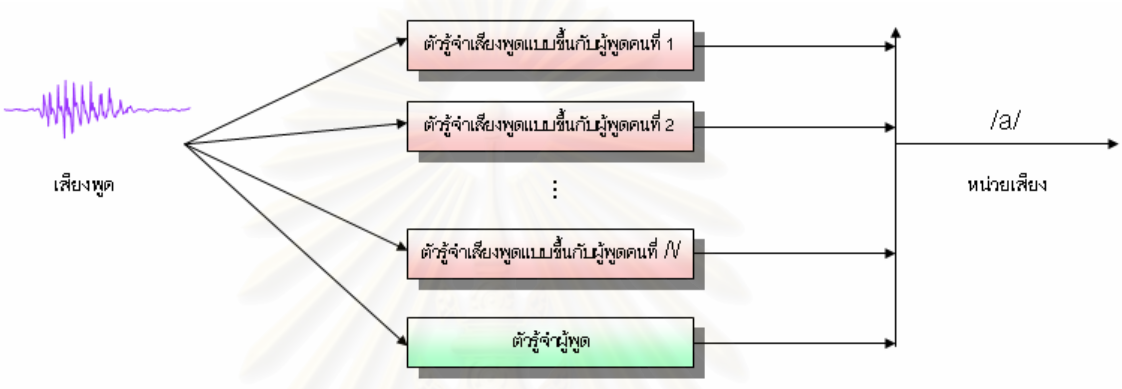
ระบบรู้จำเสียงพูดแบบไม่ขึ้นกับผู้พูดก็มีวิธีการออกแบบอยู่หลายวิธี (Tebelskis [2]) เช่น

1. ใช้ตัวรู้จำเสียงพูดตัวเดียวสำหรับรู้จำเสียงพูดจากผู้พูดทุกคน ดังรูปที่ 2.3 ซึ่งวิธีนี้ตัวรู้จำที่ใช้ต้องมีประสิทธิภาพเพียงพอที่จะจำแนกเสียงพูดที่มีความแปรผันทางเสียงจากการออกเสียงของผู้พูดหลายคนได้ นอกจากนี้ตัวรู้จำเสียงพูดแบบไม่ขึ้นกับผู้พูดยังต้องอาศัยข้อมูลจากผู้พูดจำนวนมากเพื่อเป็นตัวอย่างในการเรียนรู้ความแปรผันทางเสียงจากการออกเสียงของผู้พูดหลายคน



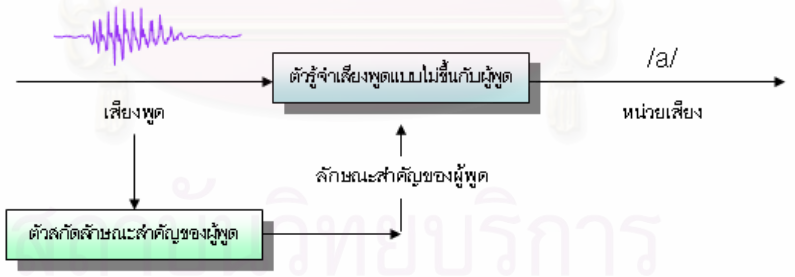
รูปที่ 2.3 การใช้ตัวรู้จำเสียงพูดตัวเดียว

2. นำเสียงพูดเข้าสู่ตัวรู้จำเสียงพูดแบบขึ้นกับผู้พูด พร้อมกันนั้นเสียงพูดจะเข้าสู่ตัวรู้จำผู้พูด และตัวรู้จำผู้พูดจะให้น้ำหนักว่าควรจะใช้ผลลัพธ์จากตัวรู้จำเสียงพูดแบบขึ้นกับผู้พูดตัวใด ดังรูปที่ 2.4



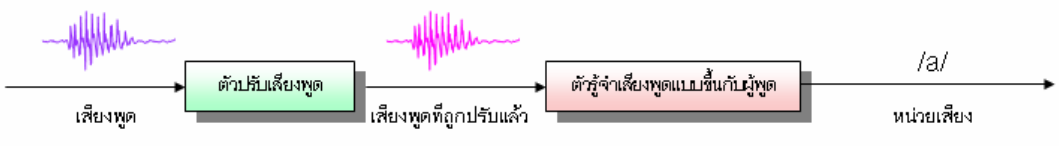
รูปที่ 2.4 การใช้ตัวรู้จำเสียงพูดแบบขึ้นกับผู้พูดพร้อมด้วยตัวรู้จำผู้พูด

3. นำเสียงพูดเข้าสู่ตัวสกัดลักษณะสำคัญของผู้พูด ซึ่งจะให้ลักษณะสำคัญของผู้พูดอย่างคร่าวๆ และนำลักษณะสำคัญนั้นเข้าสู่ตัวรู้จำเป็นข้อมูลเพิ่มเติม เพื่อเพิ่มข้อมูลให้ตัวรู้จำเสียงพูดนั้น ดังรูปที่ 2.5



รูปที่ 2.5 การใช้ตัวสกัดลักษณะสำคัญของผู้พูดเข้าช่วย

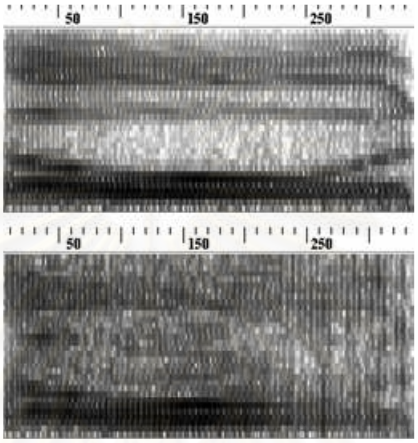
4. ทำการปรับเสียงพูดที่เข้ามาให้มีลักษณะใกล้เคียงกับเสียงพูดต้นแบบมากที่สุด ก่อนที่จะนำเข้าสู่ตัวรู้จำเสียงพูดที่ขึ้นกับต้นแบบนั้น ดังรูปที่ 2.6



รูปที่ 2.6 การปรับเสียงพูดให้ใกล้เคียงกับเสียงพูดต้นแบบ

สัญญาณรบกวน

รูปที่ 2.7 แสดงสเปกโตรแกรมของเสียงพูดคำว่าศูนย์สองเสียงพูด ซึ่งไม่มีความแปรผันทางเสียงจากสาเหตุอื่นเลย นอกเหนือจากสัญญาณรบกวนที่เข้าแทรกในเสียงพูดหลัง ทำให้ลักษณะทางเสียงของเสียงพูดผิดเพี้ยนไปมากเมื่อเทียบกับเสียงพูดที่ไม่มีสัญญาณรบกวนเข้าแทรก ซึ่งสัญญาณรบกวนอาจจะมีที่มาจากหลายสาเหตุ เช่น จากสภาพแวดล้อม หรือจากอุปกรณ์การส่งและบันทึกเสียง เป็นต้น



รูปที่ 2.7 ความแปรผันทางเสียงจากสัญญาณรบกวน

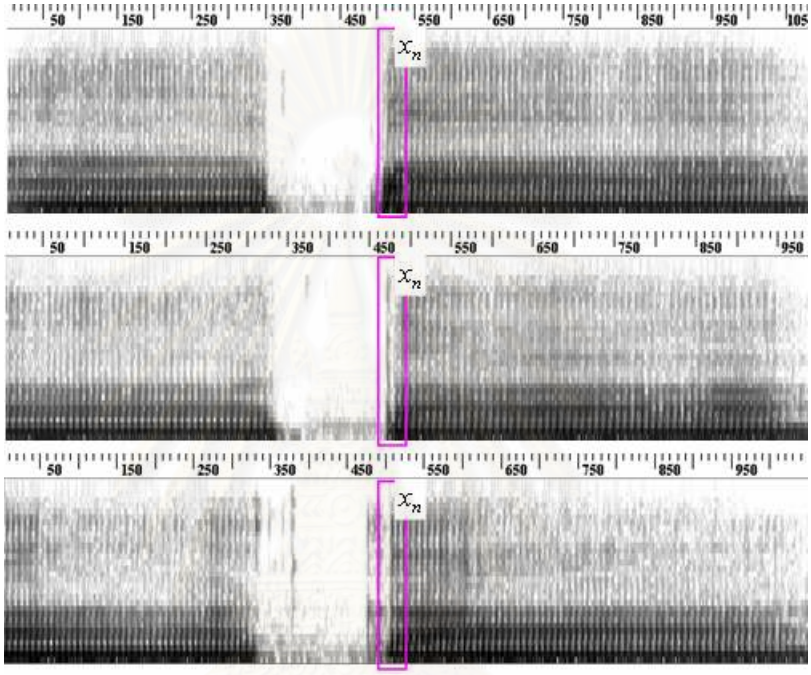
หูมนุษย์นับว่าเป็นระบบรู้จำเสียงพูดที่ทนต่อสัญญาณรบกวนได้ดีมาก สังเกตว่าแม้เราจะอยู่ในที่ที่อึกทึกครึกโครมปานใด เราก็ยังเลือกที่จะฟังเฉพาะเสียงที่อยากฟังได้ การทำให้ระบบรู้จำเสียงพูดทนต่อสัญญาณรบกวนวิธีหนึ่งคือการหาลักษณะสำคัญของเสียงให้ลักษณะที่ออกมานั้นแปรผันกับสัญญาณรบกวนน้อยที่สุด การหาลักษณะสำคัญของเสียงบางวิธีก็ได้รับแรงบันดาลใจจากการทำงานของหูมนุษย์ เช่น พีแอลพี (Hermansky [3]) และราสตาพีแอลพี (Hermansky and Morgan [4]) ที่ถือว่าเป็นวิธีการหาลักษณะสำคัญของเสียงพูดที่ทนต่อสัญญาณรบกวนได้ดีวิธีหนึ่ง

การออกเสียงร่วม

เนื่องจากเสียงพูดที่ได้เป็นผลมาจากการเคลื่อนไหวของอวัยวะการออกเสียงในช่องทางเดินเสียง ซึ่งเมื่อออกเสียงเสียงหนึ่งมาแล้ว เมื่อต้องการออกเสียงถัดไป อวัยวะการออกเสียงไม่สามารถเคลื่อนให้อยู่ในรูปที่สามารถออกเสียงถัดไปได้ในทันที แต่จะค่อยๆ เคลื่อน เปลี่ยนรูปจากการออกเสียงเสียงหนึ่งไปสู่การออกเสียงอีกเสียงหนึ่ง ลักษณะเสียงพูดที่ได้ออกมาในเวลาหนึ่งจึงขึ้นอยู่กับเสียงพูดก่อนหน้าและเสียงพูดถัดไป ทำให้มีความแปรผันทางเสียงที่เกิดจากการออกเสียงร่วม ยิ่งหน่วยย่อยทางภาษาที่เป็นผลลัพธ์จากการรู้จำเล็กลงเท่าใด ก็ยิ่งทำให้ความแปรผันทางเสียงที่เกิดจากการออกเสียงร่วมมากขึ้นเท่านั้น เพราะการออกเสียงของแต่ละหน่วยย่อยทาง

ภาษามีความกระชั้นมากขึ้น เช่น หน่วยย่อยทางภาษาที่เป็นหน่วยเสียง จะได้รับผลกระทบจากการออกเสียงร่วมมากกว่าหน่วยย่อยทางภาษาที่เป็นพยางค์ หรือคำ

ดังรูปที่ 2.8 แสดงลักษณะของหน่วยเสียง /a/ ในบริบทต่างๆ จากเสียงพูดคำว่า /#aa0paa0/, /#aa0taa0/ และ /#aa0kaa0/ ตามลำดับ เห็นได้ว่า x_n ของหน่วยเสียง /a/ มีลักษณะเปลี่ยนไปตามหน่วยเสียงที่เป็นบริบท ซึ่งได้แก่ /p/, /t/ และ /k/ อันเป็นผลมาจากความแปรผันทางเสียงที่เกิดจากการออกเสียงร่วม



รูปที่ 2.8 ความแปรผันทางเสียงจากการออกเสียงร่วม

ปัญหาการออกเสียงร่วมเป็นปัญหาที่สำคัญมากในการรู้จำเสียงพูด ระบบรู้จำเสียงพูดส่วนใหญ่แก้ปัญหานี้โดยให้ผลลัพธ์จากการรู้จำเป็นผลลัพธ์ที่ขึ้นกับบริบท เช่น ถ้าระบบรู้จำเสียงพูดต้องการที่จะรู้จำหน่วยเสียงของ x_n ในรูปที่ 2.8 จากเดิม การรู้จำทั้งเสียงพูดทั้งสามให้ผลลัพธ์เป็นหน่วยเสียง /a/ เพียงหน่วยเดียว ก็จะเปลี่ยนเป็นให้รู้จำหน่วยเสียง /p-a/, /t-a/ และ /k-a/ แทน ตามลำดับ โดย /p-a/ คือสัญลักษณ์ที่ใช้แทนหน่วยเสียงหนึ่ง ซึ่งเป็นหน่วยเสียงของ /a/ ที่นำหน้าด้วยหน่วยเสียง /p/ เช่นเดียวกับ /t-a/ ซึ่งแทนหน่วยเสียงของ /a/ ที่นำหน้าด้วยหน่วยเสียง /t/ และ /k-a/ ที่แทนหน่วยเสียงของ /a/ ที่นำหน้าด้วยหน่วยเสียง /k/ แบบจำลองที่สร้างได้เพื่อใช้อธิบายหน่วยเสียงเหล่านี้อาจเรียกว่าเป็นแบบจำลองไดโฟน

ซึ่งวิธีนี้ช่วยลดความแปรผันทางเสียงที่เกิดจากการออกเสียงร่วมของหน่วยเสียง โดยผลักภาระไปให้ระบบรู้จำเสียงพูดรู้จำหน่วยเสียงที่มากขึ้น แม้ความแปรผันของข้อมูลในแต่ละหน่วยเสียงนั้นน้อยกว่า แทนที่จะให้ระบบรู้จำเสียงพูดรู้จำหน่วยเสียงในจำนวนน้อย แต่ความแปรผันของข้อมูลในแต่ละหน่วยเสียงนั้นมีมาก

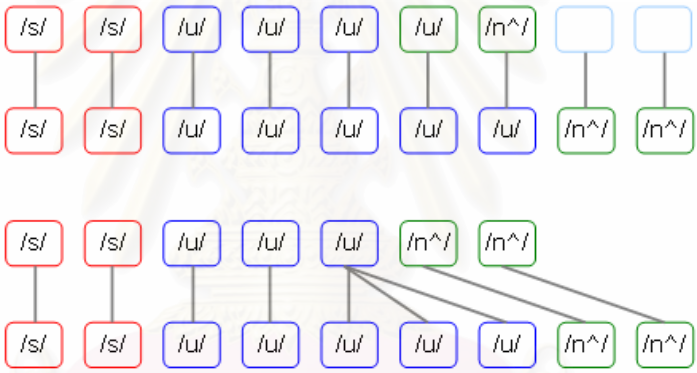
2.1.3 วิธีการที่ใช้ในการรู้จำเสียงพูดไม่ต่อเนื่อง

วิธีการทั่วไปที่ใช้สำหรับการรู้จำเสียงพูดไม่ต่อเนื่องมีดังนี้

2.1.3.1 การจับคู่แผ่นแบบ

เป็นการรู้จำเสียงพูดโดยนำเสียงพูดที่เข้ามามาเปรียบเทียบกับแผ่นแบบเสียงพูดที่มีอยู่ แต่เนื่องจากเสียงพูดมีความแปรผันทางเวลา ทำให้ไม่สามารถเปรียบเทียบ x_n ของเสียงพูดสองเสียงที่ตำแหน่ง n เดียวกันอย่างตรงไปตรงมาได้ จำเป็นต้องอาศัยการปรับแนวแบบไม่เชิงเส้นเข้ามาช่วยในการเปรียบเทียบแทน (Vintsyuk [5]) (Sakoe and Chiba [6])

เราอาจใช้ตัวอย่างคล้ายๆ กับ (2.5) และ (2.6) มาแสดงให้เห็นภาพการเปรียบเทียบระหว่างเสียงพูดสองเสียงดังรูปที่ 2.9 โดยภาพบนเป็นการเปรียบเทียบกันโดยตรง จะเห็นว่าวิธีการนี้ทำให้ผลลัพธ์ที่ออกมาบ่งชี้ว่าเสียงพูดทั้งสองต่างกันมาก ทั้งที่ถ้าใช้การปรับแนวแบบไม่เชิงเส้นดังภาพล่างแล้ว เสียงพูดทั้งสองจะถูกมองว่าเป็นเสียงพูดเดียวกัน



รูปที่ 2.9 การปรับแนวแบบตรงตัวและการปรับแนวแบบไม่เชิงเส้น

เมื่อให้เสียงพูดแผ่นแบบ x มีลักษณะสำคัญ $x=(x_1,x_2,...,x_{N_x})$ และเสียงพูดที่จะทำการรู้จำ y มีลักษณะสำคัญ $y=(y_1,y_2,...,y_{N_y})$ เราสามารถนิยามการปรับแนวระหว่าง x และ y ว่าคือ C_K โดย $C_K=(c_1,c_2,...,c_K)$ เมื่อ c_k เป็นการจับคู่ของสองลักษณะสำคัญจาก x และ y หรือเขียนได้ว่า $c_k=(x_{\phi_x(k)},y_{\phi_y(k)})$ โดย $\phi_x(k)$ และ $\phi_y(k)$ เป็นการแมพจาก $\{1,2,...,K\}$ กลับไปยัง $\{1,2,...,N_x\}$ และ $\{1,2,...,N_y\}$ ตามลำดับ ซึ่งในการปรับแนวนิยมกำหนดเงื่อนไขต่างๆ เพื่อให้สอดคล้องกับลักษณะของปัญหา ดังนี้

1. คู่แรกต้องเป็นการจับระหว่างลักษณะสำคัญแรกของทั้งสองเสียงพูด และคู่สุดท้ายต้องเป็นการจับระหว่างลักษณะสำคัญสุดท้ายของทั้งสองเสียงพูด เราอาจเรียกเงื่อนไขนี้ว่าเป็น*เงื่อนไขขอบเขต* ดังนี้

$$c_1=(x_1,y_1) \text{ และ } c_K=(x_{N_x},y_{N_y})$$

2. ไม่มีการจับคู่ย้อนหลัง ซึ่งอาจเรียกเงื่อนไขนี้ว่าเป็น*เงื่อนไขการไปในทิศทางเดียว* ดังนี้

$$\phi_x(k)-\phi_x(k-1)\geq 0 \text{ และ } \phi_y(k)-\phi_y(k-1)\geq 0$$

3. ไม่มีการข้ามลักษณะใด เราอาจเรียกเงื่อนไขนี้ว่าเป็น*เงื่อนไขความต่อเนื่อง* ดังนี้

$$\phi_x(k)-\phi_x(k-1)\leq 1 \text{ และ } \phi_y(k)-\phi_y(k-1)\leq 1$$

4. มีการกำหนดขอบเขตเพื่อไม่ให้เส้นทางการค้นหากว้างจนเกินไป ซึ่งอาจเรียกเงื่อนไขนี้ว่าเป็น*เงื่อนไขหน้าต่างการปรับแก้* ดังนี้

$$\left|\phi_x(k)-\phi_y(k)\right|\leq R$$

เมื่อ R เป็นค่าคงที่ค่าหนึ่ง

การเปรียบเทียบระหว่างการจับคู่ของสองลักษณะใดๆ อาจนิยามได้ว่าเป็นขนาดของการจับคู่นั้น หรือเป็นความผิดเพี้ยนระหว่างสองลักษณะที่จับคู่กัน ดัง

$$\|c_k\| = d(x_{\phi_x(k)}, y_{\phi_y(k)}) \tag{2.9}$$

ซึ่งฟังก์ชัน d ที่ใช้วัดค่าความผิดเพี้ยนนี้ อาจจะเป็นฟังก์ชันระยะทางแบบยุคลิด หรืออาจเป็นฟังก์ชันอื่น ซึ่งอาจจะไม่เป็นไปตามนิยามของฟังก์ชันระยะทางในปริภูมิอิงระยะทางก็ได้

จุดมุ่งหมายของการเปรียบเทียบเสียงพูดทั้งสองคือ การหาการปรับแนวแบบไม่เชิงเส้นที่ให้ขนาดของการจับคู่รวมน้อยที่สุด หรือคือการหา C^* โดย

$$C^* = \arg\min_{C_k} \sum_{k=1}^{K} \|c_k\| \tag{2.10}$$

ในบางกรณี เราอาจกำหนดให้การจับแต่ละคู่มีน้ำหนักไม่เท่ากันได้ ซึ่งจะให้ C^* เป็น

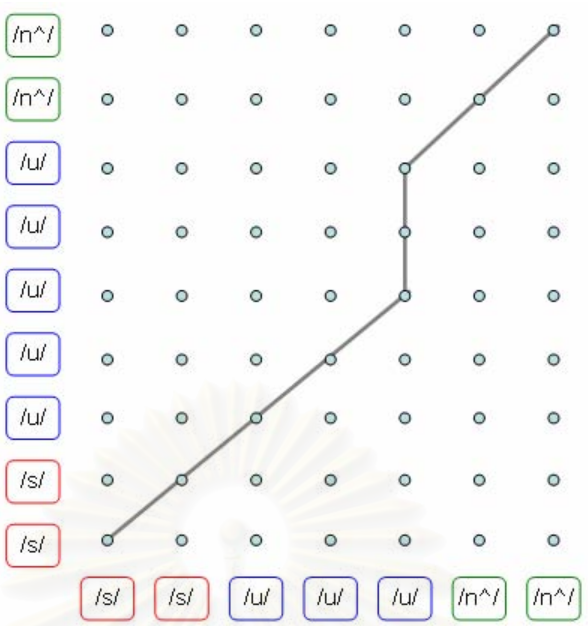
$$C^* = \arg\min_{C} \frac{\sum_{k=1}^{K} w_k \|c_k\|}{\sum_{k=1}^{K} w_k} \tag{2.11}$$

ในการหาการปรับแนวแบบไม่เชิงเส้นที่ให้ขนาดของการจับคู่รวมน้อยที่สุดตาม (2.11) นั้น สามารถทำได้โดยใช้การโปรแกรมแบบพลวัตเข้าช่วย เนื่องจาก

$$\min_{C_n} \sum_{c_k \in C_n} \|c_k\| = \left(\min_{C_{n-1}} \sum_{c_k \in C_{n-1}} \|c_k\| \right) + \|c_n\| \tag{2.12}$$

เมื่อ C_n คือ การปรับแนวที่มีสมาชิกเป็นจำนวน n

จาก (2.12) และเงื่อนไขการจับคู่ ทำให้สามารถใช้การโปรแกรมแบบพลวัตเข้าทำการค้นหาการปรับแนวแบบไม่เชิงเส้นที่ดีที่สุดได้ บางทีก็เรียกกันวิธีนี้ว่าเป็น*การวาร์ปเวลาแบบพลวัต* ดังรูปที่ 2.10 แสดงตัวอย่างการปรับแนวแบบไม่เชิงเส้นที่ดีที่สุดที่เป็นไปได้ของ (2.5) และ (2.6)

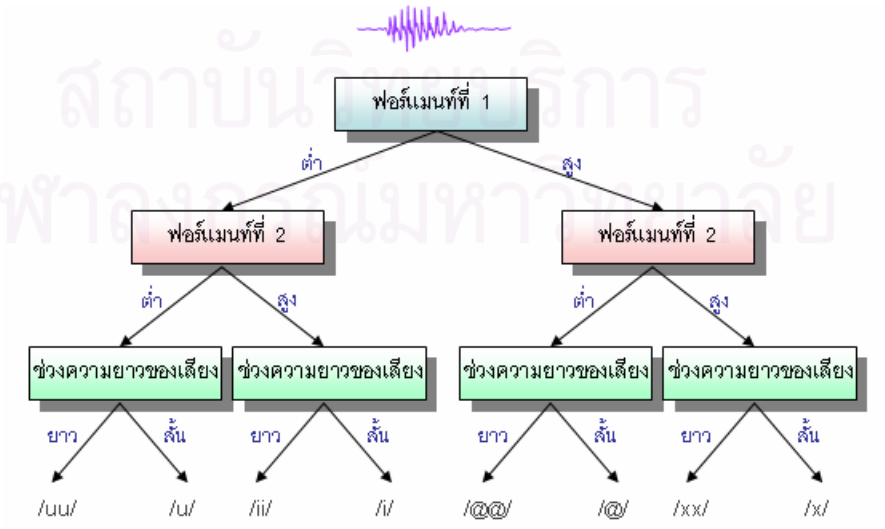


รูปที่ 2.10 ตัวอย่างการปรับแนวแบบไม่เชิงเส้นที่ดีที่สุดที่เป็นไปได้

2.1.3.2 การใช้ความรู้ทางสวนศาสตร์-สัทศาสตร์

การรู้จำเสียงพูดโดยใช้ความรู้ทางสวนศาสตร์-สัทศาสตร์ เริ่มจากการหาลักษณะสำคัญของเสียงพูด และนำลักษณะสำคัญนั้นเข้าสู่ตัวรู้จำ โดยลักษณะสำคัญของเสียงพูดที่หามานั้นเป็นลักษณะเชิงคุณภาพที่บ่งชี้เสียงพูด ซึ่งอาจจะได้แก่ ตำแหน่งของฟอร์แมนต์ ความขึ้นจมูกของเสียง ช่วงความยาวของเสียง การพ่นลม และความก้องของเสียง เป็นต้น (Rabiner and Juang [7])

จากนั้นตัวรู้จำจะนำคุณสมบัติเหล่านี้มาจำแนกเป็นหน่วยย่อยทางภาษาของเสียงพูดต่อไป ตัวรู้จำที่ใช้อาจเป็นกฎในรูปต้นไม้ตัดสินใจ เช่น ในรูปที่ 2.11 แสดงต้นไม้ตัดสินใจที่ใช้จำแนกหน่วยเสียงที่เป็นสระ



รูปที่ 2.11 ต้นไม้ตัดสินใจจำแนกหน่วยเสียงสระภาษาไทย

อย่างไรก็ตาม การใช้ความรู้ทางสวนศาสตร์-สัทศาสตร์ในการรู้จำเสียงพูดยังมีข้อจำกัดตรงที่คุณสมบัติจากการสกัดลักษณะสำคัญและกฎที่ใช้ในการจำแนกเสียงพูดที่เข้ามานั้นตายตัวและต้องกำหนดเองโดยใช้ความรู้ทางสวนศาสตร์-สัทศาสตร์ รวมทั้งการสกัดลักษณะสำคัญของเสียงพูดให้เป็นลักษณะเชิงคุณภาพนั้นค่อนข้างจะลำบาก

2.1.3.3 การสร้างแบบจำลองเฟ้นสุ่ม

แบบจำลองเฟ้นสุ่มเป็นแบบจำลองที่สร้างขึ้นเพื่ออธิบายลำดับของข้อมูลที่สังเกตได้ในช่วงระยะเวลาหนึ่ง ซึ่งข้อมูลนั้นเป็นได้ทั้งแบบต่อเนื่องและไม่ต่อเนื่อง และอาจเกิดขึ้นในช่วงเวลาที่ต่อเนื่องหรือไม่ต่อเนื่องก็ได้ แบบจำลองเฟ้นสุ่มส่วนใหญ่เกิดจากการประกอบกันของสถานะต่างๆ โดยสถานะในแบบจำลองอาจเป็นสถานะแบบต่อเนื่องหรือไม่ต่อเนื่อง และอาจมีสถานะซ่อนตัวหรือไม่ก็ได้ คาลมานฟิลเตอร์เป็นแบบจำลองเฟ้นสุ่มชนิดหนึ่งที่มีสถานะแบบต่อเนื่องและมีสถานะซ่อนตัว ขณะที่แบบจำลองมาร์คอฟซ่อนตัวมีสถานะแบบไม่ต่อเนื่องและมีสถานะซ่อนตัว

แบบจำลองมาร์คอฟซ่อนตัวถูกนำมาใช้อย่างแพร่หลายในการรู้จำเสียงพูด (Baker [8]) (Jelinek et al. [9]) (Rabiner [10]) ขณะที่คาลมานฟิลเตอร์ก็ถูกนำมาใช้เช่นกัน ในที่นี้จะขอกล่าวถึงเพียงแบบจำลองมาร์คอฟซ่อนตัวในหัวข้อถัดไป

แบบจำลองมาร์คอฟซ่อนตัว

แบบจำลองมาร์คอฟซ่อนตัวเป็นแบบจำลองที่ใช้อธิบายลำดับของข้อมูลที่เกิดขึ้นในช่วงเวลาที่ไม่ต่อเนื่อง โดยลำดับของข้อมูลนั้นจะถูกมองว่าเป็นผลลัพธ์ของกระบวนการเชิงสุ่มในแบบจำลองมาร์คอฟซ่อนตัว แบบจำลองมาร์คอฟซ่อนตัวประกอบด้วยสถานะต่างๆ จำนวนจำกัด โดยที่ ณ เวลาหนึ่ง แบบจำลองมาร์คอฟซ่อนตัวจะอยู่ที่สถานะหนึ่ง และมีโอกาสที่จะให้ผลลัพธ์ใดผลลัพธ์หนึ่ง และ ณ เวลาถัดไป แบบจำลองมาร์คอฟซ่อนตัวก็มีโอกาสที่จะย้ายสถานะไปยังสถานะอื่น และสถานะนั้นก็มีโอกาสที่จะให้ผลลัพธ์ใดผลลัพธ์หนึ่งออกมาต่อไป

เพื่อความง่าย แบบจำลองมาร์คอฟซ่อนตัวจะมีข้อกำหนดดังต่อไปนี้

1. กระบวนการเชิงสุ่มที่แบบจำลองมาร์คอฟซ่อนตัวจำลองขึ้นเป็นกระบวนการคงที่ นั่นคือโอกาสในการเปลี่ยนจากสถานะหนึ่งไปสู่อีกสถานะหนึ่งมีค่าคงที่เสมอไม่ขึ้นกับเวลา
2. โอกาสที่ ณ เวลาหนึ่ง แบบจำลองมาร์คอฟซ่อนตัวจะอยู่ที่สถานะหนึ่ง ขึ้นอยู่กับสถานะก่อนหน้านั้นสถานะเดียว หรือเรียกว่าเป็นสมมติฐานมาร์คอฟอันดับหนึ่ง

3. ผลลัพธ์จากแบบจำลองมาร์คอฟซ่อนตัว ณ เวลาหนึ่ง ขึ้นอยู่กับสถานะของแบบจำลอง ณ เวลานั้นเพียงอย่างเดียว

กล่าวโดยสรุป แบบจำลองมาร์คอฟซ่อนตัวมีส่วนประกอบต่างๆ ดังต่อไปนี้

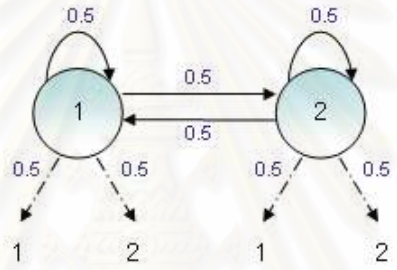
1. *สถานะ* โดยสถานะในแบบจำลองมาร์คอฟซ่อนตัวมีจำนวนจำกัด ในที่นี้จะใช้ตัวเลขจำนวนนับ $\{1,2,...,N\}$ แทนชื่อของสถานะต่างๆ ซึ่ง ณ เวลาหนึ่ง ผลลัพธ์ของแบบจำลองมาร์คอฟซ่อนตัวจะเกิดจากสถานะใดสถานะหนึ่ง โดยจะใช้สัญลักษณ์แทนสถานะที่เวลา t ว่า q_t
2. *ความน่าจะเป็นในการเปลี่ยนสถานะ* แต่ละสถานะในแบบจำลองมาร์คอฟซ่อนตัวจะมีความน่าจะเป็นในการเปลี่ยนสถานะ โดยความน่าจะเป็นที่ ณ เวลา t แบบจำลองมาร์คอฟซ่อนตัวอยู่ที่สถานะ i แล้ว ณ เวลา $t+1$ แบบจำลองมาร์คอฟซ่อนตัวไปอยู่ที่สถานะ j จะใช้สัญลักษณ์ a_{ij} โดย $a_{ij} = P(q_{t+1} = j \mid q_t = i)$ โดยความน่าจะเป็นในการเปลี่ยนสถานะทั้งหมดในแบบจำลองมาร์คอฟซ่อนตัวอาจแทนได้ด้วยเมตริกซ์ A ที่ $A_{ij} = a_{ij}$
3. *ความน่าจะเป็นในการออกผลลัพธ์* ผลลัพธ์ที่เกิดขึ้น ณ เวลาหนึ่ง เกิดจากสถานะของแบบจำลองมาร์คอฟซ่อนตัว ณ เวลานั้น โดยความน่าจะเป็นที่ ณ เวลา t สถานะ j ให้ผลลัพธ์ k จะใช้สัญลักษณ์ $b_j(k)$ โดย $b_j(k) = P(o_t = k \mid q_t = j)$ ในกรณีที่ผลลัพธ์ k เป็นผลลัพธ์แบบไม่ต่อเนื่อง เราอาจใช้ตัวเลขจำนวนนับ $\{1,2,...,M\}$ แทนผลลัพธ์ต่างๆ และอาจแทนความน่าจะเป็นในการออกผลลัพธ์ทั้งหมดในแบบจำลองมาร์คอฟซ่อนตัวด้วยเมตริกซ์ B ที่ $B_{jk} = b_j(k)$ ในกรณีที่ผลลัพธ์เป็นแบบต่อเนื่อง ความน่าจะเป็นในการออกผลลัพธ์อาจแทนด้วยฟังก์ชันความหนาแน่นของความน่าจะเป็น หรือฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบผสม
4. *ความน่าจะเป็นของสถานะเริ่มต้น* โดยจะใช้สัญลักษณ์ π_i แทนความน่าจะเป็นที่สถานะเริ่มต้นของแบบจำลองมาร์คอฟซ่อนตัวคือสถานะ i หรือ $\pi_i = P(q_1 = i)$ โดยความน่าจะเป็นของ สถานะเริ่มต้นทั้งหมดในแบบจำลองมาร์คอฟซ่อนตัวด้วยเวกเตอร์ π

จะเห็นได้ว่าในแบบจำลองมาร์คอฟซ่อนตัว หลายสถานะมีโอกาสจะให้ผลลัพธ์ o เดียวกัน ผลลัพธ์ที่ได้มาจึงยังไม่รู้แน่ชัดว่ามาจากสถานะใด ราวกับว่าสถานะนั้นซ่อนตัวอยู่ และไม่สามารถบอกได้จากผลลัพธ์ที่ออกมา จึงเป็นที่มาของชื่อแบบจำลองมาร์คอฟ*ซ่อนตัว* โดยแบบจำลองมาร์คอฟซ่อนตัวที่สร้างขึ้นอาจเขียนให้กระชับได้ว่า $\lambda = (A, B, \pi)$

ตัวอย่างเช่น ในแบบจำลองมาร์คอฟซ่อนตัวหนึ่งอาจมีส่วนประกอบต่างๆ เป็นได้ดังนี้

1. มีจำนวนสถานะเท่ากับ 2 คือ $\{1,2\}$
2. ผลลัพธ์ของแบบจำลองมีลักษณะไม่ต่อเนื่อง เป็นได้ 2 ค่า คือ $\{H,T\}$
3. ความน่าจะเป็นในการเปลี่ยนสถานะเป็นเมตริกซ์ $A=\begin{bmatrix}0.5 & 0.5\\ 0.5 & 0.5\end{bmatrix}$
4. ความน่าจะเป็นในการออกผลลัพธ์เป็นเมตริกซ์ $B=\begin{bmatrix}0.5 & 0.5\\ 0.5 & 0.5\end{bmatrix}$
5. ความน่าจะเป็นของสถานะเริ่มต้นเป็นแวกเตอร์ $\pi=\begin{bmatrix}0.5\\ 0.5\end{bmatrix}$

เราอาจแสดงแบบจำลองมาร์คอฟซ่อนตัวนี้ได้ด้วยแผนภาพ ดังรูปที่ 2.12



รูปที่ 2.12 ตัวอย่างแบบจำลองมาร์คอฟซ่อนตัว

การอนุมานในแบบจำลองมาร์คอฟซ่อนตัว โดยหลักมีอยู่ 3 แบบ คือ

1. การหาความน่าจะเป็นที่จะเกิดผลลัพธ์ $O=o_1o_2...o_T$ จากแบบจำลองมาร์คอฟซ่อนตัว λ หรือการหาค่า $P(O\mid\lambda)$
2. การหาลำดับของสถานะ $Q=q_1q_2...q_T$ ที่น่าจะเป็นที่สุดเมื่อกำหนดลำดับของผลลัพธ์ $O=o_1o_2...o_T$ และแบบจำลองมาร์คอฟซ่อนตัว λ มาให้ หรือการหาค่า $\arg\max_{Q} P(Q\mid O,\lambda)$
3. การหาแบบจำลองมาร์คอฟซ่อนตัวที่น่าจะเป็นที่สุด เมื่อกำหนดลำดับของผลลัพธ์มาให้ หรือการหาค่า $\arg\max_{\lambda} P(\lambda\mid O)$

การอนุมานแต่ละแบบมีวิธีดังนี้

<u>การหาความน่าจะเป็นของผลลัพธ์</u>

ความน่าจะเป็นของผลลัพธ์ $P(O \mid \lambda)$ คือผลรวมของความน่าจะเป็นของผลลัพธ์ที่เกิดจากลำดับของสถานะต่างๆ ที่เป็นไปได้ทั้งหมด หรือ $P(O \mid \lambda) = \sum_{Q} P(O,Q \mid \lambda)$ การคำนวณต่อจากนี้ทำได้โดยใช้กฎของความน่าจะเป็น ทำให้สามารถสรุปเป็นสูตรได้ว่า

$$P(O \mid \lambda) = \sum_{q_1,q_2,\ldots,q_T} b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \qquad (2.13)$$

แต่การหาค่าความน่าจะเป็นของผลลัพธ์โดยตรงเช่นนี้ต้องอาศัยการคำนวณมากครั้ง เนื่องจากลำดับของสถานะที่เป็นไปได้ทั้งหมดมีถึง N^T ลำดับ การคำนวณ $P(O \mid \lambda)$ ที่มีประสิทธิภาพขึ้นสามารถทำได้โดยอาศัยการโปรแกรมแบบพลวัตเข้าช่วย ซึ่งจะคำนวณความน่าจะเป็นของผลลัพธ์ในแต่ละสถานะทีละช่วงเวลาไปเรื่อยๆ โดยอาจคำนวณไปข้างหน้า จากเวลาเริ่มต้นไปจนเวลาสุดท้าย หรือคำนวณมาข้างหลัง จากเวลาสุดท้ายมาเวลาเริ่มต้นก็ได้

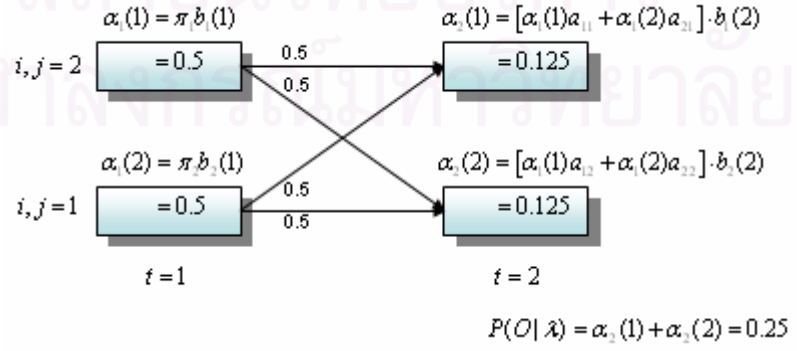
การหาความน่าจะเป็นของผลลัพธ์ไปข้างหน้าอาจเรียกว่ากระบวนการไปข้างหน้า ซึ่งจะคำนวณค่าตัวแปรไปข้างหน้า $\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = i \mid \lambda)$ ในทุกๆ สถานะ i จากเวลา $t = 1$ ถึง $t = T$ ซึ่งเมื่อใช้กฎของความน่าจะเป็น และข้อกำหนดของแบบจำลองมาร์คอฟซ่อนตัวแล้ว สามารถอนุมานได้ว่า

ค่าตัวแปรไปข้างหน้าในเวลาเริ่มต้น $\alpha_1(i) = \pi_i b_i(o_1)$

ค่าตัวแปรไปข้างหน้าในเวลาถัดไป $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1})$

และเมื่อถึงเวลาสุดท้าย จะได้ $P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$

ตัวอย่างการใช้กระบวนการไปข้างหน้าหาความน่าจะเป็นของผลลัพธ์จากแบบจำลองมาร์คอฟซ่อนตัวรูปที่ 2.12 เมื่อ $o_1 = 1$ และ $o_2 = 2$ สามารถแสดงได้ดังรูปที่ 2.13



รูปที่ 2.13 การใช้กระบวนการไปข้างหน้าหาความน่าจะเป็นของผลลัพธ์

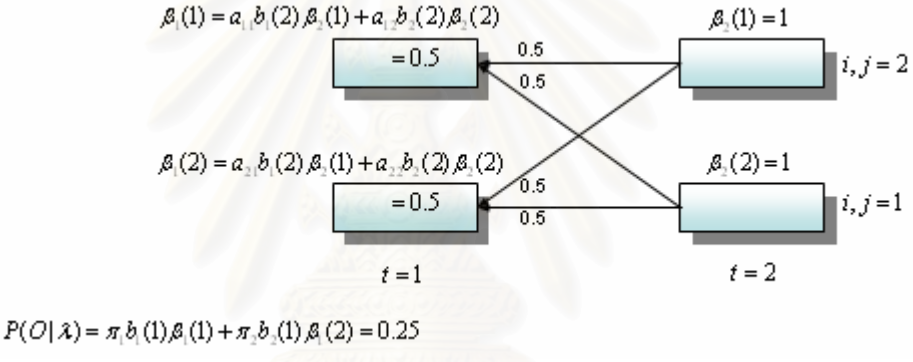
การหาความน่าจะเป็นของผลลัพธ์มาข้างหลังอาจเรียกว่ากระบวนการมาข้างหลัง ซึ่งจะคำนวณค่าตัวแปรมาข้างหลัง $\beta_t(i) = P(o_{t+1}o_{t+2}...o_T \mid q_t = i, \lambda)$ ในทุกๆ สถานะ i จากเวลา $t = T$ ถึง $t = 1$ ซึ่งเมื่อใช้กฎของความน่าจะเป็น และข้อกำหนดของแบบจำลองมาร์คอฟซ่อนตัวแล้วสามารถอนุมานได้ว่า

ค่าตัวแปรมาข้างหลังในเวลาเริ่มต้น $\beta_T(i) = 1$

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

และเมื่อถึงเวลาสุดท้าย ความน่าจะเป็นของผลลัพธ์ $P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$

ตัวอย่างการใช้กระบวนการมาข้างหลังหาความน่าจะเป็นของผลลัพธ์จากแบบจำลองมาร์คอฟซ่อนตัวรูปที่ 2.12 เมื่อ $o_1 = 1$ และ $o_2 = 2$ สามารถแสดงได้ดังรูปที่ 2.14



รูปที่ 2.14 การใช้กระบวนการมาข้างหลังหาความน่าจะเป็นของผลลัพธ์

<u>การหาลำดับของสถานะที่ดีที่สุด</u>

การหาลำดับของสถานะที่ดีที่สุด $\arg\max_Q P(Q \mid O, \lambda)$ จะเท่ากับการหาลำดับของสถานะที่ดีที่สุด $\arg\max_Q \frac{P(O, Q \mid \lambda)}{P(O \mid \lambda)}$ หรือ $\arg\max_Q P(O, Q \mid \lambda)$ นั่นเอง ซึ่งสามารถทำได้โดยคำนวณทุกลำดับของสถานะที่เป็นไปได้ และเลือกลำดับของสถานะที่ดีที่สุดเป็นคำตอบ แต่วิธีนี้ก็สิ้นเปลืองเวลาที่ใช้ในการคำนวณมาก วิธีการที่มีประสิทธิภาพกว่าเรียกว่าอัลกอริทึมวิเทอบี ซึ่งจะคำนวณหาลำดับของสถานะที่ดีที่สุดที่จะมาถึงสถานะปัจจุบันทีละช่วงเวลาไปเรื่อยๆ คล้ายๆ กับกระบวนการไปข้างหน้า เพียงแต่เปลี่ยนจากผลรวมของทุกสถานะที่แล้วเป็นค่ามากสุดจากสถานะที่แล้วแทน โดยกำหนดตัวแปรวิเทอบี $\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P(q_1 q_2 ... q_{t-1}, q_t = i, o_1 o_2 ... o_t \mid \lambda)$ สำหรับเก็บค่าความน่าจะเป็นของลำดับของสถานะที่ดีที่สุด และสามารถอนุมานในแบบจำลองมาร์คอฟซ่อนตัวได้ดังนี้

ค่าตัวแปรวิเทอบีในเวลาเริ่มต้น $\delta_1(i) = \pi_i b_i(o_1)$

ค่าตัวแปรวิเทอบีในเวลาถัดไป $\delta_t(i) = \max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]\cdot b_j(o_t)$

และเมื่อถึงเวลาสุดท้าย จะได้ $\max_{Q} P(Q \mid O,\lambda) = \max_{1\le i\le N}\left[\delta_T(i)\right]$

<u>การหาแบบจำลองมาร์คอฟซ่อนตัวที่น่าจะเป็นที่สุด</u>

การหาแบบจำลองมาร์คอฟซ่อนตัว $\lambda = (A,B,\pi)$ ที่น่าจะเป็นที่สุด เมื่อกำหนดลำดับของผลลัพธ์ O มาให้ นับเป็นปัญหาที่ยาก เนื่องจากยังไม่มีวิธีการตายตัวที่รับรองว่าจะให้คำตอบที่ดีที่สุด แต่ก็อาจทำได้โดยการกำหนดค่า (A,B,π) ขึ้นมาก่อนค่าหนึ่ง แล้ววนปรับค่านี้ไปเรื่อยๆ จนกว่าจะให้คำตอบที่ดีที่สุด ซึ่งถือเป็นการประมาณค่าควรจะเป็น $P(O \mid \lambda)$ โดยวิธีคาดหวัง-สูงสุดชนิดหนึ่ง อัลกอริทึมหนึ่งที่ใช้แก้ปัญหานี้มีชื่อเรียกว่าอัลกอริทึมไปข้างหน้า-มาข้างหลัง หรืออัลกอริทึมบอม-เวลช์ ซึ่งมีวิธีการหาค่าพารามิเตอร์ต่างๆ ของแบบจำลองมาร์คอฟซ่อนตัวดังนี้

ค่าความน่าจะเป็นของสถานะเริ่มต้นค่าใหม่ $\bar{\pi}_i$ จะเท่ากับจำนวนครั้งเฉลี่ยที่แบบจำลองมาร์คอฟซ่อนตัวเดิมจะผ่านสถานะ i ที่เวลา $t=1$

ค่าความน่าจะเป็นในการเปลี่ยนสถานะค่าใหม่ $\bar{a}_{ij}$ จะเท่ากับจำนวนครั้งเฉลี่ยที่แบบจำลองมาร์คอฟซ่อนตัวเดิม เมื่อผ่านสถานะ i แล้วจะผ่านสถานะ j หรือเท่ากับจำนวนครั้งเฉลี่ยที่แบบจำลองมาร์คอฟซ่อนตัวเดิม เมื่อผ่านสถานะ i ในเวลาหนึ่ง และผ่านสถานะ j ในเวลาถัดไป หารด้วย เท่ากับจำนวนครั้งเฉลี่ยที่แบบจำลองมาร์คอฟซ่อนตัวเดิม เมื่อผ่านสถานะ i ทั้งหมด

ค่าความน่าจะเป็นในการออกผลลัพธ์ค่าใหม่ $\bar{b}_j(k)$ จะเท่ากับจำนวนครั้งเฉลี่ยที่แบบจำลองมาร์คอฟซ่อนตัวเดิม เมื่อผ่านสถานะ i แล้วจะให้ผลลัพธ์ k หรือเท่ากับจำนวนครั้งเฉลี่ยที่ลำดับของแบบจำลองมาร์คอฟซ่อนตัวเดิม เมื่อผ่านสถานะ i และให้ผลลัพธ์ k หารด้วยจำนวนครั้งเฉลี่ยที่แบบจำลองมาร์คอฟซ่อนตัวเดิม เมื่อผ่านสถานะ i ทั้งหมด

โดยกำหนดให้ $\xi_t(i,j)$ แทนความน่าจะเป็นที่แบบจำลองมาร์คอฟซ่อนตัวจะอยู่ที่สถานะ i ณ เวลา t และอยู่ที่สถานะ j ณ เวลา $t+1$ เมื่อกำหนดลำดับของผลลัพธ์ O มาให้ หรือ $\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid O,\lambda)$ และเราสามารถคำนวณค่า $\xi_t(i,j)$ โดยใช้กฎของความน่าจะเป็น และข้อกำหนดของแบบจำลองมาร์คอฟซ่อนตัว ได้ว่า

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, O \mid \lambda)}{P(O \mid \lambda)}$$
$$= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O \mid \lambda)}$$

จะได้สูตรในการหาค่าพารามิเตอร์ค่าใหม่ของแบบจำลองมาร์คอฟซ่อนตัว ดังนี้

$$\bar{\pi}_i = \sum_{j=1}^{N} \xi_1(i,j)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \xi_t(i,j)}$$

$$\bar{b}_j(k) = \frac{\sum_{t:o_t=k}^{} \sum_{j=1}^{N} \xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \xi_t(i,j)}$$

<u>แบบจำลองมาร์คอฟซ่อนตัวสำหรับข้อมูลที่มีค่าต่อเนื่อง</u>

เมื่อผลลัพธ์จากแบบจำลองมาร์คอฟซ่อนตัวมีค่าต่อเนื่อง ความน่าจะเป็นในการออกผลลัพธ์จำเป็นต้องรองรับข้อมูลเหล่านี้ด้วย ซึ่งโดยทั่วไป จะใช้การกระจายแบบเกาส์แทนความน่าจะเป็นในการออกผลลัพธ์ ซึ่งสามารถนิยามได้ดังนี้

$$N(o;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp(-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)) \qquad (2.14)$$

เมื่อ n เป็นจำนวนมิติของผลลัพธ์ o ซึ่งการกระจายแบบเกาส์นี้มีค่าเฉลี่ยและเมตริกซ์ความแปรปรวนเป็น μ และ Σ ตามลำดับ

สำหรับการประมาณค่าพารามิเตอร์เหล่านี้ จะคล้ายๆ การหาค่าเฉลี่ยและค่าความแปรปรวนทั่วไป โดยถ่วงด้วยความน่าจะเป็นที่แบบจำลองมาร์คอฟซ่อนตัวจะตกอยู่ในสถานะต่างๆ ดังนี้

$$\bar{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)}$$

$$\bar{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t-\mu_j)(o_t-\mu_j)^T}{\sum_{t=1}^{T} L_j(t)}$$

เมื่อ $L_j(t)$ คือความน่าจะเป็นที่แบบจำลองมาร์คอฟซ่อนตัวจะอยู่ที่สถานะ j ณ เวลา t

2.1.3.4 การใช้การเรียนรู้แบบแบ่งแยก

ในการรู้จำรูปแบบโดยทั่วไป สำหรับข้อมูลเข้า x ที่ต้องการจำแนกออกเป็นคลาสต่างๆ คือ $C_1, C_2, ..., C_K$ เราอาจหาฟังก์ชันการแบ่งแยกของแต่ละคลาส คือ $y_1(x), y_2(x), ..., y_K(x)$ โดยที่ข้อมูลเข้า x จะถูกจำแนกเป็นคลาส C_k ก็ต่อเมื่อ $y_k(x) > y_j(x)$ สำหรับทุกๆ $j \neq k$

$y_k(x)$ อาจสามารถกำหนดให้เท่ากับ $P(C_k \mid x)$ ซึ่งสามารถพิสูจน์ได้ว่าฟังก์ชันการแบ่งแยกนี้ให้ค่าความน่าจะเป็นที่จะเกิดความผิดพลาดในการจำแนกน้อยสุด (Duda et al. [11]) แต่โดยทั่วไปการหา $P(C_k \mid x)$ ทำได้ยาก อาจจะต้องแยก $P(C_k \mid x)$ เป็นผลคูณของ $P(x \mid C_k)$ และ $P(C_k)$ แล้วใช้วิธีการทางสถิติในการประมาณการกระจายตัวของความน่าจะเป็นเหล่านี้

นิวรอลเน็ตเวิร์ก (Bishop [12]) เป็นวิธีการหนึ่งที่สามารถหาฟังก์ชันการแบ่งแยก $y_k(x)$ ได้โดยตรง ซึ่งจะกล่าวถึงในหัวข้อถัดไป

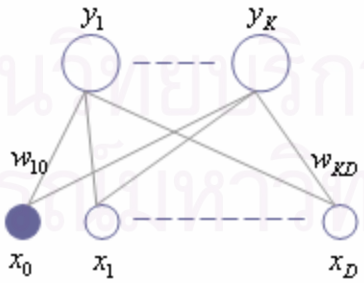
นิวรอลเน็ตเวิร์ก

<u>นิวรอลเน็ตเวิร์กชั้นเดียว</u>

นิวรอลเน็ตเวิร์กเป็นแบบจำลองการเรียนรู้ที่ประกอบด้วยพารามิเตอร์คือค่าน้ำหนักต่างๆ ซึ่งจะถูกนำไปใช้เพื่อสร้างเป็นฟังก์ชันการแบ่งแยก โดยฟังก์ชันการแบ่งแยก $y_k(x)$ ที่ง่ายที่สุดที่นิวรอลเน็ตเวิร์กสร้างได้ จะเป็นผลรวมเชิงเส้นระหว่างเวกเตอร์ค่าน้ำหนัก w_k และข้อมูลเข้า x ของนิวรอลเน็ตเวิร์ก หรือเขียนได้ว่า

$$y_k(x) = w_k^T x + w_{k0} \tag{2.15}$$

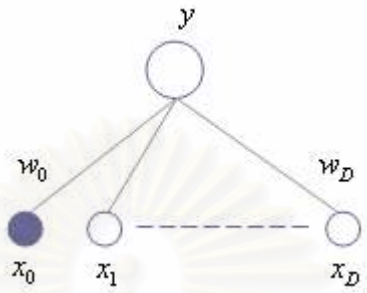
ซึ่งสมการ (2.15) สามารถเขียนได้เป็นแผนภาพดังรูปที่ 2.15



รูปที่ 2.15 นิวรอลเน็ตเวิร์กชั้นเดียว

รูปที่ 2.15 อาจเรียกว่าเป็นนิวรอลเน็ตเวิร์กชั้นเดียว เนื่องจากน้ำหนักทั้งหมดต่อเชื่อมอยู่กับเวกเตอร์ของข้อมูลเข้าเท่านั้น

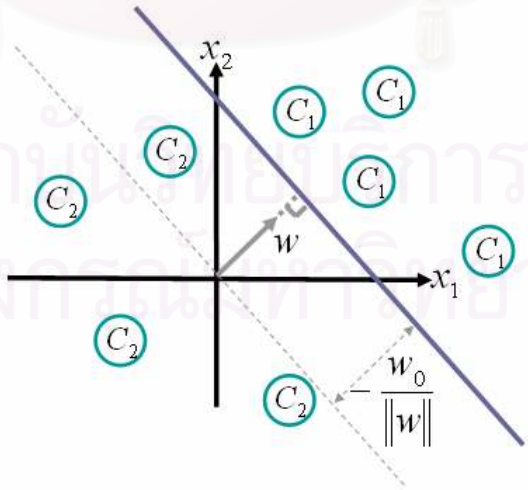
ในกรณีที่มีเพียงสองคลาส เราอาจลดรูปให้เหลือเพียง $y(x) = w^T x + w_0$ โดยกำหนดว่า ถ้า $y(x)$ มีค่ามากกว่า 0 แล้ว ข้อมูลเข้า x จะถูกจำแนกเป็นคลาส C_1 ขณะที่ถ้า $y(x)$ มีค่าน้อยกว่า 0 ข้อมูลเข้า x จะถูกจำแนกเป็นคลาส C_2 ในกรณีนี้ นิวรอลเน็ตเวิร์กอาจเขียนได้เป็นแผนภาพดังรูปที่ 2.16



รูปที่ 2.16 นิวรอลเน็ตเวิร์กชั้นเดียวในการจำแนกสองคลาส

นิวรอลเน็ตเวิร์กรูปที่ 2.16 จะให้ขอบเขตการตัดสินใจ หรือระนาบการแบ่งแยกข้อมูล ที่ $y(x) = 0$ โดยระนาบนี้จะจำแนกข้อมูลออกเป็นสองส่วน คือ ข้อมูลที่มีคลาสเป็น C_1 กับข้อมูลที่มีคลาสเป็น C_2 ขอบเขตการตัดสินใจนี้จะมีลักษณะเป็นเชิงเส้น โดยขอบเขตการตัดสินใจจะตั้งฉากกับเวกเตอร์ค่าน้ำหนัก และระยะห่างที่สั้นที่สุดของขอบเขตการตัดสินใจจากจุดกำเนิดจะเท่ากับ $-\frac{w_0}{\|w\|}$

สำหรับข้อมูลเข้าที่มีสองมิติ อาจแสดงเวกเตอร์ค่าน้ำหนักและขอบเขตการตัดสินใจได้ดังรูปที่ 2.17



รูปที่ 2.17 เวกเตอร์ค่าน้ำหนักและขอบเขตการตัดสินใจของนิวรอลเน็ตเวิร์กชั้นเดียว

ฟังก์ชันการแบ่งแยกอาจอาจขยายจากผลรวมเชิงเส้นให้เป็นฟังก์ชันไม่เชิงเส้นได้ โดยนำผลรวมเชิงเส้นนั้นผ่านฟังก์ชันไม่เชิงเส้น $g(\cdot)$ ฟังก์ชันหนึ่ง เช่น ในกรณีของนิวรอลเน็ตเวิร์กชั้นเดียวในการจำแนกสองคลาส อาจเขียนฟังก์ชันการแบ่งแยกได้เป็น

$$y(x) = g(w^T x + w_0) \tag{2.16}$$

โดย $g(\cdot)$ ในที่นี้ถูกเรียกว่าฟังก์ชันกระตุ้น ซึ่งเป็นได้หลายรูปแบบ เช่น

1. ฟังก์ชันไบโพลาร์ $g(a) = \begin{cases} -1, a < 0 \\ 1, a \geq 0 \end{cases}$
2. เฮวิไซด์สเต็ปฟังก์ชัน $g(a) = \begin{cases} 0, a < 0 \\ 1, a \geq 0 \end{cases}$
3. ฟังก์ชันซิกมอยด์ $g(a) = \dfrac{1}{1 + \exp(-a)}$

หรือฟังก์ชันอื่นๆ นอกเหนือจากนี้ เป็นต้น

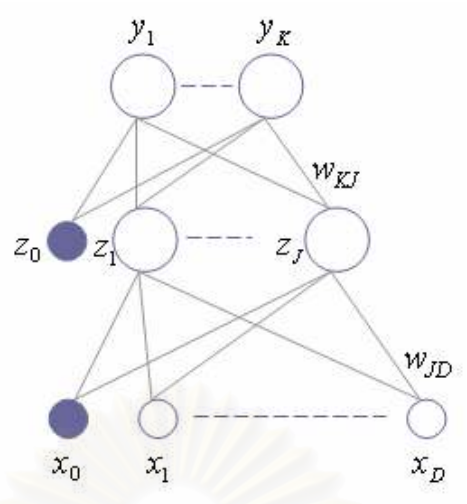
นิวรอลเน็ตเวิร์กชั้นเดียวในการจำแนกสองคลาสที่ใช้ฟังก์ชันไบโพลาร์เป็นฟังก์ชันกระตุ้นเรียกว่า*เพอร์เซพตรอน* เพอร์เซพตรอนมีความสามารถในการจำแนกข้อมูลโดยให้ขอบเขตการตัดสินใจที่เป็นเชิงเส้น ค่าน้ำหนักในเพอร์เซพตรอนสามารถหาได้โดยวนปรับค่าน้ำหนักสำหรับทุกข้อมูล n ที่เพอร์เซพตรอนจำแนกผิดพลาด โดย

$$w_i^{new} = w_i^{old} + \eta t^n x_i^n \tag{2.17}$$

โดย t^n มีค่าเป็น 1 เมื่อ x^n เป็นคลาส C_1 และมีค่าเป็น -1 เมื่อ x^n เป็นคลาส C_2 และ η เป็นค่าอัตราการเรียนรู้ ซึ่งจากทฤษฎีการลู่เข้าของเพอร์เซพตรอน สามารถพิสูจน์ได้ว่า ถ้าข้อมูลเข้าสามารถแบ่งแยกด้วยขอบเขตการตัดสินใจเชิงเส้นได้แล้ว การวนปรับค่าน้ำหนักด้วยวิธีการนี้จะทำให้ได้ค่าน้ำหนักที่สามารถแบ่งแยกข้อมูลได้ในจำนวนครั้งที่จำกัด

<u>นิวรอลเน็ตเวิร์กหลายชั้น</u>

นิวรอลเน็ตเวิร์กชั้นเดียวมีข้อจำกัดเนื่องจากขอบเขตการตัดสินใจที่ได้มีลักษณะเป็นเพียงแบบเชิงเส้น การสร้างขอบเขตการตัดสินใจที่ซับซ้อนขึ้นอาจทำได้โดยใช้นิวรอลเน็ตเวิร์กหลายชั้นดังรูปที่ 2.18



รูปที่ 2.18 นิวรอลเน็ตเวิร์กสองชั้น

ในนิวรอลเน็ตเวิร์กหลายชั้น ผลลัพธ์ที่ได้จากชั้นหนึ่ง จะถูกนำไปเป็นข้อมูลเข้าสำหรับการประมวลผลในชั้นถัดไป โดยชั้นที่อยู่ระหว่างข้อมูลเข้าและผลลัพธ์จะถูกเรียกว่าชั้นซ่อน เช่นในรูปที่ 2.18 เป็นนิวรอลเน็ตเวิร์กสองชั้นที่มีชั้นซ่อนหนึ่งชั้น ซึ่งหาฟังก์ชันการแบ่งแยกได้โดย

ที่ชั้นแรก หรือที่ชั้นซ่อน คำนวณ

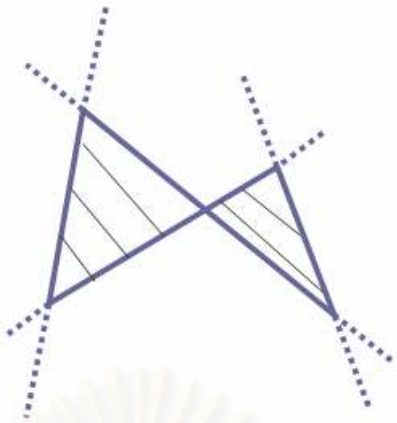
$$z_j = g_j(w_j^T x + w_{j0}) \tag{2.18}$$

ที่ชั้นที่สอง หรือที่ชั้นผลลัพธ์ คำนวณ

$$y_k = g_k(w_k^T z_j + w_{k0}) \tag{2.19}$$

ซึ่งถ้าเป็นนิวรอลเน็ตเวิร์กที่มีมากกว่าสองชั้น ก็สามารถทำการคำนวณไปเรื่อยๆ ได้ทีละชั้นต่อไป

นิวรอลเน็ตเวิร์กสองชั้นสามารถสร้างขอบเขตการตัดสินใจที่ซับซ้อนมากๆ ได้ แต่ก็มีขอบเขตการตัดสินใจบางรูปแบบเหมือนกันที่นิวรอลเน็ตเวิร์กสองชั้นไม่สามารถสร้างได้ เช่นในรูปที่ 2.19 อย่างไรก็ตาม สำหรับข้อมูลการเรียนรู้ที่มีจำนวนจำกัด Nilsson พิสูจน์ได้ว่าสำหรับข้อมูลเข้าจำนวน N ตัวใดๆ สามารถใช้นิวรอลเน็ตเวิร์กสองชั้นจำแนกได้ทั้งนั้นโดยใช้โหนดในชั้นซ่อนจำนวน $N-1$ โหนด [13] และ Baum พิสูจน์ว่าสำหรับข้อมูลเข้าจำนวน N ตัว ที่มี D มิติ สามารถใช้นิวรอลเน็ตเวิร์กสองชั้นจำแนกได้โดยใช้โหนดในชั้นซ่อนจำนวน $\lceil N/D \rceil$ โหนด [14] นอกจากนี้ ถ้าใช้ฟังก์ชันต่อเนื่องเป็นฟังก์ชันกระตุ้นในนิวรอลเน็ตเวิร์ก เช่นฟังก์ชันซิกมอยด์แล้ว Funahashi [15] Cybenko [16] และ Hornik et al. [17] พบว่า นิวรอลเน็ตเวิร์กสองชั้นสามารถประมาณฟังก์ชันต่อเนื่องใดๆ ก็ได้ โดยให้ค่าความคลาดเคลื่อนน้อยเท่าที่ต้องการ



รูปที่ 2.19 ขอบเขตการตัดสินใจที่นิวรอลเน็ตเวิร์กสองชั้นไม่สามารถสร้างได้

<u>การเรียนรู้นิวรอลเน็ตเวิร์กสองชั้น</u>

ในการเรียนรู้ค่าน้ำหนักของนิวรอลเน็ตเวิร์กสองชั้น วิธีที่มีประสิทธิภาพและใช้กันทั่วไปคือวิธีแบ็กพรอพาเกชัน ซึ่งจะวนรอบคำนวณค่าอนุพันธ์ของฟังก์ชันความผิดพลาด และนำค่าอนุพันธ์นี้ไปใช้ในกระบวนการเกรเดียนต์เดสเซนต์ เพื่อปรับค่าน้ำหนักของนิวรอลเน็ตเวิร์กต่อไป [18]

เมื่อตัวอย่างการเรียนรู้อยู่ในรูป (x,t) โดย x เป็นข้อมูลเข้า และ t เป็นเวกเตอร์เป้าหมายที่บอกค่าที่แท้จริงของข้อมูลเข้านี้ เราสามารถกำหนดฟังก์ชันความผิดพลาด E ของตัวอย่างใดๆ ที่เข้ามาทำการเรียนรู้ได้ เช่นกำหนดให้เป็นฟังก์ชันความผิดพลาดผลบวกของกำลังสอง โดยที่ $E=1/2\sum_k (y_k - t_k)^2$ เมื่อ y_k เป็นค่าของโหนดในชั้นผลลัพธ์ตัวที่ k และ t_k คือค่าที่แท้จริงของโหนดในชั้นผลลัพธ์ตัวนี้ ซึ่งก็คือค่าในมิติที่ k ของ t นั่นเอง

ในชั้นผลลัพธ์ ค่าอนุพันธ์ย่อยของ E เทียบกับ w_{kj} หรือ $\partial E/\partial w_{kj}$ สามารถเขียนได้ในรูป $\partial E/\partial a_k \cdot \partial a_k/\partial w_{kj}$ เมื่อ $a_k = \sum_j w_{kj} z_j$ หรือ a_k เป็นผลรวมเชิงเส้นของโหนดชั้นผลลัพธ์ตัวที่ k นั่นเอง ซึ่งเทอม $\partial E/\partial a_k$ นิยมเขียนโดยย่อว่า δ_k

ส่วนที่ชั้นซ่อนนั้น เราไม่สามารถหาอนุพันธ์ย่อยของ E เทียบกับ w_{jd} ได้โดยตรง แต่ก็สามารถใช้กฎลูกโซ่ โดยถือว่าความผิดพลาดจาก w_{jd} มีส่วนต่อ E โดยส่งผ่านไปยังโหนดในชั้นผลลัพธ์ทุกตัว เมื่อมองเช่นนี้ จะได้ว่า δ_j หรือ $\partial E/\partial a_j$ เท่ากับ $\sum_k \partial E/\partial a_k \cdot \partial a_k/\partial a_j$ ซึ่งในที่นี้ $\partial a_k/\partial a_j$ เท่ากับ $w_{kj} g'_j(a_j)$ เพราะฉะนั้น $\partial E/\partial a_j$ สามารถเขียนแจงรูปต่อไปได้ว่า เท่ากับ $g'_j(a_j)\sum_k w_{kj}\delta_k$

ถ้านิวรอลเน็ตเวิร์กมีฟังก์ชันกระตุ้นเป็นฟังก์ชันซิกมอยด์ ซึ่งค่าอนุพันธ์ของฟังก์ชันซิกมอยด์ $g'(a)$ เท่ากับ $g(a)(1-g(a))$ และใช้ฟังก์ชันความผิดพลาดผลบวกของกำลังสอง เมื่อนำฟังก์ชันเหล่านี้แทนเข้าไปในสูตรอนุพันธ์ย่อยข้างต้น จะได้ว่า

$$\delta_k = y_k(1-y_k)(y_k - t_k) \tag{2.20}$$

และ

$$\delta_j = z_j(1-z_j)\sum_k w_{kj}\delta_k \tag{2.21}$$

สุดท้าย ด้วยกระบวนการเกรเดียนต์เดสเซนต์ ซึ่งจะปรับค่าน้ำหนักให้ไปในทิศทางที่ฟังก์ชันความผิดพลาดลดลงไปมากที่สุด จะนำค่าอนุพันธ์ย่อยที่ได้ไปทำการปรับค่าน้ำหนัก โดย w_{kj} และ w_{jd} ใหม่ จะได้จากการนำ $\eta\delta_k z_j$ และ $\eta\delta_j x_d$ ไปลบออกจากค่าเดิมตามลำดับ

โดยสรุปแล้ว การเรียนรู้นิวรอลเน็ตเวิร์กสองชั้นที่ใช้ฟังก์ชันความผิดพลาดผลบวกของกำลังสอง และมีฟังก์ชันซิกมอยด์เป็นฟังก์ชันกระตุ้น สามารถเขียนเป็นขั้นเป็นตอนได้ดังนี้

สำหรับทุกๆ ตัวอย่างการเรียนรู้ (x,t)

1. ป้อน x เป็นข้อมูลเข้าสู่นิวรอลเน็ตเวิร์ก และคำนวณค่าผลลัพธ์ของทุกโหนดในชั้นซ่อน และทุกโหนดในชั้นผลลัพธ์ ด้วยสมการ (2.18) และ (2.19) ตามลำดับ
2. คำนวณค่า δ_k สำหรับทุกโหนดในชั้นผลลัพธ์ จากสมการ (2.20)
3. นำค่า δ_k ที่คำนวณได้ ย้อนกลับมาคำนวณค่า δ_j สำหรับทุกโหนดในชั้นซ่อน ด้วยสมการ (2.21)
4. ปรับค่าน้ำหนักของชั้นผลลัพธ์ โดย

$$w_{kj}^{new} = w_{kj}^{old} - \eta\delta_k z_j \tag{2.22}$$

และปรับค่าน้ำหนักที่เชื่อมไปยังชั้นซ่อน โดย

$$w_{jd}^{new} = w_{jd}^{old} - \eta\delta_j x_d \tag{2.23}$$

เมื่อ η เป็นค่าอัตราการเรียนรู้

2.2 การรู้จำเสียงพูดต่อเนื่องอัตโนมัติ

2.2.1 ลักษณะของปัญหา

ปัญหาการรู้จำเสียงพูดต่อเนื่องอัตโนมัติมีลักษณะดังนี้

1. เป็นการรู้จำลำดับของคำในภาษา โดยคำที่รู้จำได้อาจมีมากน้อยแตกต่างกันไป
2. เป็นการรู้จำหน่วยย่อยทางภาษา ซึ่งมีจำนวนจำกัดจำนวนหนึ่ง ก่อนที่จะประกอบหน่วยย่อยทางภาษาเหล่านั้นออกมาเป็นลำดับของคำในภาษา
3. เสียงพูดมีลักษณะต่อเนื่อง ไม่มีการแบ่งแยกแต่ละหน่วยย่อยทางภาษาออกจากกัน

2.2.2 มุมมองต่อปัญหา

การรู้จำเสียงพูดต่อเนื่องอัตโนมัติ ASR อาจมองว่าเป็นปัญหาการรู้จำรูปแบบชนิดหนึ่ง ได้เช่นเดียวกับการรู้จำเสียงพูดไม่ต่อเนื่อง แตกต่างกันตรงที่ข้อมูลเข้า $x_1, x_2, ..., x_N$ หรือ X บรรจุมากกว่าหนึ่งหน่วยย่อยทางภาษา และหน่วยย่อยทางภาษานั้นสามารถนำมาประกอบเป็นลำดับของคำ W ที่ใช้เพื่อการสื่อสาร หรือ

$$W = ASR(X) \tag{2.24}$$

การรู้จำเสียงพูดต่อเนื่องอัตโนมัติอาจมองในเชิงความน่าจะเป็นได้ว่า คือการหาลำดับของคำที่น่าจะเป็นที่สุดเมื่อกำหนดข้อมูลเสียงพูดมาให้ หรือ

$$\hat{W} = \arg\max_{W} P(W \mid X) = \arg\max_{W} P(X \mid W)P(W) \tag{2.25}$$

ในที่นี้ $P(W)$ คือความน่าจะเป็นที่ลำดับของคำ W จะเกิดขึ้นในภาษา จึงเรียก $P(W)$ ว่าความน่าจะเป็นก่อน หรือแบบจำลองทางภาษา ส่วน $P(X \mid W)$ คือความน่าจะเป็นที่ลำดับของข้อมูลเข้าคือ X เมื่อลำดับของคำที่เป็นที่มาของ X คือ W และเรียก $P(X \mid W)$ ว่าความน่าจะเป็นควรจะเป็น หรือแบบจำลองทางเสียง

ในส่วนของแบบจำลองทางเสียง เราสามารถนำวิธีการสำหรับรู้จำเสียงพูดไม่ต่อเนื่องมาใช้สร้างแบบจำลองได้ ส่วนแบบจำลองทางภาษาในที่นี้จะสนใจเฉพาะแบบจำลองทางภาษาแบบเอ็นแกรม ซึ่งจะกล่าวถึงในหัวข้อถัดไป นอกจากนั้น ในการได้มาซึ่งลำดับของคำที่น่าจะเป็นที่สุดหากทำการค้นหาโดยไล่เรียงลำดับของคำที่เป็นไปได้ทั้งหมดอาจต้องใช้ทรัพยากรจำนวนมหาศาล จึงจำเป็นจะต้องใช้วิธีการค้นหาที่มีประสิทธิภาพ เพื่อให้ระบบรู้จำเสียงพูดสามารถนำไปประยุกต์ใช้ได้จริงตามต้องการ

2.2.3 แบบจำลองทางภาษาแบบเอ็นแกรม

แบบจำลองทางภาษาเป็นแบบจำลองที่สร้างขึ้นเพื่ออธิบายลักษณะทั่วไปของภาษาธรรมชาติในขอบเขตที่กำหนด โดยแบบจำลองทางภาษาอาจเป็นได้ทั้งในเชิงภาษาศาสตร์ ซึ่งต้องใช้ความรู้ทางภาษาเข้าช่วย และในเชิงสถิติ ซึ่งสามารถสร้างได้ด้วยการเก็บสถิติจากฐานข้อมูลทางภาษา โดยทั่วไปแบบจำลองทางภาษาเชิงสถิติจะอยู่ในรูปการแจกแจงความน่าจะเป็นของหน่วยย่อยทางภาษาที่สนใจ เช่น ตัวอักษร หน่วยเสียง หรือคำ โดยแบบจำลองทางภาษาเชิงสถิติส่งผลอย่างมากต่อความถูกต้องของหลายเทคโนโลยีทางภาษา ได้แก่ การรู้จำเสียงพูด การแปลภาษาด้วยเครื่องจักร การจำแนกประเภทเอกสาร การอ่านอักขระด้วยแสง การรู้จำลายมือเขียน การค้นคืนสารสนเทศ และการตรวจสอบตัวสะกด เป็นต้น

แบบจำลองทางภาษาแบบเอ็นแกรม (Jelinek [19]) เป็นแบบจำลองทางภาษาเชิงสถิติอย่างง่าย ซึ่งประมาณความน่าจะเป็นในการเกิดลำดับของคำ $W = w_1 w_2 ... w_n = w_1^n$ ด้วยกฎผลคูณ ดังนี้

$$P(W) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}) = \prod_{i=1}^{n} P(w_i \mid w_1^{i-1}) \quad (2.26)$$

จากสูตรข้างต้นพบว่าในการคำนวณ $P(w_i \mid w_1^{i-1})$ เมื่อ i มีค่ามากเป็นไปไม่ได้ในทางปฏิบัติ จึงทำการลดรูปสมการโดยใช้สมมติฐานมาร์คอฟที่ว่า ความน่าจะเป็นของคำลำดับที่ i ขึ้นอยู่กับคำลำดับก่อนหน้าเพียง $N-1$ ตัวเท่านั้น หรือ

$$P(w_i \mid w_1^{i-1}) \approx P(w_i \mid w_{i-N+1}^{i-1}) \quad (2.27)$$

ซึ่งเรียกวิธีนี้ว่า การประมาณแบบเอ็นแกรม เช่น ในกรณีที่ค่า N เท่ากับ 2 จะได้

$$P(W) \approx P(w_1)\prod_{i=2}^{n} P(w_i \mid w_{i-1}) \quad (2.28)$$

ซึ่งเรียกว่าแบบจำลองไบแกรม

ส่วนในกรณีที่ค่า N เท่ากับ 3 จะได้

$$P(w_1^n) \approx P(w_1)P(w_2 \mid w_1)\prod_{i=3}^{n} P(w_i \mid w_{i-2}^{i-1}) \quad (2.29)$$

ซึ่งเรียกว่าแบบจำลองไตรแกรม เป็นต้น

ซึ่งความน่าจะเป็นแบบมีเงื่อนไข $P(w_i \mid w_{i-N+1}, w_{i-N+2}, ..., w_{i-1})$ สามารถประมาณได้ด้วยความถี่จากการนับคำในฐานข้อมูลทางภาษา หรือ

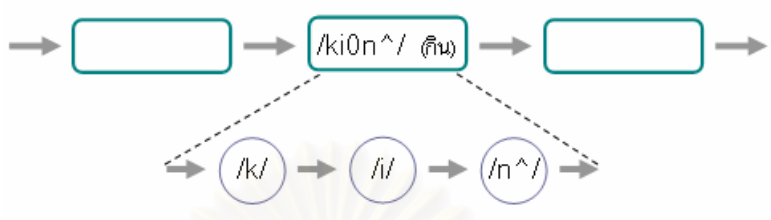
$$P(w_i \mid w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i})}{C(w_{i-N+1}^{i-1})} \quad (2.30)$$

โดยที่ C จะคืนค่าจำนวนของลำดับคำที่ปรากฏอยู่ในฐานข้อมูลทางภาษา

2.2.4 **อัลกอริทึมการค้นหาสำหรับการรู้จำเสียงพูดต่อเนื่อง**

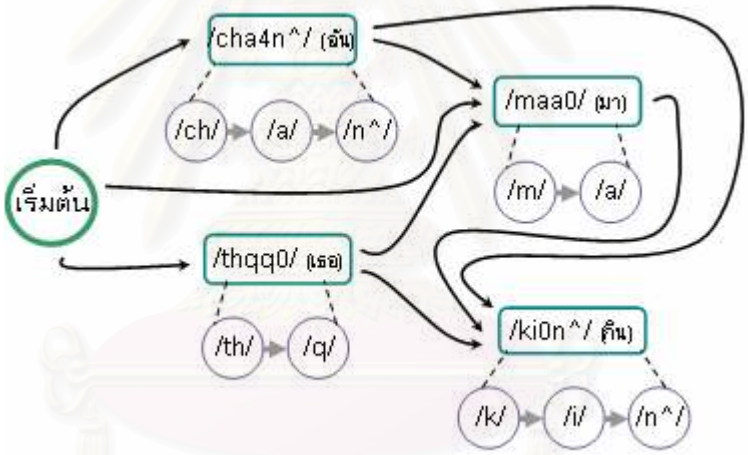
อัลกอริทึมการค้นหา หรืออัลกอริทึมการถอดรหัส คือกระบวนการในการหาลำดับของคำที่ดีที่สุดจากเสียงพูดที่กำหนดให้ โดยใช้แบบจำลองทางเสียงและแบบจำลองทางภาษาเข้าช่วย ซึ่งลำดับของคำมักจะเป็นการประกอบกันในหลายระดับ โดยระดับพื้นฐานอาจจะเป็นระดับของหน่วยเสียง หรือระดับของหน่วยย่อยทางภาษาอื่นที่เป็นผลลัพธ์จากส่วนการรู้จำเสียงพูดไม่ต่อเนื่อง

จากหน่วยย่อยทางภาษาจะรวมตัวขึ้นมาเป็นคำ จากคำก็มีการร้อยเรียงต่อกันเป็นลำดับ โดยอาจจะมีการจำกัดเงื่อนไขในการรวมตัวของหน่วยย่อยทางภาษา เรียกว่าพจนานุกรม และกำหนดลักษณะในการร้อยเรียงของคำ ด้วยแบบจำลองทางภาษา ซึ่งอาจแสดงระดับชั้นของการรู้จำเสียงพูดต่อเนื่องได้ดังรูปที่ 2.20



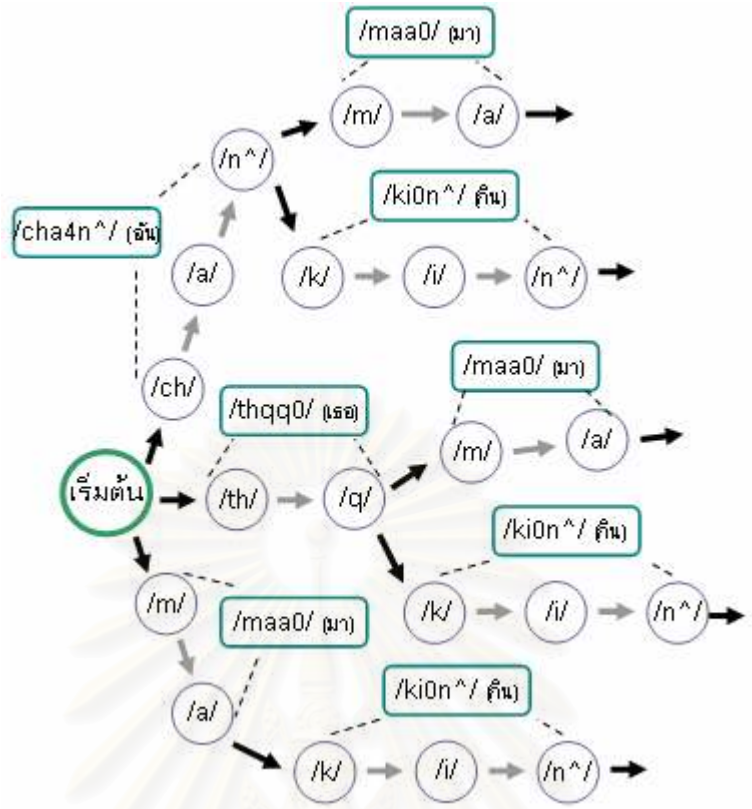
รูปที่ 2.20 ระดับชั้นของการรู้จำเสียงพูดต่อเนื่อง

เมื่อเราทำการเชื่อมลำดับของคำที่เป็นไปได้ทั้งหมดจากโหนดเริ่มต้น จะได้เป็นเน็ตเวิร์กของการรู้จำขึ้น ซึ่งทุกทางในเน็ตเวิร์กมีโอกาสที่จะเป็นผลลัพธ์ของการรู้จำเสียงพูดต่อเนื่องทั้งสิ้น โดยรูปที่ 2.21 แสดงตัวอย่างของเน็ตเวิร์กสำหรับการรู้จำเสียงพูดต่อเนื่อง



รูปที่ 2.21 เน็ตเวิร์กสำหรับการรู้จำเสียงพูดต่อเนื่อง

จากโหนดเริ่มต้น อัลกอริทึมการค้นหาจะทำการหาทางที่น่าจะใช่ที่สุดจากทางทั้งหมดที่เป็นไปได้ ซึ่งมีหลายวิธีด้วยกัน วิธีหนึ่งซึ่งรับประกันว่าจะได้ผลลัพธ์ที่ดีที่สุดนั้นคือการใช้อัลกอริทึมวิเทอบี เช่นเดียวกับการหาลำดับของสถานะที่ดีที่สุดในแบบจำลองมาร์คอฟซ่อนตัว แต่ในการรู้จำเสียงพูดต่อเนื่อง ถ้าคำศัพท์มีเป็นจำนวนมากแล้ว สเปซการค้นหาโดยรวมจะมีขนาดใหญ่ ซึ่งแม้อัลกอริทึมวิเทอบีที่ใช้เวลาการทำงานในระดับโพลิโนเมียลก็มิอาจให้ผลลัพธ์ออกมาโดยรวดเร็วได้ ในรูปที่ 2.22 เป็นเน็ตเวิร์กเดียวกับรูปที่ 2.21 แต่แผ่ออกมาให้เห็นทางทั้งหมดที่เป็นไปได้



รูปที่ 2.22 เน็ตเวิร์กสำหรับการรู้จำเสียงพูดต่อเนื่องเมื่อทำการแผ่ออกมา

อัลกอริทึมการผ่านโทเคน (Young et al. [20]) จะใช้โทเคนแทนทางที่ทำการค้นหาผ่านตั้งแต่เวลาที่ 0 ถึงเวลาที่ t โดย ณ เวลาที่ 0 นั้น โทเคนจะถูกวางไว้ที่โหนดเริ่มต้น ทุกๆ ช่วงเวลาโทเคนจะถูกผ่านไปในแต่ละโหนดของเน็ตเวิร์กตามการเชื่อมต่อต่างๆ ถ้าโหนดหนึ่งสามารถเชื่อมต่อไปได้กับหลายโหนด โทเคนก็จะถูกทำซ้ำไปยังทุกโหนดนั้น เพื่อการสำรวจทุกทางที่เป็นไปได้ และค่าความน่าจะเป็นของโทเคนก็จะถูกปรับปรุงตามค่าความน่าจะเป็นในการเปลี่ยนสถานะของการเชื่อมต่อและค่าความน่าจะเป็นในการออกผลลัพธ์ของโหนด ในแต่ละช่วงเวลา ณ โหนดหนึ่งจะมีได้เพียง N โทเคนเท่านั้นที่รอดชีวิตอยู่ โดยโทเคนที่ให้ค่าความน่าจะเป็นน้อยกว่าโทเคนที่มีค่าความน่าจะเป็นสูงสุด N ตัวแรกในโหนดนั้น จะถูกกำจัดออกไปจากกระบวนการค้นหา ซึ่งในงานทั่วไป ค่า N เป็น 1 ก็ดูว่าจะเพียงพอแล้ว

เมื่อโทเคนท่องไปตามเส้นทางต่างๆ ในเน็ตเวิร์ก โทเคนนั้นจำเป็นต้องบันทึกการเดินทางของตัวเองเอาไว้ด้วย ทั้งนี้ขึ้นอยู่กับว่าเราต้องการผลลัพธ์ในการรู้จำอยู่ในระดับใด ซึ่งโดยทั่วไป จะเป็นในระดับคำ แต่ด้วยจุดประสงค์อื่น อาจให้ผลลัพธ์การรู้จำอยู่ในระดับที่ต่ำลงไปก็ได้ เช่นในระดับหน่วยเสียง เป็นต้น

นอกจากนี้ ยังอาจสามารถใช้การค้นหาแบบบีมเข้าช่วย โดยดูโทเคนในทั้งเน็ตเวิร์ก ถ้าโทเคนไหนให้ค่าความน่าจะเป็นที่ต่ำก็จะถูกตัดออกไป ซึ่งวิธีนี้อาจก่อให้เกิดความผิดพลาดในการค้นหา แต่ก็ทำให้ได้ผลลัพธ์ที่รวดเร็วกว่า

2.3 การรู้จำเสียงพูดโดยใช้นิวรอลเน็ตเวิร์ก

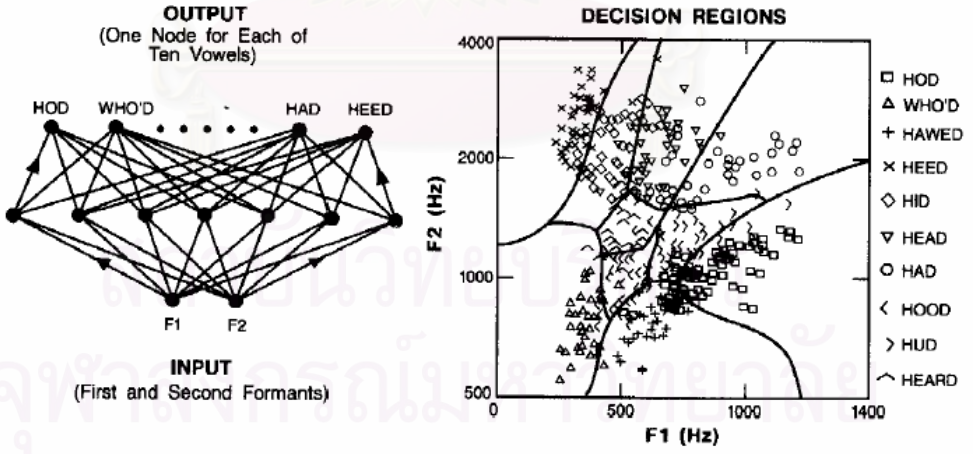
การนำนิวรอลเน็ตเวิร์กมาประยุกต์ใช้ในงานรู้จำเสียงพูดในช่วงแรกเป็นการนำนิวรอลเน็ตเวิร์กมาทำการรู้จำเสียงพูดไม่ต่อเนื่อง ขณะที่ต่อมาได้พัฒนาให้นิวรอลเน็ตเวิร์กรู้จำเสียงพูดต่อเนื่องอัตโนมัติ ซึ่งส่วนใหญ่เป็นระบบผสมผสานร่วมกับแบบจำลองมาร์คอฟซ่อนตัว โดยมีรายละเอียดดังนี้

2.3.1 การใช้นิวรอลเน็ตเวิร์กรู้จำเสียงพูดไม่ต่อเนื่อง

2.3.1.1 นิวรอลเน็ตเวิร์กชั้นเดียวและนิวรอลเน็ตเวิร์กสองชั้น

งานวิจัยที่ใช้นิวรอลเน็ตเวิร์กสองชั้นมารู้จำเสียงพูดไม่ต่อเนื่องมีดังนี้

Huang และ Lippmann แสดงให้เห็นว่านิวรอลเน็ตเวิร์กสองชั้นสามารถให้ขอบเขตการตัดสินใจที่ซับซ้อน สามารถจำแนกหน่วยเสียงสระ 10 หน่วยเสียงได้ [21] ดังรูปที่ 2.23 โดยนิวรอลเน็ตเวิร์กรับข้อมูลเข้าเป็นฟอร์แมนต์ที่หนึ่งและฟอร์แมนต์ที่สองของหน่วยเสียงสระ มีจำนวนโหนดในชั้นซ่อนเท่ากับ 50 และใช้จำนวนรอบในการเรียนรู้ทั้งหมด 50,000 รอบ



รูปที่ 2.23 นิวรอลเน็ตเวิร์กในการจำแนกเสียงสระ และขอบเขตการตัดสินใจที่สร้างขึ้น

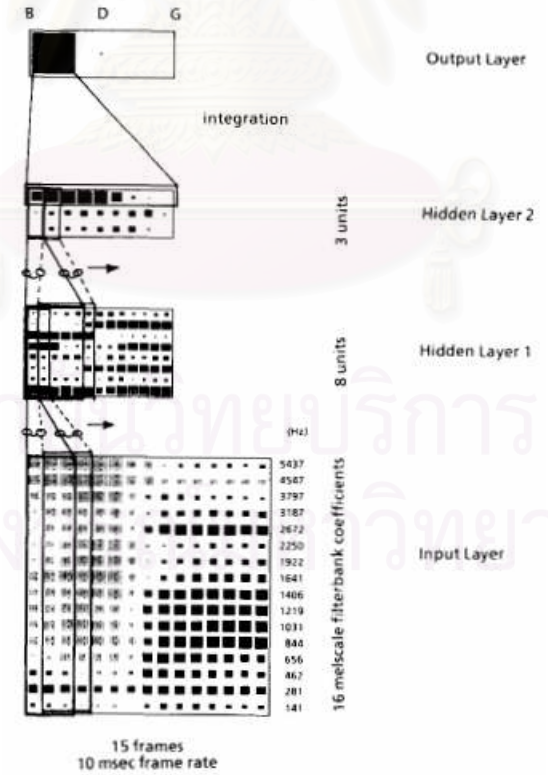
Elman และ Zipser ใช้นิวรอลเน็ตเวิร์กสองชั้นเพื่อจำแนกสามหน่วยเสียงสระ /a/, /i/ และ /u/ และสามหน่วยเสียงพยัญชนะต้น /b/, /d/ และ /g/ โดยนิวรอลเน็ตเวิร์กรับข้อมูลเข้าเป็นสัมประสิทธิ์เชิงความถี่ของ 20 เฟรม ในช่วง 64 มิลลิวินาทีของเสียงพูด ใช้จำนวนโหนดในชั้นซ่อน

จำนวน 2 ถึง 6 โหนด ได้ค่าความผิดพลาด 0.5% เมื่อจำแนกหน่วยเสียงสระ และ 5.0% สำหรับหน่วยเสียงพยัญชนะต้น [22]

Kammerer และ Kupper ใช้นิวรอลเน็ตเวิร์กชั้นเดียวรู้จำเสียงพูดในระดับคำจำนวน 20 คำ โดยใช้ข้อมูลเข้าเป็นสัมประสิทธิ์เชิงความถี่ 16 เฟรม และพบว่าผลที่ได้ดีกว่าการใช้นิวรอลเน็ตเวิร์กหลายชั้นและการจับคู่แผ่นแบบ [23]

2.3.1.2 นิวรอลเน็ตเวิร์กแบบไทม์ดีเลย์

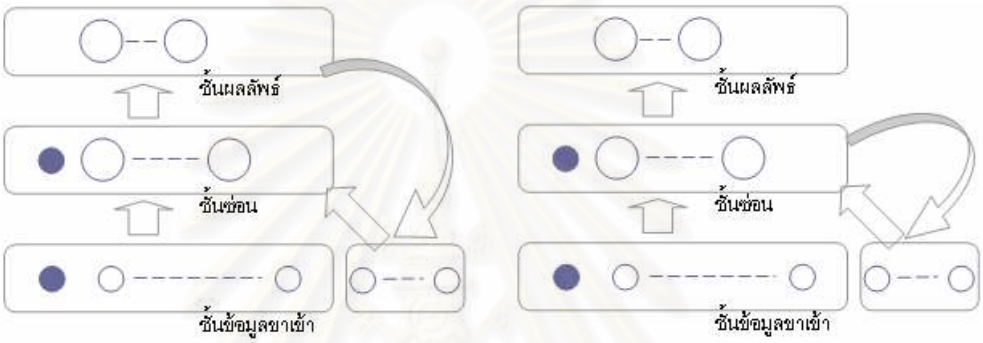
Waibel et al. ใช้นิวรอลเน็ตเวิร์กหลายชั้นที่มีรูปแบบเฉพาะชื่อว่านิวรอลเน็ตเวิร์กแบบไทม์ดีเลย์ ทำการรู้จำสามหน่วยเสียงพยัญชนะต้น /b/, /d/ และ /g/ โดยนิวรอลเน็ตเวิร์กแบบไทม์ดีเลย์เป็นนิวรอลเน็ตเวิร์กสามชั้น แต่ละชั้นจะรับข้อมูลเข้าจากชั้นก่อนหน้าเป็นช่วงๆ มาประกอบกันจนถึงชั้นสุดท้ายซึ่งเป็นชั้นผลลัพธ์ [24] ดังแผนภาพฮินตันในรูปที่ 2.24 เชื่อว่าการประกอบกันเป็นชั้นๆ ทำให้การรู้จำเสียงพูดไม่ต่อเนื่องด้วยนิวรอลเน็ตเวิร์กแบบไทม์ดีเลย์มีความทนทานต่อความแปรผันทางเวลามากขึ้น โดยเมื่อนำไปทดสอบกับฐานข้อมูลเสียงพูดคำภาษาญี่ปุ่น พบว่าให้ค่าความผิดพลาด 1.5% ขณะที่การรู้จำโดยใช้แบบจำลองมาร์คอฟซ่อนตัวให้ค่าความผิดพลาด 6.3%



รูปที่ 2.24 แผนภาพฮินตันอธิบายนิวรอลเน็ตเวิร์กแบบไทม์ดีเลย์

2.3.1.3 นิวรอลเน็ตเวิร์กแบบเวียนซ้ำ

นิวรอลเน็ตเวิร์กแบบเวียนซ้ำเป็นนิวรอลเน็ตเวิร์กที่มีการนำผลจากการคำนวณในนิวรอลเน็ตเวิร์กย้อนกลับมาเป็นข้อมูลเข้าของนิวรอลเน็ตเวิร์กอีกครั้งหนึ่ง ซึ่งต่างจากนิวรอลเน็ตเวิร์กทั่วไปซึ่งไม่มีการนำผลที่ได้ย้อนกลับมาคำนวณใหม่ หรืออาจเรียกนิวรอลเน็ตเวิร์กทั่วไปว่าเป็นนิวรอลเน็ตเวิร์กแบบป้อนไปข้างหน้า นิวรอลเน็ตเวิร์กแบบเวียนซ้ำอาจนำผลจากชั้นผลลัพธ์ย้อนกลับมาเป็นข้อมูลเข้า (Jordan [25]) หรือนำผลจากชั้นซ่อนย้อนกลับมาเป็นข้อมูลเข้าก็ได้ (Elman [26]) โดยนิวรอลเน็ตเวิร์กทั้งสองนี้มีชื่อเรียกว่า จอร์แดนเน็ตเวิร์ก และ เอลแมนเน็ตเวิร์ก ตามลำดับ ดังแสดงได้ในรูปที่ 2.25



รูปที่ 2.25 ซ้าย: จอร์แดนเน็ตเวิร์ก ขวา: เอลแมนเน็ตเวิร์ก

Elman พบว่านิวรอลเน็ตเวิร์กแบบเวียนซ้ำสามารถนำมาใช้กับข้อมูลเข้าในรูปอนุกรมเวลาได้ผลดี [26] และมีหลายงานวิจัยประยุกต์นิวรอลเน็ตเวิร์กแบบเวียนซ้ำกับการรู้จำเสียงพูดไม่ต่อเนื่อง ดังนี้

Watrous ใช้นิวรอลเน็ตเวิร์กแบบเวียนซ้ำเพื่อจำแนกสามหน่วยเสียงสระ /a/, /i/ และ /u/ และสามหน่วยเสียงพยัญชนะต้น /b/, /d/ และ /g/ โดยจะใช้พัลส์แบบเกาส์เป็นเป้าหมายของการเรียนรู้แทนค่าคงที่ ได้ค่าความผิดพลาด 0.0% เมื่อจำแนกหน่วยเสียงสระ และ 0.8% สำหรับหน่วยเสียงพยัญชนะต้น [27]

Robinson และ Fallside ใช้โครงข่ายจอร์แดนในการรู้จำหน่วยเสียง และใช้วิธีการเรียนรู้แบบแบ็กพรอพาเกชันทรูไทม์ ได้ค่าความผิดพลาด 22.7% เมื่อเทียบกับนิวรอลเน็ตเวิร์กแบบป้อนไปข้างหน้าที่มีค่าความผิดพลาด 26% [28]

2.3.2 การใช้นิวรอลเน็ตเวิร์กรู้จำเสียงพูดต่อเนื่องอัตโนมัติ

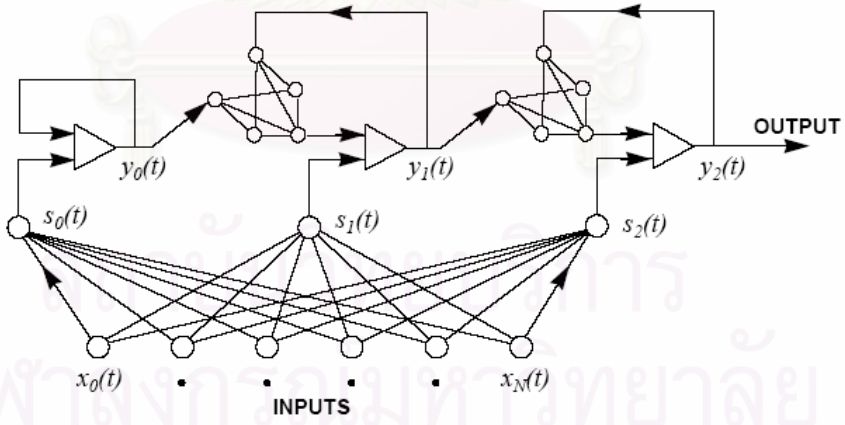
แบบจำลองมาร์คอฟซ่อนตัวถูกนำมาประยุกต์ใช้ในการรู้จำเสียงพูดตั้งแต่ในช่วงคริสตทศวรรษ 1970 และใช้ได้ผลดีจนถือได้ว่าเป็นอัลกอริทึมหลักในการรู้จำเสียงพูดมาจนถึงปัจจุบัน อย่างไรก็ตาม นักวิจัยด้านการรู้จำเสียงพูดได้เห็นถึงจุดอ่อนและข้อจำกัดในการใช้แบบจำลอง

มาร์คอฟซ่อนตัว และพยายามคิดหาวิธีที่ต่างออกไปในการรู้จำเสียงพูด ซึ่งนิวรอลเน็ตเวิร์กนับว่ามีคุณสมบัติที่น่าสนใจและน่าจะสามารถแก้ไขจุดอ่อน รวมทั้งข้อจำกัดของแบบจำลองมาร์คอฟซ่อนตัวได้

ในช่วงคริสตทศวรรษ 1980 มีหลายงานวิจัยที่ใช้นิวรอลเน็ตเวิร์กทำการรู้จำเสียงพูดไม่ต่อเนื่อง และให้ผลที่ดี แต่การรู้จำเสียงพูดต่อเนื่องอัตโนมัตินั้น ระบบต้องรู้จำลำดับของคำในเสียงพูดซึ่งมีได้จำนวนมาก ทำให้นิวรอลเน็ตเวิร์กในรูปแบบที่มีอยู่เพียงอย่างเดียวไม่สามารถรองรับงานนี้ได้ จำเป็นต้องประยุกต์ใช้ร่วมกับวิธีอื่น เช่นแบบจำลองมาร์คอฟซ่อนตัว โดยนับตั้งแต่ช่วงคริสตทศวรรษ 1990 เป็นต้นมา มีงานวิจัยจำนวนมากที่มีจุดมุ่งหมายเพื่อพัฒนาระบบผสมผสานระหว่างนิวรอลเน็ตเวิร์กและแบบจำลองมาร์คอฟซ่อนตัว (Morgan and Boulard [29]) (Trentin and Gori [30]) โดยการสร้างระบบผสมผสานระหว่างนิวรอลเน็ตเวิร์กและแบบจำลองมาร์คอฟซ่อนตัวทำได้มากมายหลายวิธี ซึ่งอาจจำแนกได้ดังนี้

2.3.2.1 การใช้นิวรอลเน็ตเวิร์กจำลองการทำงานของแบบจำลองมาร์คอฟซ่อนตัว

Lippmann และ Gold เสนอแบบจำลองชื่อว่าวิเทอบีเน็ต [31] ที่ใช้นิวรอลเน็ตเวิร์กจำลองการทำงานของอัลกอริทึมวิเทอบีในแบบจำลองมาร์คอฟซ่อนตัว โดยมีโครงสร้างเป็นดังรูปที่ 2.26 ซึ่งวิเทอบีเน็ตนี้จำลองแบบจำลองมาร์คอฟซ่อนตัวที่มีสามสถานะ เน็ตเวิร์กด้านล่างที่ต่อกับข้อมูลเข้าจะคำนวณความน่าจะเป็นในการออกผลลัพธ์ และเน็ตเวิร์กด้านบนคำนวณค่าตัวแปรวิเทอบี



รูปที่ 2.26 วิเทอบีเน็ต

เนื่องจากวิเทอบีเน็ตเป็นเพียงการใช้นิวรอลเน็ตเวิร์กจำลองแบบจำลองมาร์คอฟซ่อนตัว ประสิทธิภาพที่ได้ในการรู้จำเสียงพูดจึงเท่ากับประสิทธิภาพของแบบจำลองมาร์คอฟซ่อนตัว

2.3.2.2 การใช้นิวรอลเน็ตเวิร์กประมาณค่าพารามิเตอร์ในแบบจำลองมาร์คอฟซ่อนตัว

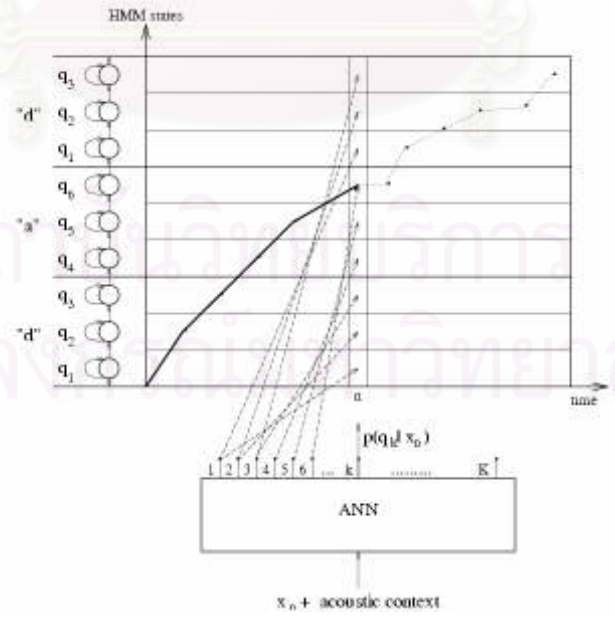
Boulard และ Wellekens พิสูจน์ว่านิวรอลเน็ตเวิร์กแบบเวียนซ้ำนั้นเป็นกรณีทั่วไปของแบบจำลองมาร์คอฟซ่อนตัว [32] นอกจากนี้ นิวรอลเน็ตเวิร์กที่มีข้อมูลเข้าเป็นลักษณะสำคัญของเสียงพูดในแต่ละเฟรมสามารถประมาณค่าความน่าจะเป็นภายหลังของสถานะในแบบจำลองมาร์คอฟซ่อนตัวได้ (Boulard and Morgan [33]) โดยเบื้องต้น ค่าความน่าจะเป็นของแบบจำลองมาร์คอฟซ่อนตัว M_i เมื่อกำหนดเสียงพูด X มาให้ คือ $P(M_i \mid X)$ ซึ่ง

$$\begin{aligned} P(M_i \mid X) &= \sum_Q P(Q, M_i \mid X) = \sum_Q P(Q \mid X) P(M_i \mid Q) \\ &= \sum_Q [P(q_1 \mid X) P(q_2 \mid X, q_1) ... P(q_L \mid X, q_1, q_2, ..., q_{L-1})] P(M_i \mid Q) \\ &= \sum_Q [\prod_{l=1}^{L} P(q_l \mid X, q_1, q_2, ..., q_{l-1})] P(M_i \mid Q) \end{aligned} \tag{2.31}$$

ซึ่งถ้ากำหนดให้ความน่าจะเป็นภายหลังของสถานะ $P(q_l \mid X, q_1, q_2, ..., q_{l-1})$ ขึ้นอยู่กับสถานะก่อนหน้าเท่านั้น และขึ้นกับเฉพาะข้อมูลเข้าจำนวน $2k+1$ ที่อยู่ล้อมรอบ จะได้สมการ (2.31) เป็น

$$P(M_i \mid X) \approx \sum_Q [\prod_{l=1}^{L} P(q_l \mid x_{l-k}, ..., x_{l+k}, q_{l-1})] P(M_i \mid Q) \tag{2.32}$$

และสามารถประมาณค่าความน่าจะเป็นภายหลังของสถานะ $P(q_l \mid x_{l-k}, ..., x_{l+k}, q_{l-1})$ ได้โดยใช้นิวรอลเน็ตเวิร์ก ดังรูปที่ 2.27



รูปที่ 2.27 การใช้นิวรอลเน็ตเวิร์กประมาณค่าความน่าจะเป็นภายหลังของสถานะต่างๆ ในแบบจำลองมาร์คอฟซ่อนตัว

Robinson นำนิวรอลเน็ตเวิร์กแบบเวียนซ้ำมาประมาณค่าความน่าจะเป็นภายหลังของสถานะในแบบจำลองมาร์คอฟซ่อนตัวแทนที่จะใช้นิวรอลเน็ตเวิร์กแบบป้อนไปข้างหน้า [34] ซึ่งภายหลังได้พัฒนาต่อเนื่องเป็นระบบรู้จำเสียงพูด ABBOT (Hochberg et al. [35])

2.3.2.3 ระบบผสมผสานระหว่างนิวรอลเน็ตเวิร์กและแบบจำลองมาร์คอฟซ่อนตัวในลักษณะอื่น ๆ

ระบบผสมผสานระหว่างนิวรอลเน็ตเวิร์กและแบบจำลองมาร์คอฟซ่อนตัวอีกแบบหนึ่งคือการใช้นิวรอลเน็ตเวิร์กสกัดลักษณะเสียงพูด ให้ได้ข้อมูลที่เหมาะสมกับการเรียนรู้ของแบบจำลองมาร์คอฟซ่อนตัว ซึ่งมีหลายงานวิจัย ดังนี้

Bengio et al. ทำการเรียนรู้นิวรอลเน็ตเวิร์กจากค่าความน่าจะเป็นไปข้างหน้าและความน่าจะเป็นมาข้างหลังของแบบจำลองมาร์คอฟซ่อนตัว เป็นการเรียนรู้ไปพร้อมๆ กัน [36] และพบว่าค่าความถูกต้องในการจำแนกหน่วยเสียง /b/, /d/, /g/, /p/, /t/, /k/, /dx/ และหน่วยเสียงอื่นๆ เพิ่มขึ้นจาก 75% เมื่อเรียนรู้นิวรอลเน็ตเวิร์กและแบบจำลองมาร์คอฟซ่อนตัวแยกกัน เป็น 86% เมื่อเรียนรู้นิวรอลเน็ตเวิร์กและแบบจำลองมาร์คอฟซ่อนตัวพร้อมกัน

Rigoll นำนิวรอลเน็ตเวิร์กมาทำการแบ่งนับลักษณะสำคัญของเสียงพูดเพื่อเข้าสู่แบบจำลองมาร์คอฟซ่อนตัว โดยใช้หลักทฤษฎีสารสนเทศในการเรียนรู้ [37] พบว่าผลที่ได้จากการแบ่งนับวิธีนี้ดีกว่าการแบ่งนับด้วยวิธีเคมีนส์

Le Cerf et al. ใช้นิวรอลเน็ตเวิร์กเพื่อจำแนกชนิดของเสียงพูดก่อนจะนำเข้าสู่แบบจำลองมาร์คอฟซ่อนตัว [38] โดยผลลัพธ์จากนิวรอลเน็ตเวิร์กจะสัมพันธ์กับคลาสต่างๆ ของเสียงพูดที่เป็นข้อมูลเข้า

นอกจากนี้อาจใช้นิวรอลเน็ตเวิร์กเพื่อให้คะแนนผลลัพธ์ที่ได้จากแบบจำลองมาร์คอฟซ่อนตัวอีกทีหนึ่ง เช่น

Zavaliagkos et al. ทำการรู้จำเสียงพูดโดยใช้แบบจำลองมาร์คอฟซ่อนตัวจนได้ลำดับของคำที่น่าจะเป็นที่สุดจำนวน N ลำดับ จากนั้นนำนิวรอลเน็ตเวิร์กที่รู้จำเสียงพูดในระดับส่วนมาให้คะแนนลำดับเหล่านี้อีกทีหนึ่ง [39] เมื่อนำวิธีนี้มาใช้กับฐานข้อมูลเสียงพูดอาร์เอ็ม พบว่าวิธีนี้ให้ผลดีกว่าการใช้แบบจำลองมาร์คอฟซ่อนตัวรู้จำเสียงพูดเพียงอย่างเดียว

บทที่ 3
การรู้จำเสียงพูดต่อเนื่องภาษาไทยโดยใช้นิวรอลเน็ตเวิร์ก

ระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยที่ได้พัฒนาขึ้น ประกอบด้วยส่วนการรู้จำหน่วยย่อยทางภาษาในแต่ละกรอบการวิเคราะห์ และการประกอบหน่วยย่อยทางภาษาเหล่านั้นจนกลายมาเป็นผลลัพธ์ ซึ่งจะขอเรียกว่าเป็นส่วนการรู้จำเสียงพูดไม่ต่อเนื่อง และส่วนการรู้จำเสียงพูดต่อเนื่องอัตโนมัติ ตามลำดับ ส่วนการรู้จำเสียงพูดไม่ต่อเนื่องจะใช้นิวรอลเน็ตเวิร์กเป็นตัวรู้จำหน่วยเสียงในแต่ละเฟรม และส่วนการรู้จำเสียงพูดต่อเนื่องอัตโนมัติจะทำการหาลำดับของคำที่เป็นไปได้มากที่สุดจากผลลัพธ์ของนิวรอลเน็ตเวิร์กในทุกเฟรม

ในบทนี้ หัวข้อแรกจะขอแนะนำฐานข้อมูลเสียงพูดที่ใช้ในการทดลอง อันได้แก่ ฐานข้อมูลเสียงพูดชื่อไทย และฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย แล้วจะกล่าวถึงส่วนการรู้จำเสียงพูดไม่ต่อเนื่อง ขั้นตอนกระบวนการ รวมทั้งผลการทดลอง จากนั้นจะเป็นส่วนการรู้จำเสียงพูดต่อเนื่องอัตโนมัติ ซึ่งต่อยอดมาจากส่วนการรู้จำเสียงพูดไม่ต่อเนื่องอีกทีหนึ่ง รวมทั้งแสดงผลการทดลอง เป็นลำดับดังนี้แล

3.1 ฐานข้อมูลเสียงพูดที่ใช้ในการทดลอง

ความยากง่ายในการพัฒนาระบบรู้จำเสียงพูดและประสิทธิภาพที่ได้นั้น ขึ้นอยู่กับปัจจัยต่างๆ มากมาย ซึ่งฐานข้อมูลเสียงพูดที่ต่างกันก็จะมีความยากง่ายในการรู้จำต่างกัน โดยขึ้นกับปัจจัยต่างๆ ดังนี้

1. *จำนวนคำศัพท์* ถ้าคำศัพท์ที่ต้องการรู้จำมีจำนวนน้อย เช่น การรู้จำเสียงพูดตัวเลขศูนย์ ถึง เก้า จะสามารถทำได้ง่ายกว่าการรู้จำเสียงพูดในกรณีที่มีคำศัพท์จำนวนมาก
2. *ความขึ้นกับผู้พูด* โดยในระบบรู้จำเสียงพูดที่ขึ้นกับผู้พูดจะรู้จำได้เฉพาะเสียงพูดของผู้ใช้ที่มีจำนวนจำกัด อาจเป็นผู้ใช้เพียงคนเดียว หรือผู้ใช้เป็นกลุ่ม ขณะที่ระบบรู้จำเสียงพูดที่ไม่ขึ้นกับผู้พูด จะรู้จำเสียงพูดโดยไม่ขึ้นอยู่กับว่าใครเป็นผู้พูด ซึ่งพัฒนาได้ยากกว่าและให้ความผิดพลาดมากกว่า
3. *สภาวะแวดล้อม* ระบบที่ทำการรู้จำในสภาพแวดล้อมที่เงียบสนิทจะพัฒนาได้ง่ายกว่าและให้ความผิดพลาดน้อยกว่าระบบที่ทำการรู้จำในสภาพแวดล้อมที่มีเสียงรบกวน
4. *รูปแบบการพูด* ระบบรู้จำเสียงพูดสามารถแบ่งตามรูปแบบการพูดที่ระบบสามารถรองรับได้ ตั้งแต่

4.1 *การพูดคำเดี่ยว* เป็นการพูดทีละคำอย่างชัดเจน จุดเริ่มต้นและจุดสิ้นสุดของคำเป็นเสียงเงียบ จึงสามารถระบุขอบเขตของคำได้อย่างแน่ชัด รวมทั้งเสียงของคำไม่เพี้ยนมากนัก ทำให้รู้จำได้ง่ายที่สุด ตัวอย่างการพูดรูปแบบนี้ได้แก่ การพูดตัวเลขเดี่ยว การพูดชื่อคน เป็นต้น

4.2 *การพูดคำต่อเนื่อง* เป็นการพูดหลายคำสั้นๆ ติดกัน โดยขอบเขตของแต่ละคำจะแยกจากกันไม่ชัดเจน นอกจากนี้แต่ละคำที่พูดมีจะมีความหลากหลายในการออกเสียง เนื่องจากมีการพูดที่เป็นธรรมชาติ และได้รับผลกระทบจากเสียงของคำอื่น ทำให้การรู้จำทำได้ยากกว่า ตัวอย่างการพูดรูปแบบนี้ได้แก่ การพูดหมายเลขโทรศัพท์ การออกคำสั่ง เป็นต้น

4.3 *การพูดแบบอ่าน* เป็นการพูดต่อเนื่องอย่างยาวนาน เสียงที่พูดมีความเป็นธรรมชาติกว่าการพูดคำต่อเนื่อง และมักมีจำนวนคำศัพท์มาก ทำให้การรู้จำทำได้ยากขึ้น ตัวอย่างการพูดรูปแบบนี้ได้แก่ การอ่านข่าวกระจายเสียง การอ่านนิทาน เป็นต้น

4.4 *การพูดสนทนา* เป็นรูปแบบการพูดที่ทำการรู้จำได้ยาก เนื่องจากเป็นการพูดที่ไม่เป็นทางการ และเป็นธรรมชาติที่สุด คำศัพท์ที่ใช้พูดอาจเป็นคำศัพท์ที่ระบบไม่รู้จัก นอกจากนี้ยังมีเสียงอื่นๆ คอยแทรก เช่น เสียงหัวเราะ และเสียงอุทาน ตัวอย่างการพูดรูปแบบนี้ได้แก่ การสนทนาทางโทรศัพท์ การพูดคุยระหว่างเพื่อนฝูง เป็นต้น

3.1.1 ฐานข้อมูลเสียงพูดชื่อไทย

ฐานข้อมูลเสียงพูดชื่อไทย (Pungprasertying and Kijsirikul [40]) เป็นการรวบรวมเสียงพูดชื่ออาจารย์ของภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย จำนวน 45 ชื่อ โดยฐานข้อมูลเสียงพูดชื่อไทยมีรายละเอียดดังตาราง 3.1

ตาราง 3.1 ลักษณะของฐานข้อมูลเสียงพูดชื่อไทย

ชื่อฐานข้อมูลเสียงพูด	ฐานข้อมูลเสียงพูดชื่อไทย
รูปแบบการพูด	การพูดคำเดี่ยว
จำนวนคำศัพท์	45
ความขึ้นกับผู้พูด	ไม่ขึ้นกับผู้พูด
สภาวะแวดล้อม	พูดทางโทรศัพท์

โดยทำการบันทึกเสียงด้วยอัตราการชักตัวอย่างเท่ากับ 11025 เฮิรตซ์ ใช้การแบ่งนับเท่ากับ 8 บิต มีผู้พูดทั้งหมด 20 คน แบ่งเป็นชาย 10 คน และหญิง 10 คน แต่ละคนทำการพูดทุกชื่อ แล้วแบ่งเสียงพูดจากชาย 7 คน และหญิง 7 คน มาใช้เป็นข้อมูลสำหรับการเรียนรู้ ที่เหลือจากนั้นเป็นข้อมูลสำหรับการทดสอบ โดยรายชื่อที่ใช้ทั้งหมดนั้นสามารถดูได้ที่ภาคผนวก ค

3.1.2 ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย (Thubthong and Kijsirikul [41]) เป็นเสียงพูดแบบอ่านในประโยคเกี่ยวกับสัตว์ จำนวน 90 ประโยค โดยฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยมีรายละเอียดดังตาราง 3.2

ตาราง 3.2 ลักษณะของฐานข้อมูลเสียงพูดชื่อไทย

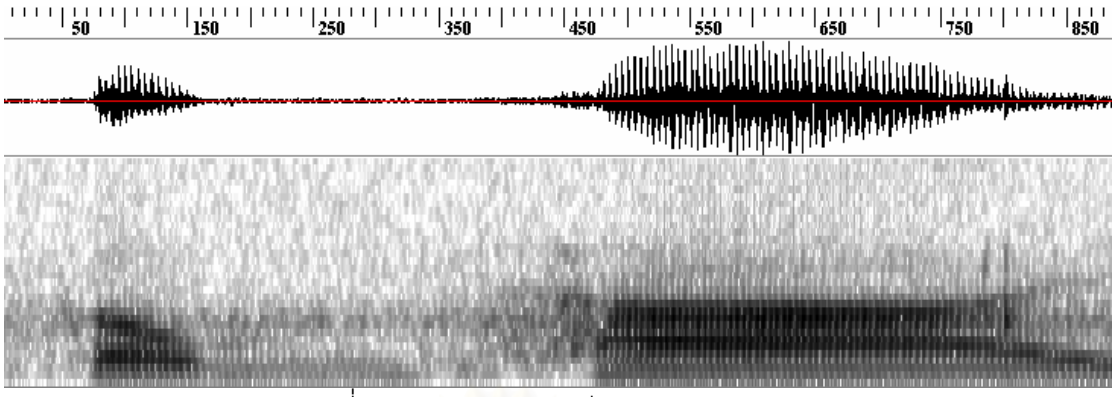
ชื่อฐานข้อมูลเสียงพูด	ฐานข้อมูลเสียงพูดชื่อไทย
รูปแบบการพูด	การพูดแบบอ่าน
จำนวนคำศัพท์	245
ความขึ้นกับผู้พูด	ไม่ขึ้นกับผู้พูด
สภาวะแวดล้อม	พูดในห้องเงียบ

โดยทำการบันทึกเสียงด้วยอัตราการชักตัวอย่างเท่ากับ 11025 เฮิรตซ์ ใช้การแบ่งนับเท่ากับ 16 บิต มีผู้พูดทั้งหมด 20 คน แบ่งเป็นชาย 10 คน และหญิง 10 คน แต่ละคนทำการพูดทุกชื่อ แล้วแบ่งเสียงพูดจากชาย 7 คน และหญิง 7 คน มาใช้เป็นข้อมูลสำหรับการเรียนรู้ ที่เหลือจากนั้นเป็นข้อมูลสำหรับการทดสอบ โดยประโยคที่ใช้ทั้งหมดนั้นสามารถดูได้ที่ภาคผนวก ค

3.2 ส่วนการรู้จำเสียงพูดไม่ต่อเนื่อง

3.2.1 รูปลักษณ์ของเสียงพูดและกรอบการวิเคราะห์

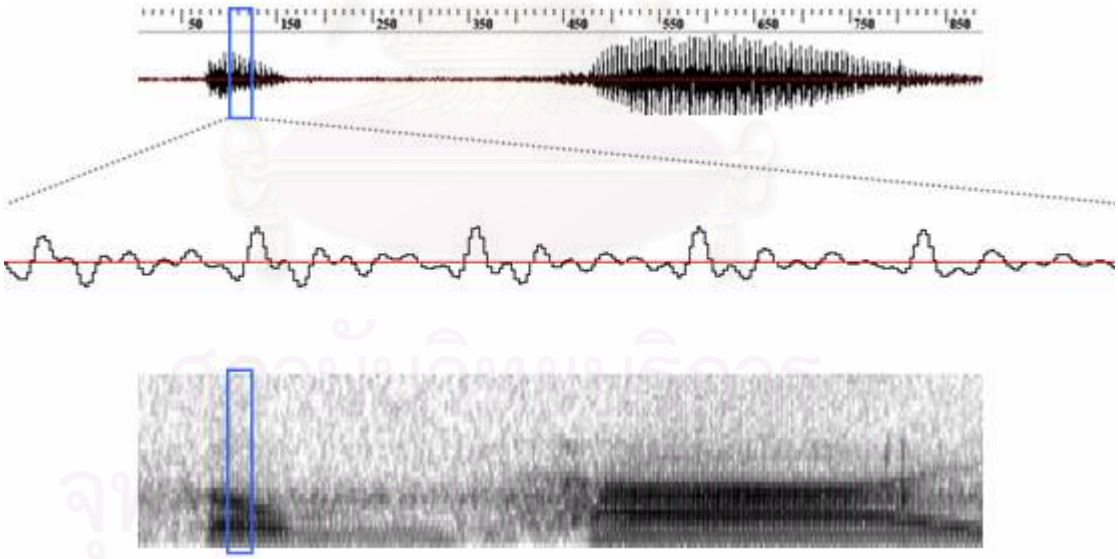
รูปที่ 3.1 แสดงถึงรูปลักษณ์เสียงพูดคำว่าสมชาย (/so4m^_chaa0j^/) ซึ่งเป็นเสียงพูดหนึ่งจากฐานข้อมูลเสียงพูดชื่อไทย โดยการออกเสียงนี้ใช้เวลาทั้งสิ้นประมาณ 900 มิลลิวินาที (0.9 วินาที) เสียงพูดนี้ประกอบด้วยหน่วยเสียงต่างๆ ต่อๆ กัน ซึ่งเมื่อดูจากสเปกโตรแกรมก็จะพบช่วงเวลาในการออกเสียงของแต่ละหน่วยเสียง โดยแต่ละหน่วยเสียงจะมีลักษณะเชิงความถี่ที่ต่างกัน สเปกโตรแกรมจึงสามารถบ่งบอกหน่วยเสียงในแต่ละช่วงเวลาได้ เช่น จากรูปที่ 3.1 จะเห็นหน่วยเสียง /aa/ อยู่ที่ช่วงเวลาประมาณ 470 ถึง 800 มิลลิวินาที



รูปที่ 3.1 รูปลักษณ์ของเสียงพูดคำว่าสมชาย

ส่วนการรู้จำเสียงพูดไม่ต่อเนื่องในที่นี้จะทำการรู้จำหน่วยย่อยทางภาษาในระดับหน่วยเสียง เพราะฉะนั้นจึงต้องกำหนดกรอบการวิเคราะห์ให้พอเหมาะพอดีกับช่วงที่เป็นหน่วยเสียงหนึ่งๆ ซึ่งกรอบการวิเคราะห์นี้จะต้องไม่กว้างเกินไปจนครอบคลุมลักษณะของหน่วยเสียงอื่นที่ไม่เกี่ยวข้อง และต้องไม่แคบเกินไปจนไม่สามารถจับลักษณะของหน่วยเสียงที่ต้องการได้

รูปที่ 3.2 แสดงกรอบการวิเคราะห์ที่มีความกว้าง 25 มิลลิวินาที ครอบคลุมระหว่างช่วง 100 มิลลิวินาที ถึง 125 มิลลิวินาที ของเสียงพูดคำว่าสมชาย (/so4m^_chaa0j^/) เมื่อขยายคลื่นเสียงออกมาดูพบว่าในช่วงนี้ประกอบด้วยลูกคลื่นประมาณ 5 คาบ ซึ่งเพียงพอต่อการวิเคราะห์ และเมื่อดูจากสเปกโตรแกรมก็เห็นว่าช่วงนี้เป็นหน่วยเสียง /o/ นั่นเอง



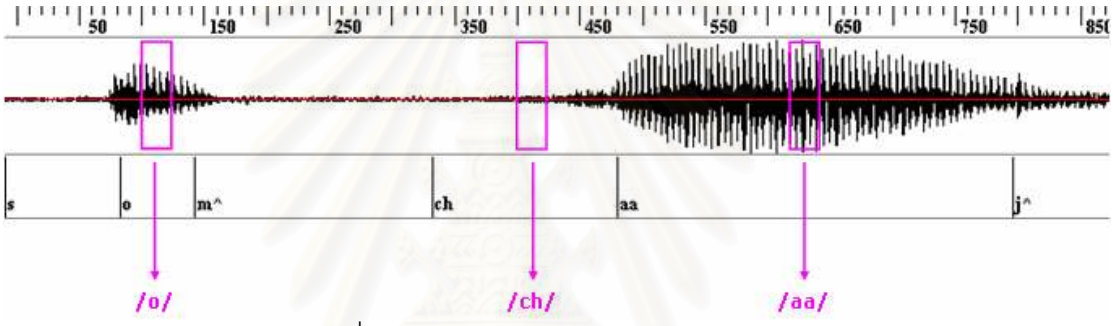
รูปที่ 3.2 กรอบการวิเคราะห์

กรอบการวิเคราะห์ในที่นี้อาจเรียกว่าเฟรม ซึ่งเฟรมที่มีความกว้าง 25 มิลลิวินาทีนั้นเป็นที่ใช้กันโดยส่วนมากในการรู้จำเสียงพูด โดยในที่นี้ก็จะใช้เฟรมขนาด 25 มิลลิวินาทีในการทดลองเช่นกัน โดยเฟรมแรกจะอยู่ที่จุดเริ่มต้นของเสียงพูด และแต่ละเฟรมจะอยู่ถัดไปทุกๆ 12.5 มิลลิวินาที นั่นคือแต่ละเฟรมจะมีส่วนที่ซ้อนทับกันอยู่ด้วย

นอกจากกรอบการวิเคราะห์ที่เป็นเฟรมแล้ว ยังมีกรอบการวิเคราะห์ระดับส่วน (Ostendorf and Roukos [42]) (Glass [43]) ซึ่งจะแบ่งช่วงที่เป็นหน่วยเสียงเดียวกันให้อยู่ในกรอบการวิเคราะห์เดียวกันไปเลย ทำให้การรู้จำทำได้ง่ายขึ้น อย่างไรก็ตาม จำเป็นต้องมีวิธีการแบ่งช่วงที่มีประสิทธิภาพประกอบด้วย

3.2.2 กระบวนการเรียนรู้

เนื่องจากข้อมูลที่จะเข้าสู่กระบวนการเรียนรู้จำเป็นต้องมีป้ายบอกว่าเป็นข้อมูลคลาสใด ซึ่งคลาสในที่นี้ก็คือหน่วยเสียงที่ต้องการรู้จำนั่นเอง จึงต้องทำการติดป้ายให้แต่ละเฟรมว่าเป็นหน่วยเสียงใด โดยในการติดป้าย จะทำการแบ่งเสียงพูดด้วยแรงงานคน ว่าช่วงไหนเป็นหน่วยเสียงใด ถ้าจุดกึ่งกลางเฟรมที่สนใจตกอยู่ ณ ช่วงไหน ก็จะทำการติดป้ายให้เป็นหน่วยเสียงนั้น ดังในรูปที่ 3.3 ซึ่งแสดงตัวอย่างการติดป้ายของสามเฟรม

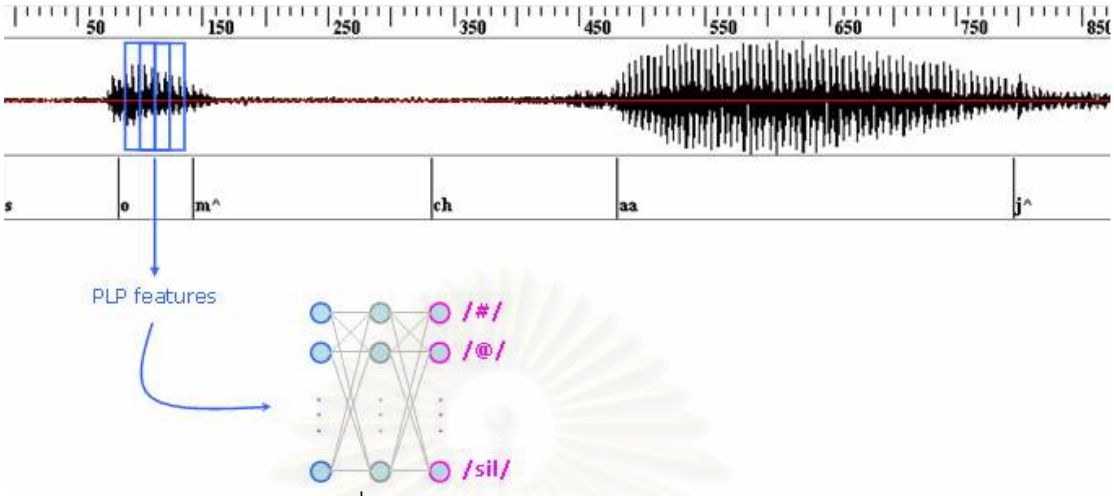


รูปที่ 3.3 ตัวอย่างการติดป้ายของสามเฟรม

ในการเรียนรู้นั้น จำเป็นต้องหาลักษณะสำคัญของเสียงเพื่อนำมาเป็นข้อมูลเข้า โดยในที่นี้ใช้วิธีพีแอลพี (Hermansky [3]) ซึ่งสามารถดูรายละเอียดได้ที่ภาคผนวก ง สำหรับโปรแกรมที่ใช้หาลักษณะของเสียงด้วยวิธีพีแอลพีนั้นมาจากการดัดแปลงโปรแกรมของ The International Computer Science Institute [44]

ข้อมูลเข้าสำหรับการเรียนรู้ในเฟรมใดๆ จะประกอบด้วยลักษณะสำคัญของเสียงในเฟรมนั้น และลักษณะสำคัญของเสียงในเฟรมข้างเคียงจำนวนหนึ่ง ด้วยเห็นว่าลักษณะสำคัญของเฟรมข้างเคียงก็มีส่วนในการจำแนกหน่วยเสียงในเฟรมที่สนใจเช่นกัน นิวรอลเน็ตเวิร์กสองชั้นจึงถูกสร้างขึ้นโดยมีจำนวนข้อมูลเข้าเท่ากับจำนวนลักษณะสำคัญของเสียงในเฟรมเฟรมหนึ่ง ซึ่งก็คืออันดับพีแอลพีที่ใช้ คูณด้วยจำนวนเฟรมที่สนใจ และมีจำนวนโหนดในชั้นผลลัพธ์เท่ากับจำนวนหน่วยเสียงที่ต้องการจะรู้จำ โดยโหนดในชั้นผลลัพธ์แต่ละโหนดจะแทนแต่ละหน่วยเสียงที่ต้องการรู้จำ หากข้อมูลเข้าได้รับการติดป้ายว่าเป็นหน่วยเสียงใด จะกำหนดให้ค่าที่แท้จริงของโหนดในชั้นผลลัพธ์ที่แทนหน่วยเสียงนี้มีค่าเท่ากับ 1 ส่วนค่าที่แท้จริงของโหนดในชั้นผลลัพธ์อื่นๆ มีค่าเท่ากับ 0 กระบวนการเรียนรู้จะใช้วิธีแบ็กพรอพาเกชันดังที่กล่าวไว้ในบทที่ 2

กระบวนการเรียนรู้สามารถเขียนเป็นแผนภาพได้ดังรูปที่ 3.4 สำหรับโปรแกรมนิวรอลเน็ตเวิร์กที่ใช้นั้นมาจากการดัดแปลงโปรแกรม NICO Toolkit [45]



รูปที่ 3.4 แผนภาพแสดงกระบวนการเรียนรู้

3.2.3 กระบวนการรู้จำ

เมื่อกระบวนการเรียนรู้เสร็จสิ้น จะได้นิวรอลเน็ตเวิร์กที่ผ่านการปรับค่าน้ำหนักแล้ว ซึ่งสามารถนำมาใช้จำแนกข้อมูลทั่วไปได้ ในกระบวนการรู้จำจะใช้กรอบการวิเคราะห์และการสกัดลักษณะสำคัญ รวมทั้งจำนวนเฟรมข้างเคียงเช่นเดียวกับในกระบวนการเรียนรู้

เมื่อข้อมูลเข้าผ่านการประมวลผลจากนิวรอลเน็ตเวิร์กมาจนถึงชั้นผลลัพธ์ หากค่าของโหนดใดในชั้นผลลัพธ์สูงที่สุด ก็จะนำหน่วยเสียงนั้นมาเป็นคำตอบของการรู้จำ

ข้อมูลสำหรับการทดสอบจะถูกติดป้ายไว้เช่นกัน ถ้าผลการรู้จำไม่ตรงกับป้ายก็แสดงว่าการรู้จำนั้นผิดพลาด ซึ่งประสิทธิภาพของระบบรู้จำเสียงพูดก็จะวัดจากความถูกต้องในการรู้จำนั้น

3.2.4 การทดลองเพื่อรู้จำหน่วยเสียงในแต่ละเฟรม

ในระบบรู้จำเสียงพูดไม่ต่อเนื่องที่ได้พัฒนาขึ้นนั้นประกอบด้วยพารามิเตอร์ต่างๆ มากมาย ในการทดลองเพื่อทดสอบประสิทธิภาพของระบบ เราจะเน้นที่ 3 พารามิเตอร์หลัก คือ

1. *หน่วยเสียงที่ต้องการรู้จำ* เป็นการกำหนดว่าระบบจะต้องเรียนรู้และรู้จำคลาสในจำนวนเท่าใด ซึ่งแม้ในภาษาไทยจะมีหน่วยเสียงที่ถูกกำหนดขึ้นมาเป็นมาตรฐาน แต่ก็พบว่าหน่วยเสียงเหล่านี้อาจจะไม่เหมาะกับการใช้นิวรอลเน็ตเวิร์กที่รับข้อมูลเป็นเฟรม จึงได้ทำการลดทอนบางหน่วยเสียงไป ดังรายละเอียดในภาคผนวก ข พร้อมทำการทดลองเพื่อทดสอบว่าจำนวนคลาสที่เหมาะสมนั้นควรจะกำหนดเช่นไร

หน่วยเสียงที่ถูกตัดออกไปสามารถแบ่งได้เป็น 3 กลุ่มใหญ่ คือ หน่วยเสียงที่เป็นพยัญชนะควบกล้ำ สระเสียงยาว และสระผสม โดยเฟรมในฐานข้อมูลเสียงพูดที่เคยถูกติดป้ายว่าเป็นพยัญชนะควบกล้ำ จะถูกติดป้ายใหม่โดยดูจากลักษณะของเสียงพูดในเฟรมนั้นว่าเป็นเสียงของพยัญชนะต้น หรือพยัญชนะที่มาควบ เช่น เฟรมที่เคยติดป้ายเป็น /pr/ อาจจะถูกเปลี่ยนเป็น /p/ หรือ /r/ เป็นต้น สำหรับเฟรมในฐานข้อมูลเสียงพูดที่เคยถูกติดป้ายว่าเป็นสระผสมก็เช่นเดียวกัน จะถูกติดป้ายใหม่โดยดูจากลักษณะของเสียงพูดในเฟรมนั้นว่าเป็นเสียงของสระใด เช่น เฟรมที่เคยติดป้ายเป็น /iia/ อาจจะถูกเปลี่ยนเป็น /i/ หรือ /a/ เป็นต้น ส่วนเฟรมในฐานข้อมูลเสียงพูดที่เคยถูกติดป้ายว่าเป็นสระเสียงยาวนั้น จะถูกติดป้ายใหม่เป็นสระเสียงสั้น เช่น /ii/ เปลี่ยนเป็น /i/ เป็นต้น

2. *อันดับของพีแอลพี* อันดับของพีแอลพีคือจำนวนลักษณะสำคัญที่วิธีพีแอลพีสกัดออกมาได้จากในแต่ละเฟรม ซึ่งก็คือจำนวนข้อมูลที่จะเข้าสู่กระบวนการเรียนรู้และกระบวนการรู้จำของแต่ละเฟรมนั่นเอง ซึ่งจะทำการทดลองว่าข้อมูลที่จะสกัดออกมานี้ควรมีจำนวนเท่าใด จึงจะให้ผลการทดสอบที่ดีที่สุด เนื่องจากถ้าใช้ข้อมูลจำนวนน้อยไป อาจจะไม่เพียงพอต่อการจำแนก ขณะที่ถ้าใช้ข้อมูลจำนวนมากไป อาจจะทำให้เกิดการโอเวอร์ฟิตขึ้น

3. *จำนวนเฟรมที่ใช้* เนื่องจากในกระบวนการเรียนรู้และกระบวนการรู้จำจะนำเฟรมข้างเคียงมาใช้เป็นข้อมูลเข้าด้วย จึงได้ทำการทดลองเพื่อทดสอบว่าควรจะใช้เฟรมข้างเคียงจำนวนเท่าใดจึงจะเหมาะ ซึ่งถ้าต้องใช้เฟรมข้างเคียงจำนวนมาก แสดงว่าลักษณะที่เป็นตัวกำหนดการจำแนกหน่วยเสียงนั้นมีความสัมพันธ์อันยาวไกล

3.2.4.1 **ผลการทดลองกับฐานข้อมูลเสียงพูดซื่อไทย**

การทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ผลการทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำสามารถแสดงได้ดังตาราง 3.3

ตาราง 3.3 ผลการทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ชุดหน่วยเสียง	ความถูกต้องในระดับเฟรม	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
มาตรฐาน	63.51 %	56.04 %
ลดทอน	**82.30 %**	**72.38 %**
จำนวนข้อมูลทั้งหมด	34796 เฟรม	16757 เฟรม

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- อันดับของพีแอลพีเท่ากับ 6
- จำนวนเฟรมที่ใช้เท่ากับ 9

จะเห็นได้ว่าการเรียนรู้ชุดหน่วยเสียงมาตรฐานนั้นทำได้ยากลำบาก เมื่อเทียบกับชุดหน่วยเสียงที่ถูกทำการลดทอน โดยการเรียนรู้ชุดหน่วยเสียงมาตรฐานให้ความถูกต้องในชุดข้อมูลสำหรับการทดสอบเพียง 21.89 % ซึ่งน้อยมาก ขณะที่ชุดหน่วยเสียงที่ถูกทำการลดทอนให้ความถูกต้องในชุดข้อมูลสำหรับการทดสอบ 72.38 % จึงเห็นว่าชุดหน่วยเสียงที่ถูกทำการลดทอนเหมาะกับใช้นิวรอลเน็ตเวิร์กที่รับข้อมูลเป็นเฟรมมากกว่า แม้ในขั้นตอนต่อไป คือการประกอบหน่วยเสียงขึ้นเป็นคำ ชุดหน่วยเสียงที่ถูกทำการลดทอนจะให้การตัดสินใจที่ยากขึ้นก็ตาม เช่น เสียงพูด /cha0j^/ และเสียงพูด /chaa0j^/ จะประกอบด้วยหน่วยเสียงเดียวกัน คือ /ch/, /a/, /j^/ ซึ่งต้องอาศัยแบบจำลองทางภาษามาช่วยในการรู้จำต่อไป

การทดลองปรับค่าอันดับของพีแอลพี

ผลการทดลองปรับค่าอันดับของพีแอลพีสามารถแสดงได้ดังตาราง 3.4

ตาราง 3.4 ผลการทดลองปรับค่าอันดับของพีแอลพี

อันดับของพีแอลพี	ความถูกต้องในระดับเฟรม	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
3	66.28 %	67.11 %
6	82.30 %	**72.38 %**
9	86.05 %	71.33 %
12	89.59 %	69.40 %
15	89.59 %	68.16 %
18	**91.18 %**	66.75 %
จำนวนข้อมูลทั้งหมด	34796 เฟรม	16757 เฟรม

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- จำนวนเฟรมที่ใช้เท่ากับ 9

การทดลองนี้ได้ปรับค่าอันดับของพีแอลพีตั้งแต่ 3 ไปเรื่อยๆ จนถึง 18 โดยเพิ่มขึ้นทีละ 3 พบว่าค่าอันดับของพีแอลพีที่มากขึ้นทำให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้มากขึ้น แต่ว่าความถูกต้องในชุดข้อมูลสำหรับการทดสอบมากที่สุดอยู่ที่ค่าอันดับของพีแอลพีเท่ากับ 6 จากนั้นก็ลดลงเป็นลำดับ นั่นแสดงว่าการสกัดลักษณะสำคัญให้มีจำนวนข้อมูลมากไปได้ทำให้เกิดการโอเวอร์ฟิตขึ้นแล้ว

การทดลองปรับจำนวนเฟรมที่ใช้

ผลการทดลองปรับจำนวนเฟรมที่ใช้สามารถแสดงได้ดังตาราง 3.5

ตาราง 3.5 ผลการทดลองปรับจำนวนเฟรมที่ใช้

จำนวนเฟรมที่ใช้	ความถูกต้องในระดับเฟรม	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
5	74.93 %	68.95 %
9	82.30 %	72.38 %
13	**86.44 %**	**73.78 %**
จำนวนข้อมูลทั้งหมด	34796 เฟรม	16757 เฟรม

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- อันดับของพีแอลพีเท่ากับ 6

ในที่นี้พบว่าถ้าใช้จำนวนเฟรมที่มากขึ้น จะทำให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้มีมากขึ้น และความถูกต้องในชุดข้อมูลสำหรับการทดสอบก็มากขึ้นด้วย ในที่นี้ได้เพิ่มจำนวนเฟรมที่ใช้จนถึง 13 เฟรม ซึ่งครอบคลุมช่วงเวลา 162.5 มิลลิวินาที และพบว่าที่ค่านี้ยังคงให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้ และความถูกต้องในชุดข้อมูลสำหรับการทดสอบสูงสุด

3.2.4.2 ผลการทดลองกับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

สำหรับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยนั้น พบว่าจำนวนเฟรมทั้งสิ้นในชุดข้อมูลสำหรับการเรียนรู้นั้นมีถึง 258341 เฟรม จึงใช้เวลานานสำหรับทำการเรียนรู้ นอกจากนั้น ในจำนวนนี้ มีบางหน่วยเสียง เช่น /sil/ ที่นิวรอลเน็ตเวิร์กเน้นการเรียนรู้ที่หน่วยเสียงเหล่านี้จนเกินไป จนทำให้จำแนกหน่วยเสียงอื่นได้ไม่ถูกต้องเลย ดังแสดงผลในตาราง 3.6

ตาราง 3.6 ผลการจำแนกหน่วยเสียงในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

ชุดข้อมูลสำหรับการเรียนรู้			ชุดข้อมูลสำหรับการทดสอบ		
หน่วยเสียง	จำนวนเฟรม	ความถูกต้อง (%)	หน่วยเสียง	จำนวนเฟรม	ความถูกต้อง (%)
/#/	791	15.68	/#/	296	19.59
/@/	6870	0.00	/@/	2846	0.00
/a/	53685	0.00	/a/	22289	0.00
/b/	713	0.00	/b/	325	0.00
/c/	1294	0.00	/c/	537	0.00
/ch/	4066	0.00	/ch/	1914	0.00
/ch^/	0	0.00	/ch^/	0	0.00
/d/	936	0.00	/d/	421	0.00
/e/	3279	0.00	/e/	1333	0.00
/f/	499	0.00	/f/	213	0.00
/f^/	0	0.00	/f^/	0	0.00
/h/	2077	0.00	/h/	927	0.00
/i/	13688	95.56	/i/	5596	94.35
/j/	1923	0.00	/j/	870	0.00
/j^/	7067	0.00	/j^/	2984	0.00
/k/	1716	0.00	/k/	747	0.00
/k^/	2057	0.00	/k^/	894	0.00
/kh/	5342	0.00	/kh/	2550	0.00
/p/	1176	0.00	/p/	512	0.00
/p^/	1297	0.00	/p^/	550	0.00
/ph/	1278	0.00	/ph/	531	0.00
/l/	4225	0.00	/l/	2071	0.00
/l^/	0	0.00	/l^/	0	0.00
/m/	4840	0.00	/m/	2315	0.00
/m^/	5251	0.00	/m^/	2344	0.00
/n/	3170	0.00	/n/	1445	0.00
/n^/	12296	97.06	/n^/	5434	94.28
/ng/	852	0.00	/ng/	386	0.00
/ng^/	9725	0.00	/ng^/	4307	0.00
/o/	4090	0.00	/o/	1648	0.00
/q/	1010	0.00	/q/	445	0.00
/r/	2167	0.00	/r/	1048	0.00
/s/	3801	0.00	/s/	1689	0.00
/s^/	0	0.00	/s^/	0	0.00
/t/	1032	0.00	/t/	458	0.00
/t^/	1624	0.00	/t^/	710	0.00
/th/	2074	0.00	/th/	946	0.00
/u/	7020	0.00	/u/	3047	0.00
/v/	3688	0.08	/v/	1521	0.13
/w/	2453	0.00	/w/	1157	0.00
/w^/	3479	0.00	/w^/	1552	0.00
/x/	5346	0.00	/x/	2215	0.00
/sil/	70444	98.06	/sil/	33648	98.32
รวมทั้งสิ้น	258341	36.47	รวมทั้งสิ้น	114721	37.96

โดยค่าพารามิเตอร์ที่กำหนดไว้มีดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- อันดับของพีแอลพีเท่ากับ 6
- จำนวนเฟรมที่ใช้เท่ากับ 9

ด้วยเหตุนี้ จึงได้พยายามทำให้จำนวนเฟรมของแต่ละหน่วยเสียงที่จะใช้ในการเรียนรู้มีค่าเท่าๆ กัน โดยดูว่าหน่วยเสียงใดที่มีจำนวนเฟรมน้อยที่สุด แล้วเลือกจำนวนเฟรมของหน่วยเสียงอื่นให้เท่ากับจำนวนเฟรมของหน่วยเสียงที่น้อยที่สุดนั้น โดยเลือกเฟรมให้กระจายครอบคลุมตลอด

ช่วงของชุดข้อมูลสำหรับการเรียนรู้ และพบว่านิวรอลเน็ตเวิร์กทำการรู้จำแต่ละหน่วยเสียงดีขึ้น ดังตาราง 3.7

ตาราง 3.7 ผลการจำแนกหน่วยเสียงเมื่อทำการปรับสมดุลแล้ว

ชุดข้อมูลสำหรับการเรียนรู้			ชุดข้อมูลสำหรับการทดสอบ		
หน่วยเสียง	จำนวนเฟรม	ความถูกต้อง (%)	หน่วยเสียง	จำนวนเฟรม	ความถูกต้อง (%)
/#/	499	20.24	/#/	296	21.62
/@/	499	68.54	/@/	2846	55.31
/a/	499	52.10	/a/	22289	41.10
/b/	499	81.96	/b/	325	72.00
/c/	499	76.35	/c/	537	75.61
/ch/	499	86.57	/ch/	1914	82.03
/ch^/	0	0.00	/ch^/	0	0.00
/d/	499	77.15	/d/	421	62.00
/e/	499	76.15	/e/	1333	69.84
/f/	499	73.95	/f/	213	67.61
/f^/	0	0.00	/f^/	0	0.00
/h/	499	51.70	/h/	927	46.06
/i/	499	69.74	/i/	5596	68.69
/j/	499	82.57	/j/	870	76.09
/j^/	499	68.54	/j^/	2984	67.96
/k/	499	58.32	/k/	747	50.33
/k^/	499	50.70	/k^/	894	44.52
/kh/	499	56.11	/kh/	2550	54.04
/p/	499	72.14	/p/	512	60.55
/p^/	499	56.51	/p^/	550	57.09
/ph/	499	42.08	/ph/	531	40.87
/l/	499	48.70	/l/	2071	36.79
/l^/	0	0.00	/l^/	0	0.00
/m/	499	38.68	/m/	2315	28.64
/m^/	499	54.31	/m^/	2344	48.46
/n/	499	41.68	/n/	1445	34.39
/n^/	499	60.52	/n^/	5434	51.34
/ng/	499	65.53	/ng/	386	39.12
/ng^/	499	26.45	/ng^/	4307	23.06
/o/	499	62.53	/o/	1648	54.25
/q/	499	58.72	/q/	445	50.56
/r/	499	57.31	/r/	1048	52.67
/s/	499	70.14	/s/	1689	61.28
/s^/	0	0.00	/s^/	0	0.00
/t/	499	69.54	/t/	458	62.88
/t^/	499	64.33	/t^/	710	58.03
/th/	499	50.10	/th/	946	44.08
/u/	499	58.12	/u/	3047	52.94
/v/	499	62.53	/v/	1521	53.19
/w/	499	61.52	/w/	1157	64.82
/w^/	499	51.10	/w^/	1552	45.55
/x/	499	67.13	/x/	2215	57.88
/sil/	499	78.36	/sil/	33648	78.86
รวมทั้งสิ้น	19461	60.74	รวมทั้งสิ้น	114721	58.24

โดยค่าพารามิเตอร์ที่กำหนดไว้มีดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- อันดับของพีแอลพีเท่ากับ 6
- จำนวนเฟรมที่ใช้เท่ากับ 9

ในการเรียนรู้ พบว่าถ้าใช้ค่าอัตราการเรียนรู้เท่ากับ 0.00001 เท่ากับในฐานข้อมูลเสียงพูดชื่อไทย นิวรอลเน็ตเวิร์กจะไม่สามารถทำการเรียนรู้ได้ จึงได้ปรับลดอัตราการเรียนรู้ลงเป็น 0.000001 ในทุกการทดลอง

การทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ผลการทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำสามารถแสดงได้ดังตาราง 3.8

ตาราง 3.8 ผลการทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ชุดหน่วยเสียง	ความถูกต้องในระดับเฟรม	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
มาตรฐาน	31.82 %	46.54 %
ลดทอน	**60.74 %**	**58.24 %**
จำนวนข้อมูลทั้งหมด	3840 เฟรม ในชุดหน่วยเสียงมาตรฐาน 19461 เฟรม ในชุดหน่วยเสียงที่ถูกทำการลดทอน	114721 เฟรม

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- อันดับของพีแอลพีเท่ากับ 6
- จำนวนเฟรมที่ใช้เท่ากับ 9

ในที่นี้ เนื่องจากจำนวนเฟรมของหน่วยเสียงที่มีจำนวนน้อยที่สุดในชุดหน่วยเสียงมาตรฐานมีจำนวนน้อย จึงส่งผลให้ข้อมูลที่ใช้ทำการเรียนรู้มีจำนวนน้อยกว่า และให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้ รวมทั้งความถูกต้องในชุดข้อมูลสำหรับการทดสอบน้อยกว่า

การทดลองปรับค่าอันดับของพีแอลพี

ผลการทดลองปรับค่าอันดับของพีแอลพีสามารถแสดงได้ดังตาราง 3.9

ตาราง 3.9 ผลการทดลองปรับค่าอันดับของพีแอลพี

อันดับของพีแอลพี	ความถูกต้องในระดับเฟรม	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
3	35.51 %	42.33 %
6	60.74 %	58.24 %
9	67.83 %	60.73 %
12	71.98 %	**61.50 %**

15	73.67 %	60.68 %
18	**74.36 %**	60.29 %
จำนวนข้อมูลทั้งหมด	19461 เฟรม	114721 เฟรม

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- จำนวนเฟรมที่ใช้เท่ากับ 9

จะเห็นว่าเมื่อเพิ่มค่าอันดับของพีแอลพีขึ้นจะทำให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้มีมากขึ้น เช่นเดียวกับการทดลองในฐานข้อมูลเสียงพูดชื่อไทย แต่ค่าอันดับของพีแอลพีที่ทำให้ความถูกต้องในชุดข้อมูลสำหรับการทดสอบสูงสุดอยู่ที่ 12 ซึ่งไม่เหมือนกับในฐานข้อมูลเสียงพูดชื่อไทยในตาราง 3.4 ซึ่งอยู่ที่ 6

การทดลองปรับจำนวนเฟรมที่ใช้

ผลการทดลองปรับจำนวนเฟรมที่ใช้สามารถแสดงได้ดังตาราง 3.10

ตาราง 3.10 ผลการทดลองปรับจำนวนเฟรมที่ใช้

จำนวนเฟรมที่ใช้	ความถูกต้องในระดับเฟรม	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
5	51.32 %	54.21 %
9	60.74 %	58.24 %
13	**65.73 %**	**60.00 %**
จำนวนข้อมูลทั้งหมด	19461 เฟรม	114721 เฟรม

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- อันดับของพีแอลพีเท่ากับ 6

ในที่นี้พบว่าถ้าใช้จำนวนเฟรมที่มากขึ้น จะทำให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้มีมากขึ้น และความถูกต้องในชุดข้อมูลสำหรับการทดสอบก็มากขึ้นด้วย เช่นเดียวกับผลการทดลองในฐานข้อมูลเสียงพูดชื่อไทยดังตาราง 3.5 ซึ่งจำนวนเฟรมที่ใช้เท่ากับ 13 ให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้และชุดข้อมูลสำหรับการรู้จำสูงสุด

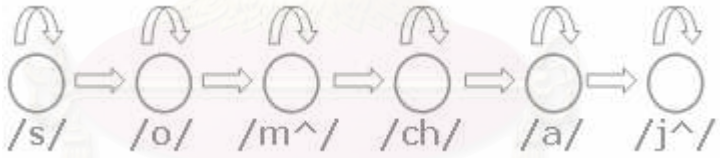
3.3 ส่วนการรู้จำเสียงพูดต่อเนื่องอัตโนมัติ

3.3.1 กระบวนการรู้จำ

จากระบบการรู้จำเสียงพูดไม่ต่อเนื่องที่พัฒนาขึ้นมานั้น ผลลัพธ์ที่ได้คือหน่วยเสียงที่น่าจะเป็นของทุกเฟรม ทว่าในงานรู้จำเสียงพูดทั่วไปเราไม่ต้องการผลลัพธ์เช่นนี้ หากแต่ต้องการเป็นลำดับของคำในภาษา จึงต้องนำผลลัพธ์จากระบบรู้จำเสียงพูดไม่ต่อเนื่องนั้นมาเป็นแบบจำลองทางเสียง ซึ่งเมื่อรวมกับแบบจำลองทางภาษาและกระบวนการค้นหาแล้ว จะได้ลำดับของคำในภาษาออกมาเป็นผลลัพธ์

ขั้นตอนแรกจะทำการเขียนพจนานุกรมขึ้นมาก่อน ซึ่งพจนานุกรมในที่นี้คือสิ่งที่บ่งบอกว่าในคำแต่ละคำประกอบด้วยหน่วยเสียงใดบ้าง เช่นคำว่าสมชาย (/so4m^_chaa0j^/) จะประกอบด้วยหน่วยเสียง /s/, /o/, /m^/, /ch/, /a/ และ /j^/ (ในที่นี้ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน) โดยรายละเอียดของพจนานุกรมในฐานข้อมูลเสียงพูดที่ใช้ทำการทดลองสามารถดูได้ที่ภาคผนวก จ

จากนั้นจะทำการสร้างแบบจำลองมาร์คอฟซ่อนตัวของทุกคำศัพท์ที่ปรากฏในฐานข้อมูลเสียงพูดขึ้นมา โดยแต่ละสถานะจะแทนแต่ละหน่วยเสียงที่ประกอบเป็นคำนั้น เช่นคำว่า สมชาย (/so4m^_chaa0j^/) ที่ประกอบด้วยหน่วยเสียง /s/, /o/, /m^/, /ch/, /a/ และ /j^/ จะสามารถสร้างแบบจำลองมาร์คอฟซ่อนตัวได้ดังรูปที่ 3.5



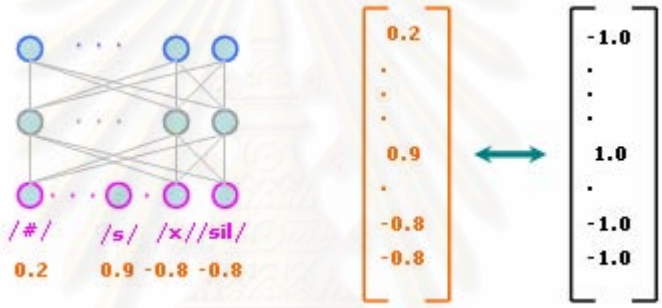
รูปที่ 3.5 แบบจำลองมาร์คอฟซ่อนตัวของคำว่าสมชาย

แบบจำลองมาร์คอฟซ่อนตัวในที่นี้ ที่แต่ละสถานะ จะมีความน่าจะเป็นในการเปลี่ยนสถานะเท่ากัน ระหว่างความน่าจะเป็นในการอยู่ที่สถานะเดิมและความน่าจะเป็นในการอยู่ที่สถานะถัดไป นั่นคือความน่าจะเป็นในการอยู่ที่สถานะเดิมเท่ากับ 0.5 และความน่าจะเป็นในการอยู่ที่สถานะถัดไปเท่ากับ 0.5 เช่นกัน

สำหรับข้อมูลเข้าของแบบจำลองมาร์คอฟซ่อนตัวนี้ ก็คือเวกเตอร์ผลลัพธ์ที่มาจากการรู้จำเสียงพูดไม่ต่อเนื่องของนิวรอลเน็ตเวิร์กนั่นเอง โดยมิติของเวกเตอร์แต่ละมิติจะสื่อถึงหน่วยเสียงต่างๆ ถ้ามิติไหนของเวกเตอร์มีค่ามาก แสดงว่าโอกาสที่จะเป็นหน่วยเสียงนั้นสูง แต่ถ้ามิติไหนของเวกเตอร์มีค่าน้อย แสดงว่าโอกาสที่จะเป็นหน่วยเสียงนั้นต่ำ เพราะฉะนั้นความน่าจะเป็นในการออกผลลัพธ์ของแต่ละสถานะจะใช้การกระจายแบบเกาส์ที่มีค่าเฉลี่ยอยู่ที่เวกเตอร์ตัวหนึ่ง ซึ่ง

เวกเตอร์นี้ ในมิติของหน่วยเสียงที่สถานะนั้นแทนอยู่จะมีค่าเท่ากับค่าสูงสุดที่เป็นไปได้จากการรู้จำ ส่วนในมิติอื่นๆ จะมีค่าเท่ากับค่าต่ำสุดที่เป็นไปได้จากการรู้จำ ยกตัวอย่างเช่น ถ้าชั้นผลลัพธ์ในนิวรอลเน็ตเวิร์กใช้ฟังก์ชันกระตุ้นที่ให้ค่าอยู่ในช่วง -1 ถึง 1 และหน่วยเสียง /s/ แทนด้วยโหนดในชั้นผลลัพธ์ตัวที่ 33 เพราะฉะนั้นจะได้ความน่าจะเป็นในการออกผลลัพธ์ของสถานะที่แทนหน่วยเสียง /s/ เป็นการกระจายแบบเกาส์ที่มีค่าเฉลี่ยอยู่ที่เวกเตอร์ $x^T = [-1,...,1,...,-1,-1]$ โดยมิติที่มีค่าเท่ากับ 1 นั้น คือมิติที่ 33

หรืออาจยกตัวอย่างเป็นแผนภาพได้ดังรูปที่ 3.6 ในที่นี้ ณ เฟรมที่เวลา t ที่ชั้นผลลัพธ์ของนิวรอลเน็ตเวิร์กคำนวณได้ค่า $o_t^T = [0.2,...,0.9,...-0.8,-0.8]$ เมื่อต้องการคำนวณหา $P(o_t \mid q_t)$ โดยที่ q_t เป็นสถานะที่แทนหน่วยเสียง /s/ จะได้ว่า $P(o_t \mid q_t) = N(o_t; \mu, \Sigma)$ เมื่อ $\mu^T = [-1.0,...,1.0,...-1.0,-1.0]$ นั่นเอง



รูปที่ 3.6 แผนภาพแสดงการเปรียบเทียบระหว่างเวกเตอร์ผลลัพธ์จากนิวรอลเน็ตเวิร์กและเวกเตอร์ค่าเฉลี่ยของการกระจายแบบเกาส์ที่ใช้เป็นความน่าจะเป็นในการออกผลลัพธ์

แบบจำลองมาร์คอฟซ่อนตัวของคำศัพท์ทุกคำที่ปรากฏในฐานข้อมูลเสียงพูดจะถูกนำมาต่อกันเป็นเน็ตเวิร์กดังรูปที่ 2.21 ซึ่งความน่าจะเป็นในการเปลี่ยนสถานะระหว่างคำจะใช้แบบจำลองทางภาษาเข้ามาร่วมด้วย กระบวนการค้นหาจะใช้อัลกอริทึมการผ่านโทเคน ดังที่กล่าวไว้ในบทที่ 2 ซึ่งทั้งกระบวนการสร้างแบบจำลองมาร์คอฟซ่อนตัว กระบวนการสร้างแบบจำลองทางภาษา และกระบวนการค้นหานี้ มาจากการดัดแปลงโปรแกรม HTK [46]

3.3.2 การทดลองเพื่อรู้จำลำดับของคำในแต่ละเสียงพูด

3.3.2.1 การทดลองกับฐานข้อมูลเสียงพูดชื่อไทย

ฐานข้อมูลเสียงพูดชื่อไทยเป็นการพูดคำเดี่ยว จึงไม่ต้องสร้างแบบจำลองทางภาษา และไม่ต้องนำคำต่างๆ มาประกอบกันเป็นเน็ตเวิร์กเพื่อใช้ในกระบวนการค้นหา เนื่องจากผลลัพธ์ที่ต้องการจะออกมาเป็นคำเดี่ยวๆ เลย แต่ทั้งนี้ทั้งนั้น ยังต้องสร้างแบบจำลองมาร์คอฟซ่อนตัวเพื่อแสดงคำศัพท์แต่ละคำอยู่ ในการทดลองจะใช้นิวรอลเน็ตเวิร์กที่ผ่านการเรียนรู้มาจากส่วนการรู้จำ

เสียงพูดไม่ต่อเนื่องมาดูว่านิวรอลเน็ตเวิร์กที่ได้รับการปรับค่าพารามิเตอร์ต่างกันจะให้ผลลัพธ์ในระดับการรู้จำเสียงพูดต่อเนื่องอัตโนมัติเป็นอย่างไร

การใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ผลการทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำสามารถแสดงได้ดังตาราง 3.11

ตาราง 3.11 ผลการใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ชุดหน่วยเสียง	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
มาตรฐาน	71.11 %	48.89 %
ลดทอน	**91.11 %**	**88.89 %**
จำนวนข้อมูลทั้งหมด	630 คำ	270 คำ

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- อันดับของพีแอลพีเท่ากับ 6
- จำนวนเฟรมที่ใช้เท่ากับ 9

จากตาราง 3.3 ที่การรู้จำชุดหน่วยเสียงมาตรฐานในระดับเฟรมมีค่าน้อยมาก เมื่อใช้ชุดหน่วยเสียงมาตรฐานมาเป็นพื้นฐานในการรู้จำเสียงพูดต่อเนื่องอัตโนมัติก็พบว่ายังให้ค่าที่น้อยอยู่ เมื่อเทียบกับชุดหน่วยเสียงที่ถูกทำการลดทอน นอกจากนี้ได้พบว่าสำหรับฐานข้อมูลเสียงพูดชื่อไทยแล้ว ความถูกต้องเมื่อวัดในระดับคำมีมากกว่าความถูกต้องเมื่อวัดในระดับหน่วยเสียงมากทีเดียว โดยในชุดข้อมูลสำหรับการทดสอบ ชุดหน่วยเสียงมาตรฐานเพิ่มความถูกต้องจาก 21.89% มาเป็น 62.22% และชุดหน่วยเสียงที่ถูกทำการลดทอนเพิ่มความถูกต้องจาก 72.38% มาเป็น 88.89%

การใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับค่าอันดับของพีแอลพี

ผลการทดลองปรับค่าอันดับของพีแอลพีสามารถแสดงได้ดังตาราง 3.12

ตาราง 3.12 ผลการใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับค่าอันดับของพีแอลพี

อันดับของพีแอลพี	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
3	55.56 %	80.00 %
6	91.11 %	**88.89 %**

9	91.11 %	**88.89 %**
12	93.33 %	**88.89 %**
15	93.33 %	86.67 %
18	**95.56 %**	82.22 %
จำนวนข้อมูลทั้งหมด	630 คำ	270 คำ

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- จำนวนเฟรมที่ใช้เท่ากับ 9

จากผลการทดลอง พบว่าความถูกต้องในการรู้จำเสียงพูดต่อเนื่องอัตโนมัติสอดคล้องกับความถูกต้องจากการใช้นิวรอลเน็ตเวิร์กรู้จำหน่วยเสียงในระดับเฟรม ดังตาราง 3.4 ซึ่งค่าอันดับของพีแอลพีที่มากขึ้นจะทำให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้มากขึ้น ทว่าความถูกต้องในชุดข้อมูลสำหรับการทดสอบกลับลดลง เป็นที่น่าสังเกตว่า ในการรู้จำหน่วยเสียงระดับเฟรมนั้น ค่าอันดับของพีแอลพีที่ให้ความถูกต้องในชุดข้อมูลสำหรับการทดสอบเท่ากับ 6 แต่ในที่นี้พบว่าค่าอันดับของพีแอลพีเท่ากับ 6, 9 และ 12 ก็ให้ความถูกต้องคำในชุดข้อมูลสำหรับการทดสอบเท่ากัน

การใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับจำนวนเฟรมที่ใช้

ผลการทดลองปรับจำนวนเฟรมที่ใช้สามารถแสดงได้ดังตาราง 3.13

ตาราง 3.13 ผลการทดลองปรับจำนวนเฟรมที่ใช้

จำนวนเฟรมที่ใช้	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
5	88.89 %	75.56 %
9	91.11 %	88.89 %
13	95.56 %	93.33 %
จำนวนข้อมูลทั้งหมด	630 คำ	270 คำ

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.00001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- อันดับของพีแอลพีเท่ากับ 6

จากผลการทดลอง พบว่าความถูกต้องในการรู้จำเสียงพูดต่อเนื่องอัตโนมัติสอดคล้องกับความถูกต้องจากการใช้นิวรอลเน็ตเวิร์กรู้จำหน่วยเสียงในระดับเฟรมอีกเช่นกัน ดังตาราง 3.5 ซึ่งจำนวนเฟรมที่ใช้ที่ให้ความถูกต้องสูงสุดสำหรับทั้งชุดข้อมูลสำหรับการเรียนรู้และชุดข้อมูลสำหรับการทดสอบคือ 13 เฟรม

3.3.2.2 การทดลองกับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

การทดลองในฐานข้อมูลเกี่ยวกับสัตว์ภาษาไทย จะเป็นการรู้จำระดับประโยค และวัดความถูกต้องในระดับคำ โดยเมื่อสร้างแบบจำลองมาร์คอฟซ่อนตัวของแต่ละคำแล้ว จะใช้แบบจำลองทางภาษาแบบไบแกรมเพื่อกำหนดการเชื่อมต่อในเน็ตเวิร์ก ซึ่งแบบจำลองทางภาษาแบบไบแกรมจะได้มาจากการเรียนรู้ประโยคต่างๆ ในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยนี้เอง

การรู้จำจะทำโดยใช้ผลจากนิวรอลเน็ตเวิร์กที่ผ่านการเรียนรู้มาจากส่วนการรู้จำเสียงพูดไม่ต่อเนื่อง มาผ่านกระบวนการค้นหา จนได้ผลลัพธ์เป็นลำดับของคำที่ดีที่สุดต่อไป และเมื่อได้ผลลัพธ์ออกมาแล้ว จะทำการวัดผลโดยนำมาเทียบกับลำดับของคำที่ถูกต้อง ซึ่งการเปรียบเทียบนี้ต้องอาศัยการปรับแนวแบบไม่เชิงเส้นที่ให้ความผิดพลาดในการปรับแนวน้อยที่สุด เช่น ประโยค แมว 10 ในภาคผนวก ค ลำดับของคำที่ถูกต้อง คือ

/wee0_laa0/ /n@@0n^/ /ch@@2p^/ /n@@0n^/ /taa0m^/ /ta0w^_fa0j^/ /rvv4/ /bo0n^/ /k@@0ng^_phaa2/

แต่ผลการรู้จำออกมาเป็น

/wee0_laa0/ /n@@0n^/ /ch@@2p^/ /la3_kh@@n0/ /ta0w^_fa0j^/ /rvv4/ /bo0n^/ /t@ng2/ /hat1/

เมื่อนำมาเทียบกันโดยให้ความผิดพลาดในการปรับแนวน้อยที่สุด จะได้การปรับแนวดังนี้

/wee0_laa0/	/n@@0n^/	/ch@@2p^/	/n@@0n^/	/taa0m^/	/ta0w^_fa0j^/	/rvv4/	/bo0n^/	/k@@0ng^_phaa2/	
/wee0_laa0/	/n@@0n^/	/ch@@2p^/	/la3_kh@@n0/		/ta0w^_fa0j^/	/rvv4/	/bo0n^/	/t@ng2/	/hat1/
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)

ก็จะพบความผิดพลาดของการรู้จำ ทั้งที่เกิดจากการแทนที่ คือในตำแหน่งที่ (4) และ (9) รวมทั้งเกิดจากการหายไป คือในตำแหน่งที่ (5) และเกิดจากการเพิ่มขึ้น คือในตำแหน่งที่ (10)

เปอร์เซนต์ความถูกต้องสามารถคำนวณได้โดยใช้สูตร

$$\text{Percentage Correct} = \frac{N - D - S}{N} \times 100$$

เมื่อ N คือจำนวนคำทั้งหมดในลำดับของคำที่ถูกต้อง

D คือจำนวนคำที่หายไป

และ S คือจำนวนคำที่แทนที่

เปอร์เซนต์ความถูกต้องในที่นี้ไม่นำความผิดพลาดจากการเพิ่มขึ้นเข้ามาคิด เนื่องจากใช้ลำดับของคำที่ถูกต้องเป็นหลัก หากนำความผิดพลาดจากการเพิ่มขึ้นเข้ามาคิดด้วย จะเรียกว่าเปอร์เซนต์ความแม่นยำ ซึ่งนิยามดังนี้

$$\text{Percentage Accuracy} = \frac{N - D - S - I}{N} \times 100$$

เมื่อ I คือจำนวนคำที่เพิ่มขึ้น

ในการเขียนผลการทดลองต่างๆ ในหัวข้อนี้ ตัวเลขที่อยู่นอกวงเล็บคือเปอร์เซนต์ความถูกต้อง ขณะที่ตัวเลขที่อยู่ข้างหลังข้างในวงเล็บจะแสดงถึงเปอร์เซนต์ความแม่นยำ ของการรู้จำที่ทำได้

การใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ผลการทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำสามารถแสดงได้ดังตาราง 3.14

ตาราง 3.14 ผลการใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับชุดหน่วยเสียงที่ต้องการรู้จำ

ชุดหน่วยเสียง	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
มาตรฐาน	21.31% (20.54%)	20.94% (19.62%)
ลดทอน	**40.81%** (**34.48%**)	**35.83%** (**28.31%**)
จำนวนข้อมูลทั้งหมด	9669 คำ	4164 คำ

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- อันดับของพีแอลพีเท่ากับ 6
- จำนวนเฟรมที่ใช้เท่ากับ 9

เมื่อใช้ชุดหน่วยเสียงมาตรฐาน ความถูกต้องทั้งในชุดข้อมูลสำหรับการเรียนรู้และในชุดข้อมูลสำหรับการทดสอบมีค่าน้อยกว่าเมื่อใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน เช่นเดียวกับผลลัพธ์จากการรู้จำในระดับเฟรมในตาราง 3.8 เห็นได้ว่าความถูกต้องจากการรู้จำในระดับคำทั้งในชุดหน่วยเสียงมาตรฐานและในชุดหน่วยเสียงที่ถูกทำการลดทอนนั้น มีค่าน้อยกว่าความถูกต้องจากการรู้จำในระดับเฟรม

การใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับค่าอันดับของพีแอลพี

ผลการทดลองปรับค่าอันดับของพีแอลพีสามารถแสดงได้ดังตาราง 3.15

ตาราง 3.15 ผลการใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับค่าอันดับของพีแอลพี

อันดับของพีแอลพี	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
3	19.70% (17.48%)	16.95% (14.17%)
6	40.81% (34.48%)	35.83% (28.31%)
9	45.48% (38.99%)	39.70% (32.08%)
12	47.83% (41.29%)	**41.31%** (32.80%)
15	48.35% (41.66%)	41.21% (**33.00%**)
18	**48.73%** (**41.93%**)	39.29% (31.22%)
จำนวนข้อมูลทั้งหมด	9669 คำ	4164 คำ

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- จำนวนเฟรมที่ใช้เท่ากับ 9

ในที่นี้ พบว่าค่าอันดับของพีแอลพีที่เพิ่มขึ้นจะให้ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้มากขึ้น แต่ความถูกต้องในชุดข้อมูลสำหรับการทดสอบอยู่ที่ 12 และ 15 ซึ่งในการรู้จำในระดับเฟรมดังตาราง 3.9 ก็ให้ผลการทดลองที่มีแนวโน้มแบบเดียวกัน

การใช้นิวรอลเน็ตเวิร์กที่ทดลองปรับจำนวนเฟรมที่ใช้

ผลการทดลองปรับจำนวนเฟรมที่ใช้สามารถแสดงได้ดังตาราง 3.16

ตาราง 3.16 ผลการทดลองปรับจำนวนเฟรมที่ใช้

จำนวนเฟรมที่ใช้	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
5	35.03% (29.93%)	31.92% (26.08%)
9	40.81% (34.48%)	35.83% (28.31%)
13	**43.71%** (**37.41%**)	**37.56%** (**29.61%**)
จำนวนข้อมูลทั้งหมด	9669 คำ	4164 คำ

โดยค่าพารามิเตอร์อื่นกำหนดไว้ดังนี้

- จำนวนโหนดในชั้นซ่อนเท่ากับ 100 และจำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000
- ค่าอัตราการเรียนรู้เท่ากับ 0.000001 และค่าโมเมนตัมเท่ากับ 0.7
- ใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
- อันดับของพีแอลพีเท่ากับ 6

พบว่าเมื่อใช้จำนวนเฟรมมากขึ้น ความถูกต้องในชุดข้อมูลสำหรับการเรียนรู้ และชุดข้อมูลสำหรับการรู้จำเพิ่มขึ้น ซึ่งสอดคล้องกับผลการรู้จำในระดับเฟรมในตาราง 3.10

3.4 วิเคราะห์ผลการทดลอง

จากการทดลองพบว่าประสิทธิภาพของระบบขึ้นอยู่กับความยากง่ายของฐานข้อมูลเสียงพูดอยู่มาก เช่นในฐานข้อมูลเสียงพูดชื่อไทย ซึ่งเป็นการพูดคำเดี่ยว และมีจำนวนคำศัพท์น้อย รวมทั้งเสียงพูดของคำแต่ละคำค่อนข้างจะต่างกัน การรู้จำจะทำได้ด้วยดี คือให้ความถูกต้องในระดับคำของข้อมูลสำหรับการทดสอบสูงสุดถึงเกือบ 90% ขณะที่ความถูกต้องในระดับเฟรมของข้อมูลสำหรับการทดสอบต่ำกว่านั้น คือสูงสุดประมาณ 70% ขณะที่ในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยให้ความถูกต้องในระดับเฟรมของข้อมูลสำหรับการทดสอบสูงสุดประมาณ 60% แต่ความถูกต้องในระดับคำของข้อมูลสำหรับการทดสอบสูงสุดอยู่ที่ประมาณ 40% เท่านั้น เมื่อวิเคราะห์ดูแล้วก็เห็นว่าเป็นเพราะในประโยคหนึ่งมีการออกเสียงที่ค่อนข้างนาน ทำให้การประกอบคำทำได้หลายแบบ นอกจากนี้ยังมีคำที่ออกเสียงคล้ายๆ กันอยู่มากมาย จึงทำให้ผลการรู้จำในระดับคำไม่ดีนัก

สิ่งหนึ่งที่เห็นได้ชัดเจนจากทุกการทดลองก็คือถ้านิวรอลเน็ตเวิร์กแบบใดให้ผลการรู้จำในระดับเฟรมที่ดี ก็ย่อมทำให้ผลการรู้จำในระดับคำดีไปด้วย เพราะฉะนั้น วิธีหนึ่งในการปรับปรุงประสิทธิภาพของระบบก็คือการทำให้การรู้จำในระดับเฟรมทำได้ดีขึ้น แต่ถึงอย่างไร ระบบนี้ก็มีข้อจำกัดอยู่มากมาย โดยข้อจำกัดที่สำคัญก็คือกรอบการวิเคราะห์ระดับเฟรมนั้นไม่สามารถครอบคลุมลักษณะของเสียงพูดที่มีระยะทางยาวไกลได้ รวมทั้งความแปรผันของแต่ละเฟรมเป็นไปได้มากแม้ในหน่วยเสียงเดียวกัน นอกจากนี้นิวรอลเน็ตเวิร์กก็ไม่ใช่วิธีสำหรับเรียนรู้ข้อมูลที่เป็นอนุกรมเวลาโดยตรง ซึ่งปัญหาเหล่านี้จะอภิปรายกันต่อในบทถัดไป

3.5 การเปรียบเทียบประสิทธิภาพกับระบบอื่น ๆ

ในที่นี้ได้ทำการเปรียบเทียบประสิทธิภาพในการรู้จำเสียงพูดต่อเนื่องอัตโนมัติของระบบที่ได้พัฒนาขึ้น กับระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำหน่วยเสียง โดยจะแบ่งเป็น 3 ระบบในการเปรียบเทียบ คือ

1. *ระบบ (1)* คือระบบที่ได้พัฒนาขึ้นซึ่งใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน โดยสกัดลักษณะสำคัญของเสียงพูดด้วยวิธีพีแอลพี ให้ค่าอันดับของพีแอลพีเท่ากับ 6 และจำนวนเฟรมที่ใช้เท่ากับ 9

   ในส่วนการเรียนรู้นั้น ได้กำหนดจำนวนโหนดในชั้นซ่อนเท่ากับ 100 จำนวนรอบในการวนปรับค่าน้ำหนักเท่ากับ 10000 ค่าอัตราการเรียนรู้เท่ากับ 0.00001 สำหรับฐานข้อมูลเสียงพูดชื่อไทย และเท่ากับ 0.000001 สำหรับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย ส่วนค่าโมเมนตัมได้ตั้งไว้ให้เท่ากับ 0.7

2. *ระบบ (2)* คือระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำชุดหน่วยเสียงมาตรฐาน มีการสกัดลักษณะสำคัญของเสียงพูดด้วยวิธีพีแอลพี ให้ค่าอันดับของพีแอลพีเท่ากับ 6 เช่นเดียวกับระบบ (1)

   โดยในส่วนการเรียนรู้ ได้กำหนดจำนวนสถานะในแต่ละแบบจำลองมาร์คอฟซ่อนตัวเท่ากับ 5 ความน่าจะเป็นในการออกผลลัพธ์ของแต่ละสถานะจะใช้การกระจายแบบเกาส์ตัวเดียว และจำนวนรอบในการวนปรับค่าพารามิเตอร์เท่ากับ 20

3. *ระบบ (3)* เหมือนกับระบบ (2) ทุกประการ หากเป็นการใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำชุดหน่วยเสียงที่ถูกทำการลดทอน แทนการใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำชุดหน่วยเสียงมาตรฐาน

3.5.1 การเปรียบเทียบประสิทธิภาพโดยทั่วไป

ในการเปรียบเทียบประสิทธิภาพโดยทั่วไปนี้ ผลการทดลองสำหรับฐานข้อมูลเสียงพูดชื่อไทยสามารถสรุปได้ดังตาราง 3.17

ตาราง 3.17 ผลการเปรียบเทียบสำหรับฐานข้อมูลเสียงพูดชื่อไทย

ระบบ	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
(1)	**91.11 %**	**88.89 %**
(2)	66.51 %	88.81 %
(3)	64.27 %	86.57 %
จำนวนข้อมูลทั้งหมด	630 คำ	270 คำ

และผลการทดลองสำหรับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยสามารถสรุปได้ดังตาราง 3.18

ตาราง 3.18 ผลการเปรียบเทียบสำหรับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

ระบบ	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
(1)	**40.81% (34.48%)**	**35.83% (28.31%)**
(2)	36.01% (31.89%)	32.83% (28.17%)
(3)	30.76% (26.58%)	27.79% (22.91%)
จำนวนข้อมูลทั้งหมด	9669 คำ	4164 คำ

จากผลการทดลองในตาราง 3.17 และตาราง 3.18 พบว่าแบบจำลองมาร์คอฟซ่อนตัวจะสามารถเรียนรู้ได้ดีกว่าหากใช้ชุดหน่วยเสียงมาตรฐาน เนื่องจากแบบจำลองมาร์คอฟซ่อนตัวมีสถานะซึ่งช่วยจับลักษณะในระยะยาว แตกต่างจากนิวรอลเน็ตเวิร์กที่เมื่อใช้ชุดหน่วยเสียงมาตรฐานแล้วไม่สามารถเรียนรู้ได้ดี อย่างไรก็ตามจะเห็นได้ว่าระบบที่พัฒนาขึ้นนี้ให้ความถูกต้องในระดับคำสูงกว่าระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำหน่วยเสียง ทั้งในฐานข้อมูลเสียงพูดชื่อไทย และฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย ทั้งในชุดข้อมูลสำหรับการเรียนรู้ และชุดข้อมูลสำหรับการทดสอบ

เนื่องจากระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำชุดหน่วยเสียงมาตรฐาน คือระบบ (2) ให้ผลดีกว่าระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำชุดหน่วยเสียงที่ถูกทำการลดทอน คือระบบ (3) อย่างเห็นได้ชัด ในการทดลองต่อๆ ไป จึงขอใช้เพียงระบบ (2) ในการเปรียบเทียบประสิทธิภาพ

3.5.2 การเปรียบเทียบประสิทธิภาพเมื่อระบบรับเฟรมเข้าประมวลผลเป็นจำนวนเท่ากัน

ในที่นี้ได้ทำการทดลองเพิ่มเติมว่า ถ้าหากให้ระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำหน่วยเสียงรับข้อมูลเข้าไปประมวลผลครั้งละ 9 เฟรม เช่นเดียวกับระบบที่ได้พัฒนาขึ้น ผลลัพธ์จะเป็นเช่นไร จึงได้เปลี่ยนแปลงค่าพารามิเตอร์ในระบบ (2) ให้จำนวนเฟรมที่ใช้ในแต่ละสถานะเท่ากับ 9 และความน่าจะเป็นในการออกผลลัพธ์ของแต่ละสถานะจะใช้การกระจายแบบเกาส์แบบผสมกัน 8 ตัว เพื่อรองรับความซับซ้อนที่เพิ่มขึ้น และให้ชื่อเรียกใหม่ว่าระบบ (2)+

จากการทดลอง พบว่าในฐานข้อมูลเสียงพูดชื่อไทย ระบบ (2)+ ไม่สามารถทำการเรียนรู้ได้ โดยจะออกผลลัพธ์เป็น /krqq1k/ เพียงอย่างเดียวในทุกเสียงพูดที่ส่งไปทำการรู้จำ ส่วนในฐานข้อมูลเสียงพูดชื่อสัตว์ภาษาไทยนั้น ให้ผลการทดลองดังตาราง 3.19

ตาราง 3.19 ผลการเปรียบเทียบสำหรับฐานข้อมูลเสียงพูดชื่อสัตว์ภาษาไทย
โดยเพิ่มจำนวนเฟรมให้ระบบที่ใช้แบบจำลองมาร์คอฟซ่อนตัวเป็นหลักในการรู้จำหน่วยเสียง

ระบบ	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
(1)	**40.81%** (**34.48%**)	**35.83%** (**28.31%**)
(2)	36.01% (31.89%)	32.83% (28.17%)
(2)+	33.54% (27.35%)	30.81% (23.41%)
จำนวนข้อมูลทั้งหมด	9669 คำ	4164 คำ

พบว่า ขณะที่ในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยนั้น ระบบให้ความถูกต้องที่ต่ำกว่าการใช้เฟรมปกติ อันน่าจะมาจากสาเหตุที่ว่าข้อมูลที่เข้ามานั้นมีจำนวนมิติสูงและมีความหลากหลายมากเกินไป ทำให้แบบจำลองมาร์คอฟซ่อนตัวเรียนรู้ได้ไม่ดีนัก

3.5.3 การเปรียบเทียบประสิทธิภาพโดยไม่ใช้แบบจำลองทางภาษา

สุดท้ายนี้ ได้ทำการทดลองโดยใช้ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยโดยไม่ใช้แบบจำลองทางภาษา เพื่อดูว่าหากไม่มีแบบจำลองทางภาษาเข้าช่วยแล้ว ผลลัพธ์ที่ได้จากแบบจำลองทางเสียงเพียงอย่างเดียวจะเป็นเช่นไร ก็เป็นเช่นตาราง 3.20

ตาราง 3.20 ผลการเปรียบเทียบโดยไม่ใช้แบบจำลองทางภาษา

ระบบ	ความถูกต้องในระดับคำ	
	ชุดข้อมูลสำหรับการเรียนรู้	ชุดข้อมูลสำหรับการทดสอบ
(1)	**34.00%** (2.52%)	**29.20%** (-4.68%)
(2)	31.92% (**24.42%**)	28.79% (**19.88%**)
จำนวนข้อมูลทั้งหมด	9669 คำ	4164 คำ

ในที่นี้ พบว่าระบบที่ได้พัฒนาขึ้นนั้นยังคงให้ความถูกต้องสูงกว่า แต่ความแม่นยำกลับน้อยมาก จึงอาจสรุปได้ว่า เมื่อไม่มีแบบจำลองทางภาษาขึ้นมาช่วยกำหนด ระบบที่ได้พัฒนาขึ้นนี้จะให้คำออกมามากเกินไป ทั้งนี้อาจจะเนื่องมาจากการที่มีเฟรมจำนวนมาก ผลการรู้จำในแต่ละเฟรมถ้าออกมาผิดเพียงเล็กน้อยก็อาจทำให้เกิดเป็นคำสั้นๆ ได้

บทที่ 4
สรุปผลการวิจัยและข้อเสนอแนะ

4.1 สรุปผลการวิจัย

งานวิจัยชิ้นนี้เสนอระบบการรู้จำเสียงพูดต่อเนื่องภาษาไทยที่ใช้นิวรอลเน็ตเวิร์กเป็นพื้นฐานในการพัฒนา โดยนิวรอลเน็ตเวิร์กจะถูกใช้สำหรับรู้จำหน่วยเสียงในระดับเฟรม ให้แบบจำลองทางเสียง ซึ่งเมื่อประกอบกับแบบจำลองทางภาษาและกระบวนการค้นหาแล้ว จะได้ผลลัพธ์เป็นลำดับของคำขึ้นมา

การทดสอบประสิทธิภาพของระบบได้ใช้ฐานข้อมูลเสียงพูดชื่อไทย อันเป็นการพูดคำเดี่ยว และฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย อันเป็นการพูดแบบอ่าน ในชุดข้อมูลสำหรับการทดสอบของฐานข้อมูลเสียงพูดชื่อไทย พบว่าความถูกต้องของระบบอยู่ที่ 70% ในระดับเฟรม และ 90% ในระดับคำ ส่วนชุดข้อมูลสำหรับการทดสอบของฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยนั้น พบว่าความถูกต้องของระบบอยู่ที่ 60% ในระดับเฟรม และ 40% ในระดับคำ

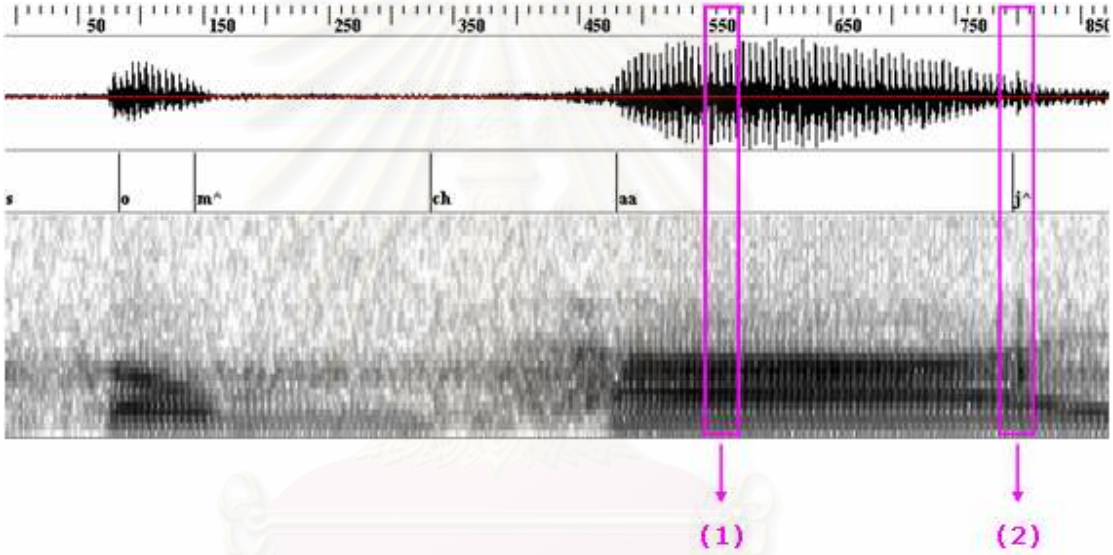
การรู้จำเสียงพูดต่อเนื่องภาษาไทยโดยทั่วไปใช้วิธีวาร์ปเวลาแบบพลวัตและแบบจำลองมาร์คอฟซ่อนตัวเป็นสำคัญ งานวิจัยนี้มุ่งสร้างระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยที่เป็นทั่วไปขึ้นมาระบบหนึ่งโดยอาศัยนิวรอลเน็ตเวิร์กเป็นหลัก ซึ่งให้ความถูกต้องพอสมควร และสามารถปรับปรุงเพื่อนำไปใช้งานในโปรแกรมประยุกต์อื่นต่อไป

4.2 ข้อคิดและข้อเสนอแนะ

ระบบรู้จำเสียงพูดที่ได้พัฒนาขึ้นนั้นยังมีข้อควรคำนึงอีกหลายประการ ดังนี้

1. *หน่วยเสียงที่ใช้* ในที่นี้ทดลองกับชุดหน่วยเสียงมาตรฐานและชุดหน่วยเสียงที่ถูกทำการลดทอนบางส่วนเท่านั้น อาจจะมีบางชุดหน่วยเสียงที่ให้ผลการรู้จำดีกว่านี้
2. *ปัญหาในตอนติดป้าย* ในการเรียนรู้แบบมีผู้สอนนั้นจำเป็นต้องมีป้ายบอกคลาสของข้อมูลสำหรับการเรียนรู้ทุกตัว จึงจำเป็นที่จะต้องระบุทุกเฟรมว่าเป็นหน่วยเสียงใด การติดป้ายในที่นี้เริ่มจากการแบ่งช่วงของหน่วยเสียงต่างๆ ซึ่งมีปัญหาอยู่ที่ว่าหน่วยเสียงเหล่านี้มีลักษณะกลืนกัน ทำให้แบ่งแยกได้ไม่ชัดเจน โดยเฉพาะหน่วยเสียงที่เป็นตัวสะกด /p^/, /t^/ และ /k^/ นั้น แทบจะสังเกตไม่เห็นเลย การติดป้ายที่ผิดพลาดส่งผลต่อกระบวนการเรียนรู้และกระบวนการรู้จำทั้งหมด ปัญหานี้จึงเป็นปัญหาใหญ่ในระบบรู้จำเสียงพูดที่ได้พัฒนาขึ้น

3. *กรอบการวิเคราะห์ระดับเฟรม* ปัญหาสำคัญของกรอบการวิเคราะห์ระดับเฟรมอยู่ที่ส่วนรอยต่อระหว่างหน่วยเสียง ซึ่งเป็นช่วงของเสียงพูดที่มีลักษณะไม่แน่นอนว่าจะเป็นหน่วยเสียงใด เช่นในรูปที่ 4.1 จะเห็นได้ว่าเฟรม (1) และเฟรมใกล้เคียง มีลักษณะของเสียงพูดที่บ่งบอกว่าเป็นหน่วยเสียง /aa/ อย่างชัดเจน ขณะที่เฟรม (2) นั้น ลักษณะของเสียงพูดคาบเกี่ยวอยู่ระหว่างหน่วยเสียง /aa/ และหน่วยเสียง /j^/ ในการติดป้ายเพื่อการเรียนรู้จำเป็นต้องระบุหน่วยเสียงที่ชัดเจนลงไป แม้ในความจริงแล้วลักษณะของเสียงพูดในเฟรมนี้ดูจะระบุให้เฉพาะเจาะจงลงไปไม่ได้ และสิ่งนี้เองที่ทำให้กระบวนการเรียนรู้ต้องพบกับความยากลำบาก เนื่องจากลักษณะของเสียงพูดในแต่ละเฟรม แม้ถูกระบุว่าเป็นหน่วยเสียงเดียวกัน ก็จะมีความแตกต่างได้หลากหลาย ซึ่งเป็นภาระให้กระบวนการเรียนรู้ต้องทำงานหนักยิ่งขึ้น



รูปที่ 4.1 ปัญหาของกรอบการวิเคราะห์ระดับเฟรม

นอกจากนี้ กรอบการวิเคราะห์ที่เป็นเฟรมนั้นครอบคลุมเสียงพูดเป็นระยะเวลาสั้นๆ ทำให้ไม่สามารถจับลักษณะที่กินระยะเวลานานได้ เช่น ความยาวของเสียงสระ และเสียงวรรณยุกต์ เป็นต้น ซึ่งเป็นปัจจัยที่มีผลต่อการรู้จำเสียงพูดทั้งสิ้น

4. *เสียงวรรณยุกต์ในภาษาไทย* เสียงวรรณยุกต์ในภาษาไทยมีความจำเป็นในการระบุเสียงพูดอยู่มาก เช่นในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย มีทั้งคำว่าคน (/kho0n^/) ข้น (/kho2n^/) และขน (/kho4n^/) ซึ่งถ้าทำการรู้จำเพียงหน่วยเสียง ทั้งสามคำนี้จะแสดงได้ด้วยหน่วยเสียงเดียวกัน คือ /kh/, /o/ และ /n^/ ซึ่งการที่จะรู้จำว่าแท้จริงแล้วเป็นคำว่าอะไรนั้น ก็ต้องอาศัยแบบจำลองทางภาษาต่อไป แต่ถ้าหากเรารู้จำเสียงวรรณยุกต์ด้วย ก็จะทำให้ผลลัพธ์ที่ได้มีความแม่นยำมากขึ้น หากแต่เสียง

วรรณยุกต์นั้นเป็นลักษณะที่กินระยะเวลานาน กรอบการวิเคราะห์ที่เป็นเฟรมจึงเล็กเกินไป ดังที่ได้กล่าวไว้แล้ว

5. *ความแปรผันในการออกเสียง* บางครั้งคำคำหนึ่งสามารถออกเสียงได้หลายแบบ ในระบบนี้จะกำหนดรูปแบบการออกเสียงของแต่ละคำไว้อย่างตายตัวด้วยพจนานุกรม เช่นรูปที่ 3.5 การรับมือกับความแปรผันในการออกเสียงได้ดีอาจจะต้องสร้างแบบจำลองทางการออกเสียงเพิ่มขึ้น

6. *การหาลักษณะสำคัญของเสียง* การหาลักษณะสำคัญของเสียงในที่นี้ใช้วิธีพีแอลพีเพียงอย่างเดียว จึงเป็นเรื่องที่น่าสนใจว่าถ้าใช้การหาลักษณะสำคัญของเสียงวิธีอื่นจะให้ความถูกต้องที่มากกว่าหรือไม่ นอกจากนี้ จากผลการทดลองพบว่าค่าอันดับของพีแอลพีที่ให้ความถูกต้องในการรู้จำสูงสุดนั้นแตกต่างกันไปตามฐานข้อมูลเสียงพูด จึงควรวิเคราะห์ต่อไปว่าอะไรเป็นปัจจัยกำหนด อาจจะเป็นความซับซ้อนของงาน หรือขึ้นอยู่กับอัตราการชักตัวอย่างหรือไม่

7. *ความสามารถของนิวรอลเน็ตเวิร์ก* นิวรอลเน็ตเวิร์กเป็นการเรียนรู้แบบแบ่งแยกที่สะดวกและมีประสิทธิภาพวิธีหนึ่ง แต่ก็ต้องตั้งค่าพารามิเตอร์ค่อนข้างมากในการเรียนรู้ เช่น จำนวนโหนดในชั้นซ่อน จำนวนรอบในการวนปรับค่าน้ำหนัก ค่าอัตราการเรียนรู้ และค่าโมเมนตัม ซึ่งในการทดลองต่างๆ จะคงค่าเหล่านี้ไว้ ซึ่งมีความเป็นไปได้ว่าถ้าลองปรับค่าพารามิเตอร์ดู อาจจะทำให้ประสิทธิภาพของระบบดีขึ้น นอกจากนี้ยังมีการเรียนรู้แบบแบ่งแยกอีกหลายวิธีที่น่าสนใจ เช่น ซัพพอร์ตเวกเตอร์แมชชีน เป็นต้น อย่างไรก็ดี วิธีการเหล่านี้ใช้สำหรับเรียนรู้ข้อมูลโดยทั่วไป ไม่ได้มีไว้สำหรับเรียนรู้ข้อมูลที่เป็นอนุกรมเวลาโดยเฉพาะ

8. *การรับมือกับข้อมูลจำนวนมาก* จากการทดลองพบว่าในฐานข้อมูลเสียงพูดที่มีข้อมูลเป็นจำนวนมาก เช่นฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย การใช้ทุกเฟรมสำหรับทำการเรียนรู้โดยใช้นิวรอลเน็ตเวิร์กนั้นแทบจะเป็นไปไม่ได้ในทางปฏิบัติ เนื่องจากกระบวนการนี้กินเวลานาน รวมทั้งจำนวนเฟรมของหน่วยเสียงต่างๆ ที่มีอยู่อย่างไม่สมดุลก็ทำให้การเรียนรู้ทำได้ไม่ดี ซึ่งในการทดลองนี้ได้เสนอการแก้ปัญหาขึ้นมาวิธีหนึ่ง คือใช้จำนวนเฟรมในการรู้จำให้น้อยลงโดยเลือกให้เฟรมของแต่ละหน่วยเสียงมีจำนวนเท่าๆ กัน เป็นที่น่าสนใจว่าถ้าเลือกเฟรมมาให้มากกว่านี้โดยที่ยังคงความสมดุลอยู่ อาจทำให้ได้การเรียนรู้ที่ดีขึ้น และเพิ่มความถูกต้องในการรู้จำต่อไป

รายการอ้างอิง

[1] T. M. Mitchell. Machine Learning. McGraw Hill, 1997.

[2] J. Tebelskis. Speech Recognition Using Neural Networks. Doctoral Dissertation Carnegie Mellon University, 1995.

[3] H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. Journal of Acoustic Society of America (1990) : 1738-1752.

[4] H. Hermansky and N. Morgan. Rasta Processing of Speech. IEEE Transactions on Speech and Audio Processing (1994) : 578-589.

[5] T. K. Vintsyuk. Element-wise Recognition of Continuous Speech Consisting of Words from a Specified Vocabulary. Kibernetika (Cybernetics) (1971) : 133-143.

[6] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics Speech and Signal Processing (1978) : 43-49.

[7] L. R. Rabiner and B.-H. Juang. Fundamental of Speech Recognition. Prentice Hall, 1993.

[8] J. Baker. The DRAGON System – An Overview. IEEE Transactions on Acoustics, Speech, and Signal Processing (1975) : 24-29.

[9] F. Jelinek, R. L. Mercer and L. R. Bahl. Continuous Recognition by Statistical Methods. Proceedings of IEEE (1975) : 250-256.

[10] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of IEEE (1989) : 257-286.

[11] R. O. Duda, P. E. Hart and D. G. Stork. Pattern Classification. Wiley, 1991.

[12] C. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 1997.

[13] N. J. Nilsson. Learning Machines. McGraw Hill, 1965.

[14] E. B. Baum. On the Capabilities of Multilayer Perceptrons. Journal of Complexity 1988 : (193-215).

[15] K. Funahashi. On the Approximate Realization of Continuous Mappings by Neural Networks. Neural Networks (1989) : 183-192.

[16] G. Cybenko. Approximation by Superpositions of Sigmoidal Function. Mathematics of Control, Signals and Systems (1989) : 304-314.

[17] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks Are Universal Approximators. Neural Networks (1989) : 359-366.

[18] D. Rumelhart, G. Hinton, and R. Williams. Learning Internal Representations by Error Propagation. in D. Rumelhart and J. McClel-land, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 318-362. MIT Press, 1986.

[19] F. Jelinek. Statistical Methods for Speech Recognition. MIT Press, 1997.

[20] S. J. Young, N. H. Russell and J. H. S. Thornton. Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems. Technical Report CUED/F-INFENG/Tr. 38 University of Cambridge, 1989.

[21] W. M. Huang and R. Lippmann. Neural Net and Traditional Classifers. Neural Information Processing Systems (1988) : 387-396.

[22] J. Elman and D. Zipser. Learning the Hidden Structure of Speech. ICS Report 8701 Institute for Cognitive Science University of California San Diego, 1986.

[23] B. R. Kammerer and W. A. Kupper. Experiments for Isolated Word Recognition with Single and Two Layer Perceptrons. Neural Networks (1990) : 693-706.

[24] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang. Phoneme Recognition using Time-Delay Neural Networks. IEEE Transactions on Acoustics Speech and Signal Processing (1989) : 328-339.

[25] M. Jordan. Serial Order: A Parallel Distributed Processing Approach. ICS Report 8604 Institute for Cognitive Science University of California San Diego, 1986.

[26] J. Elman. Finding Structure in Time. Cognitive Science (1990) :179-211.

[27] R. Watrous. Speech Recognition using Connectionist Networks. Doctoral Dissertation University of Pensylvania, 1988.

[28] A. J. Robinson and F. Fallside. Static and Dynamic Error Propagation Networks with Application to Speech Coding. Neural Information Processing Systems (1988) : 632-641.

[29] N. Morgan and H. Bourlard. An Introduction to Hybrid HMM/Connectionist Continuous Speech Recognition. IEEE Signal Processing Magazine (1995) : 24-42.

[30] E. Trentin and M. Gori. A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition. Neurocomputing (2001) : 91-126.

[31] R. P. Lippmann and B. Gold. Neural-Net Classifiers Useful for Speech Recognition. Proceedings of IEEE International Conference on Neural Networks (1987) : 417-422.

[32] H. Bourlard and C. Wellekens. Links between Markov Models and Multilayer Perceptrons. IEEE Transactions on Pattern Analysis and Machine Intelligence (1990) : 1167-1178.

[33] H. Bourlard and N. Morgan. Neural Networks for Statistical Recognition of Continuous Speech. Proceedings of the IEEE (1995) : 741-770.

[34] A. J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. IEEE Transaction on Neural Networks (1994) : 298-305.

[35] M. M. Hochberg, S. J. Renals, A. J. Robinson and G. D. Cook. Recent Improvements to the ABBOT Large Vocabulary CSR System. Proceedings of International Conference on Acoustics, Speech and Signal Processing (1995) : 69-72.

[36] Y. Bengio, R. De Mori, G. Flammia and R. Kompe. Global Optimization of a Neural Netwok-Hidden Markov Model Hybrid. IEEE Transactions on Neural Networks (1992) : 252-259.

[37] G. Rigoll. Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition System. IEEE Transactions on Speech and Audio Processing (1994) : 175-184.

[38] P. Le Cerf, W. Ma and D. Van Compernolle. Multilayer Perceptrons as Labelers for Hidden Markov Models. IEEE Transactions on Speech and Audio Processing (1994) : 185-193.

[39] G. Zavaliagkos, Y. Zhao, R. Schwartz and J. Makhoul. A Hybrid Segmental Neural Nets/Hidden Markov Model System for Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing (1994) : 151-160.

[40] P. Pungprasertying and B. Kijsirikul. An Automatic Dialing System using Speech of Thai Names. Proceedings of the 7th National Computer Science and Engineering Conference (2004) : 24-29.

[41] N. Thubthong and B. Kijsirikul. An Empirical Study for Constructing Thai Tone Models. Proceedings of the 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop (2002) : 179-186.

[42] M. Ostendorf and S. Roukos. A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition. IEEE Transactions on Acoustics Speech and Signal Processing (1989) : 1857-1869.

[43] J. Glass. A Probabilistic Framework for Segment-based Speech Recognition. Computer, Speech and Language (2003) : 137-152.

[44] The International Computer Science Institute. http://www.icsi.berkeley.edu.

[45] N. Storm. http://www.speech.kth.se/NICO.

[46] S. Young, J. Jansen, J. Odell, D. Ollasen and P. Woodland. The HTK Book (Version 2.0). Entropic Cambridge Research Labolatory University of Cambridge, 1995.

[47] P. Ladefoged. A Course in Phonetics. Harcourt Brace Jovanovich Inc., 1975.

[48] International Phonetic Association. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press, 1999.

[49] พระยาอุปกิตศิลปสาร. หลักภาษาไทย. ไทยวัฒนาพานิช, 2533.

[50] J. Gandour. Aspects of Thai Tone. Doctoral Dissertation University of California at Los Angeles, 1976.

[51] J. L. Flanagan. Speech Analysis Synthesis and Perception. Springer-Verlag, 1972.

[52] I. H. Witten. Principles of Computer Speech. Academic Press Inc., 1982.

ภาคผนวก ก
ธรรมชาติของเสียงพูด

ธรรมชาติของเสียงพูดเป็นศาสตร์หนึ่งที่มนุษย์สนใจศึกษามาโดยตลอด โดยในระยะแรกเป็นการทำความเข้าใจเกี่ยวกับรูปแบบและการเคลื่อนไหวของอวัยวะการออกเสียงที่ส่งผลให้เกิดเสียงต่างๆ ทำให้สามารถจำแนกเสียงพูดได้อย่างมีหลักเกณฑ์ ตามรูปแบบและการเคลื่อนไหวของอวัยวะการออกเสียงที่ก่อเป็นเสียงนั้น สาขาวิชาที่สนใจศึกษาเสียงพูดจากเหตุปัจจัยการออกเสียงนี้เรียกว่า*สัทศาสตร์* ซึ่งตั้งต้นจากปัญหาที่ว่า เสียงพูดมาจากไหน และทำไมเสียงพูดจึงต่างกัน

ต่อมา เมื่อความก้าวหน้าทางด้านวิทยาศาสตร์มีมากขึ้น ทำให้ได้รู้ว่าเสียงพูดนั้นเป็นคลื่นตามยาวอันเกิดจากการสั่นสะเทือนของมวลอากาศ จึงเกิดความสนใจศึกษาลักษณะของคลื่นเสียงที่เกิดจากเสียงพูดต่างๆ ก่อให้เกิดสาขาวิชาที่เรียกว่า*สวนศาสตร์ของเสียงพูด*ขึ้นมา เพื่อตอบปัญหาที่ว่า เสียงพูดเป็นอย่างไร และลักษณะใดที่ทำให้เสียงพูดต่างกัน

สุดท้าย การเติบโตของเทคโนโลยีทางด้านปัญญาประดิษฐ์ และการเรียนรู้ของเครื่อง ทำให้เกิดความตระหนักว่าการสร้างเครื่องจักรให้มีความสามารถในด้านต่างๆ เทียมเท่ามนุษย์นั้นมีความเป็นไปได้ รวมทั้งความสามารถในการรับฟัง สาขาวิชา*การรู้จำเสียงพูด*จึงอุบัติขึ้นมา เสียงพูดต่างๆ จะถูกจำแนกได้อย่างไร นี่เป็นปัญหาที่สาขาวิชาการรู้จำเสียงพูดสนใจ

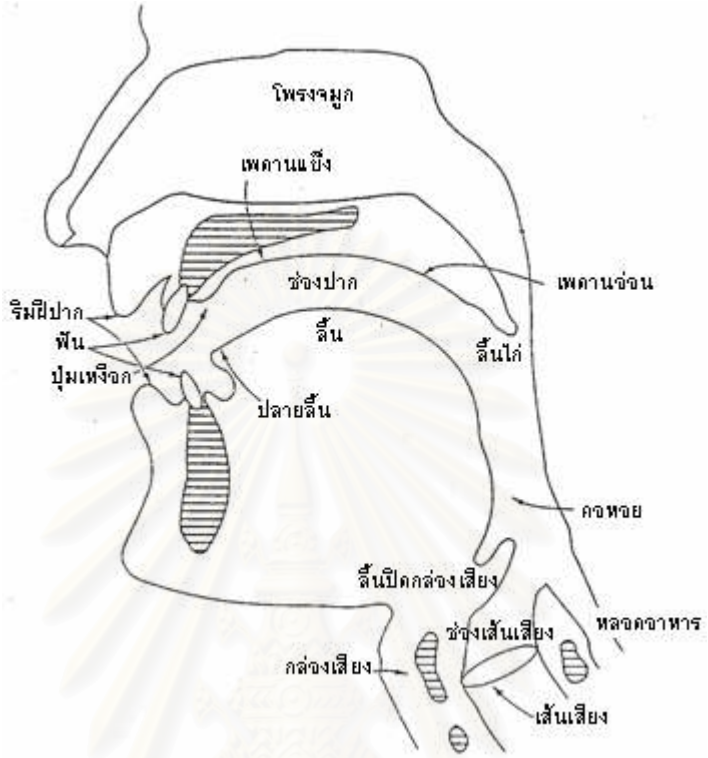
จากโครงสร้างของการสื่อสารด้วยเสียงพูด อาจกล่าวได้ว่า สัทศาสตร์มองกระบวนการนี้ในแง่ของผู้ส่งสาร ขณะที่สวนศาสตร์ของเสียงพูดและการรู้จำเสียงพูดมองกระบวนการนี้ในแง่ของสื่อและผู้รับสาร ตามลำดับ โดยรายละเอียดในส่วนของการรู้จำเสียงพูดได้ถูกอธิบายไว้ในบทก่อนหน้าแล้ว ในบทนี้จะเป็นการมองย้อนกลับไปถึงส่วนของสัทศาสตร์และสวนศาสตร์ของเสียงพูด อันเป็นความรู้พื้นฐานในการรู้จำเสียงพูดนั่นเอง

ก.1 สัทศาสตร์

ก.1.1 อวัยวะการออกเสียง

ด้วยระบบทางเดินหายใจ รวมถึงสรีระของช่องคอ ช่องปาก และโพรงจมูก ที่เหมาะสม ทำให้มนุษย์สามารถสร้างเสียงพูดขึ้นมาได้ ในการพูดนั้น แรกเริ่ม ลมที่ออกจากปอดผ่านหลอดลมจะเข้าสู่กล่องเสียง ถ้าเส้นเสียงมีอาการสั่น ก็จะทำให้เกิดความถี่ของคลื่นเสียงตามจังหวะการสั่นนั้น เป็นเสียงก้องคล้ายเสียงดนตรี แต่ถ้าเส้นเสียงไม่มีอาการสั่น ลมนั้นก็จะผ่านต่อไป ทำให้เกิดเสียงไม่ก้องคล้ายเสียงรบกวน

จากการสั่นของเส้นเสียงจะมาสู่การเปลี่ยนรูปร่างของอวัยวะในช่องปาก ณ ที่นี้เอง ที่ทำให้เกิดเสียงต่างๆ ได้หลากหลาย และนำมาใช้เพื่อการสื่อสารได้ โดยอวัยวะการออกเสียงทั้งหมดของมนุษย์สามารถแสดงได้ดังรูปผนวกที่ ก.1 [47]



รูปผนวกที่ ก.1 อวัยวะการออกเสียง

โดยแต่ละอวัยวะที่มีผลต่อการก่อให้เกิดเสียงพูด ไล่เรียงมาตามลำดับได้ดังนี้

1. *หลอดลม* เป็นทางเดินอากาศจากปอดถึงกล่องเสียง
2. *กล่องเสียง* ตั้งอยู่ตอนบนของหลอดลมตรงตำแหน่งที่เรียกว่าลูกกระเดือก ประกอบด้วยกระดูกอ่อนหลายส่วนด้วยกัน ส่วนที่อยู่ด้านหน้า คือ กระดูกอ่อนไทรอยด์ ซึ่งมีปลายด้านหนึ่งของเส้นเสียงทั้งสองเชื่อมอยู่ชิดกัน ส่วนปลายอีกด้านหนึ่งของเส้นเสียงทั้งสองจะเชื่อมอยู่กับกระดูกอ่อนอาริตินอยด์ ซึ่งเป็นกระดูกอ่อนอีกสองชิ้น กระดูกอ่อนอาริตินอยด์และกล้ามเนื้อในกล่องเสียงจะทำให้เส้นเสียงทั้งสองอยู่ชิดติดกันหรือห่างจากกันได้
3. *เส้นเสียง* เป็นอวัยวะสำคัญที่ทำให้เกิดเสียง เส้นเสียงประกอบด้วยเส้นเอ็นและกล้ามเนื้อเป็นแผ่น 2 แผ่น มีความยาวประมาณ 1.2-1.7 เซนติเมตร กว้างประมาณ 0.2-0.3 เซนติเมตร ปิดขวางอยู่ตรงปากของช่องหลอดลม โดยจะวางตัวจากด้านหลังมายังด้านหน้าอยู่ตรงกลางของกล่องเสียง เส้นเสียงทั้งสองสามารถที่จะดึงออกให้ห่างจากกัน หรือดึงเข้ามาให้ชิดกันก็ได้

4. *ช่องเส้นเสียง* จะเปิดอยู่ระหว่างที่หายใจเข้าออกตามปกติ แต่จะปิดลงเมื่อมีการเปล่งเสียง ก่อให้เกิดการสั่น และเป็นเสียงดังขึ้น

5. *ลิ้นปิดกล่องเสียง* เป็นแผ่นเนื้อเล็กๆ อยู่ต่อโคนลิ้นลงไปในคอ มีหน้าที่ปิดช่องหลอดลม เพื่อป้องกันมิให้มีอะไรตกลงไปในหลอดลมในเวลาที่กลืนอาหาร แต่เมื่อมีการพูด แผ่นเนื้อนี้จะเปิดขึ้นเพื่อให้ลมออกมา

6. *คอหอย* เป็นโพรงซึ่งอยู่ถัดปากลงไปจากช่องปากจนถึงเส้นเสียง

7. *ลิ้นไก่* เป็นก้อนเนื้อเล็กๆ อยู่ต่อจากปลายเพดานอ่อนเข้าไปข้างใน และห้อยอยู่ตรงกลางปาก สามารถสั่นรัวได้ เวลาอ้าปากมักจะเห็น ลิ้นไก่ใช้ออกเสียงในบางภาษา เช่น ภาษาเยอรมัน ฝรั่งเศส นอร์เวย์ อาหรับ อิสราเอล เป็นต้น

8. *ช่องปาก* ทำหน้าที่เป็นช่องกำทอน ซึ่งสามารถเปลี่ยนให้มีรูปร่างต่างๆ กัน ตามรูปร่างของอวัยวะภายในช่องปาก โดยอวัยวะภายในช่องปากอาจสามารถแบ่งได้เป็น

    8.1 *อวัยวะส่วนกระทำอาการ* หรือ กรณ์ หมายถึงอวัยวะส่วนที่เคลื่อนไหวเพื่อผลักหรือกักลมในที่ต่างๆ อวัยวะส่วนกระทำอาการที่สำคัญคือลิ้น ซึ่งเคลื่อนไหวได้มากที่สุด

    8.2 *อวัยวะส่วนเกิดอาการ* หรือ ฐาน หมายถึง ตำแหน่งที่อวัยวะส่วนกระทำอาการเคลื่อนไหวไป เพื่อผลักหรือกักลมไว้ ฐานภายในช่องปากที่สำคัญ ได้แก่ ริมฝีปาก ฟัน ปุ่มเหงือก เพดานแข็ง และเพดานอ่อน

9. *โพรงจมูก* คือโพรงอยู่เหนือลิ้นไก่ขึ้นไป ลมซึ่งผ่านเส้นเสียงขึ้นมาจะผ่านออกไปทางจมูกได้เมื่อเวลาหายใจและเวลาออกเสียงนาสิก ในเวลาเปล่งเสียงอื่นๆ ลิ้นไก่จะถูกยกขึ้นไปปิดโพรงจมูกเพื่อให้ลมออกมาทางช่องปาก

10. *ลิ้น* เป็นอวัยวะภายในช่องปากที่เคลื่อนไหวได้มากที่สุดสำหรับการออกเสียงพูด แต่ละส่วนที่เคลื่อนไหวของลิ้นล้วนมีผลต่อการออกเสียง จึงมีการแบ่งลิ้นออกเป็น 3 ส่วน ด้วยกัน คือ ปลายลิ้น หน้าลิ้น และหลังลิ้น

11. *เพดานอ่อน* คือส่วนที่อยู่ต่อเพดานแข็งเข้าไปข้างใน มีลักษณะเป็นกระดูกอ่อนที่ขยับขึ้นลงได้เล็กน้อย เวลาหายใจเพดานอ่อนและลิ้นไก่ซึ่งอยู่ปลายเพดานอ่อนจะลดระดับลงมาเปิดช่องให้ลมออกไปทางจมูก ฉะนั้นเวลาที่เราไม่พูด เพดานอ่อนและลิ้นไก่จะลดระดับลงมา เวลาพูดส่วนใหญ่เพดานอ่อนและลิ้นไก่จะถูกยกขึ้นไปจดกับผนังคอ จะมีแต่เวลาออกเสียงนาสิกเท่านั้นที่เพดานอ่อนจะลดระดับลงมาเพื่อให้ลมออกไปทางจมูกได้ เมื่อลิ้นแตะหรือวางใกล้เพดานอ่อนจะทำให้เกิดเสียงมุทธชะ

12. *เพดานแข็ง* หมายถึงส่วนโค้งของเพดานปากส่วนที่เป็นกระดูกแข็ง ซึ่งอยู่ถัดจากปุ่มเหงือกเข้ามา เมื่อลิ้นแตะหรือวางใกล้เพดานแข็งจะทำให้เกิดเสียงตาลุชะ

13. *ปุ่มเหงือก* เป็นส่วนที่นูนออกมาตรงบริเวณหลังฟันด้านบน มีลักษณะนูนเว้าเป็นคลื่น เมื่อลิ้นแตะหรือวางใกล้ปุ่มเหงือกทำให้เกิดเสียงมุทธชะ

14. *ฟัน* เป็นอวัยวะที่เป็นฐานหรือตำแหน่งที่เกิดของเสียงหลายชนิด เช่น เมื่อฟันบนกดลงบนริมฝีปากล่าง ลมที่ผ่านออกมาโดยแรงจะลอดช่องที่พอจะผ่านได้ออกมา ทำให้เกิดเป็นเสียงชนิดที่เรียกว่า เสียงเสียดแทรกที่เกิดระหว่างฟันกับริมฝีปาก ถ้าฟันบนกดกับฟันล่าง ลมที่ผ่านออกมาโดยแรงจะทำให้ได้เสียงเสียดแทรกที่เกิดที่ฟัน เป็นต้น นอกจากนี้ เนื่องจากปลายลิ้นอยู่ใกล้กับฟัน ปลายลิ้นจึงมักจะทำอาการต่างๆ บริเวณฟันและหลังฟันบ่อยๆ ทำให้เกิดเสียงทันตชะ

15. *ริมฝีปาก* เป็นอวัยวะส่วนที่สามารถเคลื่อนไหวได้มาก และทำให้เสียงแตกต่างกันได้มาก เราอาจจะบังคับริมฝีปากให้ปิดสนิท ให้เปิดเล็กน้อย ให้เปิดกว้างขึ้น ให้ยื่นออกมา ให้ห่อกลม หรือทำเป็นรูปรีก็ได้ ลักษณะต่างๆ ของริมฝีปากล้วนมีผลต่อการออกเสียง และทำให้เสียงแตกต่างกันไป เสียงพยัญชนะที่เกิดจากการผลักหรือกักที่ริมฝีปากเรียกว่าเสียงโอฐชะ

ก.1.2 เสียงพยัญชนะ

เสียงพยัญชนะหมายถึงเสียงของลมที่ผ่านปอดขึ้นมายังกล่องเสียงแล้วปะทะกับอวัยวะต่างๆ ในช่องปาก ทำให้ลมเพียงส่วนหนึ่งหรือทั้งหมดพบกับอุปสรรคที่อยู่เหนือช่องของเส้นเสียง โดยอุปสรรคเหล่านี้เกิดจากการทำงานประสานกันของอวัยวะในช่องปาก เสียงพยัญชนะที่เกิดขึ้นมาจึงมีหลายแบบแตกต่างกัน ซึ่งเสียงที่แตกต่างกันมักจะทำให้ความหมายในภาษาแตกต่างกันไปด้วย คุณสมบัติที่ทำให้เสียงพยัญชนะแตกต่างกันมีดังนี้

1. *ความก้องของเสียง* เป็นคุณสมบัติที่ใช้ในการแบ่งแยกเสียงพยัญชนะออกได้เป็นสองชนิด คือ

    1.1 *เสียงพยัญชนะก้อง* หรือเสียงโฆษะ เป็นเสียงพยัญชนะที่เส้นเสียงสั่นสะเทือนขณะที่เปล่งเสียง

    1.2 *เสียงพยัญชนะไม่ก้อง* หรือเสียงอโฆษะ เป็นเสียงพยัญชนะที่เส้นเสียงไม่สั่นสะเทือนขณะที่เปล่งเสียง

2. *ลักษณะของลมที่ผ่านอวัยวะการออกเสียง* เป็นคุณสมบัติที่ใช้ในการแบ่งแยกเสียงพยัญชนะออกได้ดังนี้

    2.1 *เสียงพยัญชนะหยุด* อาจแบ่งออกเป็น 2 ลักษณะย่อยๆ ได้แก่ เสียงพยัญชนะผลัก และเสียงพยัญชนะกัก เสียงพยัญชนะผลักเกิดจากการที่ลมซึ่งเปล่งออกมาถูกกักเอาไว้ ณ ที่ใดที่หนึ่งในช่องปาก แล้วช่องที่กักนั้นเปิดให้ลมพุ่งออกมา นอกจากนี้เสียงพยัญชนะผลักอาจแบ่งออกได้อีกเป็น เสียงพยัญชนะผลักมีลม หรือธนิต ซึ่งจะมีลมหายใจพ่นออกมาหลังจากเปล่งเสียง และเสียงพยัญชนะผลักไม่มีลม หรือสิถิล ซึ่งไม่มีลมหายใจพ่นออกมาหลังจากเปล่งเสียง ส่วนเสียงพยัญชนะกักเกิดจากการที่ลมซึ่งเปล่งออกมาถูกกักเอาไว้ ณ ที่ใดที่หนึ่งในช่องปาก แต่ไม่ได้ถูกปล่อยให้พุ่งออกมา โดยเสียงพยัญชนะกักนี้มักจะเป็นเสียงตัวสะกดท้ายพยางค์

    2.2 *เสียงพยัญชนะเสียดแทรก* เป็นเสียงพยัญชนะที่เมื่อออกเสียงแล้วลมที่ผ่านขึ้นมาถูกบังคับให้ต้องบีบตัวผ่านช่องแคบๆ ที่ใดที่หนึ่งในช่องปาก ซึ่งเสียงเสียดแทรกนี้เราจะทำค้างไว้นานเท่าใดก็ได้ ตราบเท่าที่ลมหายใจจะอำนวย

    2.3 *เสียงพยัญชนะนาสิก* เป็นเสียงพยัญชนะที่มีลมผ่านออกมาทางจมูก ซึ่งเกิดจากการที่ลมมากักอยู่ในช่องปาก แล้วเพดานอ่อนและลิ้นไก่ลดระดับลง ทำให้เกิดเสียงที่ขึ้นจมูกออกมา

    2.4 *เสียงพยัญชนะข้างลิ้น* เป็นเสียงที่เกิดจากการนำลิ้นปิดบริเวณปุ่มเหงือกและเพดานแข็งส่วนกลางไว้ แล้วปล่อยให้ลมผ่านออกมาทางข้างลิ้น

    2.5 *เสียงพยัญชนะรัว* เกิดจากการที่อวัยวะส่วนใดส่วนหนึ่งในช่องปากกระทบกับอวัยวะอีกส่วนหนึ่งในขณะที่ลมถูกพ่นผ่านอวัยวะนั้นออกมาอย่างรุนแรง ทำให้เกิดเสียงรัวขึ้น

    2.6 *เสียงพยัญชนะกึ่งสระ* หรืออรรธสระ เป็นเสียงเลื่อนที่เกิดขึ้นระหว่างเสียงสระสองเสียง ในการเปล่งเสียงพยัญชนะกึ่งสระ อวัยวะที่ใช้ในการออกเสียงจะอยู่ในตำแหน่งของการออกเสียงสระใดสระหนึ่งก่อน แล้วจึงเปล่งเสียงออกมาขณะที่เปลี่ยนตำแหน่งอวัยวะไปสู่การออกเสียงของอีกสระหนึ่ง

3. *ฐานที่เกิดของเสียง* ไม่ว่าลมที่ใช้ในการออกเสียงพยัญชนะนั้นจะมาถูกผลัก กัก หรือการเสียดแทรก จำเป็นต้องมีฐานที่เกิดอยู่ด้วยเสมอในช่องปาก โดยอาจจะเป็นที่เพดานอ่อน เพดานแข็ง ปุ่มเหงือก ฟัน หรือริมฝีปาก ก็ได้

ก.1.3 เสียงสระ

เสียงสระเป็นเสียงซึ่งถูกเปล่งผ่านออกมาทางช่องปากหรือโพรงจมูกโดยไม่มีอวัยวะส่วนใดในปากมาเป็นอุปสรรคปิดกั้นทางลมไว้เลย เสียงสระเกิดจากการที่ลมผ่านเส้นเสียงในตำแหน่งที่เส้นเสียงทั้งสองอยู่ชิดกันมากจนเกือบปิดสนิท ทำให้ลมต้องดันตัวออกมาอย่างรุนแรงจนเส้นเสียงเกิดการสั่นสะเทือน และส่งผลทำให้เกิดเสียงดังที่เป็นเสียงก้อง โดยคุณสมบัติที่ทำให้เสียงสระมีความแตกต่างกันมีดังนี้

1. *ส่วนของลิ้นที่ใช้ออกเสียง* จากการศึกษาภาพถ่ายเอ็กซเรย์ช่องปากมนุษย์ในขณะที่ออกเสียงสระต่างๆ พบว่ามีลิ้นหลายส่วนที่ใช้ในการออกเสียงสระ ไม่ว่าจะเป็นลิ้นส่วนหน้า ลิ้นส่วนกลาง หรือลิ้นส่วนหลัง โดยลิ้นส่วนนั้นๆ จะยกขึ้นใกล้เพดานปากในขณะที่ออกเสียงสระหนึ่งๆ ก่อให้เกิดเสียงสระที่แตกต่างกัน โดยถ้าลิ้นส่วนหน้ายกขึ้นให้จุดสูงสุดอยู่ใกล้เพดานแข็ง เราก็จะเรียกเสียงสระนั้นว่าเสียงสระส่วนเพดานแข็ง หรือสระหน้า เช่น สระอี สระเอ สระแอ เป็นต้น แต่ถ้าการออกสียงสระใดใช้ลิ้นส่วนหลัง โดยทำการยกลิ้นส่วนหลังขึ้นให้จุดสูงสุดอยู่ใกล้เพดานอ่อน เราก็จะเรียกเสียงสระนั้นว่าเป็นเสียงสระส่วนเพดานอ่อน หรือสระหลัง เช่น สระอู สระโอ สระออ เป็นต้น ส่วนถ้าในการออกเสียงสระใดลิ้นส่วนกลางถูกยกขึ้นไปยังส่วนกลางของเพดานปาก เราก็จะเรียกเสียงสระนั้นว่า สระกลาง เช่น สระอือ สระเออ สระอา เป็นต้น
2. *ระยะห่างระหว่างลิ้นและเพดานปาก* หรือความสูงของลิ้น เป็นลักษณะที่สำคัญอย่างหนึ่งในการแบ่งชนิดของเสียงสระ โดยระยะห่างนี้จะเป็นตัวกำหนดว่าเสียงสระที่เปล่งออกมาเป็นสระเปิดหรือสระปิด ถ้าหากลิ้นอยู่ห่างจากเพดานปากมาก หรือลิ้นอยู่ในระดับต่ำ ทำให้ช่องโพรงปากกว้าง ลมก็จะผ่านออกมาได้มาก เสียงสระที่ได้จะเป็นสระเปิด เช่น สระอา ในทางตรงกันข้าม ถ้าตำแหน่งของลิ้นอยู่ใกล้กับเพดานปากมาก หรือลิ้นอยู่ในระดับสูง ช่องโพรงในปากก็จะแคบ ทำให้ลมผ่านออกมาได้น้อย เสียงสระที่ได้จะเป็นสระปิด เช่น สระอี สระอู เป็นต้น แต่ถ้าระยะห่างระหว่างลิ้นกับเพดานปากอยู่ในระหว่างสระเปิดและสระปิด เช่นเสียงสระที่เปิดกว้างกว่าสระปิดเล็กน้อย เราก็จะเรียกว่าเป็นสระกลางปิด หรือสระกึ่งปิด เช่น สระเอ สระโอ เป็นต้น แต่ถ้าเปิดกว้างขึ้นอีก จะเรียกว่าเป็นสระกลางเปิด หรือสระกึ่งเปิด เช่น สระ แอ สระออ เป็นต้น
3. *การห่อริมฝีปาก* หมายถึงการที่ริมฝีปากทั้งสองเคลื่อนไหวโดยยื่นตัวไปข้างหน้า แล้วห่อกลมมากน้อยเพียงใด ถ้าริมฝีปากยื่นออกไปข้างหน้าแล้วห่อกลมมาก เสียงสระที่ได้จะเรียกว่าสระกลม เช่น สระอู สระโอ สระออ เป็นต้น แต่ถ้าริมปากทั้งสองฉีกออก

หรือไม่ห่อกลมขณะเปล่งเสียง สระที่ได้ก็จะเป็นสระไม่กลม เช่น สระอี สระเอ สระแอ สระอา เป็นต้น

4. *ลักษณะนาสิก* เป็นลักษณะในการออกเสียงสระที่ทำให้เกิดเสียงสระขึ้นจมูกหรือสระนาสิกขึ้น ซึ่งจะทำให้เสียงแตกต่างจากสระโอฐชะ กล่าวคือ ในการเปล่งเสียงสระโอฐชะนั้น เพดานอ่อนจะยกขึ้นปิดโพรงจมูก อากาศจึงไม่สามารถผ่านออกไปทางโพรงจมูกได้ แต่ออกมาทางปากทั้งหมด สำหรับสระนาสิกนั้น เพดานอ่อนจะลดต่ำลง และปล่อยให้อากาศผ่านออกทางโพรงจมูกด้วยในเวลาเดียวกัน เช่นในภาษาฝรั่งเศสจะมีหน่วยเสียงนาสิกอยู่ 4 หน่วยเสียงด้วยกัน แต่สำหรับภาษาไทย ตามปกติแล้วไม่มีการออกเสียงสระนาสิก แต่ในบางครั้งก็อาจได้รับอิทธิพลจากการเปล่งเสียงพยัญชนะนาสิกที่อยู่ใกล้เคียง เช่น คำว่า นั้น เป็นต้น

5. *ความยาวในการออกเสียง* ความสั้นยาวของการออกเสียงนั้นมีความสำคัญมากในภาษาไทย เพราะหน่วยเสียงที่ใช้ความยาวในการออกเสียงต่างกันจะทำให้ความหมายของพยางค์แตกต่างกันได้ เช่นคำว่า ขุด และคำว่า ขูด โดยสระจะถูกแบ่งออกเป็นสองประเภทตามความสั้นยาวของสระ คือ สระเสียงสั้น หรือรัสสระ และสระเสียงยาว หรือฑีฆสระ

ก.1.4 เสียงวรรณยุกต์

เสียงวรรณยุกต์นั้นคือเสียงสูงต่ำในภาษา ซึ่งเกิดจากการสั่นสะเทือนของเส้นเสียงในอัตราความถี่ที่ต่างกันไป ดังนั้นเสียงวรรณยุกต์จะปรากฏอยู่ในส่วนของเสียงสระ เพราะเสียงสระเป็นเสียงที่เกิดจากการสั่นของเส้นเสียง นอกจากนี้ยังอาจมีเสียงวรรณยุกต์ปรากฏอยู่บ้างในบางส่วนของเสียงพยัญชนะ แต่จะต้องเป็นส่วนของเสียงพยัญชนะที่เป็นเสียงก้องหรือพยัญชนะนาสิกเท่านั้น เพราะเสียงพยัญชนะไม่ก้องนั้นไม่ได้เกิดจากการสั่นของเส้นเสียง จึงไม่สามารถมีเสียงวรรณยุกต์อยู่ด้วยได้

ก.2 สัทศาสตร์ภาษาไทย

ก.2.1 เสียงพยัญชนะภาษาไทย

พยัญชนะในภาษาไทยมีทั้งหมด 44 รูป 21 หน่วยเสียง แบ่งเป็น 2 กลุ่มใหญ่ๆ คือ กลุ่มพยัญชนะหยุด 11 หน่วยเสียง และกลุ่มที่ไม่ใช่พยัญชนะหยุด 10 หน่วยเสียง ดังแสดงในตารางผนวก ก.1 ทั้งนี้หน่วยเสียงพยัญชนะทั้ง 21 หน่วยเสียง สามารถที่จะอยู่ในต้นพยางค์ได้ทุกหน่วยเสียง แต่จะมีหน่วยเสียงพยัญชนะที่ปรากฏท้ายพยางค์ได้เพียง 9 หน่วยเสียงเท่านั้น คือ เสียงพยัญชนะหยุด 4 หน่วยเสียง คือ /p/, /t/, /k/, /ʔ/ เสียงพยัญชนะนาสิก 3 หน่วยเสียง คือ /m/, /n/,

/ŋ/ และเสียงพยัญชนะกึ่งสระ 2 หน่วยเสียง คือ /w/, /j/ ส่วนพยัญชนะควบกล้ำในภาษาไทยแท้เป็นได้ 11 หน่วยเสียง คือ /pr/, /p^{h}r/, /pl/, /p^{h}l/, /tr/, /kr/, /k^{h}r/, /kl/, /k^{h}l/, /kw/, /k^{h}w/ ส่วนพยัญชนะควบกล้ำในภาษาไทยทับศัพท์อังกฤษมีได้ 6 หน่วยเสียง คือ /br/, /bl/, /dr/, /fr/, /fl/, /t^{h}r/ ส่วนคำไทยที่ยืมมาจากภาษาสันสกฤตก็ควบ /t^{h}r/ ได้เช่นกัน

ตารางผนวก ก.1 เสียงพยัญชนะภาษาไทย

หน่วยเสียง	หน่วยเสียงควบกล้ำ	ลักษณะของลม	การพ่นลม	ความก้อง	ฐานที่เกิด	รูปพยัญชนะ
พยัญชนะหยุด						
/p/[1]	/pr/, /pl/	กัก	ไม่พ่นลม	ไม่ก้อง	ริมฝีปาก	ป
/p^{h}/	/p^{h}r/, /p^{h}l/	กัก	พ่นลม	ไม่ก้อง	ริมฝีปาก	ผ พ ภ
/b/	/br/[2], /bl/[2]	กัก	ไม่พ่นลม	ก้อง	ริมฝีปาก	บ
/t/[1]	/tr/	กัก	ไม่พ่นลม	ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ฏ ต
/t^{h}/	/t^{h}r/[3]	กัก	พ่นลม	ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ฐ ฑ ฒ ถ ท ธ
/d/	/dr/[2]	กัก	ไม่พ่นลม	ก้อง	ฟัน หรือ ปุ่มเหงือก	ฎ ฑ ด
/c/		กัก	ไม่พ่นลม	ไม่ก้อง	เพดานแข็ง	จ
/c^{h}/		กัก	พ่นลม	ไม่ก้อง	เพดานแข็ง	ฉ ช ฌ
/k/[1]	/kr/, /kl/, /kw/	กัก	ไม่พ่นลม	ไม่ก้อง	เพดานอ่อน	ก
/k^{h}/	/k^{h}r/, /k^{h}l/, /k^{h}w/	กัก	พ่นลม	ไม่ก้อง	เพดานอ่อน	ข ฃ ค ฅ ฆ
/ʔ/[1]		กัก	ไม่พ่นลม	ไม่ก้อง	เส้นเสียง	อ
พยัญชนะนาสิก						
/m/[1]		นาสิก		ก้อง	ริมฝีปาก	ม
/n/[1]		นาสิก		ก้อง	ฟัน หรือ ปุ่มเหงือก	ณ น
/ŋ/[1]		นาสิก		ก้อง	เพดานอ่อน	ง
พยัญชนะเสียดแทรก						
/f/	/fr/[2], /fl/[2]	เสียดแทรก		ไม่ก้อง	ริมฝีปาก	ฝ ฟ
/s/		เสียดแทรก		ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ซ ศ ษ ส
/h/		เสียดแทรก		ไม่ก้อง	เส้นเสียง	ห ฮ
พยัญชนะรัว						
/r/		รัว		ก้อง	ฟัน หรือ ปุ่มเหงือก	ร
พยัญชนะข้างลิ้น						
/l/		ข้างลิ้น		ก้อง	ฟัน หรือ ปุ่มเหงือก	ล ฬ
พยัญชนะกึ่งสระ						
/w/[1]		กึ่งสระ		ก้อง	ริมฝีปาก – เพดานอ่อน	ว
/j/[1]		กึ่งสระ		ก้อง	เพดานแข็ง	ญ ย

หมายเหตุ

[1] ปรากฏท้ายพยางค์ได้

[2] ปรากฏเฉพาะในคำไทยทัพศัพท์อังกฤษ

[3] ปรากฏในคำไทยทับศัพท์อังกฤษ หรือคำไทยที่ยืมมาจากภาษาสันสกฤต

ใช้สัญลักษณ์ตามสัทอักษรสากล [48]

ก.2.2 เสียงสระภาษาไทย

สระในภาษาไทยตามไวยากรณ์ดั้งเดิม [49] มีทั้งหมด 21 รูป 32 หน่วยเสียง แบ่งเป็น 3 กลุ่มใหญ่ๆ คือ

1. *สระเดี่ยว* เป็นสระเสียงแท้ ซึ่งการออกเสียงสระตั้งแต่เริ่มต้นจนสิ้นสุดไม่มีการเปลี่ยนรูปร่างของลิ้นและช่องปาก สระเดี่ยวในภาษาไทยมีทั้งสิ้น 18 หน่วยเสียง เป็นสระเสียงสั้น 9 หน่วยเสียง และสระเสียงยาว 9 หน่วยเสียง
2. *สระผสม* เป็นสระที่เกิดจากการออกเสียงผสมกันของสระแท้ โดยลิ้นและช่องปากจะเปลี่ยนจากรูปร่างการออกเสียงของสระหนึ่งไปยังอีกสระหนึ่งอย่างค่อนข้างกลมกลืนและรวดเร็ว สระประสมในภาษาไทยมีทั้งสิ้น 6 หน่วยเสียง เป็นสระเสียงสั้น 3 หน่วยเสียง และสระเสียงยาว 3 หน่วยเสียง
3. *สระเกิน* ในภาษาไทยมีรูปสระที่เกิดจากการรวมของเสียงสระกับตัวสะกดหรือคำควบเข้าไว้ด้วยกัน ซึ่งมีทั้งหมด 8 หน่วยเสียง

เสียงสระในภาษาไทยสามารถสรุปได้ดังตารางผนวก ก.2

ตารางผนวก ก.2 เสียงสระภาษาไทย

หน่วยเสียง	ส่วนของลิ้นที่ใช้ออกเสียง	ความสูงของลิ้น	การห่อริมฝีปาก	ความยาวเสียง	รูปสระ
สระเดี่ยว					
/i/	หน้า	ปิด	ไม่ห่อ	สั้น	อิ
/i:/	หน้า	ปิด	ไม่ห่อ	ยาว	อี
/e/	หน้า	กึ่งปิด	ไม่ห่อ	สั้น	เอะ
/e:/	หน้า	กึ่งปิด	ไม่ห่อ	ยาว	เอ
/ɛ/	หน้า	กึ่งเปิด	ไม่ห่อ	สั้น	แอะ
/ɛ:/	หน้า	กึ่งเปิด	ไม่ห่อ	ยาว	แอ
/ɯ/	หลัง ค่อนมาทางกลาง	ปิด	ไม่ห่อ	สั้น	อึ
/ɯ:/	หลัง ค่อนมาทางกลาง	ปิด	ไม่ห่อ	ยาว	อือ
/ɤ/	หลัง ค่อนมาทางกลาง	กึ่งปิด	ไม่ห่อ	สั้น	เออะ
/ɤ:/	หลัง ค่อนมาทางกลาง	กึ่งปิด	ไม่ห่อ	ยาว	เออ
/a/	กลาง	เปิด	ไม่ห่อ	สั้น	อะ
/a:/	กลาง	เปิด	ไม่ห่อ	ยาว	อา
/u/	หลัง	ปิด	ห่อ	สั้น	อุ
/u:/	หลัง	ปิด	ห่อ	ยาว	อู
/o/	หลัง	กึ่งปิด	ห่อ	สั้น	โอะ
/o:/	หลัง	กึ่งปิด	ห่อ	ยาว	โอ
/ɔ/	หลัง	กึ่งเปิด	ห่อ	สั้น	เอาะ
/ɔ:/	หลัง	กึ่งเปิด	ห่อ	ยาว	ออ

หน่วยเสียง	ส่วนประกอบ		ความยาวเสียง	รูปสระ
สระประสม				
/ia/	/i/ + /a/		สั้น	เอียะ
/i:a/	/i:/ + /a/		ยาว	เอีย
/ɯa/	/ɯ/ + /a/		สั้น	เอือะ
/ɯ:a/	/ɯ:/ + /a/		ยาว	เอือ
/ua/	/u/ + /a/		สั้น	อัวะ
/u:a/	/u:/ + /a/		ยาว	อัว
สระเกิน				
/am/	/a/ + /m/		สั้น	อำ
/aj/	/a/ + /j/		สั้น	ไอ ใอ
/aw/	/a/ + /w/		สั้น	เอา
/ri/, /rɯ/	/r/ + /i/ , /r/ + /ɯ/		สั้น	ฤ
/ri:/, /rɯ:/	/r/ + /i:/ , /r/ + /ɯ:/		ยาว	ฤา
/li/, /lɯ/	/l/ + /i/ , /l/ + /ɯ/		สั้น	ฦ
/li:/, /lɯ:/	/l/ + /i:/ , /l/ + /ɯ:/		ยาว	ฦา

หมายเหตุ

ใช้สัญลักษณ์ตามสัทอักษรสากล [48]

ก.2.3 เสียงวรรณยุกต์ภาษาไทย

สำหรับในภาษาไทย วรรณยุกต์นั้นถือได้ว่าเป็นหน่วยเสียงที่สำคัญ เพราะสามารถใช้แยกแยะความแตกต่างทางความหมายของคำในภาษาไทยได้ ตรงกันข้ามกับบางภาษา เช่น ภาษาอังกฤษ ซึ่งไม่จัดว่าเสียงวรรณยุกต์เป็นหน่วยเสียงในภาษา เพราะไม่ว่าเราจะพูดภาษาอังกฤษด้วยเสียงสูงต่ำอย่างไร ผู้ฟังก็สามารถเข้าใจได้เหมือนกัน แต่ก็อาจจะต้องมีการใช้เสียงวรรณยุกต์ประกอบบ้าง ทั้งนี้เพื่อสื่ออารมณ์ของผู้พูดเท่านั้น ภาษาไทยจึงจัดได้ว่าเป็นภาษามีวรรณยุกต์ เสียงวรรณยุกต์ภาษาไทยสามารถแบ่งออกเป็น 2 ชนิดใหญ่ๆ คือ

1. *เสียงวรรณยุกต์ระดับ* เป็นเสียงวรรณยุกต์ที่มีระดับความถี่ค่อนข้างคงที่ตลอดพยางค์ ถึงแม้ว่าในการออกเสียงพูดโดยปกตินั้น เสียงต้นพยางค์มักจะไม่ได้มีความถี่และความดังเท่ากันกับเสียงท้ายพยางค์ โดยเสียงต้นพยางค์มักมีมีระดับความถี่สูงกว่าและดังกว่าเสียงท้ายพยางค์ แต่ในทางสัทศาสตร์แล้ว ความถี่ที่ต่างกันหรือการเปลี่ยนแปลงของระดับเสียงนี้ถือว่าเล็กน้อยมาก เมื่อเทียบกับการเปลี่ยนระดับความถี่ของเสียงในพยางค์อีกจำพวกหนึ่งซึ่งจะได้กล่าวต่อไป สำหรับเสียงวรรณยุกต์ระดับในภาษาไทยนั้น มีอยู่ด้วยกัน 3 เสียงดังนี้คือ

    1.1 *เสียงวรรณยุกต์สามัญ* เสียงวรรณยุกต์นี้มีระดับความถี่ปานกลาง ประมาณ 120 เฮิรตซ์ และคงที่อยู่ที่ระดับนั้นจนกระทั่งปลายพยางค์ จึงจะลดต่ำลงมา

จนเกือบถึงประมาณ 110 เฮิรตซ์ เสียงวรรณยุกต์สามัญนี้จะไม่ปรากฏในพยางค์ที่มีพยัญชนะหยุดเป็นพยัญชนะท้าย หรือที่เรียกกันว่าคำตาย

1.2 *เสียงวรรณยุกต์เอก* เสียงวรรณยุกต์นี้มีระดับความถี่ต้นเสียงปานกลาง ประมาณ 120 เฮิรตซ์ แล้วลดต่ำลงมาเหลือประมาณ 100 เฮิรตซ์อย่างรวดเร็ว และคงที่อยู่ในระดับนี้ สำหรับเสียงวรรณยุกต์เอกจะปรากฏกับพยางค์ได้ทุกรูปแบบ ทั้งคำเป็นและคำตาย

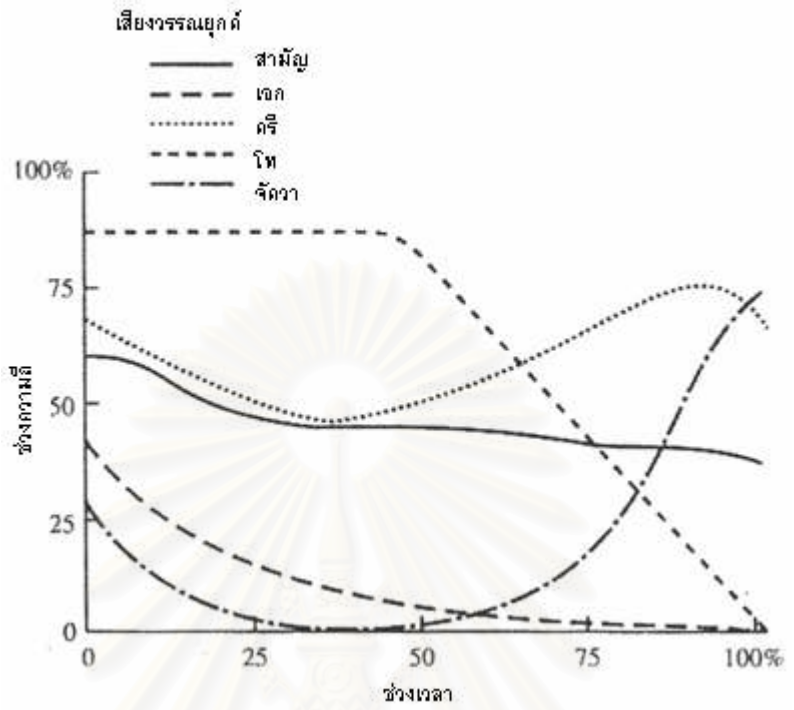
1.3 *เสียงวรรณยุกต์ตรี* เสียงวรรณยุกต์นี้มีระดับความถี่ค่อนข้างสูง โดยจะค่อยๆ สูงขึ้นทีละน้อยจากต้นพยางค์ซึ่งมีความถี่ประมาณ 125 เฮิรตซ์ ไปจนถึงประมาณ 135 – 140 เฮิรตซ์เมื่อสิ้นพยางค์ หรืออาจจะลดต่ำลงตอนปลายพยางค์มาอยู่ที่ประมาณ 130 เฮิรตซ์ก็ได้ ขึ้นอยู่กับว่าพยางค์นั้นๆ จบลงด้วยเสียงประเภทใด ถ้าพยางค์นั้นคำเป็น ระดับของเสียงตอนปลายของพยางค์จะไม่ลดต่ำลงมา แต่ถ้าพยางค์นั้นเป็นคำตาย ระดับเสียงตอนปลายจะลดต่ำลงอย่างรวดเร็ว

2. *เสียงวรรณยุกต์เปลี่ยนระดับ* เป็นเสียงวรรณยุกต์ที่มีระดับความถี่ของการออกเสียงเปลี่ยนแปลงมากในช่วงพยางค์หนึ่งๆ เช่น ต้นพยางค์ออกเสียงให้มีระดับสูง แล้วลดระดับเสียงลงอย่างรวดเร็วไปสู่ระดับต่ำที่ท้ายพยางค์ หรือต้นพยางค์ออกเสียงให้มีระดับต่ำ แล้วเพิ่มระดับเสียงอย่างรวดเร็วไปเป็นระดับสูงที่ท้ายพยางค์ นอกจากนี้ ยังอาจจะเกิดจากการเปลี่ยนระดับเสียงจากสูงแล้วไปต่ำแล้วไปสูงอีก หรือเปลี่ยนจากต่ำแล้วไปสูงแล้วไปต่ำอีกก็ได้ สำหรับในภาษาไทยนั้นมีเสียวรรณยุกต์เปลี่ยนระดับอยู่ 2 เสียงดังนี้

2.1 *เสียงวรรณยุกต์โท* ระดับเสียงจะเริ่มต้นที่ระดับความถี่ประมาณ 140 เฮิรตซ์ แต่เมื่อถึงประมาณ 1 ใน 4 ของความยาวช่วงพยางค์ ระดับความถี่จะเริ่มลดลงเรื่อยๆ จนต่ำกว่า 100 เฮิรตซ์ที่ปลายพยางค์ หรืออาจจะมีการเปลี่ยนระดับความถี่สูงขึ้นจากต้นพยางค์เล็กน้อยก่อนที่จะลดระดับเสียงลงอย่างรวดเร็วก็ได้ เสียงวรรณยุกต์โทนี้จะไม่ปรากฏในคำตาย ยกเว้นในคำเลียนเสียงธรรมชาติ หรือคำลงท้ายประโยคบางคำ เช่น "พลัก" หรือ "ละ" เป็นต้น

2.2 *เสียงวรรณยุกต์จัตวา* ระดับเสียงจะเริ่มที่ระดับความถี่ประมาณ 110 เฮิรตซ์ แล้วมักจะลดลงเล็กน้อยก่อนจะเพิ่มความถี่ขึ้นอย่างรวดเร็วจนสูงถึงประมาณ 140 เฮิรตซ์ที่ท้ายพยางค์ เสียงวรรณยุกต์จัตวานี้จะไม่ปรากฏที่คำตาย

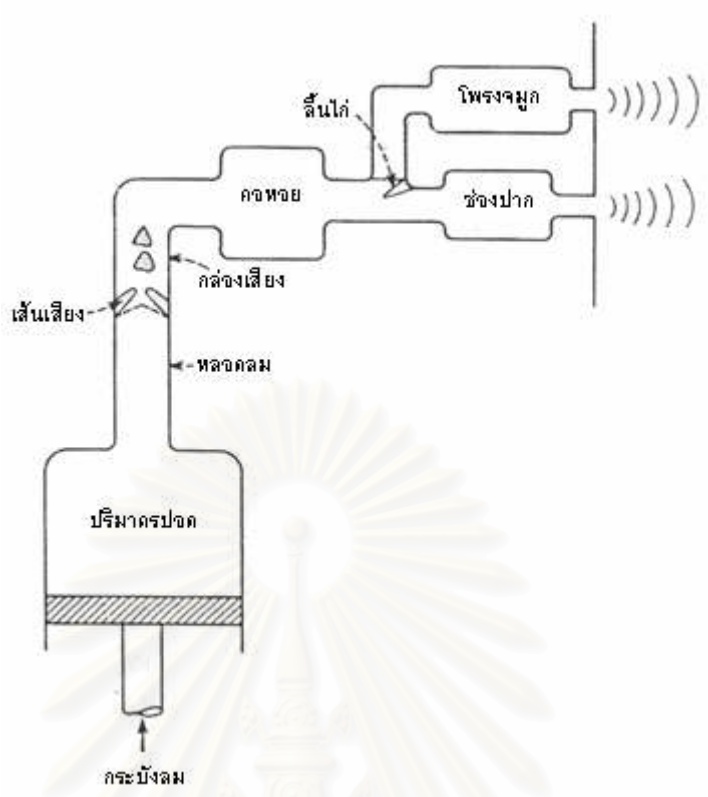
การเปลี่ยนแปลงความถี่ของเสียงในวรรณยุกต์ภาษาไทยสามารถแสดงได้ดังรูปผนวกที่ ก. 2 [50]



รูปผนวกที่ ก.2 การเปลี่ยนแปลงความถี่ของเสียงวรรณยุกต์ภาษาไทย

ก.3 สวนศาสตร์ของเสียงพูด

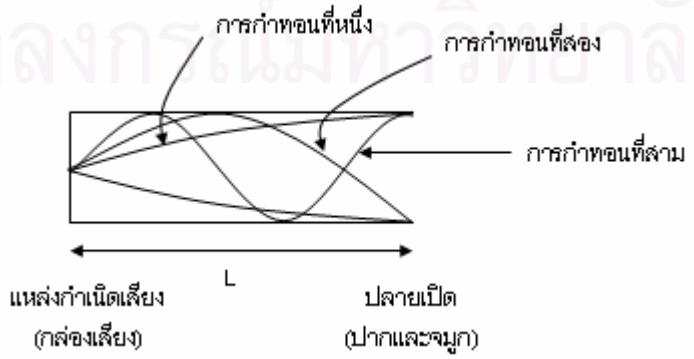
กระบวนการสร้างเสียงพูดสามารถพิจารณาในเชิงสวนศาสตร์ได้อย่างง่ายๆ ว่าประกอบไปด้วยลำดับของท่อและช่อง ซึ่งเปรียบได้กับทางเดินของเสียงจากปอดไปยังปากและจมูก ท่อและช่องนี้มีความยาวโดยรวมประมาณ 7 นิ้ว โดยเส้นเสียงจะอยู่ในตำแหน่งปลายสุด ทำหน้าที่ควบคุมการไหลของลมจากปอดให้เข้าสู่ช่องทางเดินเสียง ส่วนประกอบของช่องทางเดินเสียงที่มีลักษณะเป็นท่อจะสามารถเปลี่ยนรูปร่างได้ในอัตราสูงถึง 10 ครั้งต่อวินาที ส่วนเส้นเสียงนั้นจะสามารถเปิดและปิดได้ด้วยอัตราเร็วประมาณ 100 – 300 ครั้งต่อวินาที ซึ่งการเปลี่ยนรูปร่างของช่องทางเดินเสียง รวมทั้งการเปิดและปิดของเส้นเสียงดังกล่าวนี้ รวมเรียกว่า กระบวนการสร้างเสียงพูด แบบจำลองของกระบวนการสร้างเสียงพูดสามารถแสดงได้ดังรูปผนวกที่ ก.3 [51]



รูปผนวกที่ ก.3 แบบจำลองกระบวนการสร้างเสียงพูด

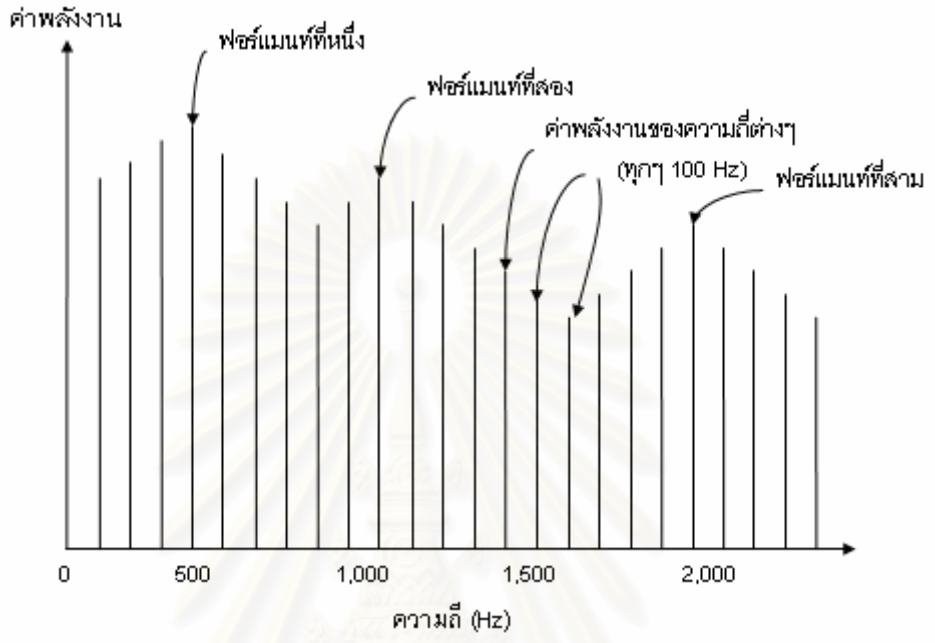
แบบจำลองอย่างง่ายของช่องทางเดินเสียงอาจมองได้เป็นลักษณะของท่อทรงกระบอกที่มีต้นกำเนิดเสียงอยู่ที่ปลายปิดข้างหนึ่ง (กล่องเสียง) ส่วนปลายอีกข้างหนึ่งจะเปิด (ปากและจมูก) ดังรูปผนวกที่ ก.4 ดังนั้นจึงเกิดการกำทอนภายในท่อได้ที่ความยาวคลื่นเท่ากับ $4L$, $4L/3$, $4L/5$, ... เมตร เมื่อ L คือ ความยาวของท่อ หรือที่ความถี่เท่ากับ $c/4L$, $3c/4L$, $5c/4L$, ... เฮิรตซ์ เมื่อ c คือค่าความเร็วของเสียงในอากาศ

ในสภาพปกติ ช่องทางเดินเสียงของมนุษย์จะมีความยาวประมาณ 7 นิ้ว หรือ 17 เซนติเมตร และ c มีค่าเท่ากับ 340 เมตรต่อวินาที ดังนั้น จึงมีการกำทอนที่ความถี่ประมาณ 500, 1,500 และ 2,500 เฮิรตซ์ เป็นต้น [52]



รูปผนวกที่ ก.4 การกำทอนภายในแบบจำลองของช่องทางเดินเสียง

เมื่อเส้นเสียงสั่น จะกระตุ้นให้เกิดคลื่นเสียงซึ่งสามารถแยกออกเป็นผลรวมของคลื่นความถี่ต่างๆ ด้วยการแปลงฟูเรียร์ การเปลี่ยนรูปร่างของช่องทางเดินเสียงทำให้เกิดการกำทอนและสร้างความถี่ที่มีพลังงานสูงเด่นขึ้นมาเมื่อดูจากสเปกตรัมพลังงาน ซึ่งเรียกว่าฟอร์แมนต์ของเสียง ดังรูปผนวกที่ ก.5



รูปผนวกที่ ก.5 สเปกตรัมพลังงาน

ฟอร์แมนต์ที่มีความถี่ต่ำที่สุดจะเรียกว่าฟอร์แมนต์ที่หนึ่ง ซึ่งจะมีค่าประมาณ 200 – 1,000 เฮิรตซ์ ทั้งนี้ขึ้นอยู่กับรูปร่างของช่องทางเดินเสียงด้วย ส่วนฟอร์แมนต์ที่สองที่อยู่ถัดไปก็จะมีค่าประมาณ 500 – 2,500 เฮิรตซ์ และฟอร์แมนต์ที่สามมีค่าประมาณ 1,500 – 3,500 เฮิรตซ์ เป็นต้น โดยฟอร์แมนต์ที่หนึ่งและสองเป็นคุณสมบัติที่สำคัญมากคุณสมบัติหนึ่งที่สามารถบ่งชี้เสียงสระ

ในการวิเคราะห์เสียงพูดเบื้องต้น สามารถทำได้โดยแบ่งเสียงพูดเป็นช่วงสั้นๆ แล้วแปลงให้อยู่ในโดเมนความถี่ จะได้สเปกตรัมพลังงานของเสียงพูดในช่วงนั้น แล้วเมื่อนำการวิเคราะห์แต่ละช่วงมาพล็อตตามแกนเวลา และแสดงค่าพลังงานด้วยความเข้มหรือสีต่างๆ จะได้ภาพที่เรียกว่าสเปกโตรแกรม ซึ่งทำให้เราสามารถรับรู้เสียงพูดได้ด้วยตา (ใครจะคิดบ้างว่าเราก็สามารถมองเห็นเสียงพูดได้) โดยเสียงพูดที่ต่างกันจะมีสเปกตรัมพลังงานที่ต่างกัน ภาพจากสเปกโตรแกรมจึงออกมาต่างกันด้วย เราจึงสามารถพิจารณาลักษณะของเสียงพูดและจำแนกเสียงพูดได้โดยดูจากสเปกโตรแกรม

ภาคผนวก ข
หน่วยเสียงที่ใช้

ข.1 หน่วยเสียงมาตรฐานในการรู้จำเสียงพูดภาษาไทย

จากเรื่องสัทศาสตร์ภาษาไทยที่ได้กล่าวไว้ในภาคผนวก ก ทำให้สามารถสรุปหน่วยเสียงมาตรฐานสำหรับการรู้จำเสียงพูดภาษาไทยได้ว่ามีทั้งหมด 75 หน่วยเสียง ดังตารางผนวก ข.1

ตารางผนวก ข.1 หน่วยเสียงมาตรฐานในการรู้จำเสียงพูดภาษาไทย

สัทอักษรสากล	ตัวอักษรแอสกี	ตัวอย่าง	สัทอักษรสากล	ตัวอักษรแอสกี	ตัวอย่าง
พยัญชนะเดี่ยว			พยัญชนะควบกล้ำ		
/p/	/p/	ปาก	/pr/	/pr/	ประสาน
/p^h/	/ph/	พบ, ภัย, ผ่าน	/pl/	/pl/	ปลา
/b/	/b/	บอก	/p^hr/	/phr/	พราน
/t/	/t/	เต้น, กุฏิ	/p^hl/	/phl/	พลาด
/t^h/	/th/	ทิ้ง, ธง, เฒ่า, ฐาน, มณโฑ	/br/	/br/	เบรน
/d/	/d/	ด้าน, ชฎา	/bl/	/bl/	บลู
/c/	/c/	จะ	/tr/	/tr/	เตรียม
/c^h/	/ch/	ชอบ, เฌอ	/t^hr/	/thr/	จันทรา
/k/	/k/	ก่อน	/dr/	/dr/	ดรากอน
/k^h/	/kh/	คน, เขิน, ฆ่า	/kr/	/kr/	กราบ
/ʔ/	/#/	อาน	/kl/	/kl/	เกลอ
/m/	/m/	ไม่	/kw/	/kw/	กวาง
/n/	/n/	นาน, เณร	/k^hr/	/khr/	คร่า
/ŋ/	/ng/	เงิน	/k^hl/	/khl/	เคลื่อน
/f/	/f/	ฝน, ฟัน	/k^hw/	/khw/	ขวา
/s/	/s/	สาย, ศิลา, รักษา, ซ่อน	/fr/	/fr/	ฟราย
/h/	/h/	โหน, เฮฮา	/fl/	/fl/	เฟลม
/r/	/r/	รอ, ฤทัย	เสียงเงียบ		
/l/	/l/	เล่น, กีฬา		/sil/	
/w/	/w/	ว่า			
/j/	/j/	ย้อน, หญิง			

สัทอักษรสากล	ตัวอักษรแอสกี	ตัวอย่าง	สัทอักษรสากล	ตัวอักษรแอสกี	ตัวอย่าง
สระเดี่ยว			สระผสม		
/i/	/i/	อิ	/ia/	/ia/	เอียะ
/iː/	/ii/	อี	/iːa/	/iia/	เอีย
/e/	/e/	เอะ	/ɯa/	/va/	เอือะ
/eː/	/ee/	เอ	/ɯːl/	/vva/	เอือ
/ɛ/	/x/	แอะ	/ua/	/ua/	อัวะ
/ɛː/	/xx/	แอ	/uːa/	/uua/	
/ɯ/	/v/	อึ	ตัวสะกด		
/ɯː/	/vv/	อือ	/pˉ/	/p^/	พบ[1]
/ɤ/	/q/	เออะ	/tˉ/	/t^/	เกร็ด[2]
/ɤː/	/qq/	เออ	/cʰˉ/	/ch^/	คลัช
/a/	/a/	อะ	/kˉ/	/k^/	ปาก[3]
/aː/	/aa/	อา	/mˉ/	/m^/	ลม
/u/	/u/	อุ	/nˉ/	/n^/	เรียน[4]
/uː/	/uu/	อู	/ŋˉ/	/ng^/	ฟาง
/o/	/o/	โอะ	/fˉ/	/f^/	กราฟ
/oː/	/oo/	โอ	/sˉ/	/s^/	เอส
/ɔ/	/@/	เอาะ	/lˉ/	/l^/	แอล
/ɔː/	/@@/	ออ	/wˉ/	/w^/	กาว
			/jˉ/	/j^/	ยาย

หมายเหตุ

[1] นอกจากนี้ยังมี กษาปณ์, เคารพ, ลาภ เป็นต้น

[2] นอกจากนี้ยังมี ตรวจ, ปราชญ์, ก๊าซ, กฎหมาย, ปรากฏ, รัฐ, ครุฑ, พัฒนา, วุฒิ, อนุญาต, ญาติ, ธาตุ, มาตร, รถ, สามารถ, มารยาท, พุทธ, สารท, พุธ, พยาธิ, ประเทศ, โทษ, ทาส เป็นต้น

[3] นอกจากนี้ยังมี จักร, โทรเลข, บริจาค, สมัคร, เมฆ เป็นต้น

[4] นอกจากนี้ยังมี บำเพ็ญ, โบราณ, อาการ, จักรวาล, ปลาวาฬ เป็นต้น

ข.2 หน่วยเสียงที่ถูกทำการลดทอนสำหรับรู้จำเสียงพูดภาษาไทย

ในการรู้จำเสียงพูดต่อเนื่องภาษาไทยนั้น พบว่าหน่วยเสียงที่เป็นพยัญชนะควบกล้ำสามารถแยกทำการรู้จำได้ เป็นหน่วยเสียงพยัญชนะต้น และหน่วยเสียงพยัญชนะที่มาควบนั้น ส่วนหน่วยเสียงที่เป็นสระเสียงสั้นและสระเสียงยาว พบว่าการวิเคราะห์ในช่วงเวลาสั้นๆ ไม่สามารถแยกความแตกต่างของทั้งสองหน่วยเสียงได้ สำหรับหน่วยเสียงที่เป็นสระผสมก็เช่นเดียวกัน จะแยกทำการรู้จำเป็นหน่วยเสียงสระที่นำมาผสมแทน

ด้วยเหตุนี้ จึงได้ทำการลดทอนหน่วยเสียงมาตรฐาน โดยตัดหน่วยเสียงที่เป็นพยัญชนะควบกล้ำ สระเสียงยาว และสระผสมออก ทำให้เหลือหน่วยเสียงทั้งหมด 43 หน่วยเสียง จาก 75 หน่วยเสียงข้างต้น ดังตารางผนวก ข.2

ตารางผนวก ข.2 หน่วยเสียงที่ถูกทำการลดทอนสำหรับรู้จำเสียงพูดภาษาไทย

สัทอักษรสากล	ตัวอักษรแอสกี	ตัวอย่าง	สัทอักษรสากล	ตัวอักษรแอสกี	ตัวอย่าง
พยัญชนะเดี่ยว			สระเดี่ยว		
/p/	/p/	ปาก	/i/	/i/	อิ
/p^h/	/ph/	พบ, ภัย, ผ่าน	/e/	/e/	เอะ
/b/	/b/	บอก	/ɛ/	/x/	แอะ
/t/	/t/	เต้น, กุฏิ	/ɯ/	/v/	อึ
/t^h/	/th/	ทิ้ง, ธง, เฒ่า, ฐาน, มณโฑ	/ɤ/	/q/	เออะ
/d/	/d/	ด้าน, ชฎา	/a/	/a/	อะ
/c/	/c/	จะ	/u/	/u/	อุ
/c^h/	/ch/	ชอบ, เฉอ	/o/	/o/	โอะ
/k/	/k/	ก่อน	/ɔ/	/@/	เอาะ
/k^h/	/kh/	คน, เขิน, ฆ่า	ตัวสะกด		
/ʔ/	/#/	อาน	/p˺/	/p^/	พบ
/m/	/m/	ไม่	/t˺/	/t^/	เกร็ด
/n/	/n/	นาน, เณร	/c^h˺/	/ch^/	คลัช
/ŋ/	/ng/	เงิน	/k˺/	/k^/	ปาก
/f/	/f/	ฝน, ฟัน	/m˺/	/m^/	ลม
/s/	/s/	สาย, ศิลา, รักษา, ซ่อน	/n˺/	/n^/	เรียน
/h/	/h/	โหน, เฮฮา	/ŋ˺/	/ng^/	ฟาง
/r/	/r/	รอ, ฤทัย	/f˺/	/f^/	กราฟ
/l/	/l/	เล่น, กีฬา	/s˺/	/s^/	เอส
/w/	/w/	ว่า	/l˺/	/l^/	แอล
/j/	/j/	ย้อน, หญิง	/w˺/	/w^/	กาว
เสียงเงียบ			/j˺/	/j^/	ยาย
	/sil/				

ทั้งนี้ เพื่อความสะดวก การเขียนหน่วยเสียงทั่วไปจะใช้สัญลักษณ์แบบตัวอักษรแอสกีเป็นหลัก

ข.3 สัญลักษณ์ที่ใช้แทนคำ

นอกจากหน่วยย่อยทางภาษาที่เป็นหน่วยเสียงแล้ว ในบางครั้งเราต้องการใช้สัญลักษณ์แทนการออกเสียงของคำ ซึ่งทำได้โดยการเขียนหน่วยเสียงที่ประกอบขึ้นเป็นคำนั้นตามลำดับ และใส่วรรณยุกต์กำกับไว้หลังหน่วยเสียงสระ โดยสัญลักษณ์ที่ใช้แทนวรรณยุกต์เป็นดังตารางผนวก ข. 3

ตารางผนวก ข.3 สัญลักษณ์ที่ใช้แทนวรรณยุกต์

เสียงวรรณยุกต์ภาษาไทย	สัญลักษณ์
สามัญ	0
เอก	1
โท	2
ตรี	3
จัตวา	4

ตัวอย่างเช่น คำว่า “ศูนย์” สามารถเขียนสัญลักษณ์การออกเสียงได้เป็น /suu4n^/ เป็นต้น

ในกรณีที่เป็นคำหลายพยางค์ จะใช้เครื่องหมาย _ แยกระหว่างพยางค์ เช่น คำว่า “สมชาย” สามารถเขียนสัญลักษณ์การออกเสียงได้เป็น /so4m^_chaa0j^/

ภาคผนวก ค
รายละเอียดของฐานข้อมูลเสียงพูด

ค.1 ฐานข้อมูลเสียงพูดชื่อไทย

ฐานข้อมูลเสียงพูดชื่อไทย หรือ Thai First Name Speech Corpus (TFNC) ประกอบด้วยรายชื่อต่างๆ ดังตารางผนวก ค.1

ตารางผนวก ค.1 คำศัพท์ต่างๆ ในฐานข้อมูลเสียงพูดชื่อไทย

รายชื่อ	เสียงพูด	รายชื่อ	เสียงพูด
อรรถวิทย์	/#a1t^_tha1_wi3t^/	วีระ	/wii0_ra3/
โปรดปราน	/proo1t^_praa0n^/	ประภาส	/pra1_phaa2t^/
ธงชัย	/tho0ng^_cha0j^/	เศรษฐา	/see1t^_thaa4/
วิษณุ	/wi3t^_sa1_nu3/	วิวัฒน์	/wi3_wa3t^/
มัณฑนา	/ma0n^_tha3_naa0/	พรศิริ	/ph@@0n^_si1_ri1/
ทักษิณา	/tha3k^_si1_naa0/	เชษฐ	/chee2t^/
ฐานิศรา	/thaa4_ni3t^_sa1_raa0/	อรรถสิทธิ์	/#a1t^_tha1_si1t^/
ญาใจ	/jaa0_ca0j^/	กอบกุล	/k@@1p^_ku0n^/
ผู้ช่วยหัวหน้าภาค	/phuu2_chuua2j^_huua4_naa2_phaa2k^/	นครทิพย์	/na3_kh@@0n^_thi3p^/
หัวหน้าภาค	/huua4_naa2_phaa2k^/	วิชาญ	/wi3_chaa0n^/
ธุรการ	/thu3_ra3_kaa0n^/	สุเมธ	/su1_mee2t^/
ธาราทิพย์	/thaa0_raa0_thi3p^/	ธนาวรรณ	/tha3_naa0_wa0n^/
บุญเสริม	/bu0n^_sqq4m^/	วันพร	/wa0n^_ph@@0n^/
อาทิตย์	/#aa0_thi3t^/	ชัยศิริ	/cha0j^_si1_ri1/
สาธิต	/saa4_thi3t^/	ทวิตีย์	/tha3_wi3t^_tii0/
เฉลิมเอก	/cha1_lqq4m^_#ee1k^/	ชัย	/cha0j^/
สืบสกุล	/svv1p^_sa1_ku0n^/	ณัฐวุฒิ	/na3t^_tha1_wu3t^/
บุญชัย	/bu0n^_cha0j^/	จารุมาตร	/caa0_ru3_maa2t^/
นงลักษณ์	/no0ng^_la3k^/	ยรรยง	/ja0n^_jo0ng^/
ฐิต	/thi1t^/	วันชัย	/wa0n^_cha0j^/
ชูชีพ	/chuu0_chii2p^/	พิษณุ	/pi3t^_sa1_nu3/
เมธี	/mee0_thii0/	เกริก	/krqq1k^/
สมชาย	/so4m^_chaa0j^/		

ค.2 ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย หรือ Thai Animal Speech Corpus (TASC) ประกอบด้วยประโยคต่างๆ ดังตารางผนวก ค.2

ตารางผนวก ค.2 ประโยคต่างๆ ในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

ประโยคเกี่ยวกับแมว	
ประโยคแมว 1	แมวมีรูปร่างคล้ายเสือ
เสียงพูด	/mxx0w^/ /mii0/ /ruu2p_raa2ng/ /khlaa3j^/ /svva4/
ประโยคแมว 2	ตีนของมันมีอุ้งเนื้อนิ่มนิ่ม
เสียงพูด	/tii0n^/ /kh@@4ng/ /ma0n^/ /mii0/ #u2ng^/ /nvva3/ /ni2m^/ /ni2m^/
ประโยคแมว 3	มันจึงเดินเบา
เสียงพูด	/ma0n^/ /cv0ng^/ /dqq0n^/ /ba0w^/
ประโยคแมว 4	ไม่มีเสียง
เสียงพูด	/ma2j^/ /mii0/ /siia4ng^/
ประโยคแมว 5	แมวมีหนวดไว้คลำทาง
เสียงพูด	/mxx0w^/ /mii0/ /nuua1t^/ /wa3j^/ /khla0m^/ /thaa0ng^/
ประโยคแมว 6	มีเล็บแหลมคมไว้ตะครุบหนูหรือจิ้งจก
เสียงพูด	/mii0/ le3p^/ /lxx4m^_kho0m^/ /wa3j^/ /ta1_khru3p^/ /nuu4/ /rvv4/ /ci2ng^_co1k^/
ประโยคแมว 7	แมวเป็นสัตว์ชอบสะอาด
เสียงพูด	/mxx0w^/ /pe0n^/ /sa1t^/ /ch@@2p^/ /sa1_#aa1t^/
ประโยคแมว 8	มันชอบเอาลิ้นเลียเนื้อตัวให้ขนเกลี้ยงสะอาดอยู่เสมอ
เสียงพูด	/ma0n^/ /ch@@2p^/ /#a0w^/ /li3n^/ /liia0/ /nvva3_tuua0/ /ha2j^/ /kho4n^/ /kliia2ng^/ /sa1_#aa1t^/ /juu1/ /sa1_mqq4/
ประโยคแมว 9	มันเป็นสัตว์ชอบอบอุ่น
เสียงพูด	/ma0n^/ /pe0n^/ /sa1t^/ /ch@@2p^/ /#o1p^_#u1n^/
ประโยคแมว 10	เวลานอนชอบนอนตามเตาไฟหรือบนกองผ้า
เสียงพูด	/wee0_laa0/ /n@@0n^/ /ch@@2p^/ /n@@0n^/ /taa0m^/ /ta0w^_fa0j^/ /rvv4/ /bo0n^/ /k@@0ng^_phaa2/
ประโยคแมว 11	แมวชอบนอนกลางวัน
เสียงพูด	/mxx0w^/ /ch@@2p^/ /n@@0n^/ /klaa0ng^_wa0n^/
ประโยคแมว 12	และออกหากินเวลากลางคืน
เสียงพูด	/lx3/ /#@@1k^/ /haa4_ki0n^/ /wee0_laa0/ /klaa0ng^_khvv0n^/
ประโยคแมว 13	มันชอบตะครุบสัตว์เล็กเล็ก

เสียงพูด	/ma0n^/ /ch@@2p^/ /ta1_khru3p^/ /sa1t^/ /le3k^/ /le3k^/
ประโยคแมว 14	เช่น หนู นก จิ้งจก
เสียงพูด	/che2n^/ /nuu4/ /no3k^/ /ci3ng^_co1k^/
ประโยคแมว 15	บางทีมันก็ขโมยปลาย่างในครัวมากิน
เสียงพูด	/baa0ng^_thii0/ /ma0n/ /k@@2/ /kha1_moo0j^/ /plaa0_jaa2ng^/ /na0j^/ /khruua0/ /maa0/ /ki0n^/
ประโยคแมว 16	แมวมีนิสัยรักเจ้าของ
เสียงพูด	/mxx0w^/ /mii0/ /ni3_sa4j^/ /ra3k^/ /ca2w^_kh@@4ng^/
ประโยคแมว 17	และชอบเคล้าเคลียอยู่กับเจ้าของ
เสียงพูด	/lx3/ /ch@@2p^/ /khla3w^_khlii0a/ /juu1/ /ka1p^/ /ca2w^_kh@@4ng^/
ประโยคแมว 18	มันจำที่อยู่ได้อย่างแม่นยำ
เสียงพูด	/ma0n^/ /ca0m^/ /thii2_juu1/ /daa2j^/ /jaa1ng^/ /mx2n^_ja0m^/
ประโยคเกี่ยวกับควาย	
ประโยคควาย 1	ควายมีลำตัวใหญ่
เสียงพูด	/khwaa0j^/ /mii0/ /la0m^_tuua0/ /ja1j^/
ประโยคควาย 2	ผิวหนังดำ
เสียงพูด	/phi4w^_na4ng^/ /da0m^/
ประโยคควาย 3	ควายที่มีผิวหนังขาว
เสียงพูด	/khwaa0j^/ /thii2/ /mii0/ /phi4w^_na4ng^/ /khaa4w^/
ประโยคควาย 4	เรียกว่า ควายเผือก
เสียงพูด	/riia2k^_waa2/ /khwaa0j^_phvva1k^/
ประโยคควาย 5	ควายมีเขายาวโค้ง
เสียงพูด	/khwaa0j^/ /mii0/ /kha4w^/ /jaa0w^/ /khoo3ng^/
ประโยคควาย 6	ปลายแหลม
เสียงพูด	/plaa0j^/ /lxx4m^/
ประโยคควาย 7	ปลายหางมีขนเป็นพวง
เสียงพูด	/plaa0j^/ /haa4ng^/ /mii0/ /kho4n^/ /pe0n^/ /phuua0ng^/
ประโยคควาย 8	ตีนเป็นกีบผ่าสองกีบ
เสียงพูด	/tii0n^/ /pe0n^/ /kii1p^/ /phaa1/ /s@@4ng^/ /kii1p^/
ประโยคควาย 9	ควายเป็นสัตว์อดทนและมีกำลังมาก
เสียงพูด	/khwaa0j^/ /pe0n^/ /sa1t^/ /#o1t^_tho0n^/ /lx3/ /mii0/ /ka0m^_la0ng^/ /maa2k^/
ประโยคควาย 10	แต่ทนแดดได้ไม่นาน
เสียงพูด	/txx1/ /tho0n^/ /dxx1t^/ /daa2j^/ /ma2j^/ /naa0n^/
ประโยคควาย 11	เวลาแดดร้อน
เสียงพูด	/wee0_laa0/ /dxx1t^/ /r@@3n^/

ประโยคควาย 12	มันเหนื่อยเร็ว
เสียงพูด	/ma0n^/ /nvva1j^/ /re0w^/
ประโยคควาย 13	มันชอบนอนแช่น้ำ
เสียงพูด	/ma0n^/ /ch@@2p^/ /n@@0n^/ /chxx2/ /naa3m^/
ประโยคควาย 14	และเกลือกโคลน
เสียงพูด	/lx3/ /klvva1k^/ /khloo0n^/
ประโยคควาย 15	ควายเป็นสัตว์ทำงานหนัก
เสียงพูด	/khwaa0j^/ /pe0n^/ /sa1t^/ /tha0m^_ngaa0n^/ /na1k^/
ประโยคควาย 16	จึงกินจุ
เสียงพูด	/cv0ng^/ /ki0n^/ /cu1/
ประโยคควาย 17	เวลากินหญ้า
เสียงพูด	/wee0_laa0/ /ki0n^/ /jaa2/
ประโยคควาย 18	ก็รีบกลืนเข้าไป
เสียงพูด	/k@@2/ /rii2p^/ /klvv0n^/ /kha2w^_pa0j^/
ประโยคควาย 19	เวลาว่างมันจึงสำรอกออกมาเคี้ยว
เสียงพูด	/wee0_laa0/ /waa2ng^/ /ma0n^/ /cv0ng^/ /sa4m^_r@@2k^/ /#@@1k^/ /maa0/ /khiia3w^/
ประโยคควาย 20	ให้ละเอียดอีกครั้งหนึ่ง
เสียงพูด	/ha2j^/ /la3_#iia1t^/ /#ii1k^/ /khra3ng^/ /nv1ng^/
ประโยคควาย 21	เรียกว่า ควายเคี้ยวเอื้อง
เสียงพูด	/riia2k^_waa2/ /khwaa0j^/ /khiia3w^_#vva2ng^/
ประโยคควาย 22	ชาวนาเลี้ยงควายไว้ไถนาและลากเกวียน
เสียงพูด	/chaa0w^_naa0/ /liia3ng^/ /khwaa0j^/ /wa3j^/ /tha4j^/ /naa0/ /lx3/ /laa2k^/ /kwiia0n^/
ประโยคควาย 23	น้ำนมควายใช้ดื่มได้ดี
เสียงพูด	/naa3m^_no0m^/ /khwaa0j^/ /cha3j^/ /dvv1m^/ /daa2j^/ /dii0/
ประโยคควาย 24	แต่ข้นและจัดกว่านมวัว
เสียงพูด	/txx1/ /kho2n^/ /lx3/ /ca1t^/ /kwaa1/ /no0m^_wuua0/
ประโยคควาย 25	เนื้อควายใช้ทำอาหารกินได้
เสียงพูด	/nvva3_khwaa0j^/ /cha3j^/ /tha0m^/ /#aa0_haa4n^/ /ki0n^/ /daa2j^/
ประโยคควาย 26	แต่ค่อนข้างเหนียว
เสียงพูด	/txx1/ /kh@2n^_khaa2ng^/ /niia4w^/
ประโยคควาย 27	และหยาบกว่าเนื้อวัว
เสียงพูด	/lx3/ /jaa1p^/ /kwaa1/ /nvva3_wuua0/
ประโยคควาย 28	ควายเป็นสัตว์กินพืชเป็นอาหาร
เสียงพูด	/khwaa0j^/ /pe0n^/ /sa1t^/ /ki0n^/ /phvv2t/ /pe0n^/ #aa0_haa4n^/

ประโยคเกี่ยวกับช้าง	
ประโยคช้าง 1	ช้างเป็นสัตว์พาหนะที่ใหญ่ที่สุด
เสียงพูด	/chaa3ng^/ /pe0n^/ /sa1t^/ /phaa0_ha1_na3/ /thii2/ /ja1j^/ /thii2_su1t^/
ประโยคช้าง 2	และเป็นสัตว์ที่ฉลาด
เสียงพูด	/lx3/ /pe0n^/ /sa1t^/ /thii2/ /cha1_laa1t^/
ประโยคช้าง 3	หัดเชื่องได้ง่าย
เสียงพูด	/ha1t^/ /chvva2ng^/ /ngaa2j^/
ประโยคช้าง 4	รักเจ้าของ
เสียงพูด	/ra3k^/ /ca2w^_kh@@4ng^/
ประโยคช้าง 5	และมีความจำดี
เสียงพูด	/lx3/ /mii0/ /khwaa0m^_ca0m^/ /dii0/
ประโยคช้าง 6	คนจึงพยายามเลี้ยงช้างไว้ให้ช่วยทำงานหนักหนัก
เสียงพูด	/kho0n^/ /cv0ng^/ /liia3ng^/ /chaa3ng^/ /wa3j^/ /ha2j^/ /chuua2j^/ /tha0m^_ngaa0n^/ /na1k^/ /na1k^/
ประโยคช้าง 7	เช่น ใช้ลากซุงในป่า
เสียงพูด	/che2n^/ /cha3j^/ /laa2k^/ /su0ng^/ /na0j^/ /paa1/
ประโยคช้าง 8	ใช้สำหรับขี่และบรรทุกของ
เสียงพูด	/cha3j^/ /sa4m^_ra1p^/ /khii1/ /lx3/ /ba0n^_thu3k^/ /kh@@4ng/
ประโยคช้าง 9	เวลาจะต้องเดินทางไกลไปในป่า
เสียงพูด	/wee0_laa0/ /ca1/ /t@2ng^/ /dqq0n^_thaa0ng^/ /kla0j^/ /na0j^/ /paa1/
ประโยคช้าง 10	ในสมัยโบราณเขาใช้ช้างสำหรับทำสงครามด้วย
เสียงพูด	/na0j^/ /sa1_ma4j^/ /boo0_raa0n^/ /kha4w^/ /cha3j^/ /chaa3ng^/ /tha0m^/ /so4ng^_khraa0m^/ /duua2j^/
ประโยคช้าง 11	ช้างเป็นสัตว์ที่มีหัวและหูใหญ่
เสียงพูด	/chaa3ng^/ /pe0n^/ /sa1t^/ /thii2/ /mii0/ /huua4/ /lx3/ /huu4/ /ja1j^/
ประโยคช้าง 12	แต่ตาเล็กมาก
เสียงพูด	/txx1/ /taa0/ /le3k^/ /maa2k^/
ประโยคช้าง 13	คอสั้นติดกับตัว
เสียงพูด	/kh@@0/ /sa2n^/ /ti1t^/ /ka1p^/ /tuua0/
ประโยคช้าง 14	มีงวงยาว
เสียงพูด	/mii0/ /nguua0ng^/ /jaa0w^/
ประโยคช้าง 15	สำหรับเหนี่ยวจับอาหารป้อนเข้าปาก
เสียงพูด	/sa4m^_ra1p^/ /niia1w^/ /ca1p^/ /#aa0_haa4n^/ /p@@2n^/ /kha2w^/ /paa1k^/
ประโยคช้าง 16	ปลายงวงมีรูจมูกสำหรับหายใจ
เสียงพูด	/plaa0j^/ /nguua0ng^/ /mii0/ /ruu0/ /ca1_muu1k^/ /sa4m^_ra1p^/ /haa4j^_ca0j^/

ประโยคช้าง 17	ช้างตัวผู้เรียกว่า ช้างพลาย
เสียงพูด	/chaa3ng^/ /tuua0_phuu2/ /riia2k^_waa2/ /chaa3ng^_phlaa0j^/
ประโยคช้าง 18	มีงางอกยาวออกมาริมปากข้างละงา
เสียงพูด	/mii0/ /ngaa0/ /ng@@2k^/ /jaa0w^/ /#@@1k^_maa0/ /ri0m^/ /paa1k^/ /khaa2ng^/ /la3/ /ngaa0/
ประโยคช้าง 19	ช้างตัวเมียเรียกว่า ช้างพัง
เสียงพูด	/chaa3ng^/ /tuua0_miia0/ /riia2k^_waa2/ /chaa3ng^_pha0ng^/
ประโยคช้าง 20	ไม่ค่อยมีงายาว
เสียงพูด	/ma2j^/ /kh@2j^/ /mii0/ /ngaa0/ /jaa0w^/
ประโยคช้าง 21	ช้างมีขากลมใหญ่มาก
เสียงพูด	/chaa3ng^/ /mii0/ /khaa4/ /klo0m^/ /ja1j^/ /maa2k^/
ประโยคช้าง 22	หนังของมันหยาบและเหนียวมาก
เสียงพูด	/na4ng^/ /kh@@4ng^/ /ma0n^/ /jaa1p^/ /lx3/ /niia4w^/ /maa2k^/
ประโยคช้าง 23	ช้างเป็นสัตว์ที่มีอายุยืน
เสียงพูด	/chaa3ng^/ /pe0n^/ /sa1t^/ /thii2/ /mii0/ /#aa0_ju3/ /jvv0n^/
ประโยคช้าง 24	บางตัวอยู่ได้ตั้งร้อยปี
เสียงพูด	/baa0ng^/ /tuua0/ /juu1/ /daa2j^/ /ta2ng^/ /r@@3j^/ /pii0/
ประโยคช้าง 25	บางทีจะเคยได้ยินคำว่า ช้างเผือก
เสียงพูด	/baa0ng^_thii0/ /ca1/ /khqq0j^/ /daa2j^_ji0n^/ /kha0m^_waa2/ /chaa3ng^_phvva1k^/
ประโยคช้าง 26	ช้างเผือกเป็นช้างที่หายาก
เสียงพูด	/chaa3ng^_phvva1k^/ /pe0n^/ /chaa3ng^/ /thii2/ /haa4_jaa2k/
ประโยคช้าง 27	ถ้าปรากฏว่ามีอยู่แห่งใด
เสียงพูด	/thaa2/ /praa0_ko1t^/ /waa2/ /mii0/ /juu1/ /hx1ng^/ /da0j^/
ประโยคช้าง 28	ก็ถือว่าเป็นช้างคู่บารมีของพระเจ้าอยู่หัว
เสียงพูด	/k@@2/ /thvv4_waa2/ /pe0n^/ /chaa3ng^/ /khuu2/ /baa0_ra3_mii0/ /kh@@4ng^/ /phra3_caa2w_juu1_huua4/
ประโยคเกี่ยวกับลิง	
ประโยคลิง 1	ลิงเป็นสัตว์ป่าที่ว่องไว
เสียงพูด	/li0ng^/ /pe0n^/ /sa1t^_paa1/ /thii2/ /w@2ng^_wa0j^/
ประโยคลิง 2	และมีนิสัยเชื่องง่าย
เสียงพูด	/lx3/ /mii0/ /ni3_sa4j^/ /chvva2ng^/ /ngaa2j^/
ประโยคลิง 3	คนจึงเอามาเลี้ยงไว้ตามบ้านและตามสวนสัตว์
เสียงพูด	/kho0n^/ /cv0ng^/ /#a0w^/ /maa0/ /liia3ng^/ /wa3j^/ /taa0m^/ /baa2n^/ /lx3/ /taa0m^/ /suua4n^_sa1t^/
ประโยคลิง 4	ในจังหวัดลพบุรี

เสียงพูด	/na0j^/ /ca0ng^_wa1t^/ /lo3p^_bu1_rii0/
ประโยคลิง 5	มีลิงรวมกันอยู่เป็นฝูงใหญ่ที่สุด
เสียงพูด	/mii0/ /li0ng^/ /ruua0m^/ /ka0n^/ /juu1/ /pe0n^/ /fuu4ng^/ /ja1j^/ /thii2_su1t^/
ประโยคลิง 6	ลิงเป็นสัตว์ชุกชนและฉลาด
เสียงพูด	/li0ng^/ /pe0n^/ /sa1t^/ /su3k^_so0n^/ /lx3/ /cha1_laa1t^/
ประโยคลิง 7	เราหัดให้ทำอะไรก็ทำได้
เสียงพูด	/ra0w^/ /ha1t^/ /ha2j^/ /tha0m^/ /#a1_ra0j^/ /k@@2/ /tha0m^/ /daa2j^/
ประโยคลิง 8	บางคนคงเคยเห็นเขาหัดลิงให้เล่นละคร
เสียงพูด	/baa0ng^/ /kho0n^/ /kho0ng^/ /khqq0j^/ /he4n^/ /kha4w^/ /ha1t^/ /li0ng^/ /ha2j^/ /le2n^/ /la3_kh@@0n^/
ประโยคลิง 9	มันก็ทำท่าต่างต่างได้คล้ายคน
เสียงพูด	/ma0n^/ /k@@2/ /tha0m^/ /thaa2/ /taa1ng^/ /taa1ng^/ /daa2j^/ /khlaa3j^/ /kho0n^/
ประโยคลิง 10	ลิงชอบกินผลไม้
เสียงพูด	/li0ng^/ /ch@@2p^/ /ki0n^/ /pho4n^_la3_maa3j/
ประโยคลิง 11	แต่บางทีมันจับแมลงกิน
เสียงพูด	/txx1/ /baa0ng^_thii0/ /ma0n^/ /ca1p^/ /ma3_lxx0ng^/ /ki0n^/
ประโยคลิง 12	เวลามันอยู่ในป่า
เสียงพูด	/wee0_laa0/ /ma0n^/ /juu1/ /na0j^/ /paa1/
ประโยคลิง 13	มันออกหากินตั้งแต่เช้าจนพลบค่ำ
เสียงพูด	/ma0n^/ /#@@1k^/ /haa4_ki0n^/ /ta2ng^_txx1/ /chaa3w^/ /co0n^/ /phlo3p^_kha2m^/
ประโยคลิง 14	เวลามันออกหากิน
เสียงพูด	/wee0_laa0/ /ma0n^/ /#@@1k^/ /haa4_ki0n^/
ประโยคลิง 15	มันจะกระโดด
เสียงพูด	/ma0n^/ /ca1/ /kra1_doo1t^/
ประโยคลิง 16	ไต่ต้นไม้ไปเป็นฝูงฝูง
เสียงพูด	/ta1j^/ /to2n^_maa3j^/ /pa0j^/ /pe0n^/ /fuu4ng^/ /fuu4ng^/

ภาคผนวก ง
พีแอลพี

พีแอลพี (PLP, Perceptual Linear Prediction) เป็นการสกัดลักษณะสำคัญของเสียงพูด โดยมีพื้นฐานมาจากการได้ยินของมนุษย์ ซึ่งมีขั้นตอนดังรูปผนวกที่ ง.1

รูปผนวกที่ ง.1 ขั้นตอนของพีแอลพี

ขั้นตอนต่างๆ ของพีแอลพี สามารถอธิบายได้ดังนี้

ง.1 การแปลงฟูเรียร์แบบเร็ว

เริ่มต้นด้วยการคำนวณค่าประมาณของสเปกตรัมกำลังสำหรับแต่ละเฟรม ซึ่งก่อนอื่นอาจนำหน้าต่างไปใส่ในเฟรมที่จะวิเคราะห์ โดยการคูณแต่ละค่าของสัญญาณในเฟรมด้วยค่าฟังก์ชันหน้าต่าง เช่น หน้าต่างแฮมมิง ดังสมการ

$$W(n) = 0.54 + 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

เมื่อ N คือ จำนวนข้อมูลในเฟรม

จากนั้น นำสัญญาณที่ได้ผ่านกระบวนการแปลงฟูเรียร์แบบเร็ว และคำนวณขนาดกำลังสองพร้อมทั้งสเปกตรัมกำลัง ดังสมการ

$$P(\omega) = \mathrm{Re}[S(\omega)]^2 + \mathrm{Im}[S(\omega)]^2$$

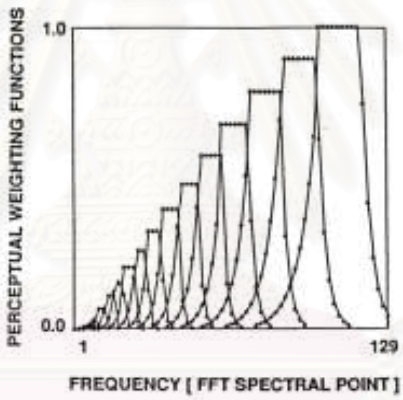
โดยที่ $S(\omega)$ คือ สัญญาณในโดเมนความถี่ที่ผ่านการแปลงฟูเรียร์

ง.2 การหาปริพันธ์ของแถบวิกฤตและการชักตัวอย่างใหม่

ในการหาปริพันธ์ของแถบวิกฤตและการชักตัวอย่างใหม่ ตัวกรองรูปสี่เหลี่ยมคางหมูดังรูปผนวกที่ ง.2 จะถูกนำมาใช้ที่ช่วงห่างประมาณ 1 บาร์ก ซึ่งแกนของบาร์กนั้นจะได้มาจากแกนความถี่ โดยใช้ฟังก์ชันวาร์ปของโชรเดอร์ และ $P(\omega)$ ในแกนความถี่เฮิรตซ์จะถูกแปลงในอยู่ในแกนความถี่บาร์ก โดยใช้สมการ

$$\Omega(\omega) = 6\ln\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right\}$$

โดยที่ ω คือ ความถี่เชิงมุม ในหน่วยของเรเดียนต่อวินาที



รูปผนวกที่ ง.2 ตัวกรองรูปสี่เหลี่ยมคางหมูของพีแอลพี

สำหรับหน้าต่างรูปสี่เหลี่ยมคางหมูนั้นก็คือการประมาณสเปกตรัมกำลังของเส้นโค้งแถบวิกฤต ซึ่งจะเป็นดังสมการ

$$\Psi(\Omega) = \begin{cases} 0 & , \quad \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & , -1.3 \le \Omega \le -0.5 \\ 1 & , \ -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & , \ 0.5 \le \Omega \le 2.5 \\ 0 & , \quad \Omega > 2.5 \end{cases}$$

$\Omega(\omega)$ จะถูกกระทำด้วยเส้นโค้งแถบวิกฤตโดยใช้สมการข้างต้น จากนั้นคำนวณ $\Theta(\Omega)$ ดังสมการ

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega)$$

$\Theta(\Omega)$ ที่ได้เรียกว่าสเปกตรัมกำลังแถบวิกฤต ซึ่งจะมีทั้งหมด 18 ค่า ครอบคลุมตั้งแต่ 0 ถึง 16.9 บาร์ก (0 ถึง 5 กิโลเฮิรตซ์) และแต่ละค่าจะปรากฏที่ตำแหน่งต่างกัน 0.994 บาร์ก

ที่ทำเช่นนี้ก็เพื่อลดความไวทางความถี่ของการประมาณค่าสเปกตรัมดั้งเดิม โดยเฉพาะอย่างยิ่งที่ความถี่สูง

ง.3 โค้งความดังเทียบเท่า

ทำการเน้นสเปกตรัมอีกครั้งหนึ่งเพื่อประมาณค่าความไวที่ไม่สมดุลของการได้ยินของมนุษย์ ณ ความถี่ต่างๆ กัน ด้วยการถ่วงน้ำหนักในของส่วนของสเปกตรัมแถบวิกฤต

$\Theta(\Omega(\omega))$ จะถูกเน้นสัญญาณโดยโค้งความดังเทียบเท่าจำลอง โดยใช้สมการ

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)]$$

โดยที่ $E(\omega)$ คือ ค่าประมาณของความไวในการรับเสียงของหูมนุษย์ที่ความถี่ต่างๆ ซึ่งถูกจำลองที่ความดัง 40 เดซิเบล และมีรูปแบบดังสมการ

$$E(\omega) = \frac{[(\omega^2 + 56.8\times10^6)\omega^4]}{[(\omega^2 + 6.3\times10^6)^2 \times (\omega^2 + 0.38\times10^9)(\omega^6 + 9.58\times10^{26})]}$$

ง.4 กฎกำลังของการได้ยิน

กฎกำลังของการได้ยินเป็นการบีบแอมปลิจูดรากที่สาม ทำการบีบอัดขนาดของสเปกตรัม ซึ่งโดยทั่วไปแล้วจะมีการหาค่าลอการิทึมหลังจากการหาปริพันธ์ โดยการทำนายเชิงเส้นแบบรับรู้จะใช้การหารากที่สามแทนการหาค่าลอการิทึม แสดงได้ดังสมการ

$$\Phi(\Omega) = \Xi(\Omega)^{0.33}$$

การคำนวณนี้เป็นการประมาณด้วยกฎกำลัง เพื่อที่จะจำลองความสัมพันธ์แบบไม่เชิงเส้นระหว่างความเข้มของเสียงและความรู้สึกถึงความดังของเสียง ซึ่งกระบวนการนี้จะช่วยลดความแปรปรวนในขนาดของสเปกตรัมแถบวิกฤต หรือการกำทอนของสเปกตรัม

ง.5 การแปลงฟูเรียร์แบบไม่ต่อเนื่องผกผัน

จะทำการแปลงฟูเรียร์แบบไม่ต่อเนื่องผกผันสำหรับการทำนายเชิงเส้นแบบรับรู้นี้ เนื่องจากค่าลอการิทึมไม่ได้ถูกคำนวณ ดังนั้นผลที่ได้จึงมักจะคล้ายกับค่าสัมประสิทธิ์อัตสหสัมพันธ์มากกว่า ถึงแม้ว่าจะมาจากสเปกตรัมที่ถูกบีบอัดก็ตาม และเนื่องจากค่าสเปกตรัมกำลังนั้นเป็นจำนวนจริงและเป็นเลขคู่ ดังนั้นจึงไม่จำเป็นที่จะต้องทำการคำนวณส่วนประกอบโคไซน์ของการแปลงฟูเรียร์แบบไม่ต่อเนื่องผกผันก็ได้

$\Phi(\Omega)$ จะถูกประมาณโดยสเปกตรัมของแบบจำลองทุกขั้วด้วยวิธีอัตสหสัมพันธ์

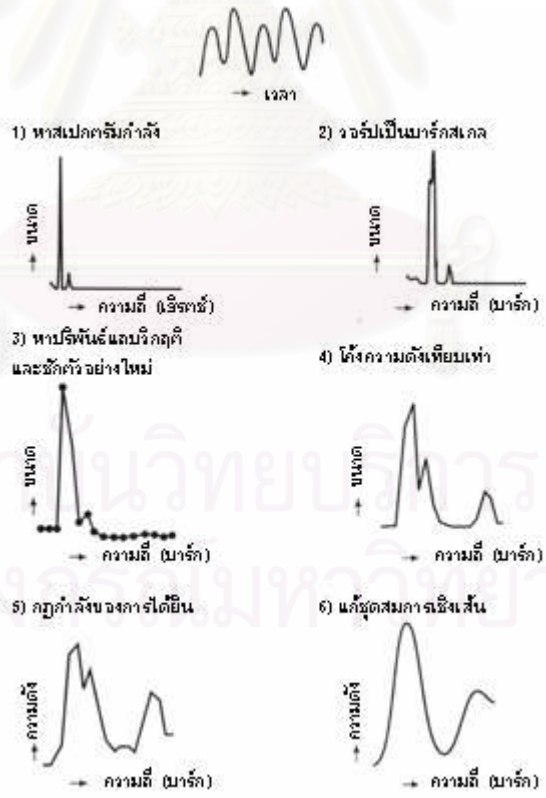
ง.6 การแก้ชุดสมการเชิงเส้น

การหาปริพันธ์นั้นจะมีประโยชน์ต่อการลดผลกระทบของต้นกำเนิดซึ่งไม่เกี่ยวข้องกับทางภาษาศาสตร์ของความแปรปรวนในสัญญาณเสียงพูดได้ แบบจำลองอัตตาถดถอยจะได้มาจากผลเฉลยของสมการเชิงเส้น ซึ่งถูกสร้างขึ้นจากอัตสหสัมพันธ์ของขั้นตอนก่อนหน้านี้นั่นเอง โดยแบบจำลองอัตตาถดถอยนี้จะถูกใช้ในการขัดเกลาสเปกตรัมแถบวิกฤตซึ่งถูกบีบอัดแล้วเหมือนกับในการทำรหัสทำนายเชิงเส้นแบบดั้งเดิม แต่สเปกตรัมผลลัพธ์ซึ่งถูกขัดเกลาแล้วจะเหมาะสมกับยอดสเปกตรัมมากกว่าท้องสเปกตรัม โดยวิธีนี้จะนำไปสู่ความทนทานต่อเสียงรบกวน และการไม่ขึ้นกับผู้พูดได้ดีกว่า

ง.7 การเวียนเกิดเซปสตรอล

การเวียนเกิดเซปสตรอลจะทำการใช้การแทนเชิงตั้งฉากสำหรับการทำนายเชิงเส้นแบบรับรู้ สัมประสิทธิ์อัตตาถดถอยจะถูกแปลงให้เป็นตัวแปรเซปสตรัมแทน

ทั้งนี้ ในแต่ละขั้นตอน จะทำให้ลักษณะของสัญญาณเปลี่ยนไปดังรูปผนวกที่ ง.3



รูปผนวกที่ ง.3 ลักษณะของสัญญาณในแต่ละขั้นตอนของพีแอลพี

ภาคผนวก จ
พจนานุกรม

จ.1 ฐานข้อมูลเสียงพูดชื่อไทย

พจนานุกรมของคำศัพท์ต่างๆ ในฐานข้อมูลเสียงพูดชื่อไทยสามารถแสดงได้ดังตารางผนวก จ.1

ตารางผนวก จ.1 พจนานุกรมสำหรับฐานข้อมูลเสียงพูดชื่อไทย

คำศัพท์	พจนานุกรม โดยใช้ชุดหน่วยเสียงมาตรฐาน	พจนานุกรม โดยใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
1. อรรถสิทธิ์	/#/ /a/ /t^/ /th/ /a/ /s/ /i/ /t^/	/#/ /a/ /t^/ /th/ /a/ /s/ /i/ /t^/
2. อรรถวิทย์	/#/ /a/ /t^/ /th/ /a/ /w/ /i/ /t^/	/#/ /a/ /t^/ /th/ /a/ /w/ /i/ /t^/
3. อาทิตย์	/#/ /aa/ /th/ /i/ /t^/	/#/ /a/ /th/ /i/ /t^/
4. บุญชัย	/b/ /u/ /n^/ /ch/ /a/ /j^/	/b/ /u/ /n^/ /ch/ /a/ /j^/
5. บุญเสริม	/b/ /u/ /n^/ /s/ /qq/ /m^/	/b/ /u/ /n^/ /s/ /q/ /m^/
6. จารุมาตร	/c/ /aa/ /r/ /u/ /m/ /aa/ /t^/	/c/ /a/ /r/ /u/ /m/ /a/ /t^/
7. ชัยศิริ	/ch/ /a/ /j^/ /s/ /i/ /r/ /i/	/ch/ /a/ /j^/ /s/ /i/ /r/ /i/
8. ชัย	/ch/ /a/ /j^/	/ch/ /a/ /j^/
9. เฉลิมเอก	/ch/ /a/ /l/ /qq/ /m^/ /#/ /ee/ /k^/	/ch/ /a/ /l/ /q/ /m^/ /#/ /e/ /k^/
10. เชษฐ	/ch/ /ee/ /t^/	/ch/ /e/ /t^/
11. ชูชีพ	/ch/ /uu/ /ch/ /ii/ /p^/	/ch/ /u/ /ch/ /i/ /p^/
12. หัวหน้าภาค	/h/ /uua/ /n/ /aa/ /ph/ /aa/ /k^/	/h/ /u/ /a/ /n/ /a/ /ph/ /a/ /k^/
13. ยรรยง	/j/ /a/ /n^/ /j/ /o/ /ng^/	/j/ /a/ /n^/ /j/ /o/ /ng^/
14. ญาใจ	/j/ /aa/ /c/ /a/ /j^/	/j/ /a/ /c/ /a/ /j^/
15. กอบกุล	/k/ /@@/ /p^/ /k/ /u/ /n^/	/k/ /@/ /p^/ /k/ /u/ /n^/
16. เกริก	/kr/ /qq/ /k^/	/k/ /r/ /q/ /k^/
17. มัณฑนา	/m/ /a/ /n^/ /th/ /a/ /n/ /aa/	/m/ /a/ /n^/ /th/ /a/ /n/ /a/
18. เมธี	/m/ /ee/ /th/ /ii/	/m/ /e/ /th/ /i/
19. นครทิพย์	/n/ /a/ /kh/ /@@/ /n^/ /th/ /i/ /p^/	/n/ /a/ /kh/ /@/ /n^/ /th/ /i/ /p^/
20. ณัฐวุฒิ	/n/ /a/ /t^/ /th/ /a/ /w/ /u/ /t^/	/n/ /a/ /t^/ /th/ /a/ /w/ /u/ /t^/
21. นงลักษณ์	/n/ /o/ /ng^/ /l/ /a/ /k^/	/n/ /o/ /ng^/ /l/ /a/ /k^/
22. พรศิริ	/ph/ /@@/ /n^/ /s/ /i/ /r/ /i/	/ph/ /@/ /n^/ /s/ /i/ /r/ /i/
23. ผู้ช่วยหัวหน้าภาค	/ph/ /uu/ /ch/ /uua/ /j^/ /h/ /uua/ /n/ /aa/ /ph/ /aa/ /k^/	/ph/ /u/ /ch/ /u/ /a/ /j^/ /h/ /u/ /a/ /n/ /a/ /ph/ /a/ /k^/
24. พิษณุ	/ph/ /i/ /t^/ /s/ /a/ /n/ /u/	/ph/ /i/ /t^/ /s/ /a/ /n/ /u/
25. ประภาส	/pr/ /a/ /ph/ /aa/ /t^/	/p/ /r/ /a/ /ph/ /a/ /t^/
26. โปรดปราน	/pr/ /oo/ /t^/ /pr/ /aa/ /n^/	/p/ /r/ /o/ /t^/ /p/ /r/ /a/ /n^/
27. สาธิต	/s/ /aa/ /th/ /i/ /t^/	/s/ /a/ /th/ /i/ /t^/
28. เศรษฐา	/s/ /ee/ /t^/ /th/ /aa/	/s/ /e/ /t^/ /th/ /a/

29. สมชาย	/s/ /o/ /m^/ /ch/ /aa/ /j^/	/s/ /o/ /m^/ /ch/ /a/ /j^/
30. สุเมธ	/s/ /u/ /m/ /ee/ /t^/	/s/ /u/ /m/ /e/ /t^/
31. สืบสกุล	/s/ /vv/ /p^/ /s/ /a/ /k/ /u/ /n^/	/s/ /v/ /p^/ /s/ /a/ /k/ /u/ /n^/
32. ทักษิณา	/th/ /a/ /k^/ /s/ /i/ /n/ /a/	/th/ /a/ /k^/ /s/ /i/ /n/ /a/
33. ธนาวรรณ	/th/ /a/ /n/ /aa/ /w/ /a/ /n^/	/th/ /a/ /n/ /a/ /w/ /a/ /n^/
34. ทวิตีย์	/th/ /a/ /w/ /i/ /t^/ /t/ /ii/	/th/ /a/ /w/ /i/ /t^/ /t/ /i/
35. ธาราทิพย์	/th/ /aa/ /r/ /aa/ /th/ /i/ /p^/	/th/ /a/ /r/ /a/ /th/ /i/ /p^/
36. ฐานิศรา	/th/ /aa/ /n/ /i/ /t^/ /s/ /a/ /r/ /aa/	/th/ /a/ /n/ /i/ /t^/ /s/ /a/ /r/ /a/
37. ฐิต	/th/ /i/ /t^/	/th/ /i/ /t^/
38. ธงชัย	/th/ /o/ /ng^/ /ch/ /a/ /j^/	/th/ /o/ /ng^/ /ch/ /a/ /j^/
39. ธุรการ	/th/ /u/ /r/ /a/ /k/ /aa/ /n^/	/th/ /u/ /r/ /a/ /k/ /a/ /n^/
40. วันชัย	/w/ /a/ /n^/ /ch/ /a/ /j^/	/w/ /a/ /n^/ /ch/ /a/ /j^/
41. วันพร	/w/ /a/ /n^/ /ph/ /@@/ /n^/	/w/ /a/ /n^/ /ph/ /@/ /n^/
42. วิชาญ	/w/ /i/ /ch/ /aa/ /n^/	/w/ /i/ /ch/ /a/ /n^/
43. วิษณุ	/w/ /i/ /t^/ /s/ /a/ /n/ /u/	/w/ /i/ /t^/ /s/ /a/ /n/ /u/
44. วิวัฒน์	/w/ /i/ /w/ /a/ /t^/	/w/ /i/ /w/ /a/ /t^/
45. วีระ	/w/ /ii/ /r/ /a/	/w/ /i/ /r/ /a/

จ.2 ฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

พจนานุกรมของคำศัพท์ต่างๆ ในฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทยสามารถแสดงได้ดังตารางผนวก จ.2

ตารางผนวก จ.2 พจนานุกรมสำหรับฐานข้อมูลเสียงพูดเกี่ยวกับสัตว์ภาษาไทย

คำศัพท์	พจนานุกรม โดยใช้ชุดหน่วยเสียงมาตรฐาน	พจนานุกรม โดยใช้ชุดหน่วยเสียงที่ถูกทำการลดทอน
1. ออก	/#/ /@@/ /k^/	/#/ /@/ /k^/
2. ออกมา	/#/ /@@/ /k^/ /m/ /aa/	/#/ /@/ /k^/ /m/ /a/
3. อะไร	/#/ /a/ /r/ /a/ /j^/	/#/ /a/ /r/ /a/ /j^/
4. อายุ	/#/ /aa/ /j/ /u/	/#/ /a/ /j/ /u/
5. อาหาร	/#/ /aa/ /h/ /aa/ /n^/	/#/ /a/ /h/ /a/ /n^/
6. เอา	/#/ /a/ /w^/	/#/ /a/ /w^/
7. อีก	/#/ /ii/ /k^/	/#/ /i/ /k^/
8. อบอุ่น	/#/ /o/ /p^/ /#/ /u/ /n^/	/#/ /o/ /p^/ /#/ /u/ /n^/
9. อดทน	/#/ /o/ /t^/ /th/ /o/ /n^/	/#/ /o/ /t^/ /th/ /o/ /n^/
10. อุ้ง	/#/ /u/ /ng^/	/#/ /u/ /ng^/
11. บารมี	/b/ /aa/ /r/ /a/ /m/ /ii/	/b/ /a/ /r/ /a/ /m/ /i/
12. บ้าน	/b/ /aa/ /n^/	/b/ /a/ /n^/
13. บาง	/b/ /aa/ /ng^/	/b/ /a/ /ng^/
14. บางที	/b/ /aa/ /ng^/ /th/ /ii/	/b/ /a/ /ng^/ /th/ /i/
15. บรรทุก	/b/ /a/ /n^/ /th/ /u/ /k^/	/b/ /a/ /n^/ /th/ /u/ /k^/

16. เบา	/b/ /a/ /w^/	/b/ /a/ /w^/
17. บน	/b/ /o/ /n^/	/b/ /o/ /n^/
18. โบราณ	/b/ /oo/ /r/ /aa/ /n^/	/b/ /o/ /r/ /a/ /n^/
19. จะ	/c/ /a/	/c/ /a/
20. จมูก	/c/ /a/ /m/ /uu/ /k^/	/c/ /a/ /m/ /u/ /k^/
21. จำ	/c/ /a/ /m^/	/c/ /a/ /m^/
22. จังหวัด	/c/ /a/ /ng^/ /w/ /a/ /t^/	/c/ /a/ /ng^/ /w/ /a/ /t^/
23. จับ	/c/ /a/ /p^/	/c/ /a/ /p^/
24. จัด	/c/ /a/ /t^/	/c/ /a/ /t^/
25. เจ้าของ	/c/ /a/ /w^/ /kh/ /@@/ /ng^/	/c/ /a/ /w^/ /kh/ /@/ /ng^/
26. ชอบ	/ch/ /@@/ /p^/	/ch/ /@/ /p^/
27. ฉลาด	/ch/ /a/ /l/ /aa/ /t^/	/ch/ /a/ /l/ /a/ /t^/
28. ช้าง	/ch/ /aa/ /ng^/	/ch/ /a/ /ng^/
29. ช้างพัง	/ch/ /aa/ /ng^/ /ph/ /a/ /ng^/	/ch/ /a/ /ng^/ /ph/ /a/ /ng^/
30. ช้างพลาย	/ch/ /aa/ /ng^/ /ph/ /l/ /aa/ /j^/	/ch/ /a/ /ng^/ /ph/ /l/ /a/ /j^/
31. ช้างเผือก	/ch/ /aa/ /ng^/ /ph/ /vva/ /k^/	/ch/ /a/ /ng^/ /ph/ /v/ /a/ /k^/
32. ชาวนา	/ch/ /aa/ /w^/ /n/ /aa/	/ch/ /a/ /w^/ /n/ /a/
33. เช้า	/ch/ /aa/ /w^/	/ch/ /a/ /w^/
34. ใช้	/ch/ /a/ /j^/	/ch/ /a/ /j^/
35. เช่น	/ch/ /e/ /n^/	/ch/ /e/ /n^/
36. ช่วย	/ch/ /uua/ /j^/	/ch/ /u/ /a/ /j^/
37. เชื่อง	/ch/ /vva/ /ng^/	/ch/ /v/ /a/ /ng^/
38. แช่	/ch/ /xx/	/ch/ /x/
39. จิ้งจก	/c/ /i/ /ng^/ /c/ /o/ /k^/	/c/ /i/ /ng^/ /c/ /o/ /k^/
40. จน	/c/ /o/ /n^/	/c/ /o/ /n^/
41. จุ	/c/ /u/	/c/ /u/
42. จึง	/c/ /v/ /ng^/	/c/ /v/ /ng^/
43. ใด	/d/ /a/ /j^/	/d/ /a/ /j^/
44. ได้	/d/ /aa/ /j^/	/d/ /a/ /j^/
45. ได้ยิน	/d/ /aa/ /j^/ /j/ /i/ /n^/	/d/ /a/ /j^/ /j/ /i/ /n^/
46. ดำ	/d/ /a/ /m^/	/d/ /a/ /m^/
47. ดี	/d/ /ii/	/d/ /i/
48. เดิน	/d/ /qq/ /n^/	/d/ /q/ /n^/
49. เดินทาง	/d/ /qq/ /n^/ /th/ /aa/ /ng^/	/d/ /q/ /n^/ /th/ /a/ /ng^/
50. ด้วย	/d/ /uua/ /j^/	/d/ /u/ /a/ /j^/
51. ดื่ม	/d/ /vv/ /m^/	/d/ /v/ /m^/
52. แดด	/d/ /xx/ /t^/	/d/ /x/ /t^/
53. ฝูง	/f/ /uu/ /ng^/	/f/ /u/ /ng^/
54. หายาก	/h/ /aa/ /j/ /aa/ /k^/	/h/ /a/ /j/ /a/ /k^/
55. หากิน	/h/ /aa/ /k/ /i/ /n^/	/h/ /a/ /k/ /i/ /n^/

56. หายใจ	/h/ /aa/ /j^/ /c/ /a/ /j^/	/h/ /a/ /j^/ /c/ /a/ /j^/
57. หาง	/h/ /aa/ /ng^/	/h/ /a/ /ng^/
58. ให้	/h/ /a/ /j^/	/h/ /a/ /j^/
59. หัด	/h/ /a/ /t^/	/h/ /a/ /t^/
60. เห็น	/h/ /e/ /n^/	/h/ /e/ /n^/
61. หู	/h/ /uu/	/h/ /u/
62. หัว	/h/ /uua/	/h/ /u/ /a/
63. แห่ง	/h/ /x/ /ng^/	/h/ /x/ /ng^/
64. หญ้า	/j/ /aa/	/j/ /a/
65. หยาบ	/j/ /aa/ /p^/	/j/ /a/ /p^/
66. ยาว	/j/ /aa/ /w^/	/j/ /a/ /w^/
67. อย่าง	/j/ /aa/ /ng^/	/j/ /a/ /ng^/
68. ใหญ่	/j/ /a/ /j^/	/j/ /a/ /j^/
69. อยู่	/j/ /uu/	/j/ /u/
70. ยืน	/j/ /vv/ /n^/	/j/ /v/ /n^/
71. ก็	/k/ /@@/	/k/ /@/
72. กองผ้า	/k/ /@@/ /ng^/ /ph/ /aa/	/k/ /@/ /ng^/ /ph/ /a/
73. กำลัง	/k/ /a/ /m^/ /l/ /a/ /ng^/	/k/ /a/ /m^/ /l/ /a/ /ng^/
74. กัน	/k/ /a/ /n^/	/k/ /a/ /n^/
75. กับ	/k/ /a/ /p^/	/k/ /a/ /p^/
76. คอ	/kh/ /@@/	/kh/ /@/
77. ของ	/kh/ /@@/ /ng^/	/kh/ /@/ /ng^/
78. ค่อย	/kh/ /@/ /j^/	/kh/ /@/ /j^/
79. ค่อนข้าง	/kh/ /@/ /n^/ /kh/ /aa/ /ng^/	/kh/ /@/ /n^/ /kh/ /a/ /ng^/
80. ขา	/kh/ /aa/	/kh/ /a/
81. ข้าง	/kh/ /aa/ /ng^/	/kh/ /a/ /ng^/
82. คำว่า	/kh/ /a/ /m^/ /w/ /aa/	/kh/ /a/ /m^/ /w/ /a/
83. เข้า	/kh/ /a/ /w^/	/kh/ /a/ /w^/
84. เข้าไป	/kh/ /a/ /w^/ /p/ /a/ /j^/	/kh/ /a/ /w^/ /p/ /a/ /j^/
85. เขา	/kh/ /a/ /w^/	/kh/ /a/ /w^/
86. ขาว	/kh/ /aa/ /w^/	/kh/ /a/ /w^/
87. ขโมย	/kh/ /a/ /m/ /oo/ /j^/	/kh/ /a/ /m/ /o/ /j^/
88. ขี้	/kh/ /ii/	/kh/ /i/
89. เคี้ยว	/kh/ /iia/ /w^/	/kh/ /i/ /a/ /w^/
90. เคี้ยวเอื้อง	/kh/ /iia/ /w^/ /#/ /vva/ /ng^/	/kh/ /i/ /a/ /w^/ /#/ /v/ /a/ /ng^/
91. คล้าย	/khl/ /aa/ /j^/	/kh/ /l/ /a/ /j^/
92. คลำ	/khl/ /a/ /m^/	/kh/ /l/ /a/ /m^/
93. เคล้าเคลีย	/khl/ /a/ /w^/ /khl/ /iia/	/kh/ /l/ /a/ /w^/ /kh/ /l/ /i/ /a/
94. โคลน	/khl/ /oo/ /n^/	/kh/ /l/ /o/ /n^/
95. โค้ง	/kh/ /oo/ /ng^/	/kh/ /o/ /ng^/

96. คน	/kh/ /o/ /n^/	/kh/ /o/ /n^/
97. ข้น	/kh/ /o/ /n^/	/kh/ /o/ /n^/
98. ขน	/kh/ /o/ /n^/	/kh/ /o/ /n^/
99. คง	/kh/ /o/ /ng^/	/kh/ /o/ /ng^/
100. เคย	/kh/ /qq/ /j^/	/kh/ /q/ /j^/
101. ครั้ง	/khr/ /a/ /ng^/	/kh/ /r/ /a/ /ng^/
102. ครัว	/khr/ /uua/	/kh/ /r/ /u/ /a/
103. คู่	/kh/ /uu/	/kh/ /u/
104. ควาย	/khw/ /aa/ /j^/	/kh/ /w/ /a/ /j^/
105. ควายเผือก	/khw/ /aa/ /j^/ /ph/ /vva/ /k^/	/kh/ /w/ /a/ /j^/ /ph/ /v/ /a/ /k^/
106. ความจำ	/khw/ /aa/ /m^/ /c/ /a/ /m^/	/kh/ /w/ /a/ /m^/ /c/ /a/ /m^/
107. กีบ	/k/ /ii/ /p^/	/k/ /i/ /p^/
108. กิน	/k/ /i/ /n^/	/k/ /i/ /n^/
109. กลางวัน	/kl/ /aa/ /ng^/ /w/ /a/ /n^/	/k/ /l/ /a/ /ng^/ /w/ /a/ /n^/
110. กลางคืน	/kl/ /aa/ /ng^/ /kh/ /vv/ /n^/	/k/ /l/ /a/ /ng^/ /kh/ /v/ /n^/
111. ไกล	/kl/ /a/ /j^/	/k/ /l/ /a/ /j^/
112. เกลี้ยง	/kl/ /iia/ /ng^/	/k/ /l/ /i/ /a/ /ng^/
113. กลม	/kl/ /o/ /m^/	/k/ /l/ /o/ /m^/
114. เกลือก	/kl/ /vva/ /k^/	/k/ /l/ /v/ /a/ /k^/
115. กลืน	/kl/ /vv/ /n^/	/k/ /l/ /v/ /n^/
116. กระโดด	/kr/ /a/ /d/ /oo/ /t^/	/k/ /r/ /a/ /d/ /o/ /t^/
117. กว่า	/kw/ /aa/	/k/ /w/ /a/
118. เกวียน	/kw/ /iia/ /n^/	/k/ /w/ /i/ /a/ /n^/
119. ละ	/l/ /a/	/l/ /a/
120. ละเอียด	/l/ /a/ /#/ /iia/ /t^/	/l/ /a/ /#/ /i/ /a/ /t^/
121. ละคร	/l/ /a/ /kh/ /@@/ /n^/	/l/ /a/ /kh/ /@/ /n^/
122. ลาก	/l/ /aa/ /k^/	/l/ /a/ /k^/
123. ลำตัว	/l/ /a/ /m^/ /t/ /uua/	/l/ /a/ /m^/ /t/ /u/ /a/
124. เล็ก	/l/ /e/ /k^/	/l/ /e/ /k^/
125. เล่น	/l/ /e/ /n^/	/l/ /e/ /n^/
126. เล็บ	/l/ /e/ /p^/	/l/ /e/ /p^/
127. เลีย	/l/ /iia/	/l/ /i/ /a/
128. เลี้ยง	/l/ /iia/ /ng^/	/l/ /i/ /a/ /ng^/
129. ลิ้น	/l/ /i/ /n^/	/l/ /i/ /n^/
130. ลิง	/l/ /i/ /ng^/	/l/ /i/ /ng^/
131. ลพบุรี	/l/ /o/ /p^/ /b/ /u/ /r/ /ii/	/l/ /o/ /p^/ /b/ /u/ /r/ /i/
132. แหลม	/l/ /xx/ /m^/	/l/ /x/ /m^/
133. แหลมคม	/l/ /xx/ /m^/ /kh/ /o/ /m^/	/l/ /x/ /m^/ /kh/ /o/ /m^/
134. และ	/l/ /x/	/l/ /x/
135. มา	/m/ /aa/	/m/ /a/

136. มาก	/m/ /aa/ /k^/	/m/ /a/ /k^/
137. ไม่	/m/ /a/ /j^/	/m/ /a/ /j^/
138. แมลง	/m/ /a/ /l/ /xx/ /ng^/	/m/ /a/ /l/ /x/ /ng^/
139. มัน	/m/ /a/ /n^/	/m/ /a/ /n^/
140. มี	/m/ /ii/	/m/ /i/
141. แม่นยำ	/m/ /x/ /n^/ /j/ /a/ /m^/	/m/ /x/ /n^/ /j/ /a/ /m^/
142. แมว	/m/ /xx/ /w^/	/m/ /x/ /w^/
143. นอน	/n/ /@@/ /n^/	/n/ /@/ /n^/
144. นา	/n/ /aa/	/n/ /a/
145. น้ำ	/n/ /aa/ /m^/	/n/ /a/ /m^/
146. น้ำนม	/n/ /aa/ /m^/ /n/ /o/ /m^/	/n/ /a/ /m^/ /n/ /o/ /m^/
147. นาน	/n/ /aa/ /n^/	/n/ /a/ /n^/
148. ใน	/n/ /a/ /j^/	/n/ /a/ /j^/
149. หนัก	/n/ /a/ /k^/	/n/ /a/ /k^/
150. หนัง	/n/ /a/ /ng^/	/n/ /a/ /ng^/
151. งอก	/ng/ /@@/ /k^/	/ng/ /@/ /k^/
152. งา	/ng/ /aa/	/ng/ /a/
153. ง่าย	/ng/ /aa/ /j^/	/ng/ /a/ /j^/
154. งวง	/ng/ /uua/ /ng^/	/ng/ /u/ /a/ /ng^/
155. เหนี่ยว	/n/ /iia/ /w^/	/n/ /i/ /a/ /w^/
156. เหนียว	/n/ /iia/ /w^/	/n/ /i/ /a/ /w^/
157. นิ่ม	/n/ /i/ /m^/	/n/ /i/ /m^/
158. นิสัย	/n/ /i/ /s/ /a/ /j^/	/n/ /i/ /s/ /a/ /j^/
159. นก	/n/ /o/ /k^/	/n/ /o/ /k^/
160. นม	/n/ /o/ /m^/	/n/ /o/ /m^/
161. นมวัว	/n/ /o/ /m^/ /w/ /uua/	/n/ /o/ /m^/ /w/ /u/ /a/
162. หนู	/n/ /uu/	/n/ /u/
163. หนวด	/n/ /uua/ /t^/	/n/ /u/ /a/ /t^/
164. หนึ่ง	/n/ /v/ /ng^/	/n/ /v/ /ng^/
165. เนื้อ	/n/ /vv/ /a/	/n/ /v/ /a/
166. เนื้อควาย	/n/ /vva/ /khw/ /aa/ /j/	/n/ /v/ /a/ /kh/ /w/ /a/ /j/
167. เนื้อตัว	/n/ /vva/ /t/ /uua/	/n/ /v/ /a/ /t/ /u/ /a/
168. เนื้อวัว	/n/ /vva/ /w/ /uua/	/n/ /v/ /a/ /w/ /u/ /a/
169. เหนื่อย	/n/ /vva/ /j^/	/n/ /v/ /a/ /j^/
170. ป้อน	/p/ /@@/ /n^/	/p/ /@/ /n^/
171. ป่า	/p/ /aa/	/p/ /a/
172. ปาก	/p/ /aa/ /k^/	/p/ /a/ /k^/
173. ไป	/p/ /a/ /j^/	/p/ /a/ /j^/
174. เป็น	/p/ /e/ /n^/	/p/ /e/ /n^/
175. พาหนะ	/ph/ /aa/ /h/ /a/ /n/ /a/	/ph/ /a/ /h/ /a/ /n/ /a/

176. ผ่า	/ph/ /aa/	/ph/ /a/
177. พยายาม	/ph/ /a/ /j/ /aa/ /j/ /aa/ /m^/	/ph/ /a/ /j/ /a/ /j/ /a/ /m^/
178. ผิวหนัง	/ph/ /i/ /w^/ /n/ /a/ /ng^/	/ph/ /i/ /w^/ /n/ /a/ /ng^/
179. พลบค่ำ	/phl/ /o/ /p^/ /kh/ /a/ /m^/	/ph/ /l/ /o/ /p^/ /kh/ /a/ /m^/
180. ผลไม้	/ph/ /o/ /n^/ /l/ /a/ /m/ /a/ /j^/	/ph/ /o/ /n^/ /l/ /a/ /m/ /a/ /j^/
181. พระเจ้าอยู่หัว	/phr/ /a/ /c/ /aa/ /w^/ /j/ /u/ /h/ /uua/	/ph/ /r/ /a/ /c/ /a/ /w^/ /j/ /u/ /h/ /u/ /a/
182. พวง	/ph/ /uua/ /ng^/	/ph/ /u/ /a/ /ng^/
183. พืช	/ph/ /vv/ /t^/	/ph/ /v/ /t^/
184. ปี	/p/ /ii/	/p/ /i/
185. ปลาย่าง	/pl/ /aa/ /j/ /aa/ /ng^/	/p/ /l/ /a/ /j/ /a/ /ng^/
186. ปลาย	/pl/ /aa/ /j^/	/p/ /l/ /a/ /j^/
187. ปรากฏ	/pr/ /aa/ /k/ /o/ /t^/	/p/ /r/ /a/ /k/ /o/ /t^/
188. ร้อย	/r/ /@@/ /j^/	/r/ /@/ /j^/
189. ร้อน	/r/ /@@/ /n^/	/r/ /@/ /n^/
190. รัก	/r/ /a/ /k^/	/r/ /a/ /k^/
191. เรา	/r/ /a/ /w^/	/r/ /a/ /w^/
192. เร็ว	/r/ /e/ /w^/	/r/ /e/ /w^/
193. เรียกว่า	/r/ /iia/ /k^/ /w/ /aa/	/r/ /i/ /a/ /k^/ /w/ /a/
194. รีบ	/r/ /ii/ /p^/	/r/ /i/ /p^/
195. ริม	/r/ /i/ /m^/	/r/ /i/ /m^/
196. รู	/r/ /uu/	/r/ /u/
197. รวม	/r/ /uua/ /m^/	/r/ /u/ /a/ /m^/
198. รูปร่าง	/r/ /uu/ /p^/ /r/ /aa/ /ng^/	/r/ /u/ /p^/ /r/ /a/ /ng^/
199. หรือ	/r/ /vv/	/r/ /v/
200. สอง	/s/ /@@/ /ng^/	/s/ /@/ /ng^/
201. สะอาด	/s/ /a/ /#/ /aa/ /t^/	/s/ /a/ /#/ /a/ /t^/
202. สมัย	/s/ /a/ /m/ /a/ /j^/	/s/ /a/ /m/ /a/ /j^/
203. เสมอ	/s/ /a/ /m/ /qq/	/s/ /a/ /m/ /q/
204. สำรอก	/s/ /a/ /m^/ /r/ /@@/ /k^/	/s/ /a/ /m^/ /r/ /@/ /k^/
205. สำหรับ	/s/ /a/ /m^/ /r/ /a/ /p^/	/s/ /a/ /m^/ /r/ /a/ /p^/
206. สั้น	/s/ /a/ /n^/	/s/ /a/ /n^/
207. สัตว์	/s/ /a/ /t^/	/s/ /a/ /t^/
208. สัตว์ป่า	/s/ /a/ /t^/ /p/ /aa/	/s/ /a/ /t^/ /p/ /a/
209. เสียง	/s/ /iia/ /ng^/	/s/ /i/ /a/ /ng^/
210. สงคราม	/s/ /o/ /ng^/ /khr/ /aa/ /m^/	/s/ /o/ /ng^/ /kh/ /r/ /a/ /m^/
211. ซุกซน	/s/ /u/ /k^/ /s/ /o/ /n^/	/s/ /u/ /k^/ /s/ /o/ /n^/
212. สวนสัตว์	/s/ /uua/ /n^/ /s/ /a/ /t^/	/s/ /u/ /a/ /n^/ /s/ /a/ /t^/
213. ซุง	/s/ /u/ /ng^/	/s/ /u/ /ng^/
214. เสือ	/s/ /vva/	/s/ /v/ /a/
215. ต้อง	/t/ /@/ /ng^/	/t/ /@/ /ng^/

216. ตะครุบ	/t/ /a/ /khr/ /u/ /p^/	/t/ /a/ /kh/ /r/ /u/ /p^/
217. ตา	/t/ /aa/	/t/ /a/
218. ตาม	/t/ /aa/ /m^/	/t/ /a/ /m^/
219. ต่าง	/t/ /aa/ /ng^/	/t/ /a/ /ng^/
220. ไต่	/t/ /a/ /j^/	/t/ /a/ /j^/
221. ตั้ง	/t/ /a/ /ng^/	/t/ /a/ /ng^/
222. ตั้งแต่	/t/ /a/ /ng^/ /t/ /xx/	/t/ /a/ /ng^/ /t/ /x/
223. เตาไฟ	/t/ /a/ /w^/ /f/ /a/ /j^/	/t/ /a/ /w^/ /f/ /a/ /j^/
224. ท่า	/th/ /aa/	/th/ /a/
225. ทาง	/th/ /aa/ /ng^/	/th/ /a/ /ng^/
226. ทำ	/th/ /a/ /m^/	/th/ /a/ /m^/
227. ทำงาน	/th/ /a/ /m^/ /ng/ /aa/ /n^/	/th/ /a/ /m^/ /ng/ /a/ /n^/
228. ไถ	/th/ /a/ /j^/	/th/ /a/ /j^/
229. ที่	/th/ /ii/	/th/ /i/
230. ที่อยู่	/th/ /ii/ /j/ /u/	/th/ /i/ /j/ /u/
231. ที่สุด	/th/ /ii/ /s/ /u/ /t^/	/th/ /i/ /s/ /u/ /t^/
232. ทน	/th/ /o/ /n^/	/th/ /o/ /n^/
233. ถือว่า	/th/ /vv/ /w/ /aa/	/th/ /v/ /w/ /a/
234. ตืน	/t/ /ii/ /n^/	/t/ /i/ /n^/
235. ติด	/t/ /i/ /t^/	/t/ /i/ /t^/
236. ต้นไม้	/t/ /o/ /n^/ /m/ /aa/ /j^/	/t/ /o/ /n^/ /m/ /a/ /j^/
237. ตัว	/t/ /uua/	/t/ /u/ /a/
238. ตัวผู้	/t/ /uua/ /ph/ /uu/	/t/ /u/ /a/ /ph/ /u/
239. ตัวเมีย	/t/ /uua/ /m/ /iia/	/t/ /u/ /a/ /m/ /i/ /a/
240. แต่	/t/ /xx/	/t/ /x/
241. ว่องไว	/w/ /@/ /ng^/ /w/ /a/ /j^/	/w/ /@/ /ng^/ /w/ /a/ /j^/
242. ว่า	/w/ /aa/	/w/ /a/
243. ว่าง	/w/ /aa/ /ng^/	/w/ /a/ /ng^/
244. ไว้	/w/ /a/ /j^/	/w/ /a/ /j^/
245. เวลา	/w/ /ee/ /l/ /aa/	/w/ /e/ /l/ /a/

ภาคผนวก ฉ
คำศัพท์ภาษาไทย-อังกฤษ

วิทยานิพนธ์ฉบับนี้มีการใช้คำศัพท์ภาษาไทย-อังกฤษ เรียงตามลำดับคำศัพท์ภาษาไทย ดังนี้

กฎกำลังของการได้ยิน	Power Law of Hearing
กระบวนการคงที่	Stationary Process
กระบวนการไปข้างหน้า	Forward Procedure
กระบวนการมาข้างหลัง	Backward Procedure
การบีบแอมปลิจูดรากที่สาม	Cubic-root Amplitude Compression
การกรอง	Filtering
การกระจายแบบเกาส์	Gaussian Distribution
การค้นหาแบบวิเทอบี	Viterbi Search
การจับคู่แผ่นแบบ	Template Matching
การปรับแนว	Alignment
การปรับแนวแบบไม่เชิงเส้น	Non-linear Alignment
การจำแนก	Classification
การชักตัวอย่างใหม่	Re-sampling
การแทนเชิงตั้งฉาก	Orthogonal Representation
การแบ่งนับ	Quantization
การออกเสียงร่วม	Co-articulation
การแปลงฟูเรียร์แบบเร็ว	Fast Fourier Transform
การแปลงฟูเรียร์แบบไม่ต่อเนื่องผกผัน	Inverse Discrete Fourier Transform
การโปรแกรมแบบพลวัต	Dynamic Programming

การรู้จำรูปแบบ	Pattern Recognition
การรู้จำเสียงพูด	Speech Recognition
การเรียนรู้ของเครื่อง	Machine Learning
การเรียนรู้แบบแบ่งแยก	Discriminative Learning
การเรียนรู้แบบมีผู้สอน	Supervised Learning
การวาร์ปเวลาแบบพลวัต	Dynamic Time Warping
การเวียนเกิดเซปสตรอล	Cepstral Recursion
การสกัดลักษณะสำคัญ	Feature Extraction
การหาปริพันธ์ของแถบวิกฤต	Critical-band Integration
การอนุมาน	Inference
เกรเดียนต์เดสเซนต์	Gradient Descent
ขอบเขตการตัดสินใจ	Decision Boundary
ข้อมูลเข้า	Input Data
ความแปรผันทางเวลา	Temporal Variability
ความแปรผันทางเสียง	Acoustic Variability
ความน่าจะเป็นก่อน	Prior Probability
ความน่าจะเป็นควรจะเป็น	Likelihood Probability
ความน่าจะเป็นในการเปลี่ยนสถานะ	Transition Probability
ความน่าจะเป็นในการออกผลลัพธ์	Emission Probability
ความน่าจะเป็นภายหลัง	Posterior Probability
ความผิดพลาดผลบวกของกำลังสอง	Sum-of-squares Error
คาลมานฟิลเตอร์	Kalman Filter
คุณสมบัติ	Attribute
โค้งความดังเทียบเท่า	Equal-loudness Curve

เงื่อนไขการไปในทิศทางเดียว	Monotonicity Condition
เงื่อนไขขอบเขต	Boundary Condition
เงื่อนไขความต่อเนื่อง	Continuity Condition
เงื่อนไขหน้าต่างการปรับแก้	Adjustment Window Condition
ป้าย	Label
ชั้นซ่อน	Hidden Layer
ชั้นผลลัพธ์	Output Layer
ฐานข้อมูลทางภาษา	Corpus
ฐานข้อมูลเสียงพูด	Speech Corpus
ต้นไม้ตัดสินใจ	Decision Trees
ตัวเชื่อมประสาน	Interface
ตัวดำเนินการคลีน	Kleen Operator
ทฤษฎีการลู่เข้าของเพอร์เซพตรอน	Perceptron Convergence Theorem
ทฤษฎีสารสนเทศ	Information Theory
นิวรอลเน็ตเวิร์ก	Neural Networks
นิวรอลเน็ตเวิร์กชั้นเดียว	Single-layer Neural Networks
นิวรอลเน็ตเวิร์กแบบไทม์ดีเลย์	Time-delay Neural Networks
นิวรอลเน็ตเวิร์กแบบเวียนซ้ำ	Recurrent Neural Networks
นิวรอลเน็ตเวิร์กสองชั้น	Two-layer Neural Networks
นิวรอลเน็ตเวิร์กหลายชั้น	Multi-layer Neural Networks
แบบจำลองเฟ้นสุ่ม	Stochastic Model
แบบจำลองทางภาษา	Langauge Model
แบบจำลองทางภาษาแบบเอ็นแกรม	N-Gram Language Model
แบบจำลองทางเสียง	Acoustic Model

แบบจำลองทุกขั้ว	All-pole Model
แบบจำลองมาร์คอฟซ่อนตัว	Hidden Markov Model
แบบจำลองอัตตาถดถอย	Autoregressive Model
ปริภูมิอิงระยะทาง	Metric Space
ผลรวมเชิงเส้น	Linear Combination
แผนภาพฮินตัน	Hinton Diagram
เพอร์เซพตรอน	Perceptron
เพื่อนบ้านใกล้สุด	Nearest Neighbor
ฟอร์แมนต์	Formant
ฟังก์ชันกระตุ้น	Activation Function
ฟังก์ชันซิกมอยด์	Sigmoid Function
ฟังก์ชันไบโพลาร์	Bipolar Function
ฟังก์ชันวาร์ป	Warping Function
ภาษาฟอร์มัล	Formal Language
ระยะทางแบบยุคลิด	Euclidean Distance
ลักษณะสำคัญ	Feature
วิธีคาดหวัง-สูงสุด	Expectation-maximization
วิธีทางสถิติ	Statistical Method
สถานะ	State
สมมติฐานมาร์คอฟ	Markov Assumption
สมมติฐานมาร์คอฟอันดับหนึ่ง	First-order Markov Assumption
สวนศาสตร์	Acoustics
สัญญาณรบกวน	Noise
สัทศาสตร์	Phonetics

สัมประสิทธิ์อัตสหสัมพันธ์	Autocorrelation Coefficients
สเปกตรัมกำลัง	Power Spectrum
สเปกตรัมกำลังแถบวิกฤต	Critical-band Power Spectrum
สเปกโตรแกรม	Spectrogram
เส้นโค้งแถบวิกฤต	Critical Band Curve
หน่วยย่อยทางภาษา	Linguistic Unit
หน่วยเสียง	Phoneme
หน้าต่างแฮมมิง	Hamming Window
อนุกรมเวลา	Time Series
อันดับของพีแอลพี	PLP Order
อัลกอริทึมการผ่านโทเคน	Token Passing Algorithm
อัลกอริทึมบอม-เวลช์	Baum-Welch Algorithm
อัลกอริทึมไปข้างหน้า-มาข้างหลัง	Forward-backward Algorithm
เฮวิไซด์สเต็ปฟังก์ชัน	Heaviside Step Function

ประวัติผู้เขียนวิทยานิพนธ์

นายประเสริฐศักดิ์ ผุงประเสริฐยิ่ง เกิดเมื่อวันพุธที่ 25 มิถุนายน พ.ศ. 2523 ที่จังหวัดสุราษฎร์ธานี สำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2545 และเข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2546