การรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง

NEW TECHNIQUE OF SPEAKER RECOGNITION BASED ON THE CLUSTERING ANNs

สุวุฒิ ตุ้มทอง

SUWUT TUMTHONG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า

คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี

พ.ศ. 2553

การรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง

สุวุฒิ ตุ้มทอง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี
พ.ศ. 2553

NEW TECHNIQUE OF SPEAKER RECOGNITION BASED ON THE CLUSTERING ANNs

SUWUT TUMTHONG

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR

THE DEGREE OF MASTER OF ENGINEERING

IN ELECTRICAL ENGINEERING DEPARTMENT OF ELECTRICAL ENGINEERING

FACULTY OF ENGINEERING

RAJAMANGALA UNIVERSITY OF TECHNOLOGY THANYABURI

2010

วิทยานิพนธ์ฉบับนี้เป็นงานวิจัยที่เกิดจากการค้นคว้าและวิจัยขณะที่ข้าพเจ้าศึกษาอยู่ในคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี ดังนั้นงานวิจัยในวิทยานิพนธ์ฉบับนี้ถือเป็นลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรีและข้อความต่างๆ ในวิทยานิพนธ์ฉบับนี้ข้าพเจ้าขอรับรองว่าไม่มีการคัดลอกหรือนำงานวิจัยของผู้อื่นมานำเสนอในชื่อของข้าพเจ้า

สุวุฒิ ตุ้มทอง
(ผู้จัดทำวิทยานิพนธ์)

ใบรับรองวิทยานิพนธ์

คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี

หัวข้อวิทยานิพนธ์	การรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง NEW TECHNIQUE OF SPEAKER RECOGNITION BASED ON THE CLUSTERING ANNs
ชื่อนักศึกษา	นายสุวุฒิ ตุ้มทอง
รหัสประจำตัว	114870402015-1
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ดร.จักรี ศรีนนท์ฉัตร
วัน เดือน ปี ที่สอบ	11 กันยายน 2553
สถานที่สอบ	ห้อง รวงข้าว ณ อาคารเฉลิมพระเกียรติ 80 พรรษา 5 ธันวาคม 2550 คณะวิศวกรรมศาสตร์

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.สมเกียรติ อุดมหรรษากุล)

..... กรรมการ
(ดร.อำนวย เรืองวารี)

..... กรรมการ
(ดร. สุรินทร์ แหงมงาม)

..... กรรมการ
(ดร.จักรี ศรีนนท์ฉัตร)

.....
(ผู้ช่วยศาสตราจารย์ ดร.สมชัย หิรัญวโรดม)
คณบดีคณะวิศวกรรมศาสตร์

หัวข้อวิทยานิพนธ์	การรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง
นักศึกษา	นายสุวุฒิ ตุ้มทอง
รหัสประจำตัว	114870402015-1
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
แขนงวิชา	วิศวกรรมอิเล็กทรอนิกส์และโทรคมนาคม
ปีการศึกษา	2553
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ดร. จักรี ศรีนนท์ฉัตร

บทคัดย่อ

ระบบการรู้จำผู้พูด คือ ระบบการรู้จำอัตโนมัติที่สามารถแยกแยะผู้พูด โดยอาศัยคุณสมบัติเฉพาะตัวที่แตกต่างกันของสัญญาณสียง การแยกแยะผู้พูดเป็นการประมวลสัญญาณเพื่อแยกผู้พูดทั้งนี้จะต้องมีฐานข้อมูลของผู้พูดนั้น เทคนิคของระบบการรู้จำผู้พูดสามารถนำไปใช้ประโยชน์สำหรับการควบคุมและสั่งงาน เช่นการใช้เสียงสั่งการโทรศัพท์เป็นต้น มีเทคนิคอยู่หลายชนิดที่ใช้ในการประมวลสัญญาณและเก็บข้อมูลเสียงสำหรับระบบการรู้จำผู้พูดเช่น การประมาณเชิงความถี่ ฮิดเดนมาคอฟ วิธีการเทียบเคียงรูปแบบ โครงข่ายประสาทเทียม เวคเตอร์ควอนไตร์เซชั่นเป็นต้น งานวิจัยนี้ได้ศึกษาและใช้โครงข่ายประสาทเทียมสำหรับบ่งชี้และแยกแยะผู้พูด

โครงข่ายประสาทเทียมเป็นการเชื่อมต่อกันของโนดในแต่ละโนด จนกลายเป็นโครงข่ายซึ่งใช้สมการทางคณิตศาสตร์ในการประมวลผล ทั้งนี้การคำนวณของฟังก์ชั่นต่างๆขึ้นอยู่กับการต่อเชื่อมกันของโนด ในงานวิจัยนี้โครงข่ายประสาทเทียมชนิด Kohonen Self-Organizing Feature Maps (KSOFM) ได้ถูกนำมาใช้ในการศึกษาระบบการรู้จำผู้พูดทั้งนี้จำนวนโนดที่ใช้ในการทดลองนี้มี 25, 36 และ 64 โนด สัญญาณเสียงสำหรับอินพุตได้บันทึกมาจากเสียงผู้ชาย 50 คนและผู้หญิง 10 คน และแต่ละคนพูดคนละ 3 วลีโดยแต่ละวลีมีไม่น้อยกว่า 3 คำ

ผลการทดลองพบว่าการใช้โนด 25 โนดใน KSOFM ระบบได้ให้ความถูกต้องในการแยกแยะกลุ่มผู้พูดได้ 64.99% และการใช้โนด 64 โนดใน KSOFM ระบบได้ให้ความถูกต้องในการแยกแยะกลุ่มผู้พูดได้ 89.99% ทั้งนี้ระบบการรู้จำผู้พูดได้ให้ความถูกต้องในการระบุผู้พูดเฉลี่ย 78.33% อย่างไรก็ตามผลการทดลองนี้ขึ้นอยู่กับวลีของคำพูดที่ใช้ในการทดลอง ฟังก์ชั่นป้อนกลับและฟังก์ชั่นการตัดสินใจของ KSOFM

คำสำคัญ : สัญญาณเสียง การรู้จำผู้พูด Kohonen Self-Organizing Feature Maps โครงข่ายประสาทเทียม

Thesis Title :	NEW TECHNIQUE OF SPEAKER RECOGNITION BASED ON THE CLUSTERING ANNs
Student Name :	Mr. Suwut Tumthong
Student ID :	114870402015-1
Degree Award :	Master of Engineering
Study Program :	Electrical Engineering (Electronic and Telecommunication Engineering)
Academic year :	2010
Thesis Advisor :	Dr. Jakkree Srinonchat

Abstract

Speaker recognition system is the process of automatically recognizing who is speaking on the basis of individual information included in speech signal. Speaker identification is the process of determining which registered speaker provides a given utterance. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing. The various technologies used to process and store speech signal include frequency estimation, hidden Markov models, pattern matching algorithms, neural networks, representation, and Vector Quantization. This research uses Artificial Neural Networks for identification and verification speaker.

An artificial neural network (ANN) is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation. This research applies the Kohonen Self-Organizing Feature Maps (KSOFM) neural network which uses 25, 36 and 64 nodes to identify and recognize the speaker. Speech input is collected from 50 male and 10 female speakers. All speakers speak the 3 phase that each phase contains at least 3 words.

The results show that the KSOFM with 25 nodes provide the minimum accuracy to classify the group of speaker approximately 64.99%. Also the KSOFM with 64 nodes provide the maximum accuracy to classify the group of speaker approximately 89.99%. Finally, this speaker recognition system provides the average accuracy to identify the speaker 78.33%. However this is depended on the phase of speech signal and the feed back and classifies function of KSOFM.

Keywords : Speech Signal, Speaker Recognition, Kohonen Self-Organizing Feature Maps, Artificial Neural Networks.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ ดร. จักรี ศรีนนท์ฉัตร อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งได้ให้คำแนะนำข้อคิดเห็นและสนับสนุนในการทำวิจัยมาด้วยดีตลอด ผู้วิจัยจึงขอกราบขอบพระคุณมา ณ ที่นี้

กราบขอบพระคุณ คณะกรรมการสอบหัวข้อและเค้าโครงวิทยานิพนธ์ ที่ให้คำแนะนำแง่มุมที่เป็นประโยชน์ในการวิจัย

และกราบขอบพระคุณ คณะกรรมการสอบวิทยานิพนธ์ ที่ให้โอกาสในการรายงานผลการวิจัย และคำแนะนำที่เป็นประโยชน์ในทำการวิจัยครั้งต่อไป

กราบขอบพระคุณ Dr. S. Danaher และทีมงานวิจัยที่ห้องวิจัยทางด้านการอิเล็กทรอนิกส์และสื่อสารของ Northumbria University, UK ที่ช่วยให้คำแนะนำทางด้านโปรแกรม

ขอขอบคุณ พี่ เพื่อน น้องนิสิตที่ห้องปฏิบัติการและวิจัยทางด้านการประมวลผลสัญญาณ ที่ได้ช่วยเหลือเกี่ยวกับข้อมูล รวมถึงคำแนะนำต่างๆ ตลอดเวลาที่การทำวิจัยอย่างดียิ่ง

และผู้วิจัยต้องขอขอบคุณ มหาลัยวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ ที่ได้สนับสนุนทุนการศึกษา

ท้ายนี้ผู้วิจัยกราบขอบพระคุณ บิดามารดาที่ให้การสนับสนุนแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สุวุฒิ ตุ้มทอง

11 กันยายน 2553

สารบัญ

สารบัญตาราง

สารบัญรูป

คำอธิบายสัญลักษณ์และคำย่อ

ANN	Artificial Neural Network
DTW	Dynamic Time Warping
GMM	Gaussian Mixture Model
Hz	Hertz
KSOFM	Kohonen Self Organizing Feature Maps
LP	Linear Predictive
LPC	Linear Predictive Coefficient
LPC-10	Linear Predictive coding order 10
LSF	Line Spectral Frequency
LSP	Line Spectrum Pair
VQ	Vector Quantization

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันงานวิจัยการรู้จำผู้พูดได้มีการวิจัยกันอย่างแพร่หลาย แต่งานวิจัยที่สามารถนำไปใช้งานได้จริงในอุตสาหกรรมหรือการใช้งานในชีวิตประจำวันโดยส่วนใหญ่ จะเป็นการรู้จำผู้พูด โดยใช้ภาษาอังกฤษซึ่งเป็นภาษาหลักที่ใช้กันอย่างแพร่หลายทั่วโลก ซึ่งในการรู้จำภาษาอังกฤษนั้น ไม่มีเรื่องเสียง วรรณยุกต์ (Tone) ทำให้การรู้จำมีความถูกแม่นยำมาก ซึ่งแตกต่างจากภาษาไทยที่มีเรื่องของเสียงวรรณยุกต์เข้ามาเกี่ยวข้อง ทำให้งานวิจัยด้านการรู้จำเสียงผู้พูด โดยใช้ภาษาไทยยังมีความถูกต้องในการรู้จำผู้พูดที่ต่ำ ทำให้ไม่สามารถนำมาใช้ชีวิตประจำวันได้

ที่ผ่านมามีการวิจัยด้านการรู้จำเสียงภาษาไทย โดยใช้เทคนิควิธีต่างๆ ในการรู้จำเสียง เช่น โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ฮิดเดนมาร์คอฟโมเดล (Hidden Markov Model: HMM) ไดนามิกส์ไทม์วาร์ปปิง (Dynamic Time Warping: DTW) และโครงข่ายนิวโรฟัซซี (Neuro Fuzzy Networks) สามารถสรุปเป็น 2 กลุ่ม คือ

1. การรู้จำเสียงภาษาไทยทั้งคำ และคำต่อเนื่อง

การรู้จำคำโดดโดยใช้โครงข่ายประสาทเทียม [1] การรู้จำทำนองเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม [2] การรู้จำเสียงพูดคำไทยต่อเนื่องจำนวนคำศัพท์ปานกลางเฉพาะบุคคล [3] การรู้จำคำไทยหลายพยางค์แบบไม่ขึ้นกับผู้พูดโดยใช้เทคนิคแบบฟัซซีและนิวรอลเน็ตเวิร์ค [4] ระบบรู้จำเสียงภาษาไทยต่อเนื่องแบบเฉพาะบุคคลสำหรับการใช้งานอีเมลล์ [5]

2. การรู้จำหน่วยเสียงภาษาไทย ได้แก่การรู้จำหน่วยเสียงพยัญชนะ หน่วยเสียงสระ

โปรแกรมฝึกออกเสียงพยัญชนะไทยสำหรับผู้บกพร่องทางการได้ยินโดยใช้โคร่งข่ายประสาทเทียม.[6] การพัฒนาการรู้จำเสียงสำหรับพยัญชนะต้นของอัฒพยางค์ [7] การแยกหน่วยเสียงสระเดี่ยวภาษาไทย โดยใช้คุณสมบัติเฉพาะเชิงความถี่ [8] การรู้จำหน่วยเสียงสระเดี่ยวสำหรับภาษาไทยโดยการใช้ ทรานส์เฟอร์ฟังก์ชันของอวัยวะกำทอนเสียงบนสเกลบาร์ก [9]

ดังนั้นงานวิจัยนี้ จึงมีแนวคิดในการค้นหาแนวทางในการแก้ปัญหาดังที่กล่าวมา โดยงานวิจัยนี้ได้นำเสนอการรู้จำผู้พูด โดยใช้เทคนิคการรู้จำผู้พูดโดยใช้โครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง เพื่อเพิ่มประสิทธิภาพการรู้จำผู้พูดให้มีความถูกต้องแม่นยำมากยิ่งขึ้น ซึ่งเทคนิคที่ใช้เป็นเทคนิคใหม่ที่มีกระบวนการรู้จำที่แตกต่างจากงานวิจัยอื่นๆ

1.2 ความมุ่งหมายและวัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อศึกษาแนวทางเดินของสัญญาณเสียงพูดและกรรมวิธีการประมวลผลของสัญญาณ

1.2.2 เพื่อศึกษาหลักการดึงค่าลักษณะสำคัญและหลักการวิเคราะห์เสียงพูด

1.2.3 เพื่อพัฒนาระบบรู้จำผู้พูดให้มีประสิทธิภาพมากยิ่งขึ้น

1.2.4 เพื่อพัฒนาแนวทางและเทคนิคใหม่ๆ ในการรู้จำผู้พูด

1.3 ขอบเขตของการศึกษา

1.3.1 ออกแบบระบบรู้จำผู้พูด โดยใช้เสียงพูดภาษาไทยไม่น้อยกว่า 3 พยางค์ แบบขึ้นกับผู้พูด

1.3.2 ใช้หลักการวิเคราะห์แนวทางเดินของสัญญาณเสียงพูดในรูปแบบของสัมประสิทธิ์คู่เส้นสเปคตรัม (LSP) ควบคู่กับโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง และอัลกอริทึมแบบเค-มีน (K-Means Algorithm)

1.4 ขั้นตอนการศึกษา

1.4.1 ศึกษาทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์สัญญาณเสียงพูด

1.4.2 ศึกษาการใช้งานโปรแกรม GoldWave และโปรแกรม MATLAB

1.4.3 เก็บรวบรวมข้อมูลสัญญาณเสียงพูดด้วยโปรแกรม GoldWave

1.4.4 ทำการประมวลผลสัญญาณเสียงพูดด้วยโปรแกรม MATLAB

1.4.5 วิเคราะห์และสรุปผลการศึกษา

1.5 ข้อจำกัดของการศึกษา

1.5.1 กำหนดเสียงพูดเป็นภาษาไทยไม่เกิน 4 พยางค์

1.5.2 อายุของกลุ่มตัวอย่างไม่นำมาพิจารณาในการศึกษา

1.5.3 เป็นเสียงพูดภาษาไทยภาคกลาง

1.5.4 สภาพแวดล้อมในขณะพูด ต้องอยู่ในสภาพแวดล้อมที่มีเสียงรบกวนต่ำ

1.6 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1.6.1 ได้เรียนรู้แนวทางเดินของสัญญาณเสียงพูดและกรรมวิธีการประมวลผลของสัญญาณ

1.6.2 ได้เรียนรู้หลักการดึงค่าลักษณะสำคัญและหลักการวิเคราะห์เสียงพูด

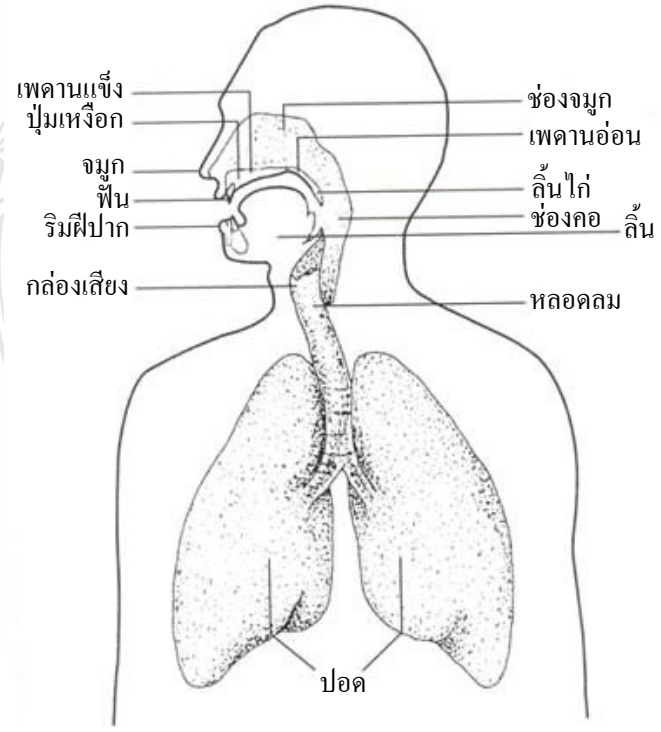
1.6.3 ได้แนวทางและเทคนิคใหม่ๆ ในการรู้จำผู้พูดสำหรับการพัฒนาระบบรู้จำผู้พูดให้มีประสิทธิภาพมากยิ่งขึ้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

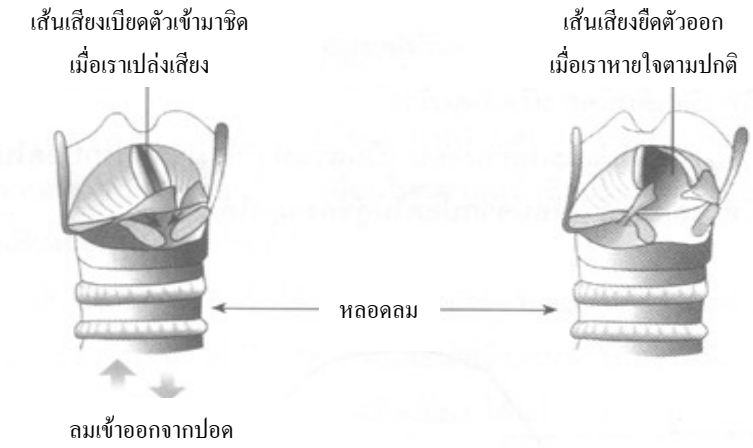
ในการดำเนินงานการศึกษาวิจัย ผู้วิจัยได้แบ่งหลักการและทฤษฎีที่เกี่ยวข้องกับวิทยานิพนธ์ออกเป็นกลุ่มๆซึ่ง ประกอบด้วยประกอบด้วยการวิเคราะห์สัญญาณเสียงพูดด้วยเทคนิคการทำนายพันธะเชิงเส้น (Linear Predictive: LPC) เทคนิคคู่เส้นสเปกตรัม (Line Spectral Pair: LSP) หรือความถี่เส้นสเปกตรัม (Line Spectral Frequency: LSF) เทคนิคการบีบอัดข้อมูลจากการเข้ารหัสสัญญาณเสียงพูดด้วยอัลกอริทึม KSOFM (Kohonen Self Organizing Feature Maps) และงานวิจัยต่างๆ ที่เกี่ยวข้องกับวิทยานิพนธ์นี้

2.1 ทฤษฎีเสียง

เสียงพูดเกิดจากการที่อวัยวะหลายส่วนในร่างกายของเราทำงานประสานกัน อวัยวะเหล่านี้จะเคลื่อนไหวตามหน้าที่ และทำให้มนุษย์สามารถเปล่งเสียงออกมาเป็นภาษาได้ [8] อวัยวะที่ช่วยในการออกเสียงดังแสดงในรูปที่ 2.1 และแสดงลักษณะของกล่องเสียงขณะออกเสียงและไม่ออกเสียงดังรูปที่ 2.2



รูปที่ 2.1 อวัยวะที่ช่วยในการออกเสียง



รูปที่ 2.2 ลักษณะของกล่องเสียงขณะออกเสียงและไม่ออกเสียง [8]

2.1.1 อวัยวะที่ใช้ในการเปล่งเสียง

การเปล่งเสียงของมนุษย์ต้องอาศัยการทำงานของอวัยวะเหล่านี้ [2] คือ

2.1.1.1 ปอดและกระบังลม ทำหน้าที่สำคัญในการหายใจ และเป็นต้นกำเนิดการไหลของอากาศในกระบวนการผลิตเสียง

2.1.1.2 หลอดลม (Larynx) ทำหน้าที่นำอากาศจากปอดผ่านกล่องเสียง และเป็นอวัยวะที่อยู่ด้านหน้าของหลอดอาหาร

2.1.1.3 กล่องเสียงและเส้นเสียง (Vocal Card) มีหน้าที่หลักในการปิดกั้นไม่ให้อาหารพลัดลงไปในหลอดลม ในการเปล่งเสียง เส้นเสียงมีหน้าที่เปลี่ยนลมจากปอดให้เป็นคลื่นเสียง เส้นเสียงทำให้เกิดข้อแตกต่างระหว่างเสียงประเภทต่างๆ

2.1.1.4 ปากและส่วนของหลอดอาหารตอนต้น อวัยวะกลุ่มนี้อยู่ต่อจากกล่องเสียง อาจเรียกว่าอวัยวะกำทอนเสียง (Vocal Tract) ทำหน้าที่กำทอนเสียง โดยทำให้กำทอนทั้งเสียงที่เกิดจากกล่องเสียงและเสียงที่เกิดภายในช่องปาก

2.1.1.5 โพรงจมูก เริ่มจากเพดานอ่อนจนถึงรูจมูกทั้งสอง ทำหน้าที่กำทอนเสียงร่วมกับช่องปากเมื่อมีการเปล่งเสียงที่ออกทางจมูก (Nasal Sounds) เช่นเสียง /ม/, /น/, และ /ง/ เป็นต้น

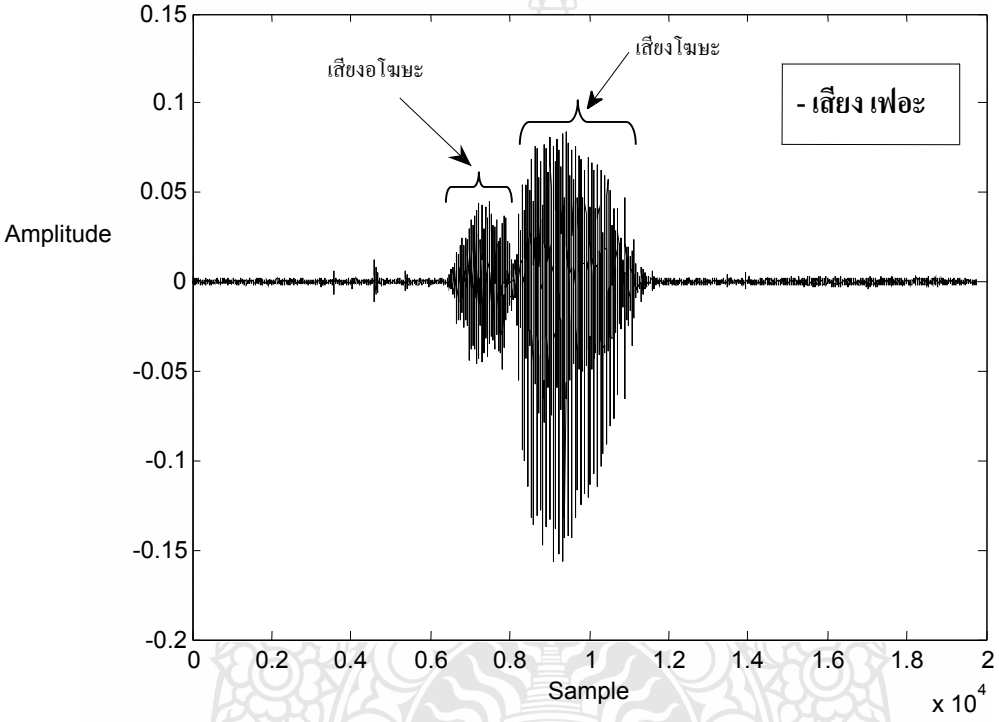
2.1.2 ลักษณะเสียงพูดของมนุษย์

เสียงพูด เป็นคลื่นตามยาว (Longitudinal Wave) เกิดจากการสั่นของอนุภาคตัวกลาง นั่นคือ อากาศ และทิศทางการสั่นของอนุภาคจะอยู่ในทิศเดียวกันกับทิศทางของการเคลื่อนที่ คลื่นเสียงเป็นคลื่นที่เปลี่ยนแปลงไปตามเวลา เสียงพูดแบ่งออกได้เป็น 2 ชนิดตามการกำเนิดเสียง (Mode) หรือการกระตุ้น คือ

2.1.2.1 เสียงก้องหรือเสียงโฆษะ (Voiced) เกิดจากการบังคับอากาศให้ผ่านช่องสายเสียงทำให้มีการเปลี่ยนแปลงความตึงหย่อนของเส้นเสียง โดยเส้นเสียงจะสั่นและเกิดเป็นพัลส์ (Pulse) ของ

อากาศไปกระตุ้นอวัยวะกำทอนเกิดเป็นเสียงก้อง ตัวอย่างเสียงก้องได้แก่ เสียงสระ เสียงพยัญชนะ ที่ต้องออกเสียงจากลำคอ (Voiced Consonants)

2.1.2.2 ไม่ก้องหรืออโฆษะ (Unvoiced หรือ Voiceless) เป็นเสียงที่ไม่เกิดจากการสั่นของเส้นเสียง แต่เกิดในช่องปากหรือโพรงจมูก โดยอวัยวะภายในช่องปาก ริมฝีปาก ขวางการไหลของอากาศให้ผ่านได้เป็นช่องเล็กๆ อากาศจึงไหลผ่านอย่างรวดเร็วและปั่นป่วนจนกระทั่งสร้างเป็นเสียงรบกวน ช่วงความถี่กว้าง (Broad-spectrum Noise) ตัวอย่างเสียงไม่ก้องได้แก่ เสียงพยัญชนะที่ไม่ได้เกิดจากลำคอ (Voiceless Consonants)



รูปที่ 2.3 ลักษณะเสียงโฆษะและเสียงอโฆษะ [8]

2.2 ลักษณะของคำไทย

สามารถพิจารณาลักษณะของเสียงต่างๆ ได้ ดังนี้

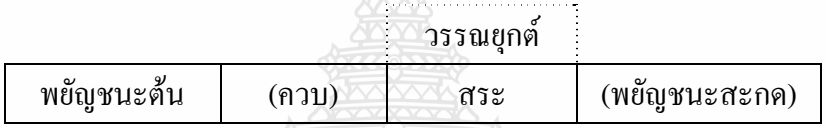
2.2.1 คำ (Word) เป็นหน่วยที่ประกอบด้วยเสียงสระ พยัญชนะ และวรรณยุกต์ เป็นอย่างน้อย และกลุ่มเสียงเหล่านี้มีความหมาย ปรากฏได้โดยลำพัง

2.2.2 พยางค์ (Syllables) พยางค์ในภาษาไทย คือ เสียงที่เปล่งออกมาครั้งหนึ่งๆ มีเสียงดังเด่น 1 เสียง และเสียงที่อยู่ข้างเคียงอย่างน้อย 2 เสียง เสียงที่ดังเด่นก็คือเสียงสระ ซึ่งลักษณะคือเป็นเสียงก้อง ดังนั้นเสียงสระจึงมักทำให้เกิดพยางค์ และพยางค์อาจจะเป็นคำได้ในกรณีที่พยางค์นั้นๆ มีความหมาย

2.2.3 อัฒพยางค์ (Demisyllables) เป็นหน่วยที่ได้จากการแบ่งครึ่งพยางค์ โดยจะตัดตรงส่วนที่เป็นเสียงสระ

2.2.4 หน่วยเสียงย่อย (Phoneme) คือ หน่วยเสียงที่ใช้สำหรับเป็นส่วนประกอบของคำ หน่วยเสียงย่อยที่สำคัญในภาษาไทย ได้แก่ หน่วยเสียงพยัญชนะ หน่วยเสียงสระ และหน่วยเสียงวรรณยุกต์

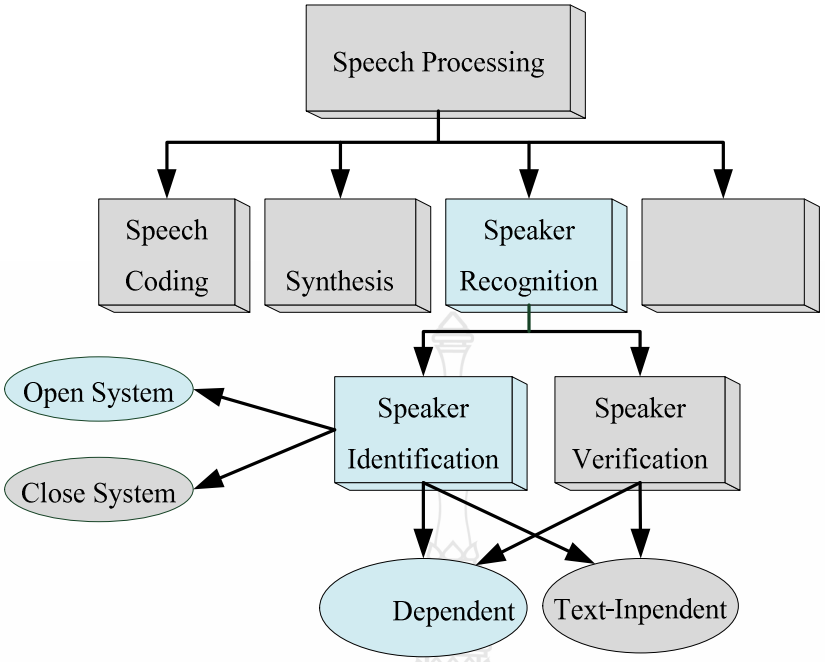
คำในภาษาไทยส่วนใหญ่จะเป็นคำพยางค์เดียว ซึ่งเป็นคำพื้นฐาน (Base Words) ของภาษา ภาษาไทยจึงจัดอยู่ในตระกูลภาษาคำโดดหรือคำพยางค์เดียว (Monosyllabic Language) [2] หน่วยเสียงที่ประกอบกันเข้าเป็นพยางค์เกิดจากการผสมกันของหน่วยเสียงหลัก 3 หน่วย คือ หน่วยเสียงพยัญชนะต้น 1 หน่วย หน่วยเสียงสระ 1 หน่วย และหน่วยเสียงวรรณยุกต์ 1 หน่วย ถ้าพยางค์นั้นมีพยัญชนะควบกล้ำและตัวสะกดหน่วยเสียงจะเป็น 5 หน่วย เพิ่มหน่วยเสียงพยัญชนะต้นที่เป็นเสียงควบกล้ำ 1 หน่วย และหน่วยเสียงพยัญชนะสะกดอีก 1 หน่วย ดังรูปที่ 2.4 [2]



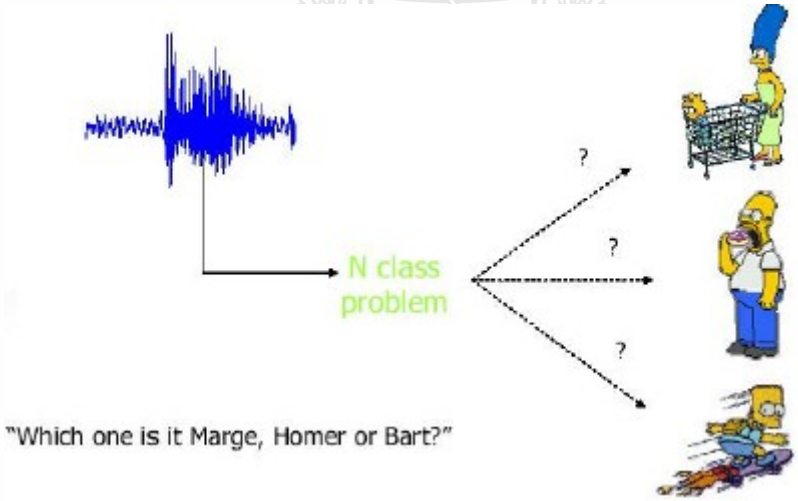
รูปที่ 2.4 องค์ประกอบของพยางค์ในภาษาไทย

2.3 การรู้จำผู้พูด (Speaker recognition)

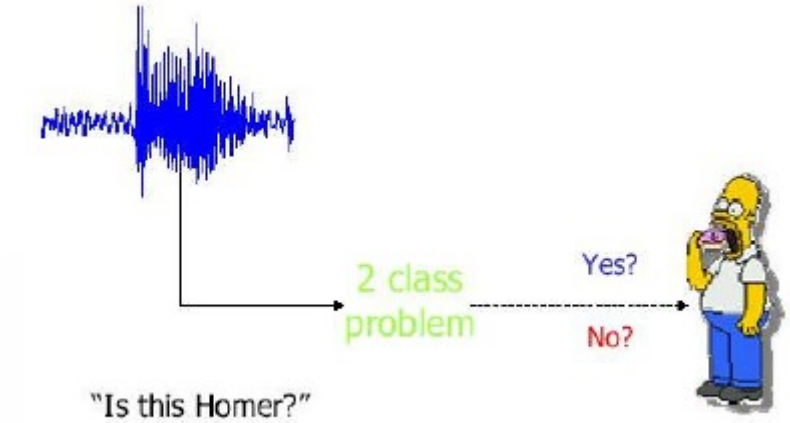
การรู้จำผู้พูดสามารถแบ่งย่อยได้เป็น 2 แบบคือการพิสูจน์ผู้พูด (Speaker Verification) ซึ่งระบบจะตอบเพียงว่าอินพุตเสียงพูดหนึ่ง ๆ เป็นเสียงของบุคคลที่กำหนดหรือไม่เท่านั้น มักจะใช้ร่วมกับรหัสของบัตรประจำตัว หรือบาร์โค้ด ส่วนอีกแบบหนึ่งเรียกว่าการบ่งชี้ผู้พูด (Speaker Identification) โดยที่ระบบจะต้องบ่งชี้ว่าเสียงอินพุตที่ได้มาเป็นเสียงใครจากข้อมูลที่มีอยู่ทั้งหมด ระบบเหล่านี้ยังแบ่งย่อยเป็นแบบกำหนดคำพูดแบบตายตัว (Text-dependent) กับแบบไม่ตายตัว (Text-independent) ซึ่งแบบแรกจะพัฒนาได้ง่ายกว่ามาก งานวิจัยด้านนี้มีปัจจัยที่นักวิจัยจะต้องทดสอบอยู่มาก เช่น คำพูดและความยาวของคำพูดที่ควรเลือกใช้เป็นอินพุตของระบบ เทคนิคการประมวลผลสัญญาณเสียงเบื้องต้น เทคนิคการบ่งชี้ผู้พูดที่เหมาะสมกับเสียงภาษาไทย ซึ่งในการคัดเลือกควรจะนำความรู้เรื่องกลสัทศาสตร์และเรื่องภาษาศาสตร์เข้ามาพิจารณาด้วย



รูปที่ 2.5 ผังระบบความเชื่อมโยง Speech Processing



รูปที่ 2.6 การบ่งชี้ผู้พูด (Speaker Identification) [6]



รูปที่ 2.7 การพิสูจน์ผู้พูด (Speaker Verification)

ระบบบ่งชี้ผู้พูดยังสามารถแบ่งย่อยได้เป็น 2 ประเภทได้แก่

1. ระบบบ่งชี้ผู้พูดแบบระบบปิด (Close system) ระบบบ่งชี้ผู้พูดแบบนี้จะบ่งชี้ว่าเสียงพูดที่เข้ามาในระบบเป็นเสียงพูดของบุคคลใดในระบบ โดยระบบจะบังคับว่าคำตอบของระบบจะเป็นเสียงของบุคคลใดบุคคลหนึ่งในระบบ (ถ้าระบบมีจำนวนผู้พูดในระบบ N คน คำตอบของระบบจะได้ N คำตอบ คือ 1,2,…,N)

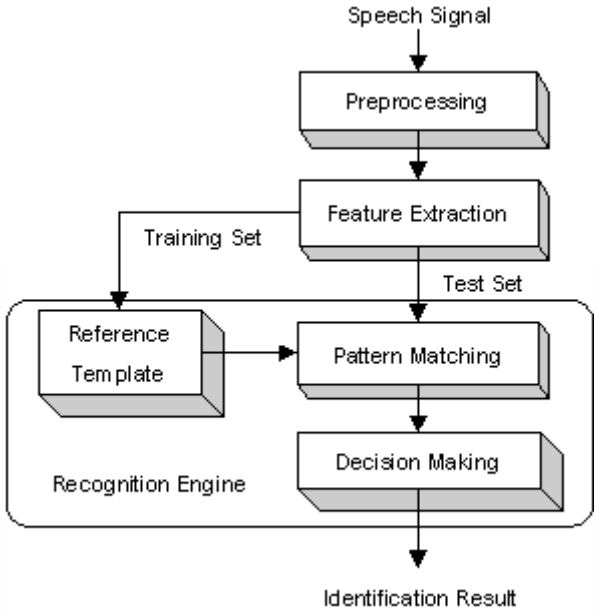
2. ระบบบ่งชี้ผู้พูดแบบระบบเปิด (Open System) ระบบบ่งชี้ผู้พูดแบบนี้จะบ่งชี้เสียงพูดที่เข้ามาในระบบเป็นเสียงพูดของบุคคลใดในระบบ แต่ก่อนตัดสินใจว่าเป็นเสียงของผู้พูดใด จะนำเสียงพูดที่เข้ามาในระบบไปผ่านขั้นตอนการตรวจสอบ (Verification) เพื่อตรวจสอบว่าเสียงที่ระบบบ่งชี้นั้นเป็นเสียงบุคคลดังกล่าวจริงหรือไม่ (ถ้าระบบมีจำนวนผู้พูดในระบบ N คน คำตอบจะมี N+1 คำตอบ คือ 1,2,…,N และไม่ใช่ผู้พูดในระบบ)

ขั้นตอน/กระบวนการในการสร้างระบบบ่งชี้ผู้พูดโดยทั่วไป

ส่วนประกอบหลักของระบบบ่งชี้ผู้พูดนั้นแสดงได้ดังในรูปข้างล่าง ซึ่งประกอบด้วย

- ส่วนประมวลผลเบื้องต้น (Preprocessing)
- การสกัดลักษณะเด่น (Feature Extraction) ของสัญญาณเสียง
- การจับคู่เพื่อบ่งชี้ผู้พูด (Pattern Matching) และการตัดสินใจเพื่อคัดเลือกเอาท์พุต (Decision Making)

สำหรับใน 2 ส่วนหลังมักจะถูกรวมเรียกว่าส่วนระบบรู้จำ (Recognition Engine) ซึ่งจะเห็นได้ว่าคล้ายคลึงและใกล้เคียงกับระบบรู้จำเสียงพูด (Speech Recognition) เป็นอย่างมาก



รูปที่ 2.8 ส่วนประกอบหลักของระบบบ่งชี้ผู้พูด

สิ่งที่ต้องศึกษาค้นคว้าวิจัยเกี่ยวกับการบ่งชี้ผู้พูดด้วยเสียงภาษาไทย ได้แก่

1. คำพูดและความยาวของคำพูดภาษาไทย

อินพุตเสียงพูดภาษาไทยสำหรับระบบบ่งชี้ผู้พูดไทยนั้น ยังต้องการการศึกษาอีก ในหลายแง่มุม เช่น พยัญชนะและสระเสียงใดที่ให้ผลดีในการบ่งชี้ผู้พูด เสียงวรรณยุกต์มีผลต่องานวิจัยแบบนี้มากน้อย อย่างไร ประโยคที่ควรเลือกใช้ควรจะมีความยาวสักเท่าไร ถ้าสั้นไปข้อมูลอาจมีน้อยจนไม่พอต่อการประมวลผล แต่ถ้ายาวเกินไปก็อาจมากเกินความจำเป็นหรือบางครั้งส่งผลให้เกิดความสับสนกับเสียงพูดของบุคคลอื่นได้ง่ายขึ้น เป็นต้น

ซึ่งในปัจจุบันได้มีงานวิจัยที่ได้วิจัยในบางเรื่องแล้วเช่น เรื่องของวรรณยุกต์มีผลต่อการรู้จำเพียงใดแต่เป็นการวิจัยที่ไม่ละเอียดเจาะลึกก็ยังต้องการการทดลองเพิ่มเติมอยู่ หรืองานวิจัยที่เกี่ยวกับความยาวของคำก็มีงานวิจัยแล้วเช่นกัน แต่เป็นการวิจัยบ่งชี้ผู้พูดด้วยเสียงตัวเลขภาษาไทย ซึ่งใช้เสียงตัวเลขภาษาไทยซึ่งเป็นเสียงโดดทำการต่อกัน (Concatenate) ยังมีงานวิจัยที่เกี่ยวกับการใช้คำภาษาไทยในการบ่งชี้ผู้พูดยังน้อยอยู่

รายละเอียดของวิธีการอัดเสียงพร้อมทั้งอุปกรณ์ที่เกี่ยวข้อง ข้อมูลเสียงพูดของเสียง ส่วนใหญ่จะแบ่งได้เป็น 2 กลุ่มคือ เสียงพูดที่จัดเก็บในสภาพแวดล้อมการทำงานทั่วไป (Office Environment) กับเสียงพูดที่จัดเก็บผ่านสายโทรศัพท์ (Via Telephone) ทั้งนี้เพราะมีงานประยุกต์ใช้หลายอย่างที่รับอินพุตจากสายโทรศัพท์ ซึ่งมักเกี่ยวเนื่องกับงานธนาคาร งานเครดิตการ์ด เช่น การโทรสอบถามยอดเงินในบัญชี หรือสอบถามยอดการใช้จ่ายเงิน เป็นต้น ดังนั้นอินพุตจากสายโทรศัพท์ก็เป็นอีกหนึ่งแนวทางซึ่งนักวิจัยในประเทศไทยที่สนใจในการทำงานวิจัยด้านนี้

2. เทคนิคการประมวลผลสัญญาณเสียงเบื้องต้น

สำหรับเสียงที่ได้จัดเก็บมาเรียบร้อยแล้ว จะถูกนำมาประมวลผลก่อนนำเข้าสู่ส่วนการบ่งชี้ผู้พูด ขั้นตอนเหล่านี้ประกอบไปด้วย การตัดหัวและท้ายของสัญญาณเสียง (End-point Detection) การกรองสัญญาณรบกวนทางความถี่ (Frequency Filtering) และการปรับความยาวเสียง (Time Normalization)

3. เทคนิคการสกัดค่าลักษณะสำคัญ

การสกัดค่าลักษณะเด่น (Feature Extraction) ซึ่งมีอยู่หลากหลายวิธี เช่น LPC, MFCC, PFL เป็นต้น ซึ่งเทคนิคต่างๆ ก็จะเหมือนๆกับการสกัดค่าลักษณะเด่นในการรู้จำเสียงพูด ทั้งนี้จะเลือกวิธีไหนนั้นก็ต้องดูด้วยว่าใช้เทคนิคการรู้จำแบบไหน

4. เทคนิคการรู้จำ

โครงข่ายประสาทเทียม (ANN) ไดนามิกไทม์วาร์ปปิง (DTW) รูปแบบฮิดเดนมาร์คอฟ (Hidden Markov Model: HMM) แบบจำลองส่วนผสมแบบเกาส์เซียน (Gaussian Mixture Model: GMM) เวกเตอร์ควอนไตซ์เซชัน (Vector Quantization: VQ) เหล่านี้ล้วนเป็นเทคนิคที่ได้รับความนิยมในการนำมาใช้ในส่วนการบ่งชี้ผู้พูดของอินพุตภาษาอังกฤษแต่สำหรับภาษาไทยก็ได้นำวิธีเหล่านี้มาทดสอบกับการบ่งชี้ผู้พูดด้วยเสียงภาษาไทย ปรากฏว่าได้ผลเป็นที่น่าพอใจ และมีข้อแนะนำที่ได้มาคือ การเลือกใช้วิธีรู้จำ ขึ้นอยู่กับข้อกำหนดของงาน เช่น DTW และ ANN เหมาะสมกับแบบกำหนดคำพูดตายตัว ในขณะที่วิธี VQ และ HMM จะเหมาะสมกับระบบงานที่เป็นแบบไม่กำหนดคำพูดมากกว่า

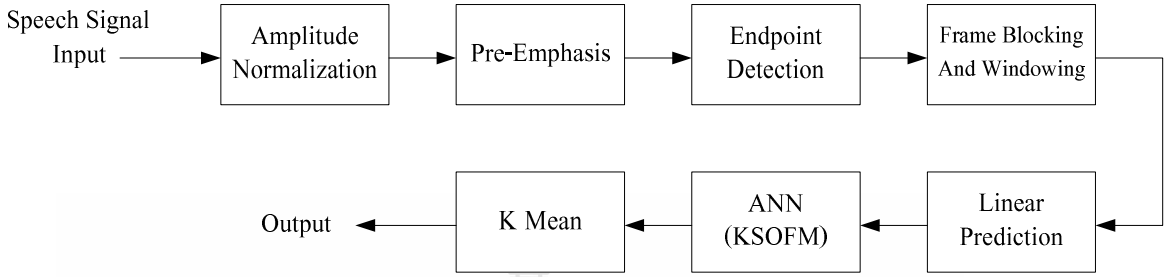
5. ปริมาณผู้พูด

เทคนิคต่าง ๆ ที่พิจารณาเลือกมาใช้กับระบบที่พัฒนา บางครั้งอาจให้ผลดีมากเมื่อทดสอบกับผู้พูดเพียง 10-20 คน แต่เมื่อนำไปทดสอบกับผู้พูดสัก 50-100 คน อาจให้ผลการบ่งชี้ที่ต่ำกว่าเดิมมากก็ได้ ทั้งนี้เนื่องจากเทคนิคบางอย่างสามารถแยกเสียงผู้พูดได้ในปริมาณจำกัดเท่านั้น ถ้ามากถึงระดับหนึ่งจะเกิดความคล้ายคลึงกันจนบ่งชี้ผู้พูดผิดพลาดไป

2.4 การทำงานของระบบรู้จำผู้พูด

ในงานวิจัยที่ดำเนินไปแล้วนั้นแบ่งออกเป็น 3 ส่วน ดังต่อไปนี้

1. การประมวลผลสัญญาณเสียงพูดเบื้องต้น
2. การดึงลักษณะสำคัญของสัญญาณเสียงพูด
3. การทดสอบความคล้ายคลึงกันของรูปแบบและกฎเกณฑ์การตัดสินใจ โดยมีการทำงานในภาพรวมของระบบรู้จำผู้พูดดังรูปที่ 2.9



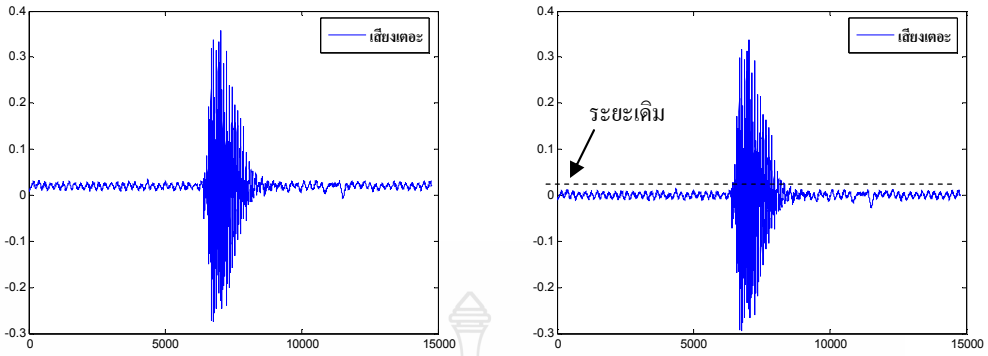
รูปที่ 2.9 การทำงานในภาพรวมของระบบรู้จำผู้พูด

2.4.1 การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

ในงานวิจัยนี้ในขั้นตอนการประมวลผลสัญญาณเบื้องต้นประกอบด้วยขั้นตอนย่อย คือ ส่วนของการปรับบรรทัดฐานแอมปลิจูด เป็นการเพิ่มลดขนาดของสัญญาณเสียงพูดเพื่อให้ขนาดของสัญญาณเสียงพูดมีความเหมาะสม ส่วนขั้นตอนถัดมา คือ กรรมวิธีการเน้นล่วงหน้าเป็นส่วนของขั้นตอนเพื่อให้สัญญาณเสียงมีความชัดเจนกว่าสัญญาณรบกวน ขั้นตอนถัดมาจะเข้ากระบวนการตัดหัวท้ายของสัญญาณเสียง เพื่อตัดสัญญาณในส่วนที่ไม่ต้องการออกไป หรือเรียกว่าเป็นการตัดส่วนที่ไม่ใช่สัญญาณเสียง (Unvoiced) ออกจากส่วนที่เป็นสัญญาณเสียง (Voice) เพื่อทำให้การประมวลผลเร็วขึ้น กรรมวิธีการตัดหัวท้ายของสัญญาณเสียงมีอยู่หลายวิธีในงานวิจัยนี้ เลือกใช้วิธีการใช้ค่าพลังงาน (Energy) เนื่องจากเป็นวิธีการคำนวณที่ไม่ยุ่งยาก ใช้เวลาน้อยและสัญญาณเสียงพูดที่นำมาวิจัย มีสัญญาณรบกวนที่มีค่าแอมปลิจูดไม่สูง โดยที่การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing) เป็นขั้นตอน เพื่อจัดเตรียมข้อมูลให้เหมาะต่อการประมวลผลในขั้นตอนต่อๆ ไป กระบวนการประมวลผลสัญญาณเบื้องต้น จะต้องอาศัยหลักการต่างๆ ดังนี้

2.4.1.1 การปรับสัญญาณสู่แกนศูนย์ ในการบันทึกเสียงแต่ละครั้งจะได้ระดับเสียงเฉลี่ยต่างกันออกไปไม่เท่ากันจำเป็นต้องปรับระดับเสียงเฉลี่ยให้เท่ากันก่อนคือปรับให้เข้าสู่แกนศูนย์ได้ตามสมการ (2.1) และรูปที่ 2.10

$$signal = signal - mean(signal) \tag{2.1}$$



(ก) (ข)

รูปที่ 2.10 การปรับสัญญาณสู่แกนศูนย์

(ก) ก่อนปรับ

(ข) หลังปรับ

2.4.1.2 ขั้นตอนกรรมวิธีการเน้นล่วงหน้า (Reemphasis)

ขั้นตอนกรรมวิธีการเน้นล่วงหน้าเป็นการบีบอัดสัญญาณเสียงพูดในช่วงพิสัยพลวัต (Dynamic Range) มีผลทำให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio) มีค่าสูงขึ้นโดยนำสัญญาณเสียงพูดผ่านวงจรกรองดิจิตอลอันดับที่หนึ่ง (First-Order Digital Filter) ที่มีฟังก์ชันถ่ายโอนดังแสดงในสมการที่ (2.2) และสมการที่ (2.3) [10] และภาพที่ 2.11 แสดงการตอบสนองต่อความถี่ของสมการ (2.3) ในการกำหนดค่า a ต่างๆ ซึ่งโดยทั่วไปค่า a จะมีค่าอยู่ระหว่าง 0.9 ถึง 1 [11]

$$H(z) = 1 - aZ^{-1} \tag{2.2}$$

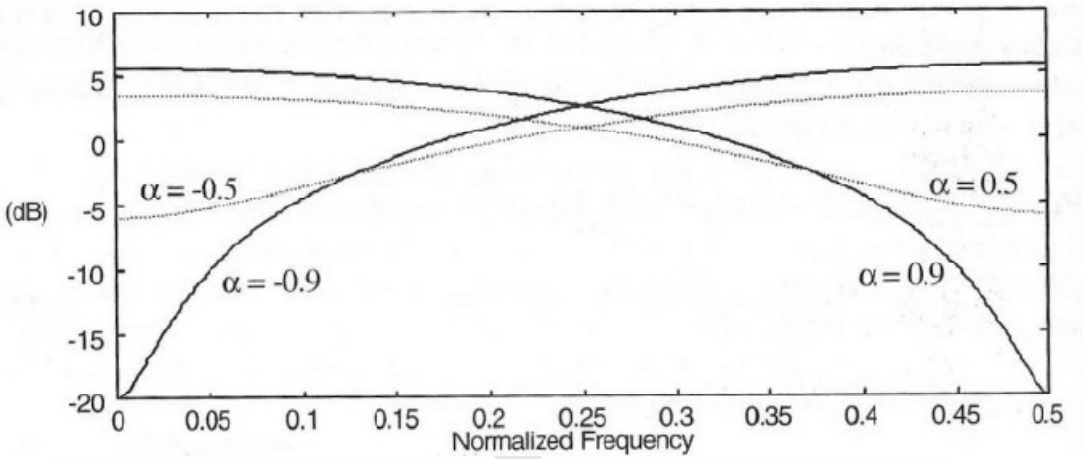
$$\widetilde{s}(n) = s(n) - as(n-1) \tag{2.3}$$

เมื่อ a เป็นค่าสัมประสิทธิ์ของวงจรกรอง

$\widetilde{s}(n)$ เป็นค่าของสัญญาณเสียงที่พูดขาออกที่ผ่านกรรมวิธีการเน้นล่วงหน้า

$s(n)$ เป็นค่าของสัญญาณเสียงพูดขาเข้าที่ n

และ $s(n-1)$ เป็นค่าของสัญญาณเสียงพูดขาเข้าที่ $n-1$



รูปที่ 2.11 การตอบสนองต่อความถี่ เมื่อเลือกใช้ค่า a ต่างๆ กัน

2.4.1.3 การวางกรอบสัญญาณ (Windowing) เนื่องจากเสียงพูดโดยรวมมีลักษณะไม่คงที่และมีการเปลี่ยนแปลงอย่างช้าๆไปตามเวลา [2] จึงจำเป็นต้องวิเคราะห์สัญญาณเสียงทีละช่วงสั้นๆ เฉพาะที่อยู่ภายในกรอบสัญญาณที่กำหนดขึ้น ขนาดของกรอบสัญญาณที่นิยมใช้ในการวิเคราะห์สัญญาณเสียงเพื่อการรู้จำมีค่าประมาณ 10-30 มิลลิวินาทีและควรวางกรอบสัญญาณถัดไปให้มีความเหลื่อมกันประมาณ 1/2 ถึง 1/3 ของขนาดกรอบสัญญาณเพื่อให้ข้อมูลที่วิเคราะห์มีความต่อเนื่องกัน [12] กรอบสัญญาณหลักมี 3 ชนิด

- กรอบสัญญาณสี่เหลี่ยม (Rectangular Window) ดังสมการที่ (2.4)

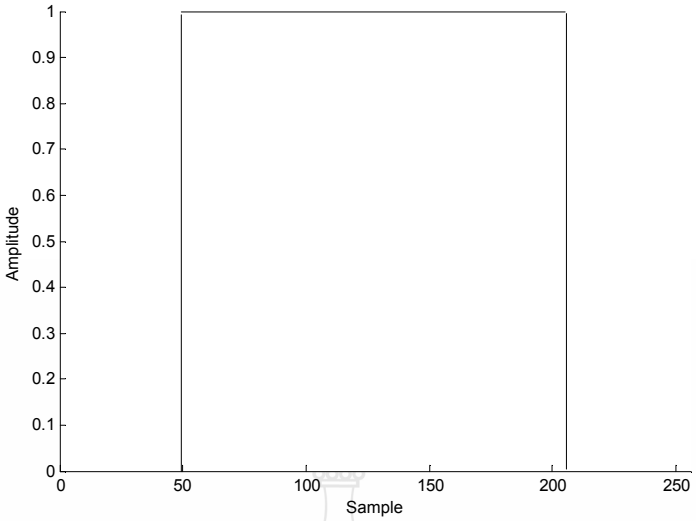
$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & Otherwise \end{cases} \tag{2.4}$$

เมื่อ $w(n)$ คือผลลัพธ์ของฟังก์ชันกรอบตำแหน่งที่ n

N คือความกว้างหน้าต่าง

n คือข้อมูลในหน้าต่าง มีค่าตั้งแต่ 0 จนถึง N -1

ซึ่งมีลักษณะดังรูปที่ 2.8



รูปที่ 2.12 ฟังก์ชันกรอบสัญญาณสี่เหลี่ยม

- กรอบสัญญาณแฮมมิง (Hamming Window) ดังสมการที่ (2.5)

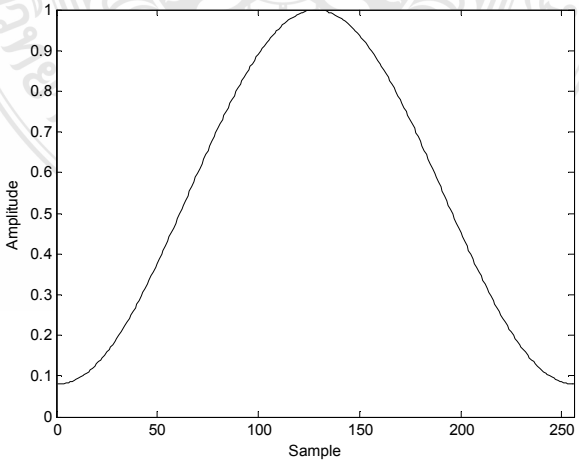
$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \qquad (2.5)$$

เมื่อ $w(n)$ คือผลลัพธ์ของฟังก์ชันกรอบตำแหน่งที่ n

N คือความกว้างหน้าต่าง

n คือข้อมูลในหน้าต่าง มีค่าตั้งแต่ 0 จนถึง N -1

ซึ่งมีลักษณะดังรูปที่ 2.9



รูปที่ 2.13 ฟังก์ชันกรอบสัญญาณแฮมมิง

- กรอบสัญญาณแฮมมิง (Hamming Window) ดังสมการที่ (2.6)

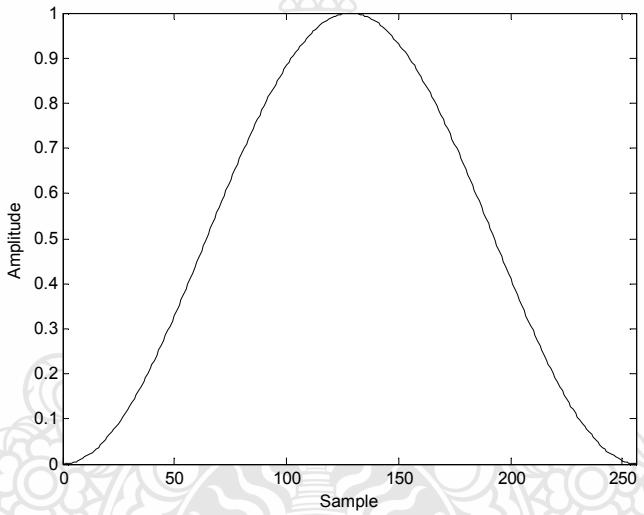
$$w(n) = \begin{cases} 0.5\left[1-\cos\left(\frac{2\pi n}{N-1}\right)\right], & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \quad (2.6)$$

เมื่อ $w(n)$ คือผลลัพธ์ของฟังก์ชันกรอบตำแหน่งที่ n

N คือความกว้างหน้าต่าง

n คือข้อมูลในหน้าต่าง มีค่าตั้งแต่ 0 จนถึง N -1

ซึ่งมีลักษณะดังรูปที่ 2.10



รูปที่ 2.14 ฟังก์ชันกรอบสัญญาณแฮมมิง

กรอบสัญญาณในแบบแฮมมิง และแฮมมิงเหมาะสำหรับการประมวลผลสัญญาณเสียง เพราะสามารถเน้นสัญญาณเสียงในกรอบที่กำลังพิจารณาให้มีความสำคัญสูงสุด โดยลดความสำคัญของสัญญาณเสียงที่อยู่ในกรอบรอบข้าง แต่ยังคงความต่อเนื่องของสัญญาณเสียงให้มีความต่อเนื่องกันในแต่ละกรอบสัญญาณทำให้เสียงที่ผ่านการวางกรอบสัญญาณยังคงข้อมูลครบถ้วนอยู่ [12]

2.4.1.4 การหาจุดสิ้นสุดของเสียงพูด (Endpoint Detection) การหาจุดสิ้นสุดของเสียงพูด ที่ใช้ในขบวนการรู้จำเป็นกระบวนการที่แยกส่วนคำพูดออกจากส่วนที่ไม่ใช่คำพูดหรือส่วนที่เป็นเสียงพื้นหลัง (Background Sound) การตัดหัวท้ายคำถือว่าเป็นกระบวนการที่สำคัญ มีผลต่ออัตราการรู้จำค่อนข้างมาก [13] และมีผลต่อเวลาในการคำนวณโดยตรง โดยเฉพาะอย่างยิ่ง การออกหน่วยเสียงพยัญชนะซึ่งเป็นเวลาช่วงสั้นๆ วิธีในการหาจุดสิ้นสุดของเสียงพูดมี 2 วิธีหลักๆ ได้แก่

- การตัดหัวท้ายคำโดยใช้ค่าพลังงาน เป็นการวิเคราะห์หาพลังงานในช่วงเวลาสั้นๆ (Short-Time Energy) เนื่องจากช่วงเสียงพูดมีค่าพลังงานมากกว่าช่วงที่ไม่ใช่คำพูด การหาค่าพลังงานสามารถเขียนเป็นสมการได้ดังสมการที่ (2.7) [14]

$$E_n = \sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2 \tag{2.7}$$

หรือเขียนเป็น $E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m)$ เมื่อ $h(n) = w^2(n)$ โดย E_n คือค่าพลังงานที่เวลา n

การหาค่าพลังงานตามวิธีการนี้จะมีความไวต่อสัญญาณที่มีแอมปลิจูดขนาดใหญ่ๆ และใช้เวลาคำนวณมากเนื่องจากการยกกำลังสอง ดังนั้นจึงแก้ปัญหาโดยการเปลี่ยนมาใช้ค่าสัมบรูณ์ (Absolute) แทนค่ากำลังสองแทนตามสมการ (2.8) ผลการวิเคราะห์ที่ได้จะเรียกว่า พลังงานสัมบรูณ์ [20]

$$M_n = \sum_{m=-\infty}^{\infty}|x(m)|w(n-m) \tag{2.8}$$

เมื่อ M_n คือพลังงานสัมบรูณ์ที่เฟรม n

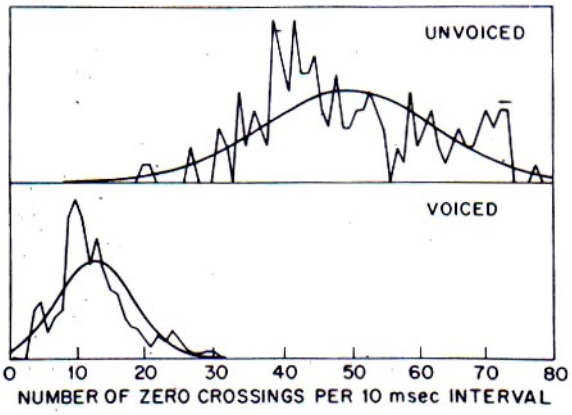
$x(m)$ คือสัญญาณเสียงที่ผ่านการปรับสู่แกนศูนย์แล้ว

$w(n)$ คือหน้าต่างกรอบสัญญาณ

โดยทั่วไปนิยมวางฟังก์ชันกรอบสัญญาณแบบสี่เหลี่ยม และแบบแฮมมิง สำหรับวิทยานิพนธ์นี้เลือกใช้การวางกรอบแบบสี่เหลี่ยม [8]

วิธีการตัดหัวท้ายคำโดยใช้พลังงานนั้นจะมีการหาค่าพลังงานสูงสุด $E_{\max}$ และค่าพลังงาน $E_{\min}$ ของเสียงพูด กำหนดค่า Upper Energy Threshold (T_u) และ Lower Energy Threshold (T_l) จากผลต่างของ $E_{\max}$ - $E_{\min}$ ดูภาพที่ 2-12 วิธีการกำหนด T_u และ T_l อยู่ในหัวข้อ 3.3.1.3

- อัตราการตัดศูนย์ (Zero Crossing Rate) คือ จำนวนครั้งของสัญญาณเสียงที่ตัดแกนศูนย์ในช่วงเวลาใดๆ ภายในหน้าต่างที่กำหนด ปกติสัญญาณเสียงที่มีค่าจุดตัดสูงจะเป็นเสียงไม่ก้องและสัญญาณเสียงที่มีค่าจุดตัดต่ำจะเป็นเสียงก้อง [14] แต่อย่างไรก็ตามการกำหนดขนาดของค่าจุดตัดศูนย์ที่แน่นอนเพื่อจำแนกชนิดของเสียงนั้น จะต้องอาศัยผลจากการทดลองเป็นหลักในภาพที่ 2.15 [14] จะแสดงให้เห็นถึงการกระจายของค่าจุดตัดศูนย์ของเสียงไม่ก้องและก้อง



รูปที่ 2.15 การกระจายของค่าจุดตัดศูนย์ของเสียงไม่ก้องและเสียงก้อง [14]

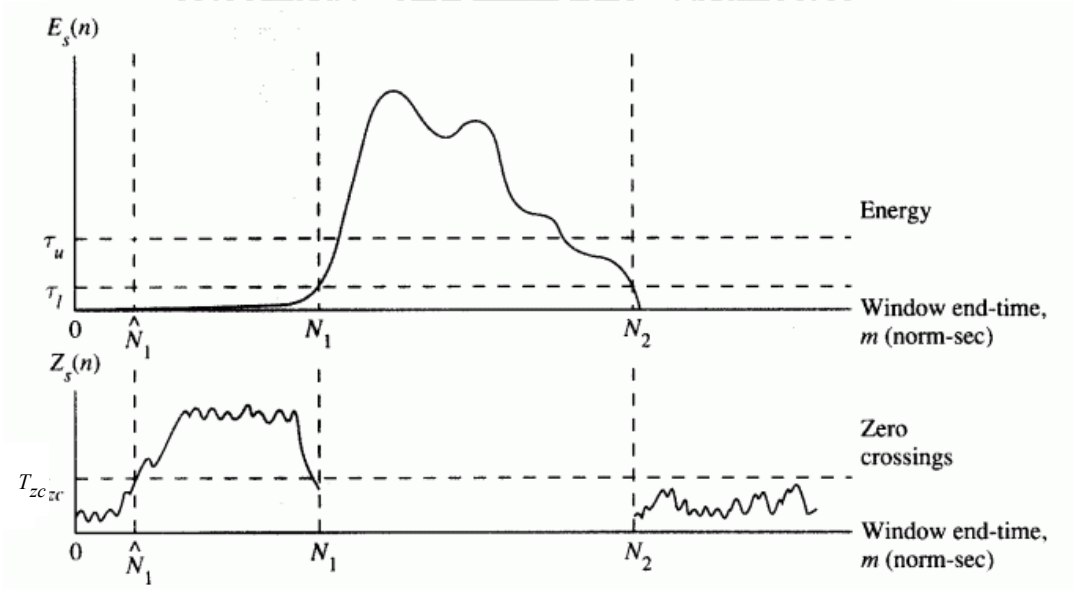
จุดตัดศูนย์สามารถนิยามได้ดังสมการที่ (2.9)

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\text{sgn}[x(m)] - \text{sgn}[x(m-1)]\right| w(n-m) \tag{2.9}$$

เมื่อ

$$\text{sgn}[x(n)] = \begin{cases} 1 & , x(n) \geq 0 \\ -1 & , x(n) < 0 \end{cases} \tag{2.10}$$

โดยที่ $w(n)$ คือฟังก์ชันกรอบสัญญาณสี่เหลี่ยม ที่ใช้กำหนดรูปร่างในการพิจารณาของสัญญาณเสียง $x(n)$ ในหนึ่งเฟรมและ N คือจำนวนตัวอย่างทั้งหมดของสัญญาณเสียงที่อยู่ในเฟรม



รูปที่ 2.16 การตัดหัวท้ายคำโดยใช้ค่าพลังงาน และอัตราการตัดศูนย์ร่วมกัน [14]

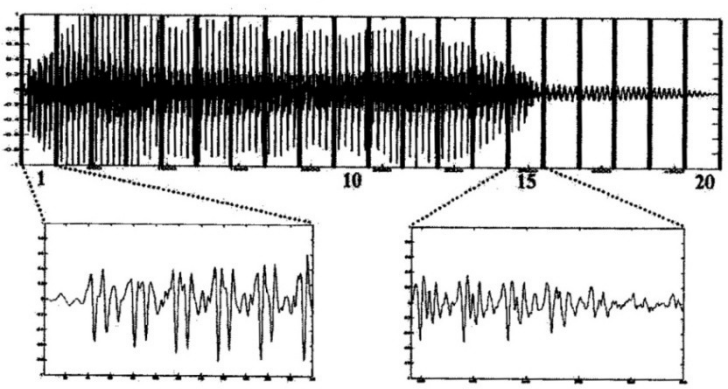
รูปที่ 2.16 แสดงการตัดหัวท้ายคำโดยใช้ค่าพลังงาน และอัตราการตัดศูนย์ร่วมกัน เพื่อหาจุดเริ่มต้นและสิ้นสุดของเสียงพูด การหาตำแหน่งเริ่มต้นและสิ้นสุดใช้พลังงานสัมบูรณ์ โดยใช้ระดับอ้างอิง T_u และ T_l เป็นตัวกำหนดตำแหน่งเริ่มต้นและสิ้นสุดของเสียงพูด ตำแหน่งแรกของพลังงานที่เริ่มมากกว่า T_u ถือเป็นตำแหน่งเริ่มต้นของคำ หลังจากนั้นถ้าตำแหน่งแรกของพลังงานที่เริ่มต้นต่ำกว่าตำแหน่ง T_l ถือเป็นตำแหน่งสิ้นสุดของคำ (จากรูปที่ 2.12 คือจุด N_1 และ N_2) ขั้นต่อไปพิจารณาช่วงต้นของสัญญาณ (จากรูปที่ 2.12 คือตำแหน่ง $\hat{N}_1$ ถึง N_1) โดยใช้อัตราการตัดศูนย์ของสัญญาณมาช่วยวิเคราะห์เพื่อให้ครอบคลุมสัญญาณเสียงที่พลังงานต่ำกว่าค่าที่กำหนดโดย T_u ด้วยการเผื่อสัญญาณไปทางซ้ายมือและใช้การพิจารณาจากอัตราการตัดศูนย์ของสัญญาณตั้งแต่ช่วง $\hat{N}_1$ ถึง N_1 หากอัตราการตัดศูนย์ของสัญญาณที่ตำแหน่งใดมีค่ามากกว่าหรือเท่ากับค่า Track Zero Crossing (T_{zc}) ก็จะกำหนดตำแหน่งนั้นเป็นตำแหน่งเริ่มต้นของเสียงพูดแทนตำแหน่งเดิม ค่า T_{zc} สามารถคำนวณได้ โดยสมมติให้ 10 กรอบหน้าต่างแรก เป็นช่วงเสียงเงียบและกำหนดให้มีอัตราการตัดศูนย์คงที่ในช่วงเวลาใดๆ ภายในหน้าต่าง [14] เช่น กำหนดอัตราการตัดศูนย์เป็น 25 ครั้งในช่วงเวลา 10 ms จากนั้นหาคำนวณค่าเฉลี่ย ($I_{zc'}$) และส่วนเบี่ยงเบนมาตรฐาน (σ) ของอัตราการตัดศูนย์ในช่วง 10 กรอบหน้าต่างแรก แล้วคำนวณ T_{zc} ได้จาก

$$T_{zc} = \min(25/10ms, I_{zc'} + 2\sigma)$$

2.4.2 การดึงลักษณะสำคัญของสัญญาณเสียงพูด

ในขั้นตอนนี้ประกอบด้วยการส่วนการแบ่งส่วนย่อย และการวางกรอบสัญญาณ (Frame Blocking and Windowing) เนื่องจากสัญญาณเสียงพูดนั้นมีการเปลี่ยนแปลงตามเวลาและไม่เสถียร ดังนั้นการประยุกต์ใช้สัญญาณเสียงพูด จึงจำเป็นจะต้องแบ่งสัญญาณเสียงพูดนั้นออกเป็นส่วนย่อย เรียกว่า เฟรม (Frame) โดยมีความยาวที่เหมาะสมประมาณ 10 ถึง 30 มิลลิวินาที

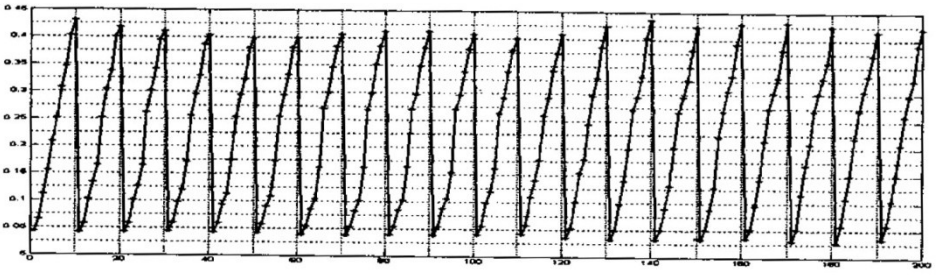
ในงานวิจัยนี้ จะใช้เฟรมที่มีความยาวดังกล่าว จึงได้มีการแบ่งเฟรมสัญญาณเสียงพูด ทุกสัญญาณออกเป็นสัญญาณละ 20 เฟรมเท่าๆ กัน ตัวอย่างการแบ่งเฟรมสัญญาณเสียงดังแสดงในรูปที่ 2.17



รูปที่ 2.17 ตัวอย่างการแบ่งเฟรมของสัญญาณเสียงพูดออกเป็น 20 เฟรมเท่าๆ กัน

หลังจากนั้นจะนำสัญญาณเสียงพูดแต่ละเฟรม คูณกับฟังก์ชั่นกรอบชนิด Hamming Window เพื่อเป็นการลดทอนแอมปลิจูดที่บริเวณปลายแต่ละข้างของเฟรม ข้อมูลเสียงพูดอย่างช้าๆ เพื่อหลีกเลี่ยงความไม่ต่อเนื่อง

หลังจากนั้นจะเป็นส่วนของขั้นตอนการหาสัมประสิทธิ์ทำนายพันธะเชิงเส้น (Linear Prediction Coefficient : LPC) ซึ่งเป็นเทคนิคสำคัญทางด้านการวิเคราะห์เสียง เนื่องจากมีความแม่นยำสูงในการประมาณค่าพารามิเตอร์ของเสียงพูด เมื่อเทียบกับความเร็วในการประมวลผล ในงานวิจัยนี้ จะใช้วิธีอัตสหสัมพันธ์ของวิธีการ Least-Squares และจะใช้ขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin ในการคำนวณหาค่าสัมประสิทธิ์ โดยจะใช้สัมประสิทธิ์ทำนายพันธะเชิงเส้น ที่ 10 อันดับ เมื่อได้สัมประสิทธิ์ LPC นำไปหาค่าสัมประสิทธิ์คู่เส้นสเปคตรัม (Line Spectral Pairs : LSP) ซึ่งในขั้นตอนนี้เป็นการนำสัมประสิทธิ์ LPC มาพัฒนาเพื่อให้ขั้นตอนการประมาณค่าพารามิเตอร์ มีความเสถียรของสัญญาณมากขึ้น นอกจากนี้ คู่เส้นสเปคตรัมยังแสดงในรูปเชิงความถี่จึงสามารถนำไปใช้ในการหาคุณสมบัติที่แน่นอนในระบบการรับรู้ของคนได้ ค่าสัมประสิทธิ์ LPC 10 อันดับที่คำนวณได้ในขั้นตอนก่อนหน้านี้ จะถูกนำมาคำนวณหาสัมประสิทธิ์คู่เส้นสเปคตรัม โดยในแต่ละเฟรมของสัญญาณเสียงพูดจะประกอบด้วยค่าสัมประสิทธิ์ LSP จำนวน 10 ค่า และเนื่องจากในแต่ละสัญญาณเสียงพูดนั้น จะถูกแบ่งออกเป็น 20 เฟรม ดังนั้นในแต่ละสัญญาณเสียงพูด จะมีค่าสัมประสิทธิ์ LSP ทั้งหมด 200 ค่า รูปที่ 2.18



รูปที่ 2.18 ค่าสัมประสิทธิ์ LSP ทั้ง 200 ค่า จากสัญญาณเสียงพูด 1 เสียง

2.4.2.1 การทำนายพันธะเชิงเส้น (Linear Predictive)

การทำนายพันธะเชิงเส้นเป็นเทคนิคที่สำคัญทางด้านการวิเคราะห์เสียงเนื่องจากมีความแม่นยำสูงในการประมาณค่าพารามิเตอร์ของเสียงพูดเมื่อเทียบกับความเร็วในการประมวลผล หลักการพื้นฐานของการทำนายพันธะเชิงเส้นอาศัยแนวความคิดว่าตัวอย่างสัญญาณเสียงพูดสามารถประมาณค่าได้จากผลรวมของตัวอย่างสัญญาณเสียงพูดจากอดีต [2] การวิเคราะห์หาพารามิเตอร์เพื่อใช้ในการทำนายโดยทั่วไปเรียกว่าการเข้ารหัสการทำนายพันธะเชิงเส้น (Linear Predictive Coding: LPC) ในด้านการประมวลผลสัญญาณเสียง การเข้ารหัสการทำนายพันธะเชิงเส้นนี้ถูกนำไปใช้ในสองแนวทาง [1] ได้แก่

- การเข้ารหัสสัญญาณเสียง โดยถูกนำไปใช้เป็นวงจรกรองวิเคราะห์การทำนายพันธะเชิงเส้น (LP Analysis Filter) เพื่อแยกส่วนซ้ำซ้อน (Redundancy) ของสัญญาณเสียงออกส่วนที่เหลือเรียกว่าสัญญาณตกค้าง (Residual Signal)

- การสังเคราะห์สัญญาณเสียงพูด โดยถูกนำไปใช้เป็นวงจรกรองการทำนายพันธะเชิงเส้นผกผัน (Inverse LP Filter) หรือวงจรกรองสังเคราะห์การทำนายพันธะเชิงเส้น (LP Synthesis Filter) โดยที่ฟังก์ชั่นถ่ายโอนของวงจรกรองดังกล่าวแสดงกรอบสเปกตรัมของสัญญาณเสียงพูด วงจรกรองสังเคราะห์การทำนายพันธะเชิงเส้นถูกใช้แสดงแทนช่องทางเดินเสียงของมนุษย์และใช้หาสัญญาณกระตุ้นที่เหมาะสม

ในการวิเคราะห์การเข้ารหัสการทำนายพันธะเชิงเส้นเริ่มต้นจากพิจารณากรอบสัญญาณเสียงที่มีตัวอย่าง N ตัวอย่าง คือ s_1 , s_2 , ... , s_N โดยอ้างว่าตัวอย่างสัญญาณปัจจุบันสามารถทำนายได้จากผลรวมของตัวอย่างสัญญาณในอดีต p ตัวอย่าง ดังสมการที่ (2.11)

$$\tilde{s}_n = -a_1 s_{n-1} - a_2 s_{n-2} - a_3 s_{n-3} - \ldots - a_p s_{n-p} = -\sum_{k=1}^{p} a_k s_{n-k} \tag{2.11}$$

เมื่อ p คืออันดับของการวิเคราะห์การทำนายพันธะเชิงเส้น และ a_1 , a_2 , ... , a_p คือสัมประสิทธิ์การเข้ารหัสการทำนายพันธะเชิงเส้น กำหนด e_n แทนค่าผิดพลาดระหว่าค่าจริงและค่าที่ทำนายได้ จะได้ตามสมการที่ (2.12) และสมการที่ (2.13)

$$e_n = s_n - \tilde{s}_n \tag{2.12}$$

$$e_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{2.13}$$

สัญญาณ e_n เรียกว่าสัญญาณตกค้าง เนื่องจากสัญญาณ e_n ได้จากผลการลบสัญญาณ s_n ด้วย $\tilde{s}_n$ และเนื่องด้วยค่าสหสัมพันธ์ช่วงสั้น (Short-Term Correlation) ระหว่างตัวอย่างของสัญญาณ

ตกค้างมีค่าต่ำ ดังนั้นประมาณได้ว่ากรอบสเปกตรัมกำลังของสัญญาณตกค้างมีลักษณะเรียบ เมื่อทำการแปลงแซด (Z-Transform) ของสมการที่ (2.14) ได้ค่าดังสมการที่ (2.15)

$$E(z) = A(z) \cdot S(z) \tag{2.14}$$

โดยที่ *S(z)* เป็นผลการแปลงแซดของสัญญาณเสียงและ *E(z)* และเป็นผลการแปลงแซดของสัญญาณตกค้างตามลำดับ

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{2.15}$$

โดยที่ *A(z)* หรือวงจรกรองไวเทนนิง (Whitening Filter) มีหน้าที่แยกความสหสัมพันธ์ช่วงสั้นที่ปรากฏในสัญญาณเสียงพูด ซึ่งก็คือการทำให้สเปกตรัมเรียบ เนื่องจาก *E(z)* ประมาณได้ว่ามีสเปกตรัมเรียบ ดังนั้นสามารถออกแบบกรอบสเปกตรัมของสัญญาณช่วงสั้นได้จากการวิเคราะห์การทำนายพันธะเชิงเส้นในรูปแบบออลโพล (All-poll Model) หรือรูปแบบออโตรีเกรสซีฟ (Autoregressive Model) ดังสมการที่ (2.16)

$$H(z) = \frac{1}{A(z)} \tag{2.16}$$

วงจรกรอง *A(z)* เรียกได้อีกชื่อหนึ่งว่าวงจรกรองผกผัน (Inverse Filter) เนื่องจากเป็นส่วนผกผันของรูปแบบออลโพล *H(z)* ของสัญญาณเสียงพูด และรากของ *A(z)* ทำให้เกิดโพลใน *H(z)* นั่นคือตำแหน่งฟอร์แมนต์ของเสียงที่ได้จากช่องทางเดินเสียงที่มีฟังก์ชั่นถ่ายโอน *H(z)*

การหากรอบสเปกตรัมกำลังช่วงสั้นของเสียงพูดด้วยวิธีการวิเคราะห์การทำนายพันธะเชิงเส้นคำนวณได้จาก *H(z)* บนวงกลมหนึ่งหน่วย (Unit Circle) โดยในขั้นแรกต้องหาสัมประสิทธิ์การทำนายพันธะเชิงเส้นของสัญญาณเสียงพูดก่อน โดยปกติหาได้จากการทำให้ค่าผิดพลาดการทำนายพันธะเชิงเส้นทั้งหมดยกกำลังสอง ดังสมการที่ (2.17) มีค่าต่ำที่สุด

$$E = \sum_{n=n_1}^{n_2} e_n^2 \tag{2.17}$$

โดยที่ผลรวมของช่วง n_1 ถึง n_2 ที่คำนวณได้ขึ้นอยู่กับวิธีการที่ใช้ ซึ่งมีอยู่ 2 วิธี [1] ได้แก่

1. วิธีอัตสหสัมพันธ์ (Autocorrelation) ในการวิเคราะห์การทำนายพันธะเชิงเส้นช่วงสั้นสามารถหาได้โดยใช้การวิเคราะห์แบบหน้าต่างสัญญาณเสียงพูดและอ้างว่าตัวอย่างสัญญาณภายนอกหน้าต่างนี้มีค่าเท่ากับศูนย์ ตามสมการที่ (2.18) แล้วจึงทำให้ได้ค่าผิดพลาดตามสมการที่ (2.17) มีค่าต่ำสุด

$$\sum_{k=1}^{p} r_{|i-k|} a_k = -r_i \qquad \text{เมื่อ } 1 \le \mathrm{i} \le \mathrm{p} \tag{2.18}$$

โดยที่ r_k คือค่าสัมประสิทธิ์อัตสหสัมพันธ์อันดับที่ k ของหน้าต่างสัญญาณเสียง โดยที่

$$r_k = \frac{1}{N} \sum_{n=k}^{N} w_n s_n w_{n-k} s_{n-k} \tag{2.19}$$

เมื่อ w_n คือฟังก์ชั่นหน้าต่างที่มีระยะเวลา N ตัวอย่าง

การหาค่าสัมประสิทธิ์การเข้ารหัสการทำนายพันธะเชิงเส้นสามารถหาได้จากการแก้สมการที่ (2.18) ซึ่งมีจำนวน p สมการ สมการดังกล่าวเรียกว่าสมการยูล-วอล์กเกอร์ (Yule-Walker) สมการทั้งหมดสามารถเขียนในรูปของเมตริกได้ดังนี้

$$Ra = -r \tag{2.20}$$

โดยที่

$$R = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots & r_{p-2} \\ r_2 & r_1 & r_0 & \cdots & r_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & r_0 \end{bmatrix} \tag{2.21}$$

$$a = [\, a_1, a_2, ..., a_p \,]^T \tag{2.22}$$

$$r = [\, r_1, r_2, ..., r_p \,]^T \tag{2.23}$$

เมตริก R เรียกว่า เมตริกอัตสหสัมพันธ์ (Autocorrelation Matrix) ซึ่งมีลักษณะโครงสร้างแบบโทพลิทซ์ (Toeplitz) โครงสร้างนี้รับรองว่าโพลของวงจรกรองสังเคราะห์การทำนายพันธะเชิงเส้น $H(z)$ จะอยู่ภายในวงกลมหนึ่งหน่วย (Unit Circle) ดังนั้นจึงรับรองได้ว่าวงจรกรองสังเคราะห์ $H(z)$ ที่ได้จากวิธีอัตสหสัมพันธ์นี้จะเสถียรเสมอ

สำหรับวิธีการคำนวณหาสัมประสิทธิ์การทำนายพันธะเชิงเส้นในสมการที่ (2.20) มีอยู่หลายวิธี และวิธีการหนึ่งที่นิยมใช้คือ วิธีการวนซ้ำของเลวินสัน-เดอบิน (Levinson-Durbin Algorithm)

ขั้นตอนวิธีการวนซ้ำของเลวินสัน-เดอบิน แบ่งออกเป็น 4 ขั้นตอน ดังนี้

ขั้นที่ 1 กำหนดค่าเริ่มต้น : $E_0 = R(0)$ และ $a_0 = 0$

ขั้นที่ 2 คำนวณหาค่าสัมประสิทธิ์การสะท้อน (Reflection coefficient)

$$k_m = \frac{R(m) - \sum_{i=1}^{m-1} a_{m-1} R(m-i)}{E_{m-1}} \quad ; \quad m = 1,2,3,...,p$$

เมื่อ *R(m)* และ *R(m-i)* คำนวณได้จากสมการ

$$R(m) = \sum_{n=m}^{N-1} x(n)x(n-m)$$

ขั้นที่ 3 คำนวณค่าสัมประสิทธิ์ของการทำนายพันธะเชิงเส้น

ให้ $a_m(m) = k_m$

และ $a_m(i) = a_{m-1}(i) - k_m a_{m-1}(m-i) \quad ; \quad 1 \le i < m$

ขั้นที่ 4 คำนวณค่าผิดพลาดใหม่

$$E_m = (1-k_m^2)E_{m-1}$$

$$m = m + 1$$

วนซ้ำขั้นที่ 2 ถึง 4 เมื่อ $m < p$

เมื่อ $m = p$ แล้ว $a_i = a_p(i)$

โดยที่ p คืออันดับของค่าสัมประสิทธิ์การทำนายพันธะเชิงเส้น

2. วิธีโคแวเรียนต์ (Covariance Method) ในการวิเคราะห์การทำนายพันธะเชิงเส้น ช่วงของการรวมอยู่ในช่วง *(p+1, N)* ดังนั้นจึงไม่จำเป็นต้องใช้หน้าต่าง การทำให้ค่าผิดพลาดทั้งหมดยกกำลังสองมีค่าต่ำที่สุดหาได้จากสมการจำนวน p สมการ ดังต่อไปนี้

$$\sum_{k=1}^{p} c_{ik} a_k = -c_{i0} \qquad \text{เมื่อ } 1 \le i \le p \tag{2.24}$$

โดยที่

$$c_{ik} = \sum_{n=p+1}^{N} s_{n-i} s_{n-k} \tag{2.25}$$

สมการจำนวน p สมการ สามารถเขียนในรูปของเมตริก ได้ดังนี้

$$Ca = -c \tag{2.26}$$

โดยที่

$$C = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1p} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2p} \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & c_{p3} & \cdots & c_{pp} \end{bmatrix} \quad (2.27)$$

$$a = [\, a_1, a_2, ..., a_p \,]^T \quad (2.28)$$

$$c = [\, c_{10}, c_{20}, ..., c_{p0} \,]^T \quad (2.29)$$

เมตริก C เรียกว่า เมตริกโคแวเรียนต์ (Covariance Matrix) และเป็นเมตริกสมมาตร นั่นคือ $c_{1p} = c_{p1}$ แต่เมตริกไม่ได้มีโครงสร้างเป็นแบบโทพลิทซ์ ดังนั้นการหาสัมประสิทธิ์การเข้ารหัสการทำนายพันธะเชิงเส้นด้วยวิธีนี้จึงมีประสิทธิภาพต่ำกว่าวิธีอัตสหสัมพันธ์ และสัมประสิทธิ์การเข้ารหัสการทำนายพันธะเชิงเส้นที่ได้จากวิธีนี้จึงรับรองไม่ได้ว่าจะให้วงจรกรองสังเคราะห์ที่มีความเสถียรเสมอ นอกจากนี้โครงสร้างที่สมมาตรทำให้การคำนวณบางส่วนสามารถใช้เทคนิคการคำนวณให้เร็วขึ้นได้ แต่ยังไม่เร็วเทียบเท่าวิธีเลวินสัน-เดอบิน

2.4.2.2 คู่เส้นสเปกตรัม (Line Spectral Pairs)

คู่เส้นสเปกตรัมหรือความถี่เส้นสเปกตรัม (Line Spectral Frequency : LSF) เป็นพารามิเตอร์รูปแบบหนึ่งที่พัฒนามาจากพารามิเตอร์การทำนายพันธะเชิงเส้น เนื่องจากพารามิเตอร์การทำนายพันธะเชิงเส้นในขั้นตอนการประมาณค่าพารามิเตอร์อาจทำให้เกิดความไม่เสถียรของสัญญาณได้ ซึ่งส่งผลต่อคุณภาพของเสียง ในขณะที่พารามิเตอร์คู่เส้นสเปกตรัมมีคุณสมบัติที่เด่นคือค่าพารามิเตอร์อยู่ภายในขอบเขตที่จำกัด มีการเรียงลำดับของค่าพารามิเตอร์ และสามารถตรวจสอบเสถียรภาพของวงจรกรองได้ง่าย นอกจากนี้คู่เส้นสเปกตรัมยังแสดงในรูปเชิงความถี่จึงสามารถนำไปใช้ในการหาคุณสมบัติที่แน่นอนในระบบการรับรู้ของคนได้ [1]

ในการคำนวณหาคู่เส้นสเปกตรัมเริ่มต้นจากพหุนามอันดับ M ของวงจรกรองผกผันในเชิงแซดดังสมการ (2.15) โดยทำการแยกส่วนสมการดังกล่าวเป็นพหุนามอันดับ $M+1$ จำนวน 2 พหุนามดังสมการที่ (2.30) และสมการที่ (2.31)

$$P(z) = A(z) + z^{-(M+1)} A(z^{-1}) \quad (2.30)$$

$$Q(z) = A(z) - z^{-(M+1)} A(z^{-1}) \quad (2.31)$$

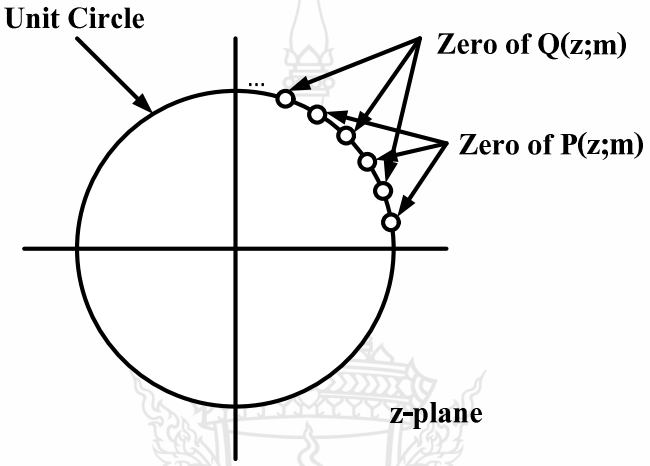
โดยพหุนาม $P(z)$ และ $Q(z)$ มีความสัมพันธ์กับ $A(z)$ ตามสมการที่ 2.32

$$A(z) = \frac{P(z) + Q(z)}{2} \quad (2.32)$$

พหุนาม *P(z)* และ *Q(z)* สอดคล้องกับรูปแบบช่องทางเดินเสียงที่ไร้การสูญเสียขณะที่ช่องระหว่างเส้นเสียง (Glottis) ปิดและเปิดตามลำดับ [4] และรากของพหุนาม *P(z)* และ *Q(z)* เรียกว่าความถี่เส้นสเปกตรัม โดยพหุนามทั้งสองมีคุณสมบัติดังต่อไปนี้

1. ราก (Zeroes) ทั้งหมดของพหุนาม *P(z)* และ *Q(z)* นั้นจะตั้งอยู่บนวงกลมหนึ่งหน่วยเสมอ

2. ราก (Zeroes) ของพหุนาม *P(z)* และ *Q(z)* จะวางเรียงสลับกันตามลำดับจากน้อยไปหามาก ดังแสดงในรูปที่ 2.19

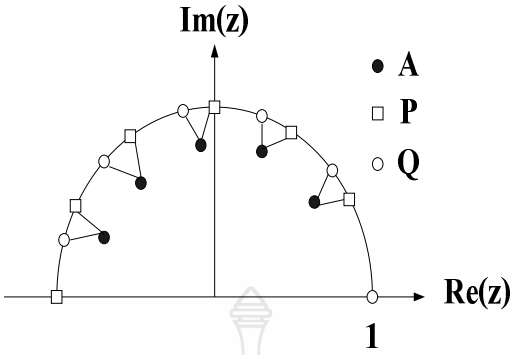


รูปที่ 2.19 การวางเรียงสลับของรากของพหุนามคู่เส้นสเปกตรัม *P(z)* และ *Q(z)*

โดยสามารถแสดงให้เห็นได้ว่าวงจรกรองผกผัน *A(z)* จะมีเฟสต่ำสุด ถ้าคู่เส้นสเปกตรัมของ *A(z)* สอดคล้องกับคุณสมบัติทั้งสองนี้ ดังนั้นวงจรสังเคราะห์การเข้ารหัสการทำนายพันธะเชิงเส้นสามารถยืนยันได้ว่ามีเสถียรภาพ โดยการทำการประมวลพารามิเตอร์การเข้ารหัสการทำนายพันธะเชิงเส้นในรูปแบบคู่เส้นสเปกตรัม [1]

เมื่อพิจารณารากของพหุนามทั้งสองพบว่าพหุนาม *P(z)* และ *Q(z)* มีรากจริงอยู่ที่ -1 และ 1 ตามลำดับ สำหรับรากอื่นๆ อยู่บนวงกลมหนึ่งหน่วยโดยวางเรียงสลับกันตามคุณสมบัติของพหุนามทั้งสอง และรากทั้งสองของพหุนามมีลักษณะเป็นคู่เชิงซ้อนสังยุค ดังนั้นในการเก็บรากของพหุนามเพื่อใช้เป็นพารามิเตอร์จึงเก็บเพียง *M* ค่า

เนื่องจากรากของพหุนาม *A(z)* แสดงตำแหน่งฟอร์แมนต์ของเสียงพูด และพหุนาม *P(z)* และ *Q(z)* สัมพันธ์กับ *A(z)* ตามสมการที่ (2.30) , สมการที่ (2.31) และสมการที่ (2.32) ดังนั้นรากของพหุนามทั้งสองจึงสัมพันธ์กับฟอร์แมนต์ด้วย โดยรากของพหุนาม *A(z)* แต่ละอันจะจับคู่กับรากของพหุนาม *P(z)* และ *Q(z)* อย่างละหนึ่งราก [4] ดังแสดงในรูปที่ 2.20



รูปที่ 2.20 ความสัมพันธ์ระหว่างรากของ *A(z)* กับรากของคู่เส้นสเปกตรัม *P(z)* และ *Q(z)*

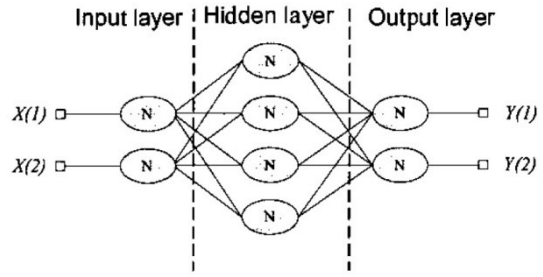
รูปที่ 2.20 รูปวงกลมสีดำจะเป็นรากของพหุนาม *A(z)*, ส่วนรูปสี่เหลี่ยมและรูปวงกลมสีขาวเป็นรากของพหุนาม *P(z)* และ *Q(z)* ตามลำดับ

การเกาะกลุ่มของรากหรือความถี่เส้นสเปกตรัมจำนวน 2 หรือ 3 รากจะกำหนดลักษณะความถี่ฟอร์แมนต์และขนาดแบนด์วิดท์ของฟอร์แมนต์ โดยขึ้นกับความชิดของกลุ่มรากดังกล่าว [1] ถ้ารากชิดกันมากจะทำให้เกิดฟอร์แมนต์และมีแบนด์วิดท์แคบ ในทางตรงข้ามกลุ่มรากจะแสดงถึงสเปกตรัมที่มีแบนด์วิดท์กว้างคือไม่เกิดฟอร์แมนต์ [3] นอกจากนี้คุณสมบัติอีกอย่างหนึ่งของคู่เส้นสเปกตรัมคือความไวทางสเปกตรัมของคู่เส้นสเปกตรัมมีลักษณะเฉพาะที่ คือเมื่อมีการเปลี่ยนแปลงคู่เส้นสเปกตรัมใดๆ จะทำให้เกิดการเปลี่ยนแปลงของสเปกตรัมกำลังของการเข้ารหัสการทำนายพันธะเชิงเส้นเฉพาะบริเวณรอบๆ เท่านั้น ทำให้สามารถทำการประเมินค่า (Quantization) ได้อย่างอิสระโดยไม่มีผลกระทบจากการลดทอนเนื่องจากการประเมินค่าจากสเปกตรัมหนึ่งไปสู่สเปกตรัมอื่นๆ [1]

2.4.3 การทดสอบความคล้ายคลึงกันของรูปแบบและกฎเกณฑ์การตัดสินใจ

2.4.3.1 โครงข่ายประสาทเทียมอัจฉริยะ (Artificial neural networks)

ขั้นตอนนี้ ในงานวิจัยนี้ได้ใช้โครงข่ายประสาทเทียมแบบก่อตัวด้วยตนเอง โดยโครงข่ายประสาทเทียมที่ใช้นี้ เป็นโครงข่ายที่มีลักษณะการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) จุดมุ่งหมายของการเรียนรู้คือ การค้นพบโครงสร้างของข้อมูล โดยในงานวิจัยนี้ใช้สถาปัตยกรรมของโครงข่ายประสาทเทียม ซึ่งแบ่งออกเป็น 3 ชั้น ประกอบด้วย ชั้นอินพุต (Input Layer) ชั้นซ่อน (Hidden Layer) ชั้นเอาต์พุต (Output Layer) ซึ่งแสดงดังรูปที่ รูปที่ 2.21



รูปที่ 2.21 สถาปัตยกรรมโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Neural networks) เป็นระบบประมวลผลสัญญาณข้อมูลต่างๆ ซึ่งประกอบด้วยจำนวนหน่วยประมวลผลพื้นฐานมากมายที่เรียกว่าปมประสาท (neurons) มีคุณลักษณะเด่นคือการกระจายข้อมูลแบบขนานไม่เป็นเชิงเส้น สามารถเชื่อมโยงกับโครงข่ายภายนอก จัดการตัวเองได้ และประมวลผลข้อมูลได้รวดเร็ว

รูปแบบการสังเคราะห์ปมประสาท เป็นพื้นฐานเบื้องต้นของการออกแบบโครงข่ายประสาทเทียม โดยค่าผลลัพธ์หรือเอาต์พุต (Output: $Y_{(t)}$) นั้นจะเกิดจากผลรวมของการคูณระหว่างค่าเวกเตอร์อินพุต $x(t)$ และค่าน้ำหนัก $w(t)$

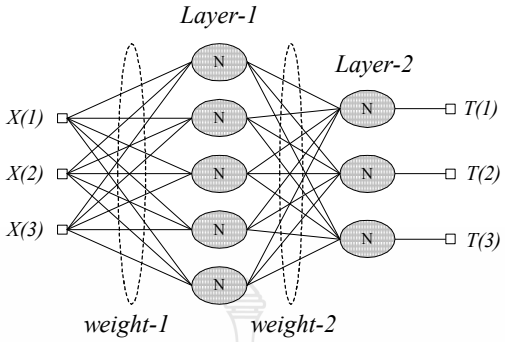
$$Y(t) = f(\sum_{i-1}^{t} x_i w_i) \tag{2.33}$$

แต่ละสมาชิกของ t ซึ่งเป็นค่าของเวกเตอร์อินพุต $X_{(t)}$ ถูกถ่วงด้วยค่า $W_{(t)}$ ซึ่งเป็นค่าน้ำหนักของเวกเตอร์ W ค่าน้ำหนักอินพุตจะถูกรวมกับฟังก์ชั่นถ่ายโอนขาเข้า พบว่าค่าฟังก์ชั่นถ่ายโอนจะให้ค่าขนาดอยู่ในขอบเขตที่จำกัด ฉะนั้นค่าเอาต์พุตของปมประสาทจึงหาค่าได้ตามขอบเขตที่กำหนด

2.4.3.2 สถาปัตยกรรมโครงข่าย (Network architectures)

สถาปัตยกรรมของโครงข่ายประสาทเทียม เป็นการเชื่อมต่อกันระหว่างปมประสาทหนึ่งกับปมประสาทหนึ่ง เพื่อที่จะกระจายหรือส่งสัญญาณ และรวมถึงการเชื่อมโยงกับโครงข่ายอื่นๆ โดยปกติแล้วปมประสาทจะถูกจัดเป็นชั้นๆ (Layer) ปมประสาทที่อยู่ในระดับชั้นเดียวกันจะส่งข้อมูลและคุณลักษณะร่วมกัน

สถาปัตยกรรมโครงข่ายประสาทเทียมพื้นฐานประกอบด้วยชั้นของข้อมูลนำเข้า (Input Layer)หนึ่งชั้นและเชื่อมต่อกับไปข้างหน้า (feed forward) ยังชั้นของข้อมูลขาออก (Output Layer) มากมาย ในกรณีที่โครงข่ายประสาทเทียมที่มีความซับซ้อนจะมีชั้นการทำงานภายใน (Hidden Layer) ตั้งแต่ 1 ชั้นขึ้นไปเพื่อช่วยในการประมวลผลและเชื่อมต่อกับโครงข่ายอื่นๆ ที่อยู่สูงขึ้นไป ทำให้มีประโยชน์มากเมื่อมีจำนวน Input Layer หลายๆ ชั้น



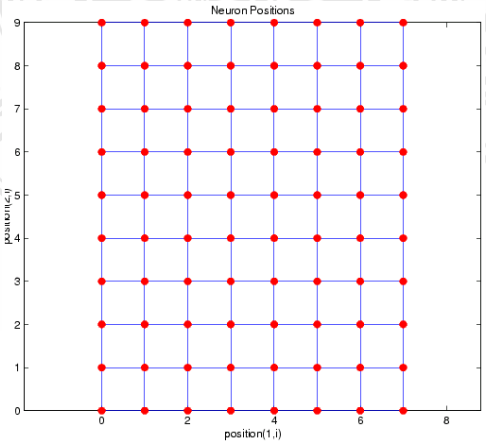
รูปที่ 2.22 Multilayer Feed Forward

โดยปกติแล้วโครงข่ายประสาทเทียมจะมีจำนวนชั้นข้อมูลมากมายหลายชั้น เมื่อนำชั้นมารวมกันก็จะทำให้ได้จำนวนปมประสาทที่แตกต่างกันมาก ในรูปที่ 2.22 ได้แสดงให้เห็นถึงสถาปัตยกรรมโครงข่ายประสาทเทียม 2 ชั้น ซึ่งแต่ละองค์ประกอบมี Input Vector $x(t)$ เป็นองค์ประกอบ ค่า $t=1,2,3$ ถูกเชื่อมต่อกับ Layer 1 และ Output ของ Layer 1 จะกลายเป็น Input ของ Layer 2 ที่เรียกว่า Hidden Layers ส่วนชั้นที่อยู่ในตำแหน่งสุดท้ายก็คือ Output Layer ดังนั้นโครงข่ายประสาทเทียมในรูปที่ 2.22 จึงมี 3 เลเยอร์

2.4.3.3 ลักษณะฟังก์ชั่นโครงสร้าง (Topology Function)

รูปแบบการเชื่อมต่อเซลล์ประสาทหรือการจัดเรียงของเซลล์ประสาทของการจัดการตนเอง (KSOFM) ถูกแบ่งตามลักษณะฟังก์ชั่นโครงสร้าง (Topology Function) ได้ 3 ลักษณะ ได้แก่ ฟังก์ชั่นโครงสร้างในแบบตาราง (Grid Topology Function) ฟังก์ชั่นโครงสร้างในแบบหกเหลี่ยม (Hexagonal Topology Function) และฟังก์ชั่นโครงสร้างในแบบสุ่ม (Random Topology Function)

1. ฟังก์ชั่นโครงสร้างแบบตาราง (Grid Topology Function)

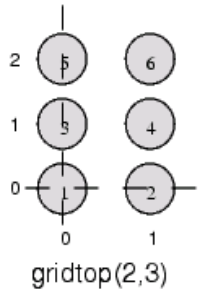


รูปที่ 2.23 การเชื่อมต่อเซลล์ประสาทตามฟังก์ชั่นโครงสร้างแบบตาราง

ฟังก์ชั่นโครงสร้างลักษณะตารางจะใช้ปมประสาทที่มีการจัดวางในลักษณะตารางสี่เหลี่ยม เช่น ปมประสาทแบบชุดเมตริกซ์ขนาด 2x3 มีจำนวนสมาชิก 6 ค่า เขียนด้วยคำสั่ง MATLAB ตามได้ดังนี้

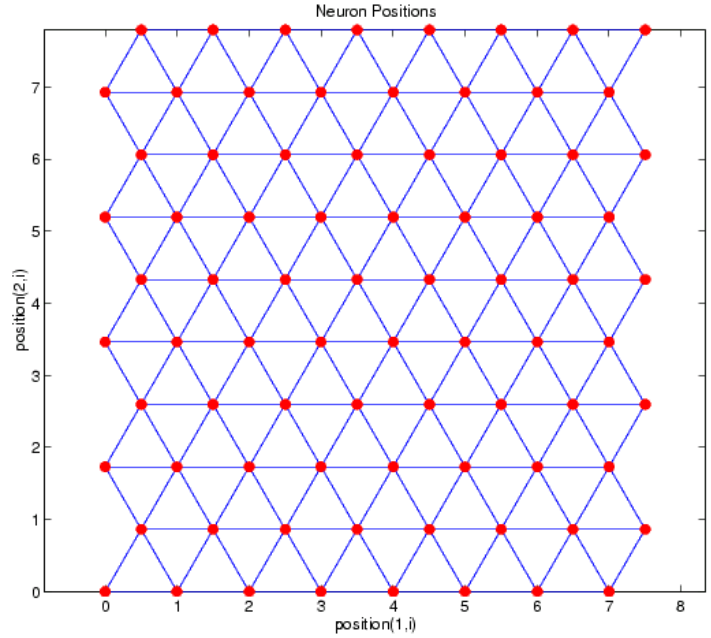
คำสั่ง	pos =	gridtop (2,3)					
แสดงผล	pos =	0	1	0	1	0	1
		0	0	1	1	2	2

เช่น ปมประสาทที่ 1 อยู่ตำแหน่ง (0,0) ปมประสาทที่ 2 อยู่ตำแหน่ง (1,0)และปมประสาทที่ 3 อยู่ที่ตำแหน่ง (0,1) แสดงได้ดังรูปที่ 2.24



รูปที่ 2.24 ปมประสาทเมตริกซ์ขนาด 2*3

2. ฟังก์ชั่นโครงสร้างลักษณะหกเหลี่ยม (Hexagonal Topology Function)

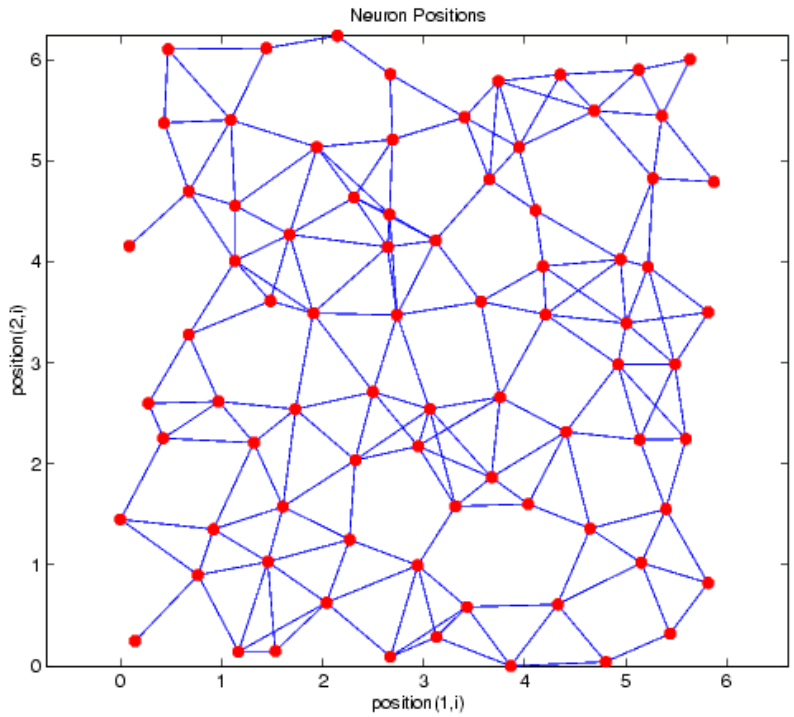


รูปที่ 2.25 การเชื่อมต่อเซลล์ประสาทตามฟังก์ชั่นโครงสร้างแบบหกเหลี่ยม

ฟังก์ชั่นโครงสร้างลักษณะหกเหลี่ยมจะใช้จัดการปมประสาทในลักษณะหกเหลี่ยม เช่น การเขียนคำสั่ง hextop เพื่อแสดงสมาชิกของปมประสาทขนาด 2x3 แบบหกเหลี่ยม

คำสั่ง	pos =	hextop(2,3)					
แสดงผล	pos =	0	1.0000	0.5000	1.5000	0	1.0000
		0	0	0.8660	0.8660	1.7321	1.7321

3. ฟังก์ชั่นโครงสร้างลักษณะสุ่ม (random topology function)



รูปที่ 2.26 การเชื่อมต่อเซลล์ประสาทตามฟังก์ชั่นโครงสร้างแบบสุ่ม

ฟังก์ชั่นโครงสร้างลักษณะสุ่มจะใช้จัดวางปมประสาทในลักษณะสุ่มค่าโดยการใช้คำสั่ง randtop เช่น การเขียนคำสั่ง randtop เพื่อแสดงสมาชิกของปมประสาทขนาด 2*3 แบบสุ่ม

คำสั่ง	pos =	randtop (2,3)					
แสดงผล	pos =	0	0.7620	0.6268	1.4218	0.0663	0.7862
		0.0925	0	0.4984	0.6007	1.1222	1.4228

2.4.4 การหาค่าระยะห่างของปมประสาท

ในโครงข่ายประสาทเทียมจะประกอบไปด้วยปมประสาท 2 ลักษณะ ได้แก่

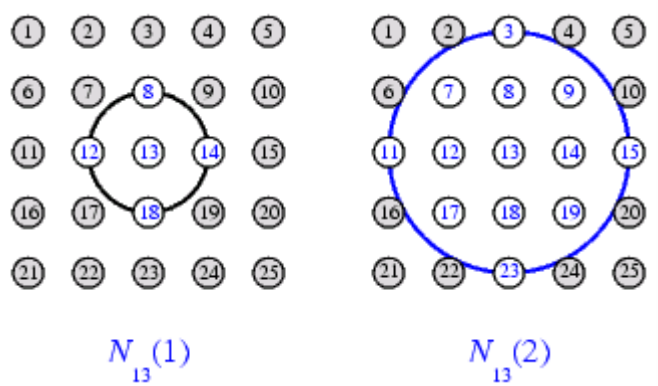
1. ปมประสาทจุดกำเนิด (Home Neuron) คือ ปมประสาทที่ใช้เป็นจุดเริ่มต้นหรือจุดศูนย์ กลางของการจัดกลุ่มการจัดการตนเอง

2. ปมประสาทข้างเคียง (Neighborhood Neuron) คือ ปมประสาทที่อยู่รอบ ๆ ปมประสาทจุดกำเนิด

แนวคิดปมประสาทจุดกำเนิดและปมประสาทข้างเคียงแสดงได้ดังรูปที่ 2.27 (ก) และ (ข) กำหนดให้ N_{13} เป็นปมประสาทจุดกำเนิด จะเห็นว่าปมประสาทข้างเคียง 1 มิติ ได้แก่ ปมประสาทที่ 8, 12, 14, 18 ปมประสาทข้างเคียง 2 มิติของ N_{13} ได้แก่ ปมประสาทที่ 3, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 23 เขียนในรูปแบบของเซตได้ ดังนี้

$$N_{13}(1) = \{8, 12, 14, 18\}$$

$$N_{13}(2) = \{3, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 23\}$$



$N_{13}(1)$ $N_{13}(2)$

(ก) (ข)

รูปที่ 2.27 ลักษณะระยะห่างของปมประสาท
(ก) แสดงปมประสาทข้างเคียง 1 มิติ
(ข) แสดงปมประสาทข้างเคียง 2 มิติ

การหาค่าระยะห่างของปมประสาทจะมีวิธีการคำนวณค่าระยะห่างอยู่ 2 วิธี ได้แก่ การหาระยะห่างด้วยวิธี Euclidean และการหาระยะห่างด้วยวิธี Manhattan ซึ่งมีวิธีการดังต่อไปนี้

1. การหาระยะห่างด้วยวิธี Euclidean จะใช้ฟังก์ชั่น dist ใน MATLAB ในการหาค่าปมประสาทจากตำแหน่งจุดกำเนิด (Home Neuron) ไปยังจุดข้างเคียง (Neighborhood) เช่น กำหนดให้ปมประสาทมี 3 ปม เขียนเป็นคำสั่งได้ดังนี้

คำสั่ง Pos2 = [0 1 2 ; 0 1 2]

แสดงผล Pos2 = 0 1 2

0 1 2

จากนั้นหาระยะห่างจากปมประสาทแต่ละค่าจากความสัมพันธ์ตามคำสั่งต่อไปนี้

คำสั่ง D2 = dist (pos2)

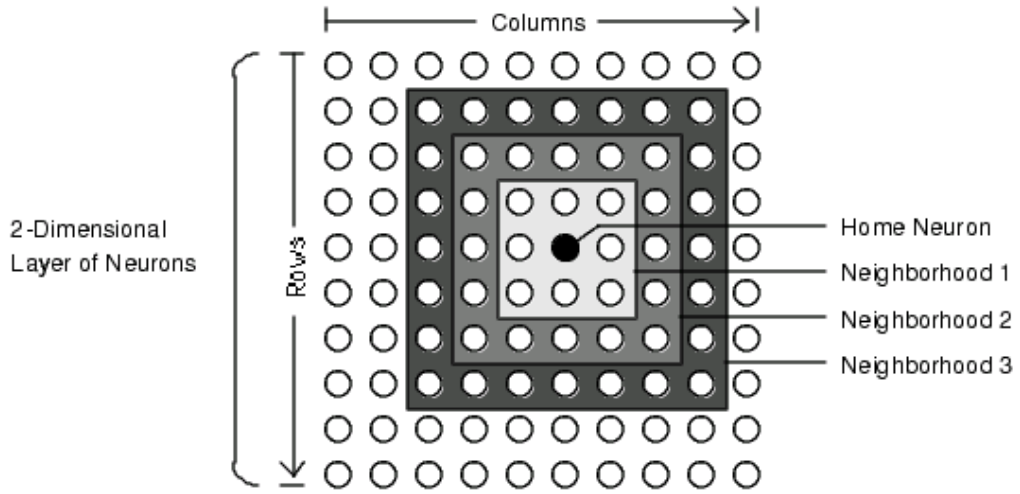
แสดงผล D2 = 0 1.4142 2.8284

1.4142 0 1.4142

2.8284 1.4142 0

จะเห็นว่าระยะห่างจากปมประสาทที่ 1 ซึ่งเป็นของตนเองมีค่าเป็น 0 และระยะห่างจากปมประสาทที่ 1 ถึงปมประสาทที่ 2 มีค่าเป็น 1.414 ค่าระยะห่างระหว่างปมประสาทนี้เรียกว่า Euclidean Distance

การแสดงค่าระยะห่างของปมประสาทใน 2 มิติ จะวัดระยะจากปมประสาทจุดกำเนิดไปยังจุดปมประสาทที่ต้องการ โดยวิธีเพิ่มค่าระยะทางไปยังปมประสาทข้างเคียงโดยรอบ



รูปที่ 2.28 ระยะห่างของปมประสาท 2 มิติ

จะเห็นว่าการใช้คำสั่ง dist จะใช้หาค่าระยะทางปมประสาท 1 มิติ แต่ในปมประสาท 2 มิติ จะใช้คำสั่ง boxdist เนื่องจากค่าระยะทางมีค่าสัมพันธ์กับปมประสาทข้างเคียงในลักษณะเมตริกซ์ เช่น กำหนดให้ปมประสาท 2 มิติ และใช้โครงสร้างแบบตารางขนาด 2x3 จะใช้คำสั่งดังนี้

คำสั่ง Pos = gridtop (2,3)

d = boxdist (pos)

แสดงผล pos = 0 1 0 1 0 1

0 0 1 1 2 2

d = 0 1 1 1 2 2
1 0 1 1 2 2
1 1 0 1 1 1
1 1 1 0 1 1
2 2 1 1 0 1
2 2 1 1 1 0

จากผลการทำงานของโปรแกรมจะพบว่าระยะทางจากปมประสาทที่ 1 ไป 2, 3 และ 4 มีค่าเท่ากับ 1 และระยะทางจากปมประสาทที่ 1 ไป 5 และ 6 มีค่าเท่ากับ 2 ส่วนระยะทางจากปมประสาท 3 และ 4 ไปยังปมประสาทอื่นๆ มีค่าเท่ากับ 1

ในการหาค่าระยะทางเชื่อมโยง (link distance) จากปมประสาทหนึ่งๆ ไปยังปมประสาทอื่นจะต้องใช้วิธีการคำนวณค่าจากชุดของปมประสาทด้วยคำสั่ง linkdist เช่น จงหาค่าของระยะทางเชื่อมโยงของปมประสาท 2 มิติ ขนาด 2*3

<u>คำสั่ง</u> pos = gridtop (2,3)
dlink = linkdist(pos)

<u>แสดงผล</u> dlink = 0 1 1 2 2 3
1 0 2 1 3 2
1 2 0 1 1 2
2 1 1 0 2 1
2 3 1 2 0 1
3 2 2 1 1 0

2. การหาค่าระยะห่างด้วยวิธี Manhattan เป็นการคำนวณค่าระยะทางระหว่างเวคเตอร์ x และ y ซึ่งสามารถหาได้จากความสัมพันธ์ดังต่อไปนี้

D = sum(abs(x-y))

ถ้ากำหนดให้ W1 = [1 2; 3 4; 5 6] และ P1 = [1;1] ค่าระยะทาง Manhattan ที่ได้จะมีค่าดังต่อไปนี้

<u>คำสั่ง</u> W1 = [1 2; 3 4; 5 6]
P1 = [1;1]
Z1 = mandist(W1,P1)

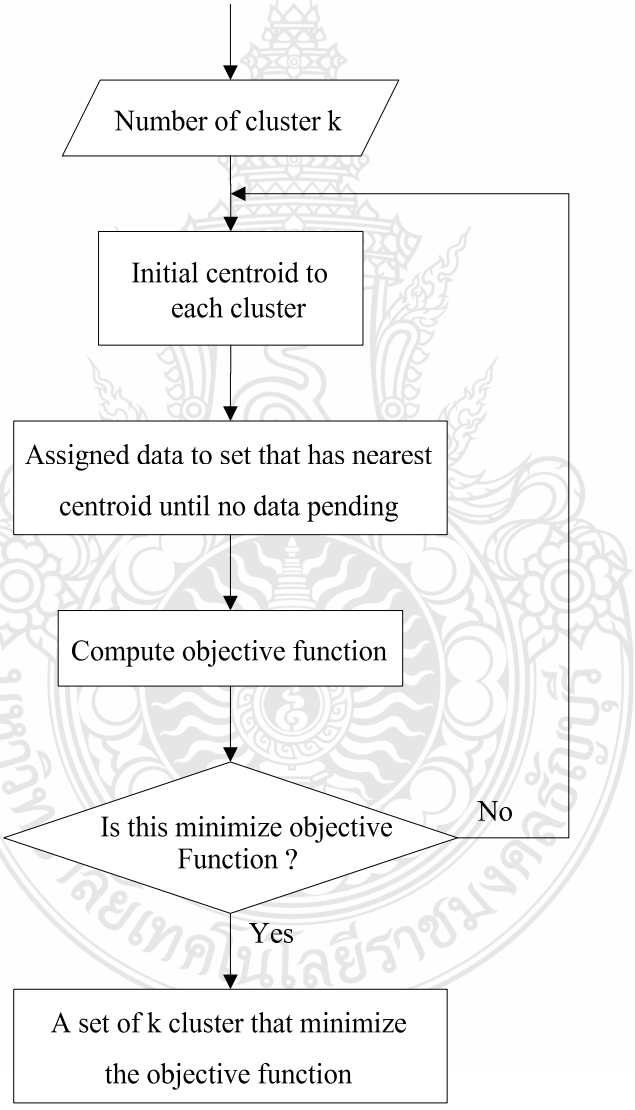
<u>แสดงผล</u> W1 = 1 2
3 4
5 6

$$P1 = \begin{matrix} 1 \\ 1 \end{matrix}$$

$$Z1 = \begin{matrix} 1 \\ 5 \end{matrix}$$

2.4.5 Kohonen Self Organizing Feature Maps (KSOFM)

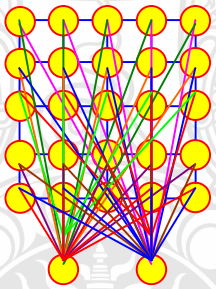
หลังจากการประมวลผลของโครงข่ายประสาทเทียมแบบก่อตัวด้วยตนเอง ผลที่ได้จะนำเข้าสู่อัลกอริทึมการแบ่งกลุ่มข้อมูลแบบเค-มีน เพื่อเพิ่มการรู้จำผู้พูดให้มากยิ่งขึ้น โดยขั้นตอนการทำงานของเค-มีน มีขั้นตอนดังรูปที่ 2.29



รูปที่ 2.29 ขั้นตอนการทำงานของอัลกอริทึม K-Means

KSOFM (Kohonen Self Organizing Feature Maps) เป็นอัลกอริทึมหนึ่งของโครงข่ายประสาทเทียม (Neuron Network) ซึ่งมีการเรียนรู้ที่สามารถจัดการตนเองโดยจะทำการประมวลผลจัดจำแนก Input Codevectors เป็นกลุ่มๆ หรือการทำซ้ำข้อมูลเพื่อหาค่าของน้ำหนักของข้อมูลที่มีอยู่ทั้งหมดตามจำนวนกลุ่มที่ต้องการ เมื่อมี Input Codevector ชุดใหม่เข้ามา ระบบก็จะประมวลผลค่าน้ำหนักของ Input Codevector ใหม่ ปมประสาทที่มีค่าใกล้เคียงกับ Input Codevector มากที่สุดจะเรียกว่า Winning Neuron โดย Winning Neuron นี้จะถูกปรับค่าหรือปรับแต่งให้มีการตอบสนองต่อโครงข่ายมากที่สุด และจะส่งผลให้ปมประสาทข้างเคียงหรือ Codevector ใกล้เคียงถูกปรับค่าเช่นกัน ซึ่งคุณลักษณะสมบัติการปรับแต่งค่าเหล่านี้ของ KSOFM จะเป็นประโยชน์อย่างมากต่อการประมวลผลข้อมูล

โครงสร้างการทำงานของ KSOFM ชั้นแรก ทำหน้าที่นำเข้าข้อมูลและจัดส่งข้อมูลให้แก่ Neurons ชั้นที่สองทุกๆ ปมประสาท ระหว่างชั้นจะเชื่อมต่อกันด้วยค่าน้ำหนัก (Weight Vector) จากนั้นข้อมูลจะถูกส่งไปยัง Neurons ในชั้นที่สอง เพื่อทำการเปรียบเทียบว่าใกล้เคียงกับค่ากลางกลุ่มใดมากที่สุด แต่ละปมประสาทในชั้นนี้จะมีความสัมพันธ์กันแบบ Neighborhood Relation การก่อรูปแบบ 2 มิติ จึงเกิดขึ้นดังรูปที่ 2.21



รูปที่ 2.30 แผนผังเรียนรู้การจัดตัวเอง Self-Organizing Map

ขั้นตอนการทำงานของ KSOFM

1. กำหนดชุดของข้อมูล ที่จะเป็นอินพุต
2. สุ่มค่าเริ่มต้นให้กับค่ากลางของกลุ่ม (Cluster Center) และกำหนดอัตราการเรียนรู้
3. วัดค่าความคล้ายด้วยวิธี แบบ ยูคลิด โดยการทำซ้ำทุกข้อมูลนำเข้า $1 \le i \le N$ ให้หาผู้ชนะจากการเลือกกลุ่ม ที่มีค่ายูคลิเดียนน้อยที่สุด ค่าความต่างระหว่างข้อมูลนำเข้ากับเวกเตอร์น้ำหนัก
4. คำนวณค่ากลางสำหรับกลุ่มที่เป็น the winner ใหม่
5. ทำซ้ำขั้นตอนที่ 3-4
6. หาระยะห่างระหว่างโหนดนั้นๆ กับ โหนดที่เป็น the winner

7. หาโหนดเพื่อนบ้านที่ใกล้เคียงที่สุด โดยขอบเขตจะลดลงตามเวลา
8. ปรับค่าน้ำหนัก ของแต่ละโหนด

2.5 การประเมินคุณภาพของข้อมูลเสียงพูด

การประเมินคุณภาพของข้อมูลเสียงพูดโดยใช้วิธีการของ MOS (Mean Opinion Score) [13] เป็นวิธีการหนึ่งที่นิยมใช้เปรียบเทียบระหว่างเสียงพูดต้นแบบกับเสียงพูดที่ผ่านการบีบอัดข้อมูล โดยใช้การรับรู้และความรู้สึกของมนุษย์เป็นเกณฑ์ในการตัดสิน (Subjective Measurement)

ตารางที่ 2.1 ค่า MOS ที่เหมาะสมกับการใช้งานในระบบต่าง ๆ

MOS	การใช้งาน
4.5-5.0	Broadcast Quality
4.0-4.5	Network or Toll Quality
3.5-4.0	Communication Quality
2.5-3.5	Synthetic Quality

วิธีการประเมินหรือวัดคุณภาพสียงนั้นจะใช้คนประมาณ 12-24 คน ทดสอบคุณภาพเสียงด้วยการฟัง โดยที่แต่ละคนจะให้คะแนนที่มีค่าอยู่ระหว่าง 1-5 ตามคุณภาพของสัญญาณที่ตัวเองรู้สึก จากนั้นหาค่าเฉลี่ยแต่ละเสียงพูดว่าอยู่ในระดับใด

ตารางที่ 2.2 รายละเอียดวิธีการให้คะแนนในการวัด MOS

คะแนน	**คุณภาพของเสียง**
5	ดีมาก (คุณภาพเสียงชัดเจนและเข้าใจง่าย)
4	ดี (คุณภาพเสียงดีและเข้าใจง่าย แต่อาจได้ยินเสียงรบกวนบ้าง)
3	พอใช้ (คุณภาพเสียงดีและเข้าใจ ได้แต่อาจต้องการอาศัยความตั้งใจ หรือบางทีต้องขอให้พูดซ้ำ)
2	เลว (คุณภาพเสียงดีและเข้าใจได้ก็ต่อเมื่อมีความตั้งใจมาก ๆ และบ่อยครั้งที่ต้องขอให้พูดซ้ำ)
1	เลวมาก (ฟังไม่รู้เรื่องเลย)

2.6 งานวิจัยที่เกี่ยวข้อง

เนื่องจากงานวิจัยทางด้านการประมวลผลสัญญาณเสียงพูดได้มีการพัฒนามาอย่างหลากหลายในหลายแนวทาง ความแม่นยำในการรู้จำ, ความยุ่งยากของการประมวลผล, ความรวดเร็วในการตัดสินใจ รวมไปถึงจำนวนของสัญญาณที่นำมาทดสอบ จึงแตกต่างกันออกไปตามวัตถุประสงค์ของการวิจัยนั้นๆ ตัวอย่างงานวิจัยที่ผ่านมาที่นำมาเสนอนี้ จึงได้คัดเลือกเฉพาะงานที่ใกล้เคียงกับงานวิจัยที่กำลังทำอยู่เท่านั้น อาทิเช่น

1) J. Srinonchat [6] ได้ศึกษาเกี่ยวกับโครงสร้างการบีบอัดข้อมูลให้มีค่า Bit Rate ต่ำลงด้วยการใช้ LPC-10 แบบใหม่ โดยพบว่าค่าพารามิเตอร์ LPC ที่ใช้ทั่วไปจะให้ค่าการควอนไตซ์ไม่ค่อยมีประสิทธิภาพ เนื่องจากจะเกิดค่าผิดพลาดขึ้นในกรณีที่ระดับความแตกต่างของสัญญาณเสียงพูดมีระดับต่ำ เพื่อแก้ปัญหาเรื่องความผิดพลาดของการควอนไตซ์ จึงได้นำเสนอวิธีการโดยเปลี่ยนค่าพารามิเตอร์ LPC เป็นพารามิเตอร์ LSP เพื่อจะสร้าง Codevector ของการแบ่งลำดับชั้นใน Vector Quantization เพื่อที่จะใช้เข้ารหัสสัญญาณเสียงพูดใหม่ เรียกว่า LPC-10 จากผลทดลองพบว่าสามารถลดจำนวนบิตของสัญญาณเสียงพูดในบิตสัญญาณเสียงพูด p1-p4 ได้ 4 บิต และอัตราการส่งข้อมูลลดลง 2-66%

2) Srinonchat, J. and other. [15] ได้ศึกษาเกี่ยวกับการใช้ตำแหน่งเวกเตอร์ ควอนไตซ์ (VQ) ในการบีบอัดสัญญาณเสียงพูด โดยการแบ่งเสียงพูดออกเป็นเฟรม แต่ละเฟรมมีค่า 30 ms. แล้วนำมาสกัดค่าเป็นสัมประสิทธิ์ LPC และ LSP เพื่อใช้เป็นพารามิเตอร์ในการสร้าง Codebook ด้วย KSOFM ซึ่งทำหน้าที่คำนวณค่าและแทนค่าสัมประสิทธ์สัญญาณเสียงพูดลงใน Codebook ให้มีความผิดเพี้ยนน้อยสุด ผลจากการทดลองเสียงพูดที่เป็นชายจำนวน 2 คนและหญิง 2 คน พบว่า ค่าผิดพลาดสูงสุดของการบีบอัดค่า LSP ในเสียงพูดชาย 35% และในเสียงพูดหญิง 40% ส่วนค่าผิดพลาดสูงสุดของการบีบอัดค่า LPC ในเสียงพูดชาย 35% และในเสียงพูดหญิง 30% สรุปแล้วการบีบอัดสัญญาณเสียงพูดด้วยเทคนิค LPC-VQ จะใช้ KSOFM และการใช้ Address-Codebook จะลดค่า Bit Rate ได้ 33%

3) สุทธี ทับทองดี ได้วิจัยและคิดค้นเทคนิคใหม่เพื่อที่จะนำมาบีบอัดสัญญาณเสียงพูดภาษาไทยบนพื้นฐานของการเข้ารหัสด้วยวิธี LPC-10 (Linear Predictive Coding Order 10) โดยใช้เสียงภาษาไทยมาจากบุคคล 4 คนโดยเป็นผู้หญิง 2 คน และเป็นผู้ชาย 2 คน พูดคนละ 30 นาทีบันทึกลงในคอมพิวเตอร์ สัญญาณเสียงถูกแบ่งออกเป็นเฟรมโดยความยาวของเฟรมที่ใช้อยู่ที่ 240 ตัวอย่าง (Samples) และใช้โปรแกรม MATLAB ออกแบบโมดูลของต้นฉบับ LPC-10 และ K-means จากการทดลองพบว่าเมื่อเปรียบเทียบวิธีบีบอัดสัญญาณเสียงพูดต้นฉบับ LPC-10 และวิธีใหม่ของ LPC-10 จะพบว่าสัญญาณเสียงสังเคราะห์ของวิธีใหม่ LPC-10 สามารถที่จะสร้างเสียงสังเคราะห์ได้ใกล้เคียงกับทางวิธีการต้นฉบับ LPC-10 ซึ่งใช้จำนวนบิตน้อยกว่าวิธีการต้นฉบับ LPC-10

4) ปรีดาวรรณ [7] พัฒนาการรู้จำเสียงสำหรับพยัญชนะต้นของอัฒพยางค์ แบบขึ้นกับผู้พูด เพื่อที่จะลดขนาดฐานข้อมูลของการเปลี่ยนตัวอักขระไปเป็นสัญญาณเสียง ลักษณะของหน่วยเสียงอัฒพยางค์ เช่น

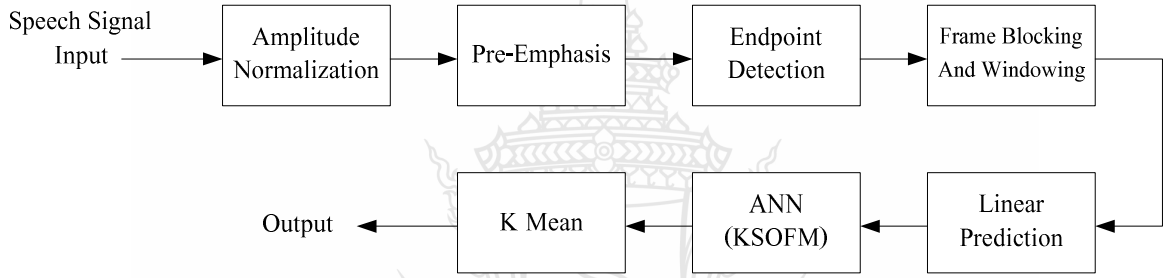
กิน ประกอบด้วย กิ (หน่วยเสียง่วนต้น) + อิน (หน่วยเสียงส่วนท้าย)

ซึ่งการหาจุดเริ่มต้นและสิ้นสุดของสัญญาณเสียงอัฒพยางค์เพื่อใช้สำหรับวิเคราะห์หาลักษณะสำคัญได้จากการตัดแบ่งสัญญาณเสียงโดยพิจารณาจากรูปคลื่นเสียง ขั้นตอนการสกัดลักษณะสำคัญใช้ค่าสัมประสิทธิ์ LPC และสร้างแบบจำลองฮิดเดนมาร์คอฟเพื่อรู้จำ ผลการทดลอง สามารถรู้จำเสียงทดสอบแบบอัฒพยางค์ได้ถูกต้องร้อยละ 84% เนื่องจากการทดลองนี้หาจุดเริ่มต้นและสิ้นสุดของหน่วยเสียงอัฒพยางค์โดยพิจารณาจากรูปคลื่นเสียง ดังนั้นสัญญาณเสียงที่ได้มีการคลาดเคลื่อน ทำให้เสียงพยัญชนะต้นเดียวกันอาจมีความคลาดเคลื่อนเกิดขึ้นได้

5) ชัย วุฒิวิวัฒน์ชัย และคณะ [2] นำเสนอระบบระบุผู้พูดสำหรับภาษาไทยแบบกำหนดคำพูด โดยใช้โครงข่ายประสาทเทียมเป็นระบบในการรู้จำ การทดลองใช้ประโยคในการพูด 6 ประโยค โดย 5 ประโยคแรกแต่ละประโยคมีเสียงวรรณยุกต์เดียวจาก 5 ระดับเสียงวรรณยุกต์ในภาษาไทย ส่วนประโยคสุดท้ายมีทั้ง 5 เสียงวรรณยุกต์ผสมกัน ในขั้นตอนการสกัดลักษณะสำคัญใช้ค่าสัมประสิทธิ์ LPC อันดับ 10 ส่งเป็นอินพุทให้โครงข่ายประสาทเทียมประเภทเพอเซปตรอนหลายชั้น (MLP) โดยใช้วิธีการเรียนรู้แบบแพร่กระจายกลับ รู้จำเสียงผู้พูด ผลการทดลอง เมื่อใช้ประโยคที่มีเสียงวรรณยุกต์ผสมจะได้ผลการระบุผู้พูดสูงที่สุดร้อยละ 95.56% และผลการรู้จำต่ำที่สุดเมื่อใช้ประโยคที่มีเสียงวรรณยุกต์เอกซึ่งเป็นเสียงที่ต่ำและไม่ชัดเจน

บทที่ 3
วิธีการดำเนินงานวิจัย

งานวิจัยนี้เป็นระบบ การรู้จำผู้พูด โดยใช้เทคนิคการรู้จำผู้พูดโดยใช้โครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง เป็นระบบที่มีขั้นตอนการดึงลักษณะสำคัญของสัญญาณเสียงได้เลือกใช้วิธีการวิเคราะห์แนวทางเดินของสัญญาณเสียงพูดในรูปแบบของสัมประสิทธิ์คู่เส้นสเปคตรัม (LSP) ส่วนในขั้นตอนการทดสอบความคล้ายคลึงกันของรูปแบบและกฎเกณฑ์การตัดสินใจ ได้เลือกใช้โครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง ควบคู่กับอัลกอริทึมแบบเค-มีน (K-Means)



รูปที่ 3.1 การทำงานในภาพรวมของระบบรู้จำผู้พูดโดยใช้โครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง

3.1 การบันทึกเสียงพูด

การบันทึกสัญญาณเสียงพูดใช้การบันทึกเสียงผ่านไมโครโฟนชนิดคอนเดนเซอร์ลงในคอมพิวเตอร์ ส่วนบุคคลในห้องทำงานที่มีสภาพแวดล้อมปกติ (สัญญาณรบกวนโดยรวมไม่เกิน 0 - 0.5 dB) โดยบันทึกเสียงแบบดิจิตอลในระบบโมโน กำหนดให้มีอัตราการสุ่มตัวอย่าง 8,000 Hz/s และมีการแบ่งระดับที่ 8 บิต บันทึกข้อมูลอยู่ในรูปของไฟล์ *.wav โดยมีขั้นตอนการบันทึกเสียงพูด ดังนี้

- เตรียมคำพูดภาษาไทย 3 วลี เพื่อใช้ในการสอนระบบรู้จำผู้พูด (Training Set) ดังนี้
  - กินข้าวกัน
  - อร่อยจังเลย
  - มีอะไรทาน
- เตรียมอาสาสมัครผู้พูด 60 คน แบ่งเป็นชาย 50 คน หญิง 10 คน
- ในการบันทึกเสียงพูดจะใช้สถานที่บันทึกเสียงที่เดียวกัน ใช้ความถี่ในการบันทึก 8 KHz 8 bit Mono บันทึกในรูปแบบ Wav ไฟล์
- ให้ผู้พูดแต่ละคนพูดวลีละ 3 ครั้งและบันทึกลงในคอมพิวเตอร์

3.1.1 อุปกรณ์ที่ใช้ในการจัดเก็บข้อมูลเสียง

อุปกรณ์ที่ใช้ในการจัดเก็บข้อมูลเสียงนั้น จะเป็นอุปกรณ์ที่ได้มาตรฐานสามารถนำมาทำการบันทึกใช้เป็นข้อมูลจริงได้ และทำการทดสอบได้จริง และอุปกรณ์ทั้งหมดอยู่ในสภาพที่สมบูรณ์สามารถใช้งานได้ ซึ่งจะมีรายละเอียดของอุปกรณ์ต่างๆ ดังนี้

3.1.1.1 เครื่องคอมพิวเตอร์ Notebooks รุ่น Intel(R) Pentium(R) core 2 Dual ความเร็วในการประมวลผล 2 GHz ที่มีหน่วยความจำหลัก 2 GB และหน่วยความจำสำรอง 250 GB

3.1.1.2 การ์ดเสียง (Sound Card) SoundMAX ติดตั้งมาพร้อมกับเครื่องคอมพิวเตอร์ Notebooks

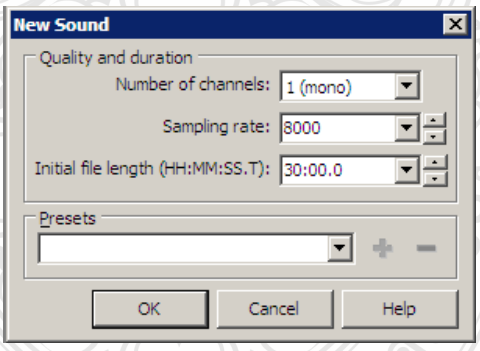
3.1.1.3 ไมโครโฟน ยี่ห้อ AIWA รุ่น DM-H200

3.1.1.4 GoldWave Version 5.23 ติดตั้งบนระบบปฏิบัติการ WindowsXP

3.1.2 ขั้นตอนในการบันทึกข้อมูลเสียงพูด

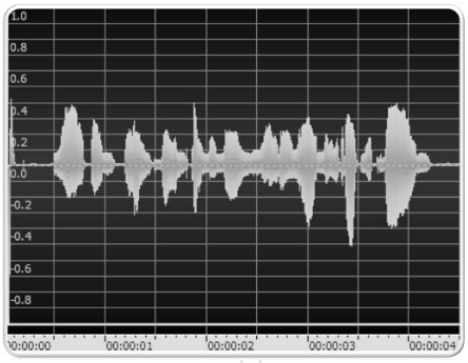
ขั้นตอนการบันทึกสัญญาณเสียงพูดด้วยโปรแกรม GoldWave version 5.5 มีดังนี้

3.1.2.1 กำหนดค่าการบันทึกสัญญาณเสียงพูด กำหนดให้ความแรงของสัญญาณเสียงพูดมีขนาด -1.0 ถึง 1.0 โดยเลือกเมนูคำสั่ง File>New จะปรากฏหน้าต่าง New Sound เป็นการกำหนดคุณภาพเสียงและระยะเวลาการบันทึก กำหนดค่าที่จำนวนช่องสัญญาณเสียงเป็น 1 (mono), จำนวนของการสุ่มค่าตัวอย่าง ตั้งค่าเป็น 8,000 และช่วงระยะเวลาการบันทึก ตั้งค่าเป็น 30:00.0



รูปที่ 3.2 การตั้งค่าเริ่มต้นบันทึกเสียงในโปรแกรม GoldWave

3.1.2.2 การบันทึกสัญญาณเสียงพูดนำเมาส์คลิกที่ปุ่มการเริ่มต้นบันทึก  Starts Recording เพื่อทำการบันทึกไฟล์เสียงใน Directory ที่กำหนดโดยใช้ผู้ทดสอบอายุระหว่าง 19 ถึง 40 ปี เป็นผู้ชาย 50 คน ผู้หญิง 10 คน



รูปที่ 3.3 ตัวอย่างสัญญาณเสียงพูดในกลุ่มผู้พูดที่ใช้ในการวิเคราะห์

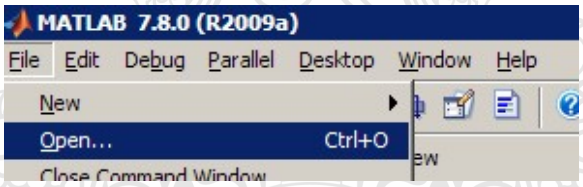
3.2 การสร้างระบบการบีบอัดสัญญาณเสียงพูด

ในการสร้างระบบการบีบอัดสัญญาณเสียงพูด จะต้องหาค่าลักษณะสำคัญของสัญญาณเสียงพูดแต่ละสัญญาณ โดยการสกัดค่าพารามิเตอร์ LPC และ LSP เพื่อนำค่าสัมประสิทธิ์ตัวแทนเสียงที่สกัดได้ไปใช้ในโครงข่ายจัดการตนเองอัลกอริทึม KSOFM ต่อไป

3.2.1 การสกัดค่าพารามิเตอร์ LPC

การสกัดค่าพารามิเตอร์ LPC มีขั้นตอนการปฏิบัติการดังนี้

3.2.1.1 เปิดโปรแกรม MatLab เรียกไฟล์ *.m โดยเลือกเมนูคำสั่ง File>Open



รูปที่ 3.4 การใช้คำสั่งเปิด *.m ไฟล์

3.2.1.2 การอ่านไฟล์เสียงด้วยคำสั่ง wavread เปิดไฟล์ main.m จะปรากฏรหัสคำสั่งในบรรทัดที่ 4 [in_speech,fs,bits] = wavread('ชื่อไฟล์เสียง'); ให้ใส่ไดเร็คทอรี่และชื่อไฟล์

```
clear
clc
% get wave-file
[in_speech,fs,bits] = wavread('org.wav')
% prepare and preprocess wave-file
aux_speech=in_speech;
in_speech = in_speech - mean(in_speech(:
speechnorm = resample(in_speech,8000,fs)
speech_filt=preprocess(speechnorm,fs);
wavwrite(speech_filt,'original.wav');
```

รูปที่ 3.5 การใช้คำสั่งการอ่านไฟล์เสียง

3.2.2 การสกัดค่าพารามิเตอร์ LSP

การสกัดค่าพารามิเตอร์ LSP มีขั้นตอนการปฏิบัติการดังนี้

3.2.2.1 การสกัดค่าพารามิเตอร์ LSP จะใช้ฟังก์ชันในการแปลงค่าสัมประสิทธิ์จาก LPC เป็น LSP (LPC/LSP Conversion) ซึ่งเขียนไว้ในไฟล์ main.m โดยเรียกใช้งานผ่านฟังก์ชันเรียกใช้ (Callback Function) คือ ฟังก์ชัน lpc_lsp ไว้ใน m-file อีกไฟล์หนึ่งคือไฟล์ LPC_LSP.m

```
% get wave-file
[in_speech,fs,bits] = wavread('org.wav');    % getting from speech.m
% prepare and preprocess wave-file
aux_speech=in_speech;
in_speech = in_speech - mean(in_speech(:));
speechnorm = resample(in_speech,8000,fs);
speech_filt=preprocess(speechnorm,fs);
wavwrite(speech_filt,'original.wav');
speech = speech_filt';

% Frame variables
sp_size=size(speech);                              %speech size
frame=240;                                         %frame size
frame_numb=fix(sp_size(2)/frame)+1;                %number of frames
speech(1,(frame_numb*frame))=0;                    %cutting the end
speechy=speech;                                    %speech for windowing
sp_matrix=(reshape(speech,frame,frame_numb));      %speech in matrix
sp_matrix=sp_matrix';                              %transposed speech
matrix
LPC_order=10;                                      %order of LPC-Analysis
LPC_coeff=zeros(frame_numb,LPC_order+1);           %LPC coefficients
matrix

% window variables
overlap=100;                                       %50 samples on each
side
win=hamming(frame+overlap);                        %Hamming window

% Analysis
LPC_coeff=get_lpc(speechy,frame_numb,frame,win,overlap,LPC_order);

% lpc/lsp conversion
LSP_coeff=lpc_lsp(LPC_coeff);
```

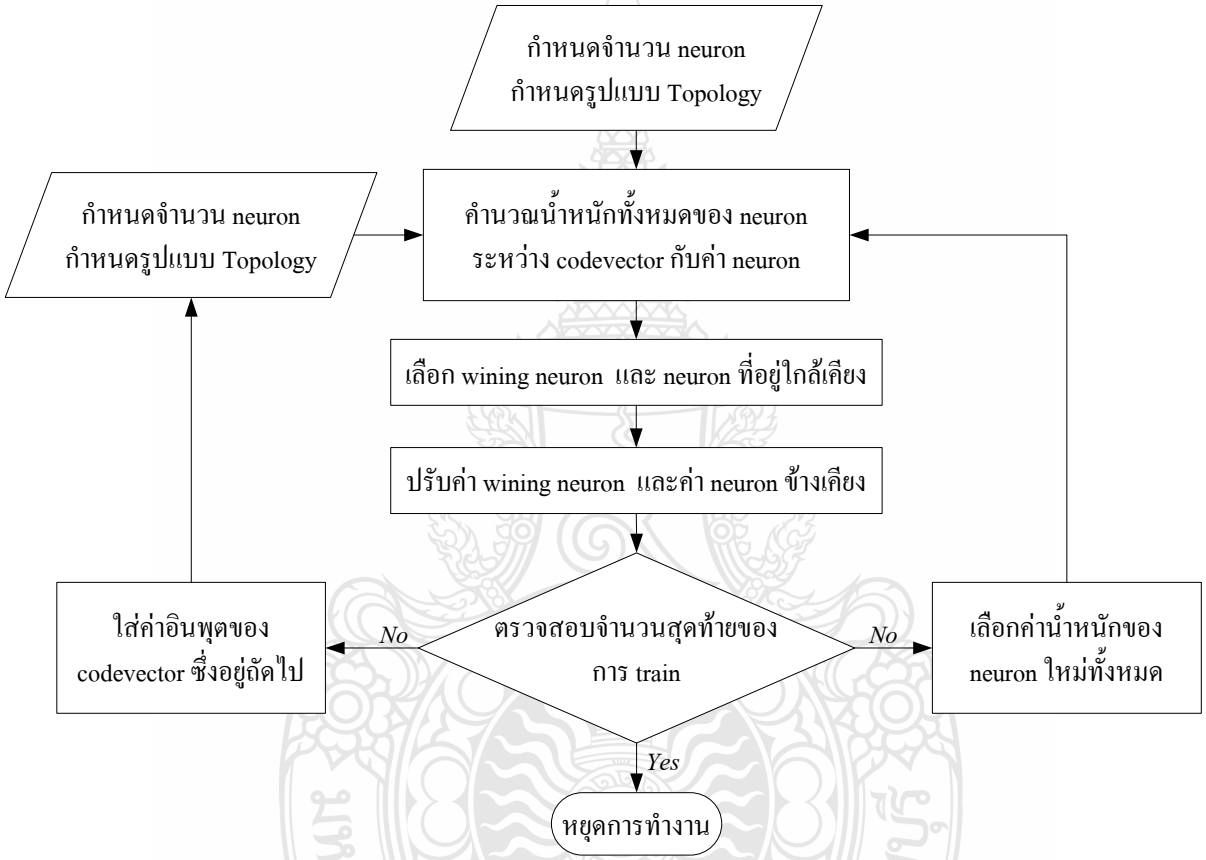
ฟังก์ชันแปลงค่า LPC เป็น LSP

รูปที่ 3.6 ฟังก์ชันแปลงค่า LPC เป็น LSP ในไฟล์ main.m

3.3 การออกแบบโครงข่ายประสาทเทียม

ในการแบ่งกลุ่มของโค้ดเวกเตอร์อินพุท จะถูกจัดแบ่งเป็นกลุ่มๆ ตามลักษณะรูปแบบที่กำหนดไว้ เวกเตอร์โค้ดอินพุทที่มีลักษณะคล้ายกันจะถูกวางในตำแหน่งเดียวกัน ซึ่งอาจจะใช้ปมประสาทเดียวกันหรือปมประสาทที่ใกล้เคียงกันทำให้โค้ดเวกเตอร์อินพุทซ้ำแล้วซ้ำเล่า เมื่อมีโค้ดเวกเตอร์อินพุทใหม่เข้ามายังระบบและค่าเฟสก็จะเริ่มถูกประมวลใหม่เช่นกัน ส่งผลให้ค่าน้ำหนักของโค้ดเวกเตอร์อินพุทที่ถูกเลือก (Winning neuron) จะถูกปรับค่าหรือปรับแต่งให้มีการตอบสนองต่อโครงข่ายมากที่สุด

3.3.1 แผนผังโครงข่ายการจัดการตนเอง



รูปที่ 3.7 ผังการทำงานของโครงข่ายการจัดการตนเอง

การสร้างไฟล์สัญญาณเสียงบีบอัดจากโครงข่ายปมประสาทแบบจัดการตนเองมีขั้นตอนดังนี้

3.3.1.1 เรียกสัมประสิทธิ์สัญญาณเสียงต้นฉบับในแต่ละเฟรมและสัมประสิทธิ์ของปมประสาทในแต่ละเฟรมมาเป็นใช้ในโปรแกรม

3.3.1.2 แปลงค่าสัมประสิทธิ์ปมประสาทในแต่ละเฟรมให้เป็นค่าเมตริกซ์ m x n เพื่อจะนำไปสร้างค่าน้ำหนักเมตริกซ์

3.3.1.3 สร้างสัมประสิทธิ์สัญญาณเสียงบีบอัดในแต่ละเฟรมโดยการนำค่าน้ำหนักเมตริกซ์ในแต่ละเฟรมคูณกับสัมประสิทธิ์สัญญาณเสียงต้นฉบับในแต่ละเฟรม

3.3.1.4 รวมค่าสัมประสิทธิ์สัญญาณเสียงบีบอัดในแต่ละเฟรมเพื่อสร้างสัญญาณเสียงบีบอัดที่ภายในบรรจุข้อมูลข่าวสารที่เหมือนกับสัญญาณเสียงต้นฉบับโดยสัมประสิทธิ์รวมนี้จะใช้เป็นตัวแปรในการเข้ารหัสเสียงแบบ Microsoft Wave

3.3.2 ขั้นตอนการสร้างโครงข่ายจัดการตนเอง

ขั้นตอนการทำงานของโครงข่ายจัดการตนเองมีขั้นตอนการปฏิบัติการดังนี้

3.3.2.1 อ่านไฟล์เสียงต้นฉบับจัดแบ่งเฟรมเสียงและสกัดค่าสัมประสิทธิ์ LPC–10 เพื่อใช้สร้างอินพุตเวกเตอร์ที่ใช้ในการสร้างโครงข่ายจัดการตนเอง โดยใช้โปรแกรม main.m

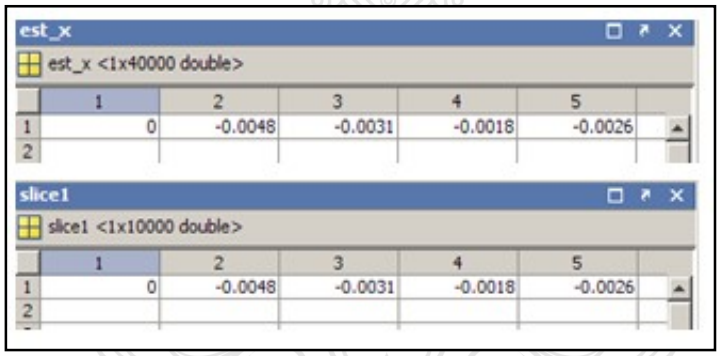
```
clc;                        %ล้างจอ
partition = -1:1/64:1;      %แบ่งระดับความแรงเป็น 128 ระดับ
codebook(1,1) = -1;         %แบ่งกำหนดขนาดสูงสุดด้านลบ
codebook(1,130) = 1;        %แบ่งกำหนดขนาดสูงสุดด้านบวก
for i=1:1:128
   codebook(i+1) = (partition(1,i)+partition(1,(i+1)))/2;   %สร้าง
codebook
end
[samp, Fs, nbits] = wavread('org.wav'); %อ่านไฟล์เสียง
[index,quant] = quantiz(samp,partition,codebook); %ปรับระดับขนาดเสียง 128
ระดับ
s = filter(1,[1 1/2 1/3 1/4],quant);        %กรองความถี่เสียง
a = lpc(s,10);                              %สกัดค่าสัมประสิทธิ์ LPC-10
est_x = filter([0 -a(2:end)],1,s);          %ค่าประมาณเสียงด้วย LPC

for(i=1:1:10000)          % แบ่งเฟรมเป็น 8 เฟรม
    slice1(1,i)    = est_x(1,i);            %เฟรม 1 ถึง 10000
    slice2(1,i)    = est_x(1,(i+10000));    %เฟรม 10001 ถึง 20000
    slice3(1,i)    = est_x(1,(i+20000));    %เฟรม 20001 ถึง 30000
    slice4(1,i)    = est_x(1,(i+30000));    %เฟรม 30001 ถึง 40000
  slice5(1,i)    = est_x(1,(i+40000));      %เฟรม 40001 ถึง 50000
  slice6(1,i)    = est_x(1,(i+50000));      %เฟรม 50001 ถึง 60000
  slice7(1,i)    = est_x(1,(i+60000));      %เฟรม 60001 ถึง 70000
  slice8(1,i)    = est_x(1,(i+70000));      %เฟรม 70001 ถึง 80000
end
```

รูปที่ 3.8 โปรแกรมในการจัดแบ่งเฟรมเสียง.wav'); %สัญญาณเสียงที่ได้จากการประมาณค่า LPC-10

```
sum = [slice1,slice2,slice3,slice4,...
        slice5,slice6,slice7,slice8];     %รวมเฟรมเสียงที่เปลี่ยนอัตราสุ่ม
save('slice1.mat', 'slice1');             %บันทึกตัวแปรเฟรมเสียงที่ 1
save('slice2.mat', 'slice2');             %บันทึกตัวแปรเฟรมเสียงที่ 2
save('slice3.mat', 'slice3');             %บันทึกตัวแปรเฟรมเสียงที่ 3
save('slice4.mat', 'slice4');             %บันทึกตัวแปรเฟรมเสียงที่ 4
save('slice5.mat', 'slice5');             %บันทึกตัวแปรเฟรมเสียงที่ 5
save('slice6.mat', 'slice6');             %บันทึกตัวแปรเฟรมเสียงที่ 6
save('slice7.mat', 'slice7');             %บันทึกตัวแปรเฟรมเสียงที่ 7
save('slice8.mat', 'slice8');             %บันทึกตัวแปรเฟรมเสียงที่ 8
wavwrite(sum,8000,8,'sumorg
```

รูปที่ 3.8 (ต่อ) โปรแกรมในการจัดแบ่งเฟรมเสียง

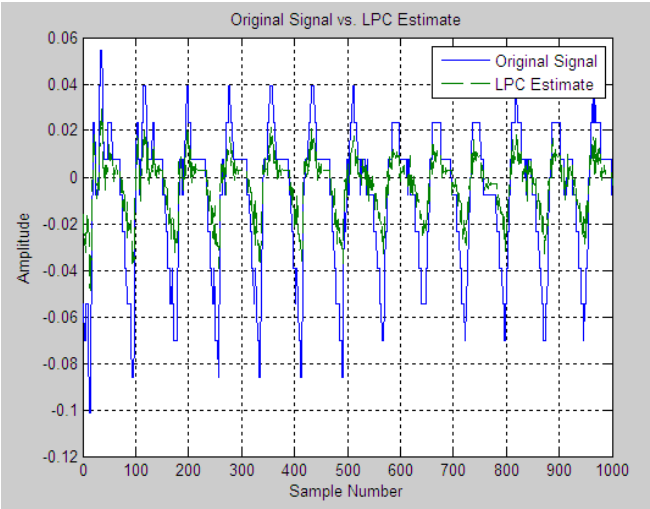


รูปที่ 3.9 สัญญาณเสียงสังเคราะห์จากสัมประสิทธิ์ LPC-10

ค่า est_x เป็นค่าสัมประสิทธิ์สัญญาณเสียงสังเคราะห์ LPC-10 ทั้งเฟรมขนาด 1x 80000 ส่วนค่า Slice1 เป็นค่าสัมประสิทธิ์สัญญาณเสียงสังเคราะห์ LPC-10 ในเฟรมที่1 ขนาด 1x 10000 และการเปรียบเทียบสัญญาณเสียงต้นฉบับกับสัญญาณเสียงที่สกัดค่าด้วย LPC-10 จำนวน 1000 ตัวอย่างที่ตำแหน่ง 4001 ถึง 5000 จะแสดงด้วยคำสั่ง plot

```
plot(1:1000,quant(4001:5000),1:1000,est_x(4001:5000),'--');
title('Original Signal vs. LPC Estimate');
xlabel('Sample Number'); ylabel('Amplitude'); grid;
legend('Original Signal','LPC Estimate')
```

รูปที่ 3.10 แสดงคำสั่งในการพล็อตกราฟ



รูปที่ 3.11 สัญญาณเสียงต้นฉบับกับสัญญาณเสียงสังเคราะห์ LPC-10

3.3.2.2 กำหนดรูปแบบโครงสร้างของโครงข่ายโดยใช้แบบหกเหลี่ยมและจำนวนปมประสาทมี 128 ปม คำสั่งที่ใช้ในการสร้างโครงข่ายคือคำสั่ง newsom , การเรียนรู้โครงข่ายด้วยคำสั่ง train และสร้างปมประสาทด้วยคำสั่ง sim โดยโปรแกรมที่ใช้ในการสร้างโครงข่ายจัดการตนเองด้วยไฟล์ outksofm.m

```
clc;                                %ล้างจอ
disp('Frame1')                      %แสดงข้อความ
load  slice1.mat;                   %เรียกข้อมูลในเฟรมที่ 1ใส่ในตัวแปร
reslice2 = reshape(slice1,25,400)   %เปลี่ยนขนาดเมตริกซ์เสียง 25x400
net = newsom(reslice1,[32,4]);      %สร้างโครงข่ายจัดการตนเองขนาด
128
net.trainParam.epochs = 1000;       %กำหนดการเรียนรู้ 1000 ครั้ง
net = train(net,reslice1);          %ฝึกฝนโครงข่าย
a1 = sim(net,reslice1);             %หาค่าเอาท์พุทปมประสาท
target = full(a1);      %แปลงค่าเอาต์พุตปมประสาทให้เป็นเมตริกซ์ 25x400
for(i=1:1:25)
   result1 = 0
  for(j=1:1:400)
   result1 = result1 + target1(i,j); %หาอัตราการใช้ข้อมูลในลำดับชั้น
  end
 num1(i,1) = result1;                %แสดงผลการจัดข้อมูลในลำดับชั้น
end

for(i=1:1:25)
```

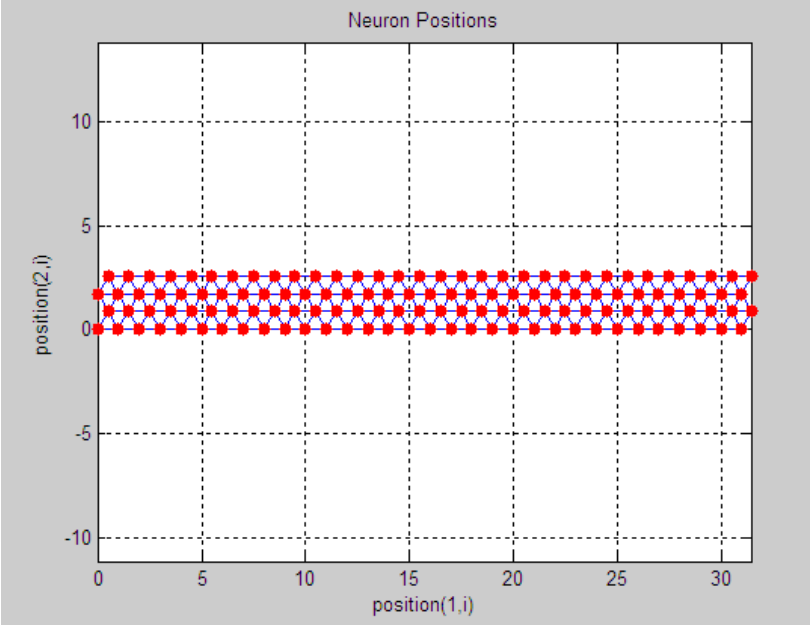
รูปที่ 3.12 โปรแกรมการสร้างปมประสาทในชั้นลำดับจัดการตนเอง

```
    if (num1(i,1)== 1)
        test1(i,1) = i;          %หาตำแหน่งที่ต้องการลบข้อมูล
    end
end
bus1 = size(test1)
RT = bus1(1,1).*bus1(1,2);  %กำหนดตำแหน่งที่ต้องการลบ

for(i=1:1:(RT-10))
 if (test1(i,1) ~= 0)
       n1(i,1) = test1(i,1);
       layer(n1(i,1),:) = []; %ลบข้อมูลในส่วนที่ซ้ำซ้อน
 end
end
len1 = size(layer1); % ขนาดข้อมูลแถวในข้อมลเสียง
layer11 = reshape(layer1,1,(len1(1,1)*len1(1,2))); %
สร้างเมตริกซ์ 1xn
wavwrite(4.0*layer11,8000,8,'layer1.wav') % สร้างไฟล์เสียง
เฟรมที่ 1
```

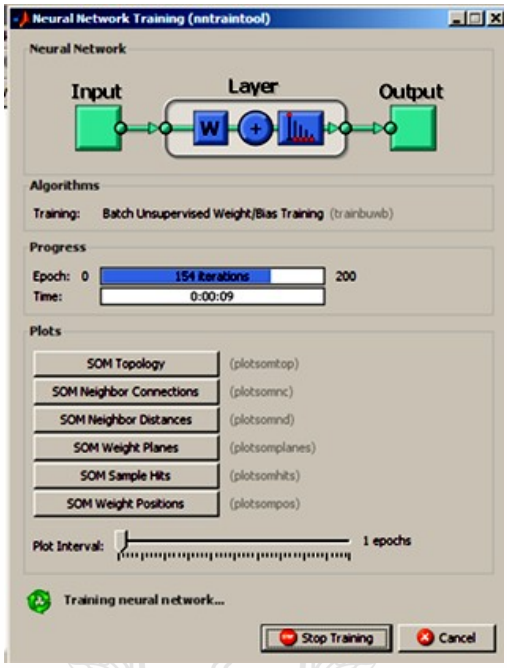
รูปที่ 3.12 (ต่อ) โปรแกรมการสร้างปมประสาทในชั้นลำดับจัดการตนเอง

1) เปิดโปรแกรม MATLAB เรียกไฟล์ outksofm.m แล้วเลือก Debug> Run โปรแกรมจะทำการจัดการสร้างโครงข่ายจัดการตนเองแบบหกเหลี่ยมขนาด 32x4 จำนวน 128 ปมประสาท



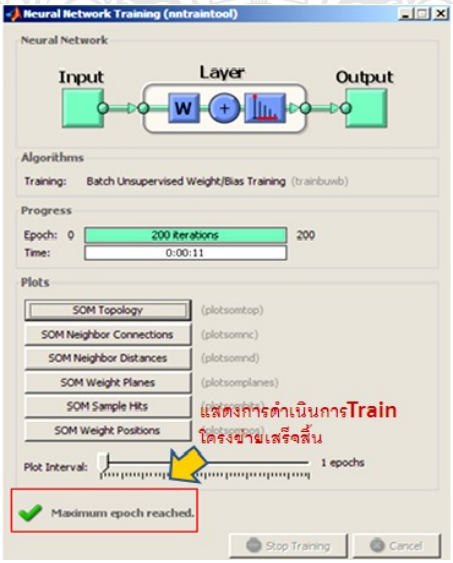
รูปที่ 3.13 โครงข่ายจัดการตนเองขนาด 32x4 จำนวน 128 ปมประสาท

2) กำหนดจำนวนการเรียนรู้โครงข่ายจัดการตนเองจากคำสั่ง train ในโปรแกรม outksofm.m จะดำเนินฝึกฝนจำนวน 200 ครั้ง เพื่อปรับค่าน้ำหนัก, หาระยะทางระหว่างปมประสาท



รูปที่ 3.14 การดำเนินการฝึกฝนข้อมูลโครงข่ายจัดการตนเอง

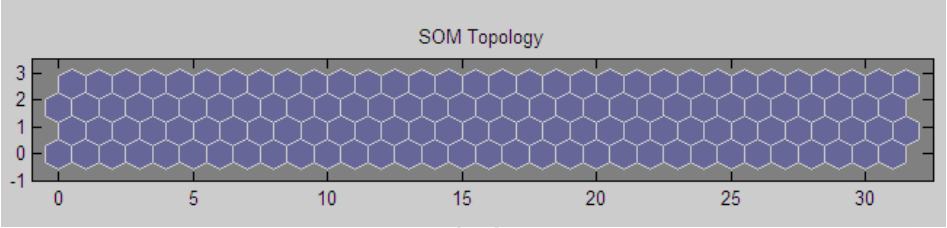
3) การแสดงสถานะการเรียนรู้โครงข่ายโดยเปิดหน้าต่าง Neural Network Training โปรแกรมจะดำเนินการฝึกฝนโครงข่ายรอจนขบวนการสิ้นสุดโดยสังเกตจากตำแหน่งล่างสุดจะปรากฏข้อความว่า Maximum Epoch Reached



รูปที่ 3.15 จุดสิ้นสุดของขบวนการฝึกฝนข้อมูลโครงข่าย

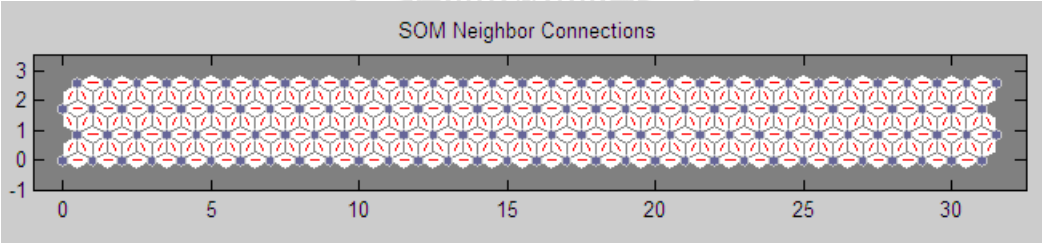
4) การแสดงผลการทำงานของโครงข่ายจัดการตนเองภายหลังการฝึกฝนเสร็จสิ้น มีรายละเอียดดังนี้

4.1) การจัดรูปแบบโครงข่ายจะเปิดหน้าต่าง Neural Network Training และคลิกที่ตำแหน่ง Som Topology



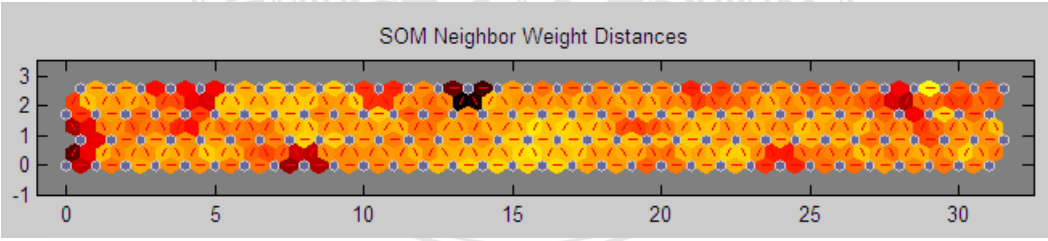
รูปที่ 3.16 การจัดรูปแบบโครงข่ายแบบหกเหลี่ยม

4.2) การเชื่อมต่อระหว่างปมประสาทข้างเคียงจะเปิดหน้าต่าง Neural Network Training และคลิกที่ตำแหน่งของคำสั่ง Som Neighbor Connection



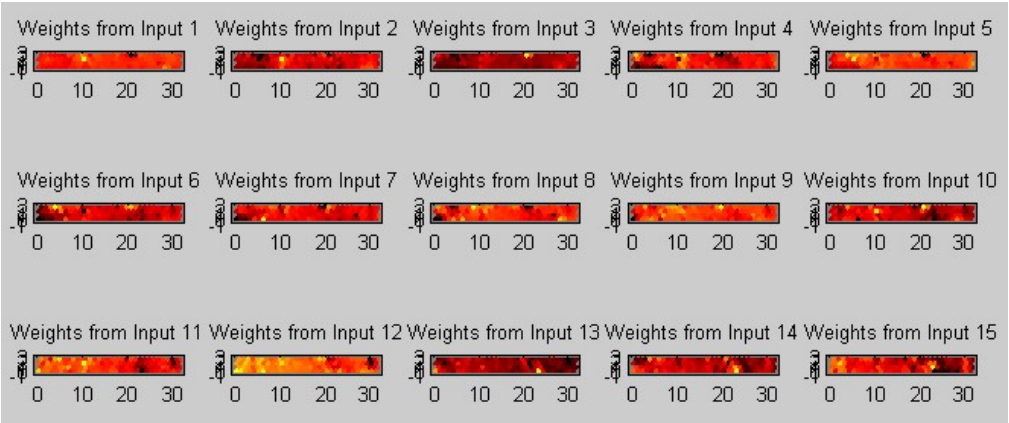
รูปที่ 3.17 การเชื่อมต่อของปมประสาทข้างเคียง

4.3) ค่าระยะทางการเชื่อมโยงระหว่างปมประสาทจะเปิดหน้าต่าง Neural Network Training และคลิกที่ตำแหน่งของคำสั่ง Som Neighbor Distance



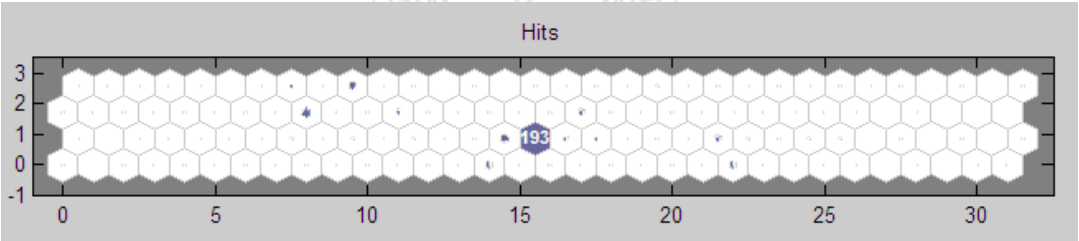
รูปที่ 3.18 ค่าระยะทางเชื่อมโยงระหว่างปมประสาท

4.4) ค่าน้ำหนักแต่ละปมประสาทจะเปิดหน้าต่าง Neural Network Train และคลิกที่ตำแหน่งของคำสั่ง Som Weight Planes



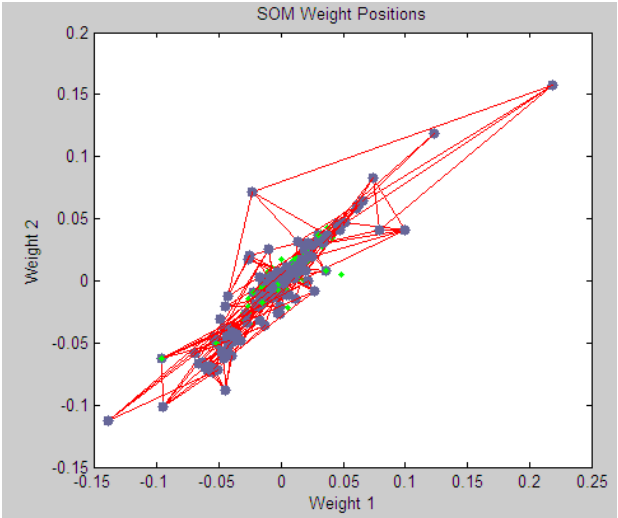
รูปที่ 3.19 ค่าน้ำหนักของปมประสาท

4.5) ค่าการกระจายตัวของข้อมูลเวกเตอร์ในแต่ละปมประสาทจะเปิดหน้าต่าง Neural Network Training และคลิกที่ตำแหน่งของคำสั่ง Som Sample Hits



รูปที่ 3.20 ค่าการกระจายตัวของเวกเตอร์ภายในปมประสาท

4.6) ค่าน้ำหนักตำแหน่งของปมประสาทจะเปิดหน้าต่าง Neural Network Training และคลิกที่ตำแหน่งของคำสั่ง Som Sample Hits



รูปที่ 3.21 ตำแหน่งของค่าน้ำหนักของโครงข่ายปมประสาท

บทที่ 4
ผลการวิจัย

ในบทนี้จะกล่าวถึงรายละเอียดของผลการวิจัยและผลการทดสอบต่างๆที่ได้จากการดำเนินการวิจัย

4.1 ผลการทดลองใช้โครงข่ายประสาทเทียมแบบ KSOFM (Kohonen Self-organizing Feature Maps)

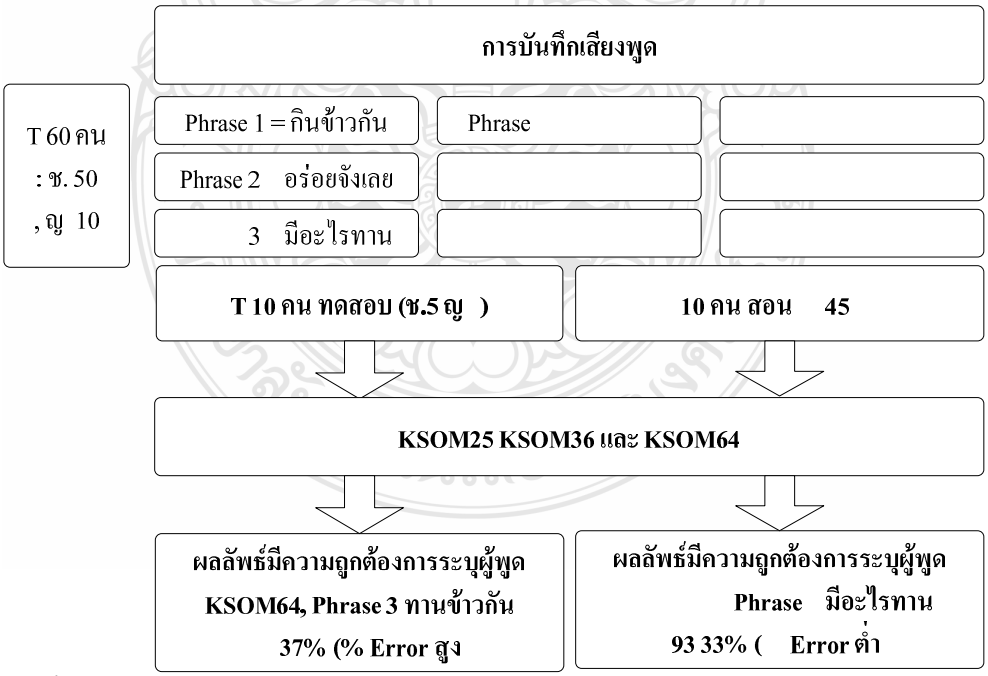
จากการทดลองได้ใช้เสียงในการสอนระบบรู้จำผู้พูด (Training Set) ซึ่งประกอบด้วยผู้พูด (T) จำนวน 60 คน ประกอบด้วยชาย (ช) จำนวน 50 คน และ หญิง (ญ) จำนวน 10 คน อายุระหว่าง 19-40 ปี โดยใช้คำพูดภาษาไทยไม่ต่ำกว่า 3 พยางค์ จำนวน 3 วลี (Phrase) ประกอบด้วย

Phrase 1 กินข้าวกัน

Phrase 2 อร่อยจังเลย

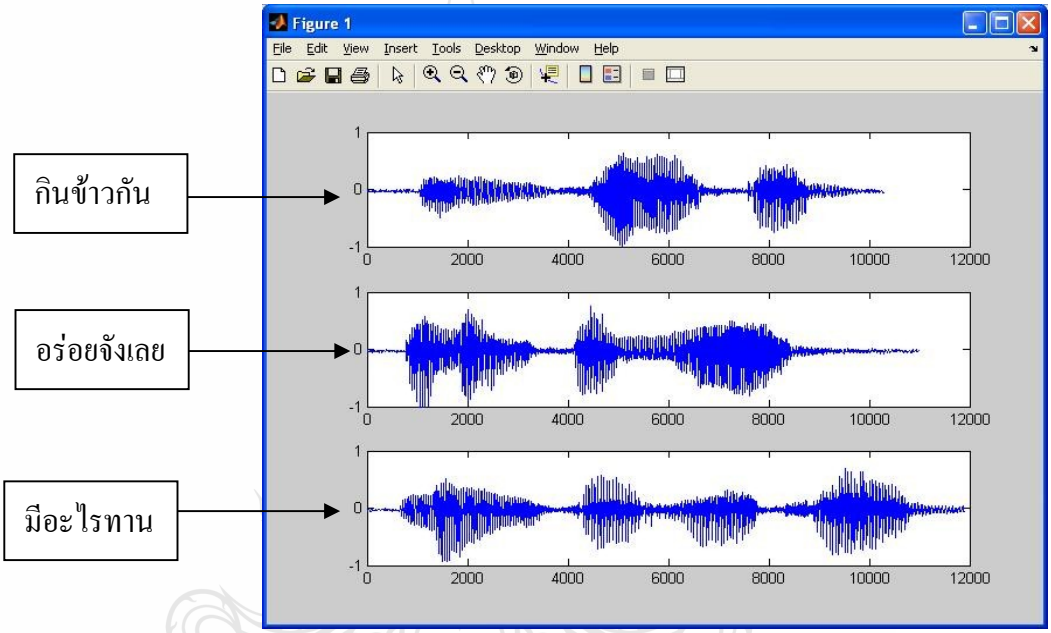
Phrase 3 มีอะไรทาน

ในการบันทึกเสียงพูดจะใช้สถานที่บันทึกเสียงที่เดียวกัน ใช้ความถี่ในการบันทึก 8 KHz 8 bit Mono บันทึกในรูปแบบ Wav ไฟล์



รูปที่ 4.1 ขั้นตอนการบันทึกเสียงพูดและทดสอบระบบการรู้จำในการบันทึกเสียงพูดผู้พูดแต่ละคนจะบันทึกเสียงพูด วลีละ 3 ครั้ง จนครบ 3 วลี เช่น

ผู้พูดที่ 1 (T1) บันทึกเสียงพูดวลีที่ 1 (Phrase 1) กินข้าวกัน กินข้าวกัน กินข้าวกัน
บันทึกเสียงพูดวลีที่ 2 (Phrase 2) อร่อยจังเลย อร่อยจังเลย อร่อยจังเลย
บันทึกเสียงพูดวลีที่ 3 (Phrase 3) มีอะไรทาน มีอะไรทาน มีอะไรทาน
โดยที่จะทำการบันทึกเสียงพูดจนครบ 60 คน(60T) ซึ่งเป็นผู้พูดที่บันทึกจะประกอบด้วยเพศชาย 50 คน และ เพศหญิง 10 คน จากข้อมูลที่บันทึกก็จะเข้าสู่ขั้นตอนการวิเคราะห์เสียงพูดตาม รูปที่ 3.1 การทำงานในภาพรวมของระบบรู้จำผู้พูดโดยใช้โครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง



รูปที่ 4.2 สัญญาณเสียงพูดของ 3 วลี

ในการทดลองใช้โครงข่ายประสาทเทียมแบบ KSOFM (Kohonen Self-organizing Feature Maps) โดยใช้จำนวนโนดต่างๆ ดังนี้คือ KSOFM 25 KSOFM 36 และ KSOFM 64 ซึ่งการทดลองได้กำหนดค่าพารามิเตอร์ของ KSOFM ดังตารางที่ 4.1

ตารางที่ 4.1 การกำหนดพารามิเตอร์ของ KSOFM และการกำหนดพารามิเตอร์ในการสอนระบบรู้จำ

Type	Neural	Lattice	η_0	epochs
KSOFM 25	25	5 x 5	0.1	250
KSOFM 36	36	6 x 6	0.1	200
KSOFM 64	64	8 x 8	0.1	150

โดยในการทดลองพบว่าการกำหนดค่าพารามิเตอร์ของโครงข่ายประสาทเทียมแบบ KSOFM และการกำหนดพารามิเตอร์ในการสอนระบบรู้จำ ที่เหมาะสมที่จะใช้การวิจัยครั้งนี้ควรกำหนดค่าพารามิเตอร์ให้กับโครงข่ายประสาทเทียมแบบ KSOFM ดังนี้ KSOFM 25 ประกอบด้วยจำนวน 25 นิวรอน (Neuron) ขนาด 5x5 จำนวนรอบ (Epochs)ในการเรียนรู้ (Train) 250 รอบ โดยกำหนดค่าผิดพลาดการรู้จำ η_0 ไม่เกิน 0.1 KSOFM 36 ประกอบด้วยจำนวน 36 นิวรอน (Neuron) ขนาด 6x6 จำนวนรอบ(epochs)ในการเรียนรู้ (Train) 200 รอบ โดยกำหนดค่าผิดพลาดการรู้จำ η_0 ไม่เกิน 0.1 และ KSOFM 64 ประกอบด้วยจำนวน 64นิวรอน (Neuron) ขนาด 8x8 จำนวนรอบ(epochs)ในการเรียนรู้ (Train) 150 รอบ โดยกำหนดค่าผิดพลาดการรู้จำ η_0 ไม่เกิน 0.1

ในขั้นตอนการเรียนรู้(Train) ของโครงข่ายประสาทเทียมแบบ KSOFM ในการทดลองได้แบ่งกลุ่มข้อมูลเสียงพูดออกเป็น 2 ชุด ประกอบด้วยชุดการเรียนรู้ของโครงข่ายประสาทเทียม จำนวน 50 คน (T=50) แบ่งเป็นเพศชาย 45 คน เพศหญิง 5 คน และ ชุดของการทดสอบ (Test) การรู้จำจำนวน 10 คน (T=10) แบ่งเป็นเพศชาย 5 คน เพศหญิง 5 คน

ในการทดสอบการรู้จำของโครงข่ายประสาทเทียมแบบ KSOFM ได้ทำการทดสอบระบบดังนี้ นำข้อมูลเสียงพูดจำนวน 50 คน (T=50) แบ่งเป็นเพศชาย 45 คน เพศหญิง 5 คน เข้าสู่ระบบของการเรียนรู้โดยที่แต่คนจะบันทึก วลี (Phrase) ละ 3 ครั้ง ในการนำข้อมูลเข้าการเรียนรู้ข้อมูลแต่ละชุดจะถูกการป้อนเข้าระบบจำนวน 30 ครั้งแต่ละวลี (Phrase) ในแต่ละชนิดของโครงข่ายประสาทเทียมแบบ KSOFM คือ KSOFM 25 KSOFM 36 KSOFM 64 ในตารางที่ 4.2

ตารางที่ 4.2 อัตราความถูกต้องของการระบุผู้พูด

Type	Data	T=50 Average 30 ครั้ง	T=50 Average 30 ครั้ง	T=50 Average 30 ครั้ง	T=50 Total Average
KSOFM 25	Phrase 1 = กินข้าวกัน	56.66%	63.33%	59.99%	59.99%
	Phrase 2 = อร่อยจังเลย	60.00%	66.66%	53.33%	60.00%
	Phrase 3 = มีอะไรทาน	73.33%	76.66%	74.99%	74.99%
KSOFM 36	Phrase 1 = กินข้าวกัน	73.33%	86.66%	79.99%	79.99%
	Phrase 2 = อร่อยจังเลย	70.00%	80.00%	75.00%	75.00%
	Phrase 3 = มีอะไรทาน	83.33%	86.66%	84.99%	84.99%
KSOFM 64	Phrase 1 = กินข้าวกัน	83.33%	90.00%	86.66%	86.66%
	Phrase 2 = อร่อยจังเลย	86.66%	93.33%	89.99%	89.99%
	Phrase 3 = มีอะไรทาน	93.33%	93.33%	93.33%	93.33%

จากตารางที่ 4.2 อัตราความถูกต้องของการระบุผู้พูดได้มาจากค่าเฉลี่ยของความถูกต้องในการทดสอบผลการเรียนรู้ กล่าวคือ ในการทดสอบความถูกต้องได้นำข้อมูลของการเรียนรู้ (T =50) ดึงออกมาทดสอบการรู้จำจำนวน 20 คน (T (Test) = 20) และทำการทดสอบความถูกต้องในการระบุผู้พูดรายละเอียดผลการทดลองดังนี้

KSOFM 25 ใน **Phrase 1 = กินข้าวกัน** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 56.66 %) ครั้งที่ 2 (T=50 Average 30 times = 63.33 %) ครั้งที่ 3 (T=50 Average 30 times = 59.99 %) และความถูกต้องเฉลี่ย (Total Average = 59.99 %)

KSOFM 25 ใน **Phrase 2 = อร่อยจังเลย** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times =60.00%) ครั้งที่ 2 (T=50 Average 30 times = 66.66 %) ครั้งที่ 3 (T=50 Average 30 times = 53.33 %) และความถูกต้องเฉลี่ย (Total Average = 60.00 %)

KSOFM 25 ใน **Phrase 3 = มีอะไรทาน** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times =73.33%) ครั้งที่ 2 (T=50 Average 30 times = 76.66 %) ครั้งที่ 3 (T=50 Average 30 times = 74.99 %) และความถูกต้องเฉลี่ย (Total Average = 74.99 %)

KSOFM 36 ใน **Phrase 1 = กินข้าวกัน** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 73.33%) ครั้งที่ 2 (T=50 Average 30 times = 86.66 %) ครั้งที่ 3 (T=50 Average 30 times = 79.99 %) และความถูกต้องเฉลี่ย (Total Average = 79.99 %)

KSOFM 36 ใน **Phrase 2 = อร่อยจังเลย** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 70.00 %) ครั้งที่ 2 (T=50 Average 30 times = 80.00 %) ครั้งที่ 3 (T=50 Average 30 times = 75.00 %) และความถูกต้องเฉลี่ย (Total Average = 75.00 %)

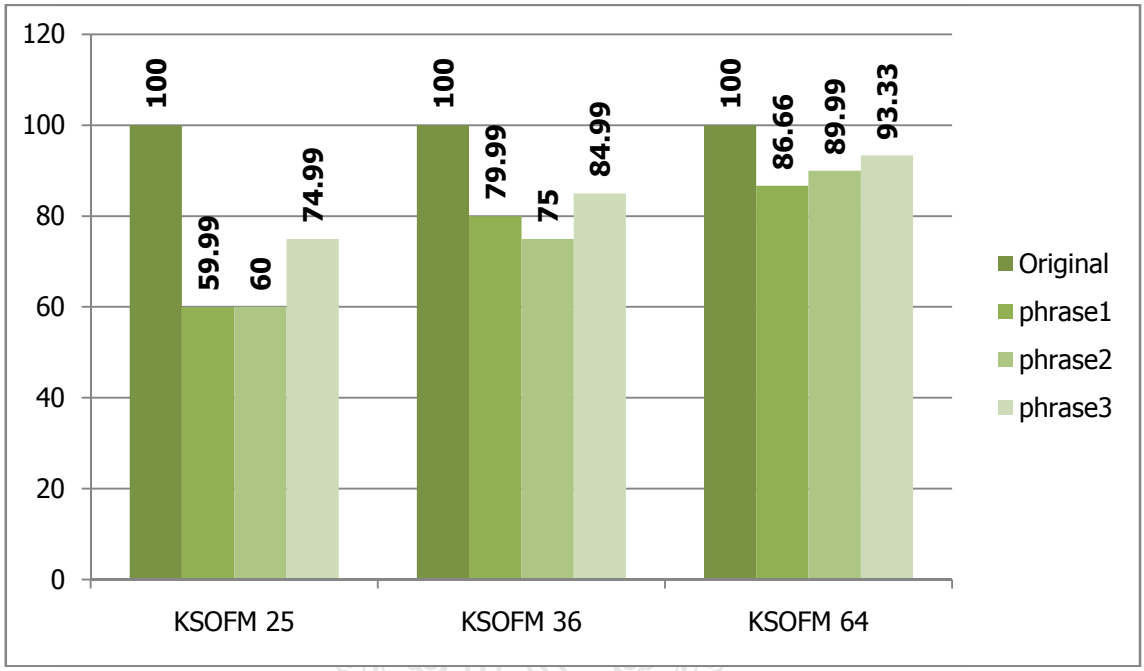
KSOFM 36 ใน **Phrase 3 = มีอะไรทาน** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 83.33 %) ครั้งที่ 2 (T=50 Average 30 times = 86.66 %) ครั้งที่ 3 (T=50 Average 30 times = 84.99 %) และความถูกต้องเฉลี่ย (Total Average = 84.99 %)

KSOFM 64 ใน **Phrase 1 = กินข้าวกัน** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 83.33 %) ครั้งที่ 2 (T=50 Average 30 times = 90.00 %) ครั้งที่ 3 (T=50 Average 30 times = 86.66 %) และความถูกต้องเฉลี่ย (Total Average = 86.66 %)

KSOFM 64 ใน **Phrase 2 = อร่อยจังเลย** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 86.66 %) ครั้งที่ 2 (T=50 Average 30 times = 93.33 %) ครั้งที่ 3 (T=50 Average 30 times = 89.99 %) และความถูกต้องเฉลี่ย (Total Average = 89.99 %)

KSOFM 64 ใน **Phrase 3 = มีอะไรทาน** มีความถูกต้องในการระบุผู้พูดเฉลี่ยครั้งที่ 1(T=50 Average 30 times = 93.33 %) ครั้งที่ 2 (T=50 Average 30 times = 93.33 %) ครั้งที่ 3 (T=50 Average 30 times = 93.33 %) และความถูกต้องเฉลี่ย (Total Average = 93.33 %)

ซึ่งในการทดลองพบว่า KSOFM 64 ในวลีที่ 3 (Phrase 3 = มีอะไรทาน) ได้ผลการรู้จำผู้พูดสูงที่สุดเฉลี่ย (Total Average) 93.33 เปอร์เซ็นต์ และพบว่า KSOFM 25 ในวลีที่ 1 (Phrase 1 = กินข้าวกัน) ได้ผลการรู้จำผู้พูดต่ำที่สุดเฉลี่ย (Total Average) 59.99 เปอร์เซ็นต์



รูปที่ 4.3 กราฟการเปรียบเทียบอัตราความถูกต้องของระบุผู้พูด

4.2 ผลของอัตราการผิดพลาดในการระบุผู้พูด

จากการทดลองอัตราความผิดพลาดในการระบุผู้พูดของระบบ โดยนำข้อมูลจำนวน 10 คน (T=10) ประกอบด้วยเพศชาย 5 คน และ เพศหญิง 5 คนที่เลือกไว้จากชุดข้อมูล 60 คนโดยข้อมูลที่นำมาทดสอบเป็นข้อมูลที่ยังไม่ผ่านกระบวนการรู้จำของโครงข่ายประสาทเทียมแบบ KSOFM มาก่อน

จากตารางที่ ตารางที่ 4.3 อัตราความถูกต้องของการระบุผู้พูดพบว่า พบว่า KSOFM 64 ในวลีที่ 3 (Phrase 3 = มีอะไรทาน) ได้ผลการระบุผู้พูดต่ำสุด (Average) 10.37 เปอร์เซ็นต์ และพบว่า KSOFM 25 ในวลีที่ 1 (Phrase 1 = กินข้าวกัน) และ KSOFM 25 ในวลีที่ 2 (Phrase 2 = อร่อยจังเลย) ได้ผลการระบุผู้พูดสูงสุดที่สุดเฉลี่ย (Average) 30.00 เปอร์เซ็นต์

ตารางที่ 4.3 อัตราความถูกต้องของการระบุผู้พูดเมื่อทำการทดสอบผู้พูด (T=10)

Type	Data	T=50 Average 30 ครั้ง	T=50 Average 30 ครั้ง	T=50 Average 30 ครั้ง	T=50 Total Average	T=10 Average 30 ครั้ง
KSOFM 25	Phrase 1 = กินข้าวกัน	56.66%	63.33%	59.99%	59.99%	30.00%
	Phrase 2 = อร่อยจังเลย	60.00%	66.66%	53.33%	60.00%	30.00%
	Phrase 3 = มีอะไรทาน	73.33%	76.66%	74.99%	74.99%	10.71%
KSOFM 36	Phrase 1 = กินข้าวกัน	73.33%	86.66%	79.99%	79.99%	11.43%
	Phrase 2 = อร่อยจังเลย	70.00%	80.00%	75.00%	75.00%	10.71%
	Phrase 3 = มีอะไรทาน	83.33%	86.66%	84.99%	84.99%	10.62%
KSOFM 64	Phrase 1 = กินข้าวกัน	83.33%	90.00%	86.66%	86.66%	10.83%
	Phrase 2 = อร่อยจังเลย	86.66%	93.33%	89.99%	89.99%	11.25%
	Phrase 3 = มีอะไรทาน	93.33%	93.33%	93.33%	93.33%	10.37%

จากตารางที่ 4.3 ยังพบว่าอัตราการผิดพลาดในการระบุผู้พูดของ KSOFM 64 ในวลีที่ 3 (Phrase 3 = มีอะไรทาน) มีอัตราการผิดพลาดในการระบุผู้พูดสูง จากการทดลองสามารถระบุผู้พูดถูก เท่ากับ 10.37 เปอร์เซ็นต์ หมายความว่า สามารถระบุผู้พูดได้ถูกต้องได้น้อยมาก และยังพบว่า KSOFM 25 ในวลีที่ 1 (Phrase 1 = กินข้าวกัน) และ KSOFM 25 ในวลีที่ 2 (Phrase 2 = อร่อยจังเลย) มีอัตราการผิดพลาดในการระบุผู้พูดต่ำ จากการทดลองสามารถระบุผู้พูดถูกต้องสูงกว่า KSOFM 64 ในวลีที่ 3 เท่ากับ 30.00 เปอร์เซ็นต์ หมายความว่า สามารถระบุผู้พูดได้ถูกมากกว่า KSOFM 64 ในวลีที่ 3

อาจกล่าวได้ว่าในทดสอบการระบุผู้พูดโดยการนำข้อมูล(T=10) ที่ยังไม่ผ่านกระบวนการรู้จำของโครงข่ายประสาทเทียม KSOFM มาให้ทำการทดสอบระบุผู้พูด KSOFM ซึ่งจะต้องไม่สามารถระบุผู้พูดได้ค่าความถูกต้องเฉลี่ยในกรณีการทดสอบ (T=10) ต้องเป็น 0.00 เปอร์เซ็นต์ คือไม่สามารถระบุผู้พูดได้เพราะ KSOFM ไม่รู้จัก แต่ในการทดลองพบว่า ค่าความถูกต้องของ KSOFM ยังสามารถระบุผู้พูดได้แต่ก็มีเปอร์เซ็นต์ไม่สูงนัก สาเหตุอาจมาจากค่าผิดพลาด (Error) ของการทำงาน KSOFM กระบวนการบันทึกเสียงพูด หรือการวิเคราะห์สัญญาณพูดก็เป็นได้

จากตารางที่ 4.4 อัตราความผิดพลาดในการระบุผู้พูดที่ได้จากการทดลองในวลี ต่างๆการทดลองในวลีที่ 1 (Phrase 1 = กินข้าวกัน) มีเปอร์เซ็นต์การผิดพลาดมากที่สุด 24.44 และในวลีที่ 3 (Phrase 3 = มีอะไรทาน) มีเปอร์เซ็นต์การผิดพลาดน้อยที่สุด 15.56 เปอร์เซ็นต์ และจากตารางข้อมูลพบว่าค่าเฉลี่ยของความถูกต้องในการระบุผู้พูดเป็น 78.33 เปอร์เซ็นต์ แสดงดังตารางที่ 4.4

ตารางที่ 4.4 อัตราความผิดพลาดในการระบุผู้พูดที่ได้จากการทดลองในวลีต่างๆ

Phrases	**Average**	**Duration Standard Deviation**	**Errors %**	**Correct Rate**
Phrase 1 = กินข้าวกัน	0.45 s	0.44 s	24.45 %	75.55%
Phrase 2 = อร่อยจังเลย	1.12 s	0.51 s	25.01 %	74.99 %
Phrase 3 = มีอะไรทาน	2.55 s	0.24 s	15.56%	84.44%
เฉลี่ย			21.67 %	78.33 %

นอกจากอัตราการผิดพลาดในการระบุผู้พูดซึ่งมีอัตราผิดพลาดเฉลี่ยเท่ากับ 21.67 เปอร์เซ็นต์ และระบบมีความถูกต้องเฉลี่ยเท่ากับ 78.33 เปอร์เซ็นต์ ซึ่งเมื่อทำการวิเคราะห์ข้อมูลในเรื่องของเวลาในการบันทึกเสียงพูดพบว่าใน Phrase 1 = **กินข้าวกัน** มีค่าเวลาเฉลี่ย 0.45 วินาที Phrase 2 = **อร่อยจังเลย** มีค่าเวลาเฉลี่ย 1.12 วินาที และ Phrase 3 = **มีอะไรทาน** มีค่าเวลาเฉลี่ย 2.55 วินาที โดยที่เวลาเฉลี่ยของ วลีที่ 1 (Phrase 1)น้อยกว่า วลีที่ 2 (Phrase 2)และ วลีที่ 3 (Phrase 3) เนื่องจากเป็น วลีที่1 (Phrase 1) เป็นคำ 3 พยางค์ ขณะที่ วลีที 2 (Phrase 2) และ วลีที่ 3 (Phrase 3) เป็นคำ 4 พยางค์ ในส่วนของ Duration Standard Deviation เป็นค่าเบี่ยงเบนมาตรฐานซึ่งสามารถศึกษาการกระจายของข้อมูลโดยมีสมการ

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

เมื่อ N คือจำนวนสมาชิกของเซตข้อมูล จากนั้นจึงสามารถคำนวณค่าเบี่ยงเบนมาตรฐานได้จาก

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

ในทางปฏิบัติ การคำนวณค่าเบี่ยงเบนมาตรฐานของตัวแปรสุ่มชนิดไม่ต่อเนื่องข้างต้น สามารถสรุปได้ดังนี้

1. สำหรับแต่ละค่าของ x_i ให้คำนวณผลต่างของ $x_i - \overline{x}$
2. นำผลต่างแต่ละตัวมายกกำลังสอง
3. บวกผลลัพธ์ทั้งหมดเข้าด้วยกันแล้วหารด้วย N ค่าที่ได้นี้คือความแปรปรวน σ^2
4. คำนวณหารากที่สองที่เป็นบวกของความแปรปรวน จะได้ค่าเบี่ยงเบนมาตรฐาน

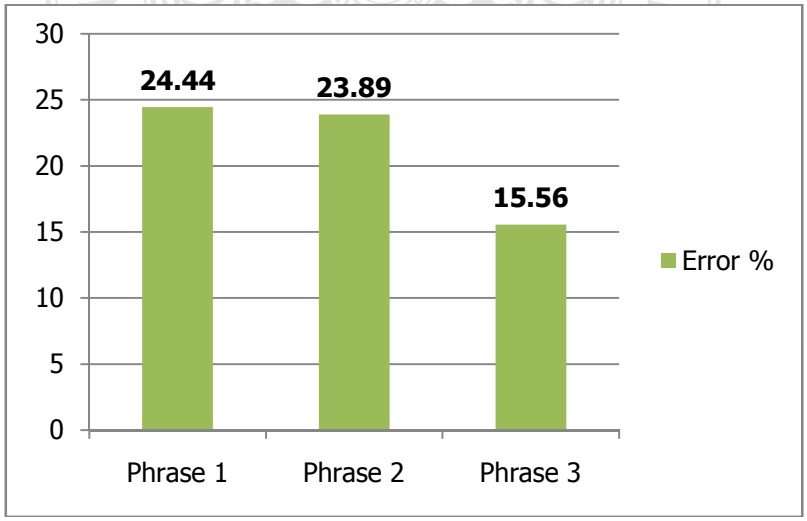
นอกจากนั้นสูตรดังกล่าวสามารถดัดแปลงให้เป็นอีกรูปแบบหนึ่งได้ดังนี้

$$\sigma = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right)}$$

ซึ่งความเท่ากันของทั้งสองสูตร สามารถพิสูจน์ได้ด้วยความรู้ทางพีชคณิต

$$\begin{aligned}
\sum_{i=1}^{N}(x_i - \overline{x})^2 &= \sum_{i=1}^{N}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\
&= \left(\sum_{i=1}^{N} x_i^2\right) - \left(2\overline{x}\sum_{i=1}^{N} x_i\right) + N\overline{x}^2 \\
&= \left(\sum_{i=1}^{N} x_i^2\right) - 2\overline{x}(N\overline{x}) + N\overline{x}^2 \\
&= \left(\sum_{i=1}^{N} x_i^2\right) - 2N\overline{x}^2 + N\overline{x}^2 \\
&= \left(\sum_{i=1}^{N} x_i^2\right) - N\overline{x}^2
\end{aligned}$$

จากการวิเคราะห์ข้อมูลด้วยสมการ Duration Standard Deviation พบว่า ใน Phrase 1(กินข้าวกัน) มีค่า Duration Standard Deviation สูง กว่า Phrase 2 (อร่อยจังเลย) Phrase 3 (มีอะไรทาน) หมายความว่า ในการกระจายของข้อมูลพบว่า Phrase 3 (มีอะไรทาน) เวลาของข้อมูลเกาะกลุ่มกัน มากกว่า Phrase 1 (กินข้าวกัน) ที่มีการกระจายข้อมูลมากกว่า เมื่อเทียบกันทั้ง 3 Phrase ซึ่งจากการทดลอง อาจทำให้มีอัตราการผิดพลาดในการระบุผู้พูดของ Phrase 1(กินข้าวกัน) สูงกว่า จึงทำให้ไม่ สามารถระบุตัวผู้พูดถูกน้อย



รูปที่ 4.4 กราฟอัตราความผิดพลาดในการระบุผู้พูด

4.3 การประเมินคุณภาพเสียงพูด MOS (Mean Opinion Score)

การประเมินคุณภาพของข้อมูลเสียงพูดโดยใช้วิธีการของ MOS (Mean Opinion Score) เป็นวิธีการหนึ่งที่นิยมใช้เปรียบเทียบระหว่างเสียงพูดต้นแบบกับเสียงพูดที่ผ่านการบีบอัดข้อมูล โดยใช้การรับรู้และความรู้สึกของมนุษย์เป็นเกณฑ์ในการตัดสิน (Subjective Measurement)

วิธีการประเมินหรือวัดคุณภาพเสียงนั้นจะใช้คนประมาณ 12-24 คน ทดสอบคุณภาพเสียงด้วยการฟัง โดยที่แต่ละคนจะให้คะแนนที่มีค่าอยู่ระหว่าง 1-5 ตามคุณภาพของสัญญาณที่ตัวเองรู้สึก จากนั้นหาค่าเฉลี่ยแต่ละเสียงพูดว่าอยู่ในระดับใด

ตารางที่ 4.5 วิธีการให้คะแนนในการวัด MOS

คะแนน	คุณภาพเสียง
5	ดีมาก (คุณภาพเสียงชัดเจนและเข้าใจง่าย)
4	ดี (คุณภาพเสียงดีและเข้าใจง่าย แต่อาจได้ยินเสียงรบกวนบ้าง)
3	พอใช้ (คุณภาพเสียงดีและเข้าใจ ได้แต่อาจต้องการอาศัยความตั้งใจ หรือบางทีต้องขอให้พูดซ้ำ)
2	เลว (คุณภาพเสียงดีและเข้าใจได้ก็ต่อเมื่อมีความตั้งใจมาก ๆ และบ่อยครั้งที่ต้องขอให้พูดซ้ำ)
1	เลวมาก (ฟังไม่รู้เรื่องเลย)

ตารางที่ 4.6 ผลการประเมินคุณภาพเสียงโดยใช้วิธีการของ MOS (Mean Opinion Score)

Type	Data	MOS
KSOFM 25	Phrase 1 = กินข้าวกัน	2.8
	Phrase 2 = อร่อยจังเลย	**2.6**
	Phrase 3 = มีอะไรทาน	2.9
KSOFM 36	Phrase 1 = กินข้าวกัน	3.0
	Phrase 2 = อร่อยจังเลย	3.2
	Phrase 3 = มีอะไรทาน	3.3
KSOFM 64	Phrase 1 = กินข้าวกัน	3.5
	Phrase 2 = อร่อยจังเลย	3.6
	Phrase 3 = มีอะไรทาน	**3.7**

จากผลการวิจัยโดยวิธีการประเมินคุณภาพพบว่า KSOFM 64 ในวลีที่ 3 ประเมินโดยวิธี MOS มีคุณภาพของอยู่ในระดับ 3.7 และ ใน KSOFM 25 ในวลีที่ 2 มีค่าอยู่ในระดับต่ำสุด 2.6

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในบทนี้จะกล่าวถึงสรุปผลการวิจัยระบบการรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง และข้อเสนอแนะ

5.1 สรุปผลการวิจัยระบบรู้จำผู้พูด

งานวิจัยในวิทยานิพนธ์ฉบับนี้เป็นระบบการรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง โดยปกติเสียงพูดของแต่ละบุคคลจะถูกจำลองโดย Line spectral pairs : LSP ซึ่งจะทำการแยกคุณลักษณะเฉพาะ โดยในการวิจัยได้ปรับปรุงโครงข่ายประสาทเทียมจะถูกเรียนรู้โดยใช้คุณลักษณะที่ได้ทำการแยกคุณลักษณะเฉพาะ ในการใช้อัลกอริทึม K-means สำหรับในการเพิ่มความสามารถในการรู้จำ โดยงานวิจัยนี้ได้ทำการทดสอบคำไม่ต่ำกว่า 3 พยางค์ จำนวน 3 สำนวน ใช้โครงข่ายประสาทเทียมแบบ KSOFM อัลกอริทึม K-means ซึ่งในการทดลองได้กำหนดโหนดแบ่งเป็น KSOFM 25 KSOFM 36 และ KSOFM 64

ระบบการรับจำเสียงผู้พูดบนฐานการปรับปรุงของ Kohonen Self-Organizing (KSOFM) โครงข่ายประสาทเทียมแบบ KSOFM อัลกอริทึม K-means ที่ได้ทำการวิจัยนี้โดยที่ระบบนี้บรรลุเป้าหมายประมาณ 93.33 % ในค่าความถูกต้องทั้ง 3 วลี เมื่อผ่านการฝึกกับ KSOFM 64

เสียงในการพูดของแต่ละบุคคลได้จำลองโดยโครงข่ายระบบประสาทเทียมแบบ KSOFM โดยระบบการรู้จำทางด้านเสียงใช้ สัมประสิทธิ์เวกเตอร์กระบวนการ Quantization ของ LSF ซึ่งถูกใช้เป็นคุณลักษณะที่สุ่มจากส่วนของเสียงตัวอย่าง เมื่อมีเสียงตัวอย่างใหม่เข้ามาก็จะเกิดการแข่งขันกันสำหรับผู้พูด ผลลัพธ์ที่ได้ออกมาจาก KSOFM จะมีการ Error ของกระบวนการ Quantization น้อยที่สุด จึงนับได้ว่าระบบการรู้จำเสียงพูด KSOFM สามารถเรียนรู้ได้ถูกต้อง ส่วนค่า K หมายถึง Algorithm ใช้ในการแยกแยะออกจาก KSOFM ในกลุ่มเวกเตอร์ที่ใกล้ที่สุด โดยเทคนิคนี้สามารถเพิ่มความถูกต้องของระบบ

จากการวิจัยพบว่า สถาปัตยกรรมคอมพิวเตอร์มีส่วนอย่างมากในการเพิ่มประสิทธิ์ภาพในการจำแนก ระบุข้อผิดพลาดของ กระบวนการ Quantization ในการฝึก และมีค่าความถูกต้องดีที่สุด มันอาจใช้ระยะเวลาที่นาน และการออกเสียงที่สมดุลกับการฝึกมีค่าความถูกต้องสูง

5.2 ข้อเสนอแนะ

1. ควรทดลองกับผู้พูดขนาดใหญ่ เพื่อเพิ่มประสิทธิภาพ
2. ควรลดเวลาในระบบการรู้จำเพื่อลดความผิดพลาด

เอกสารอ้างอิง

[1] กาญจนา ทองบุญนาค. การรู้จำเสียงคำโดดด้วยโครงข่ายประสาทเทียม. การค้นคว้าอิสระเชิงวิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยเชียงใหม่, 2544.

[2] ปฐวี ชาญไววิทย์. ระบบรู้จำทำนองเสียงพูดสำหรับเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2546.

[3] พิชัย ชูกาญจนพิทักษ์. การรู้จำเสียงพูดคำไทยต่อเนื่องจำนวนคำศัพท์ปานกลางเฉพาะบุคคล. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2545.

[4] ชัย วุฒิวิวัฒน์ชัย. การรู้จำเสียงคำไทยหลายพยางค์แบบไม่ขึ้นกับผู้พูดโดยใช้เทคนิคแบบฟัซซีและนิวรอลเน็ตเวิร์ก. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2540.

[5] ฐนียา สัตยพานิช. ระบบรู้จำเสียงภาษาไทยต่อเนื่องแบบเฉพาะบุคคลสำหรับการใช้งานอีเมล์. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์, 2546.

[6] เอกชัย เนาวนิช. โปรแกรมฝึกออกเสียงพยัญชนะไทยสำหรับผู้บกพร่องทางการได้ยินโดยใช้โคร่งข่ายประสาทเทียม. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2550.

[7] ปรีดาวรรณ เกษเมธีการุณ. การพัฒนาการรู้จำเสียงสำหรับพยัญชนะต้นของอัฒพยางค์. วิทยานิพนธ์ครุศาสตรอุตสาหกรรมมหาบัณฑิต สาขาวิชาเทคโนโลยีคอมพิวเตอร์ ภาควิชาคอมพิวเตอร์ศึกษา บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2548.

[8] อนุรักษ์ โนสาน และธันวา ศรีประมง. “การแยกหน่วยเสียงสระเดี่ยวภาษาไทย โดยใช้คุณสมบัติเฉพาะเชิงความถี่.” Engineer Transactions (Group A). Volume 9 (July-Dec 2006) : 48-56.

[9] นริศ บุญศักดิ์เฉลิม วรา คงคาวิทูร และ ไกรสิน ส่งวัฒนา. “การรู้จำหน่วยเสียงสระเดี่ยวสำหรับภาษาไทยโดยการใช้ ทรานส์เฟอร์ฟังก์ชันของอวัยวะกำทอนเสียงบนสเกลบาร์ก.” NECTEC Technic

[10] J. Srinonchat, Investigation and Explotation of the Repetiveness of Speech Signals in a Speaker Dependent Coding System, Ph.D Thesis, University of Northumbria, 2005.

[11] Wei HAN, Cheong-Fat CHAN,Chiu-Sing CHOY and Kong-Pang PUN. An Efficient MFCC Extraction Method in Speech Recognition. [Online] 2006. Available from: http://ieeexplore.ieee.org/iel5/8698/27542/01232085.pdf.

[12] ศวิต กาสุริยะ. การบ่งชี้ผู้พูดแบบขึ้นกับบทคำพูด. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2542.

[13] เอกรัฐ ปัญญาเทพ. ระบบการตรวจรู้เสียงสระภาษาไทยหนึ่งพยางค์โดยใช้โครงข่ายนิวโรฟัซซี. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2548.

[14] T. Kohonen, O. Simula and J. Kangas, “Self-organizing maps: Optimization approaches,” Journal of Artificial Neural Networks, 1991, pp. 981-990.

[15] ไชยันต์ สุวรรณชีวะศิริ. การรู้จำเสียงพูดภาษาไทยแบบไม่ขึ้นกับผู้พูด โดยใช้ลักษณะเด่นบ่งความแตกต่างของหน่วยเสียง. รายงานการวิจัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2542.

[16] สุธีร์ ครบุรี. บทเรียนคอมพิวเตอร์ช่วยสอนการฝึกฟัง เรื่องพยัญชนะ สระ และตัวเลขในภาษาไทยสำหรับผู้บกพร่องทางการได้ยิน. สารนิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2548.

[17] ศิฬาณี นุชิตประสิทธิ์ชัย. โปรแกรมแปลงเสียงพูดเป็นภาษามือเพื่อการติดต่อสื่อสารกับผู้บกพร่องทางการได้ยิน. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2548.

[18] วารินทร์ อัจฉริยะกุลพร ชัย วุฒิวิวัฒน์ชัย และ จุฬารัตน์ ตันประเสริฐ. “ระบบระบุผู้พูดภาษาไทยด้วยวิธีไดนามิกส์ไทม์วอร์ปปิง.” NECTEC Technical Journal. Vol II, No.8, (July-October 2000) : 108-118.

[19] เยาวลักษณ์ ชาติสุขศิริเดช. การใช้เสียงในภาษาไทย. กรุงเทพฯ : สำนักพิมพ์อักษรเจริญทัศน์, 2548.

[20] อัจฉรา ภูมิชูชิต บทเรียนช่วยสอนวิชาวิทยาศาสตร์ เรื่องบรรยากาศสำหรับผู้บกพร่องทางการได้ยิน. สารนิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2549.

[21] Fry, D, B. The Physics of Speech, Cambridge University Press, 1979.

[22] Furui, Sadaoki, “Recent Advances in Speaker Recognition.” Audio and Video based biometrics Person Authentication, 1997.

[23] S. Limpanakorn and C, Tanprasert. “Voice Articulator for Thai Speaker Recognition System,” ICONIP’02, vol.5, p.2396-2400.

[24] A. T. Mafra and M. G. Simoes, “Text Independent Automatic Speaker Recognition using Self-Organizing Maps”, IAS 2004, p.1503-1510.

[25] C. Wutiwiwatchai and S. Furui, “Thai speech processing technology: A review,” Speech Communication, Volume 49, Issue 1 , January 2007, p.8-27.

[26] Haykin, S., Neural Networks: A comprehensive foundation. 1999:Prentice Hall, pp. 1-379.

[27] T. Kohonen,Self-Organizing Maps. 2nd edition. Springer, 1997.

[28] J.Srinonchat et al., An Efficient Codebook Design for Speaker Dependent Coding System, 4th International Symposium Communication Systems, Networks and Digital Signal Processing, 2004, p.484-486.

[29] A. M. Kondoz, “Digital Speech: coding for low bit rate communication systems,” John Wiley & sons Publishers, 1998.

[30] IEEE recommended practice for speech quality measurements. IEEE Transactions on Audio and Electroacoustics, pp.227-246, 1969.

[31] Furui S. “Speaker-independent isolated word recognition using dynamic features of speech spectrum.” IEEE Transaction on Acoustics, Speech and Signal Proceesing. 1986.

[32] O’shaunghnessy, D. Speech Communication Human and Machine. Addison—Wesley Publicshing Company. 1987.

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

[34] Amarin Deemagarn, Asanee Kawtrakul. Thai Connected Digit Speech Recognition Using Hidden Markov Models. [Online] 2004. Available from: http://www.isca-speech.org/archive/specom_04/spc4_731.pdf.

[35] K. Songwatana, S. Sriratanapaprat, P. Kultap, K. Sittiprasert and N. Suktangman. Recognition of 24 Thai spoken Vowels Using the coefficients of 3rd Order Polynomial Regression on the Voice Energy and Spectrum of LPC on the Bark Scale. [Online] 2006. Available from: http://ieeexplore.ieee.org/iel5/10746/33870/01613615.pdf.

[36] Marius Zbancioc, Mihaela Costin. Using Neural Network and LPCC to Improve Speech Recognition. [Online] 2003. Available from: http://ieeexplore.ieee.org/iel5/8698/27542/01227085.pdf.

[37] Xu Wang.,Lifang Xue., and Dan Yang. Speech Plot Display for Deaf-mute base on Combined Characters Encoding of Speech signal . [Online] 2007. Available from: http://ieeexplore.ieee.org/iel5/4381672/4381673/04381935.pdf.

[38] Chularat Tanprasert, Chai Wutiwiwatchai, Sutat Sae-tang. "Text-dependent Speaker Identification Using Neural Network On Distinctive Thai Tone Marks." NECTEC Technical Journal. Vol I , No.6, (January-February 2000) : 249-253.

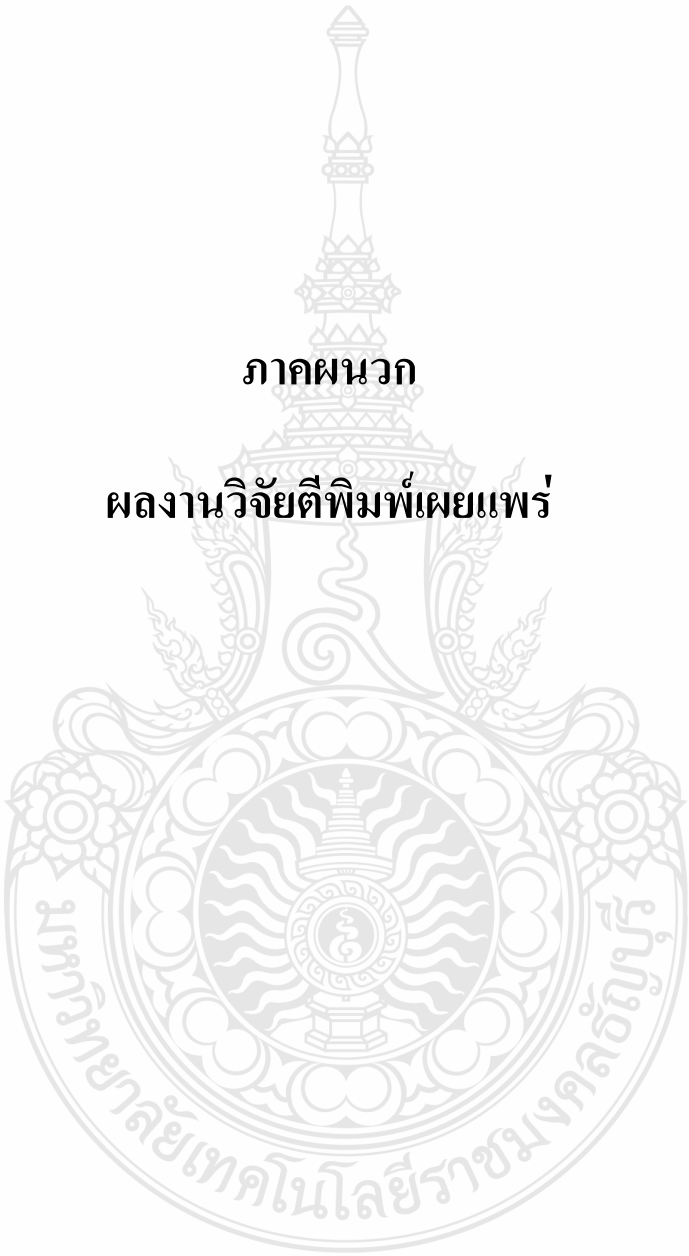
[39] Yun-Hsuan Sung Speech Recognition and Synthesis. [Online] 2006. Available from: http://www.stanford.edu/class/cs224s/.

[40] Jyh-Shing Roger Jang Audio Signal Processing and Recognition. [Online] 2006 Available from: http://www.cs.nthu.edu.tw/~jang.

[41] Marie Roch. Cepstral Processing. San DiegoState University [Online] Available from: http://www-rohan.sdsu.edu/~mroch/cs682/slides/06Cepstral.

ภาคผนวก

ผลงานวิจัยตีพิมพ์เผยแพร่



ประวัติผู้เขียน

ชื่อ - นามสกุล	นายสุวุฒิ ตุ้มทอง
วัน เดือน ปีเกิด	12 ธันวาคม 2511
ที่อยู่	42/4 หมู่ 2 ต.บางเขน อ.เมือง จ.นนทบุรี 11000
ประวัติการศึกษา	สำเร็จการศึกษาระดับครุศาสตรอุตสาหกรรมมหาบัณฑิต (ไฟฟ้า) จากสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ เมื่อ พ.ศ.2543
ประวัติการทำงาน	
พ.ศ. 2543 – ปัจจุบัน	อาจารย์ประจำสาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ (นนทบุรี)

ผลงานวิจัยที่ตีพิมพ์

สุวุฒิ ตุ้มทอง และ จักรี ศรีนนท์ฉัตร ,“การรู้จำผู้พูดโดยใช้เทคนิคโครงข่ายประสาทเทียมแบบคลัสเตอร์ริ่ง”, The 2007 Joint International Conference on Information Communication Technology (JICT), 19-22 December 2007, Vientiane, Lao PDR, 2007. หน้า 153-157