การประยุกต์ใช้ระบบรู้จำเสียงพูดคำไทยสำหรับงานพิมพ์เอกสาร โดยใช้เทคนิควิเคราะห์สเปกตรัมและโครงข่ายประสาทเทียม

ADAPTATION OF THAI SPEECH RECOGNITION SYSTEM FOR DOCUMENT TYPING USING SPECTRUM ANALYSIS AND NEURAL NETWORK

เอกรินทร์ แซ่เฮ้ง

EGKARIN SAEHENG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีคมนาคม

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2547

ISBN 974-9680-59-6

การประยุกต์ใช้ระบบรู้จำเสียงพูดคำไทยสำหรับงานพิมพ์เอกสาร โดยใช้เทคนิควิเคราะห์สเปกตรัมและโครงข่ายประสาทเทียม

ADAPTATION OF THAI SPEECH RECOGNITION SYSTEM FOR DOCUMENT TYPING USING SPECTRUM ANALYSIS AND NEURAL NETWORK

เอกรินทร์ แซ่เฮ้ง
EGKARIN SAEHENG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ.2547

ISBN 974-9680-59-6

ADAPTATION OF THAI SPEECH RECOGNITION SYSTEM FOR DOCUMENT TYPING USING SPECTRUM ANALYSIS AND NEURAL NETWORK

EGKARIN SAEHENG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUREMENT FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2004

ISBN 974-9680-59-6

หัวข้อวิทยานิพนธ์	การประยุกต์ใช้ระบบรู้จำเสียงพูดคำไทยสำหรับงานพิมพ์เอกสารโดยใช้เทคนิควิเคราะห์สเปกตรัมและโครงข่ายประสาทเทียม
ชื่อนักศึกษา	นายเอกรินทร์ แซ่เฮ้ง
รหัสประจำตัว	43067137
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2547
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รองศาสตราจารย์ ดร. วิเชียร เปรมชัยสวัสดิ์
อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร. วรพจน์ กรีสุระเดช

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการรู้จำเสียงพูดไทยโดยการอาศัยหลักการจัดกลุ่มของโครงข่ายประสาทเทียมแบบส่งค่าย้อนกลับ ร่วมกับการวิเคราะห์ค่าเซปสตรัมจากสเปกตรัมของความถี่ที่เป็นองค์ประกอบของเสียงพูดภาษาไทยในรูปแบบของสเปกตรัม เพื่อใช้กับงานพิมพ์เอกสาร โดยจัดระดับพลังงานของเสียงที่นำมาวิเคราะห์ให้อยู่ในรูปแบบมาตรฐานเดียวกันก่อนส่งมาประมวลผลในโครงข่ายประสาทเทียม เพื่อสอนให้รู้จักเสียงดังกล่าวและให้พัฒนาความสามารถของการเรียนรู้โดยอ้างอิงกับข้อมูลที่ใช้ในการทดสอบและเรียนรู้ ซึ่งโครงข่ายประสาทเทียมที่ใช้ในการประมวลผลจะมีอยู่ 3 ชุด ชุดแรกใช้วิเคราะห์คำที่มีขนาด 1 พยางค์ ส่วนชุดที่สองใช้วิเคราะห์คำที่มี 2 พยางค์ และชุดสุดท้ายสำหรับวิเคราะห์คำที่มีมากกว่า 2 พยางค์

Thesis Title	Adaptation of Thai Speech Recognition System for Document Typing using Spectrum Analysis and Neural Network
Student	Mr.Egkarin Saeheng
Student ID.	43067137
Degree	Master of Science
Programme	Information Technology
Year	2004
Thesis Advisor	Assoc.Prof.Dr. Wichian Premchaiswadi
Thesis Co-Advisor	Asst.Prof.Dr. Worapoj Kreesuradej

ABSTRACT

This thesis presents the method of Thai speech recognition by classification for Backpropagation Neural Network with the cepstral of frequency, an element of voice in spectrum form for typing. Speech signal is convert to standard form of energy power level before sent to processing for Neural Network, recognize the speech signal and develop capacity in learning, as referenced to testing data and learning. The Processing has three neural network sets, for 1, 2 and more than 2 syllables.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความกรุณาจาก รศ.ดร.วิเชียร เปรมชัยสวัสดิ์ ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่าน ตลอดจนคำชี้แนะแนวทางในการทำงานวิจัยทั้งหมด และขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ควบคุมวิทยานิพนธ์ร่วม ที่คอยให้คำปรึกษา และชี้แนะแนวทางการใช้งานโครงข่ายประสาทเทียม ซึ่งเป็นประโยชน์อย่างมากสำหรับงานวิจัยนี้

ขอขอบคุณเจ้าหน้าที่หน่วยงานราชบัณฑิตยสถาน ที่ให้ความอนุเคราะห์ข้อมูลคำศัพท์ในพจนานุกรม ซึ่งใช้เป็นฐานข้อมูลในการทดลองระบบการรู้จำ

ขอขอบคุณเจ้าหน้าที่คณะเทคโนโลยีสารสนเทศทุกท่าน ที่ไดเช่วยดำเนินการจัดการสอบและให้ข้อมูลที่เป็นประโยชน์ต่อการทำงานวิจัย

ขอขอบคุณญาติพี่น้องทุกท่านที่ให้กำลังใจที่อยู่เคียงข้างกันเสมอมา และขอขอบคุณเพื่อนๆ IS10 ที่เสียสละเวลาเพื่อบันทึกเสียง และทดลองจนงานวิจัยสำเร็จลงได้ด้วยดี

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้ และถ่ายทอดประสบการที่ดีแก่ข้าพเจ้า

เอกรินทร์ แซ่เฮ้ง

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

ในการใช้งานคอมพิวเตอร์จะใช้แป้นพิมพ์หรือเมาส์เป็นอุปกรณ์รับข้อมูล (Input Device) เพื่อสั่งงานให้คอมพิวเตอร์ทำงาน หรือรับข้อมูลตามที่ผู้ใช้งานต้องการ แต่ในบางสถานการณ์ หรือบางสภาวะผู้ใช้งานไม่สามารถใช้อุปกรณ์ดังกล่าวในการสั่งงานคอมพิวเตอร์ได้ เช่น ผู้พิการทางแขน มือ หรือนิ้ว หรืออาจมีความจำเป็นในด้านอื่น ๆ ไม่สามารถใช้มือในการสั่งงานคอมพิวเตอร์ได้ นอกจากนั้นในบางระบบงานอาจจำเป็นต้องใช้เสียงในการปฏิบัติงาน เช่น ระบบรักษาความปลอดภัยด้วยเสียง งานควบคุมด้านจักรกล เช่น หุ่นยนต์ ดังนั้นการศึกษาขบวนการรู้จำจากสัญญาณเสียงพูด (Speech) จึงเป็นขั้นเริ่มต้นของการพัฒนาและส่งเสริมให้มีการใช้งานเทคโนโลยีนี้อย่างแพร่หลายมากขึ้น เพื่อก่อให้เกิดประโยชน์และเพิ่มความสะดวกสบายแก่ผู้ใช้งานทั้งแก่ผู้พิการหรือคนปกติโดยทั่วไป

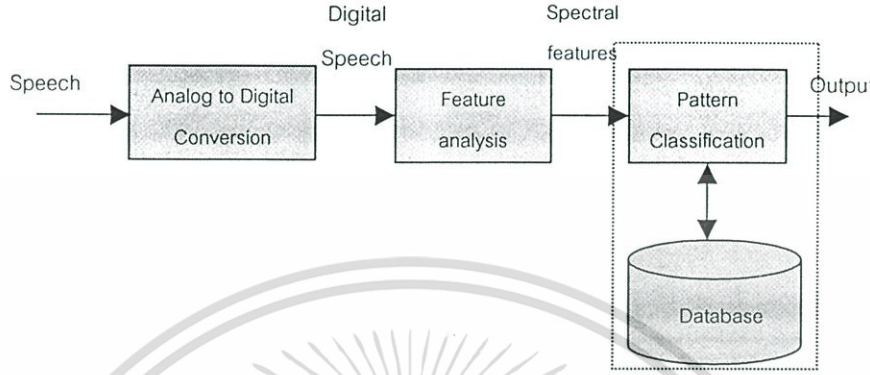
ในการศึกษาในเรื่องของระบบรู้จำเสียงพูด (Speech Recognition System) ตั้งอยู่บนพื้นฐาน 3 ประการคือ

1. ข้อมูลในสัญญาณเสียงพูด (The information in the speech signal) เราสามารถจัดข้อมูลให้อยู่ในรูปแบบต่างๆ ที่สามารถแสดงถึงรายละเอียดสำคัญของสัญญาณเสียงได้ โดยมากมักจะใช้เทคนิคที่เรียกว่า "Short time amplitude spectrum of the speech waveform" ซึ่งเทคนิคนี้จะสามารถหาข้อมูลสำคัญในข้อมูลสัญญาณเสียง เมื่อได้ข้อมูลสำคัญแล้วจึงนำไปสู่ขั้นตอนการเทียบเคียงรูปแบบ (Pattern matching)

2. การจำกัดสัญญาณเสียง (The contents of the speech signal) คือขบวนการจัดการกับข้อมูลสำคัญที่ได้จากการวิเคราะห์ เพื่อการอธิบายรูปแบบของสัญญาณต้นฉบับ ซึ่งสามารถนำกลับมาเพื่ออธิบายสัญญาณได้ ซึ่งอาจใช้สัญลักษณ์หรือตัวอักษรต่างๆ เพื่อแทนข้อมูลและลำดับของสัญญาณเสียง เช่นในระบบ Text to speech Synthesis ซึ่งสามารถเก็บเสียงที่ต้องสร้างโดยใช้สัญลักษณ์แทนเสียง และลักษณะการออกเสียง (Phonetic) หรือในพจนานุกรมใช้ตัวอักษรเพื่อแสดงลักษณะของคำต่างๆ

3. การรู้จำเสียงพูด (Speech Recognition) คือขบวนการจดจำและรู้จักรูปแบบของข้อมูล (Pattern) ซึ่งการทำความเข้าใจคำพูดของมนุษย์ได้ ถือว่าเป็นเรื่องที่ยากมาก เพราะต้องเข้าใจในหลักไวยากรณ์ ความหมาย และความยุ่งยากในโครงสร้างภาษา เพราะถ้าใช้เพียงอย่างใดอย่างหนึ่งเพื่อให้เข้าใจคำพูดนั้น อาจให้ความหมายที่ไม่ถูกต้องได้

การศึกษาและวิจัยด้านรู้จำเสียงพูดนั้นมีการพัฒนาออกมาอย่างต่อเนื่อง มีการเลือกใช้เทคนิคในการรู้จำมากมายหลากหลายวิธีการเช่น โครงข่ายประสาทเทียม, ฮิดเดนมาร์คอฟโมเดล และไดนามิกไทม์วาร์ปปิง เป็นต้น ซึ่งระบบการรู้จำเสียงพูดสามารถแสดงได้ดังรูปที่ 1.1



รูปที่ 1.1 บล็อกไดอะแกรมของระบบรู้จำเสียงพูด

ระบบการรู้จำโดยทั่วไปนั้นประกอบด้วยส่วนสำคัญ 3 ส่วน คือการแปลงสัญญาณอนาล็อกเป็นดิจิตอล (Analog to Digital Conversion) ซึ่งทำหน้าที่ในการแปลงสัญญาณเสียงพูดที่อยู่ในรูปของสัญญาณทางไฟฟ้าหรืออนาล็อก (Analog Signal) ให้กลายเป็นสัญญาณดิจิตอล (Digital Signal) โดยการสุ่มสัญญาณตัวอย่าง (Sampling Signal) ด้วยอัตราคงที่เพื่อให้ง่ายต่อการคำนวณและวิเคราะห์ ปัจจุบันขั้นตอนนี้ได้รวมอยู่ภายในอุปกรณ์การ์ดรับเสียง (Sound Card) ซึ่งเป็นอุปกรณ์ที่จำเป็นและสำคัญมากในการวิเคราะห์สัญญาณด้านเสียงจากนั้นจะส่งต่อไปยัง Feature Analysis หรือ Feature Extraction เพื่อวิเคราะห์หาคุณลักษณะที่สำคัญของข้อมูลสัญญาณเสียง (Speech Signal) สำหรับการจัดแบ่งกลุ่ม (Classification) ด้วยกระบวนการ Pattern Classification โดยตรวจสอบกับฐานข้อมูลหรือข้อมูลอ้างอิงว่ารูปแบบของข้อมูลที่ส่งเข้ามาวิเคราะห์นั้นตรงกับคำตอบอะไร [1]

ระบบการรู้จำที่มีอยู่ในปัจจุบันนี้สามารถแบ่งออกได้ 2 กลุ่มหลักๆ คือระบบรู้จำที่ขึ้นกับผู้พูด (Speaker-Dependent) และไม่ขึ้นกับผู้พูด (Speaker-Independent) กล่าวคือระบบที่ขึ้นกับผู้พูดความถูกต้องจะขึ้นตรงกับผู้พูดเป็นหลักเพราะระบบดังกล่าวนี้จะใช้ข้อมูลในการฝึกสอนรู้จำด้วยเสียงบุคคลนั้นเพียงคนเดียว ส่วนระบบที่ไม่ขึ้นกับผู้พูดนี้จะอาศัยพื้นฐานการหาคุณสมบัติร่วมของข้อมูลเสียงในกลุ่มข้อมูลเสียง (จากหลายๆ คน) เป็นข้อมูลอ้างอิง แม้ว่าระบบที่ไม่ขึ้นต่อผู้พูดจะสามารถรู้จำคำพูดบุคคลได้มากแต่ในทางปฏิบัติแล้วหลักการนี้ไม่สามารถใช้ได้กับทุกบุคคล ทั้งนี้เนื่องจากคุณสมบัติของเสียงพูดนั้นมีปัจจัยแวดล้อมที่เกี่ยวข้องอยู่มากมาย แม้จะพูดจาก

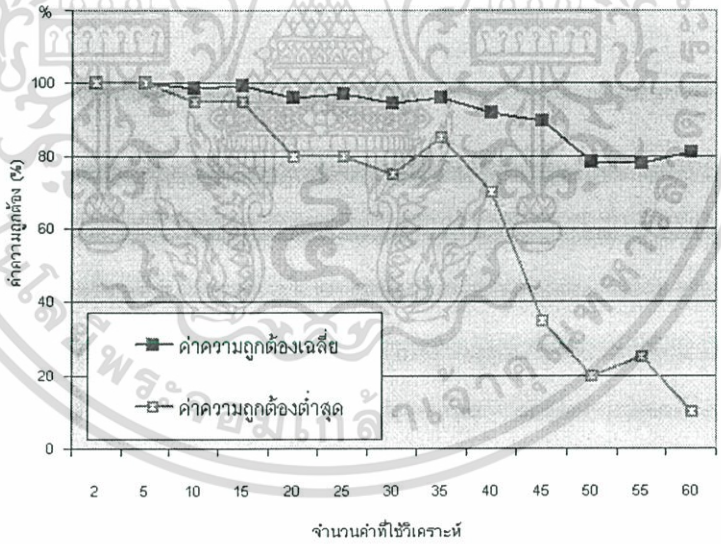
บุคคลคนเดียวกันและใช้คำๆเดียวกันก็ตาม ซึ่งปัจจัยดังกล่าวได้แก่ เพศ อายุ อารมณ์ ความดัง และเสียงรบกวน เป็นต้น

ในปัจจุบันมีการพัฒนาโปรแกรมที่เรียกว่า Speech Recognition ออกมามากพอสมควร เช่น บริษัท IBM กับผลิตภัณฑ์ที่ชื่อ ViaVoice แม้ว่าโปรแกรมที่มีออกวางจำหน่ายนี้สามารถตอบสนองการทำงานได้ถูกต้องกว่า 90% (ตามผู้ผลิตกล่าวอ้าง) แต่เนื่องจากต้นแบบของผู้ผลิตเป็นชาวต่างประเทศจึงไม่สามารถนำมาประยุกต์ใช้กับคนไทยได้ ด้วยเหตุเพราะลักษณะของสำเนียงและข้อกำหนดในเรื่องของสระและพยัญชนะที่แตกต่างกัน อย่างไรก็ตามได้มีงานวิจัยในประเทศไทยที่เกี่ยวข้องกันนี้ออกมาอย่างต่อเนื่องดังต่อไปนี้

ตารางที่ 1.1 แสดงงานวิจัยเกี่ยวกับระบบรู้จำเสียงพูดภาษาไทย

กรรมวิธี	ลักษณะสำคัญ	ประเภทคำ	จ.น.คำ	ถูกต้อง
การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นกับผู้พูดโดยใช้ไดนามิกไทม์วาร์ปปิง (ระพีพัฒน์ เพ็ญศิริ, 2538)	Normalized HDT Parameters	ตัวเลขไทย	10	79.25%
การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดนมาร์คอฟโมเดลและเวกเตอร์ควอนไตซ์เซชั่น(เสาวลักษณ์ อารีพงศา, 2538)	10 LP Coefficients 64 Vector codebook	ตัวเลขไทย	10	82.00%
ระบบรู้จำคำไทยหลายพยางค์แบบไม่ขึ้นกับผู้พูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟ(วิศรุต อาขุบุตร, 2539)	10 LP Coeffecients 128 and 256 vectors	คำ 1, 2 และ 3 พยางค์	70	89.91%
การรู้จำตัวเลขภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยใช้แอลพีซีและโครงข่ายประสาทเทียมแบบแบ็กพรออาเกชั่น (วุฒิพงษ์ พรสุขจันทรา, 2539)	10 LP Coefficients	ตัวเลขไทย	10	89.40%
การรู้จำคำพูดภาษาไทยหลายพยางค์แบบไม่ขึ้นต่อผู้พูดโดยใช้เทคนิคแบบฟัซซีและนิวรอลเน็ตเวอร์ค (ชัย วุฒิวิวัฒน์ชัย, 2540)	10 LP Coefficients	คำ 1, 2 และ 3 พยางค์	70	91.00%
Tone Recognition for Thai (Tungthang-thum, 1998)	ความถี่มูลฐาน	สระ	10	91.00%
การรู้จำเสียงพูดตัวเลขไทยแบบหลายผู้พูดด้วยนิวรอลเน็ตเวอร์ค (ไชยันต์ สุวรรณชีวะศิริ, 2541)	กลุ่มความถี่ฟอร์แมนซ์ที่ 1 และ 2	ตัวเลขไทย	10	91.60%
การรู้จำหน่วยเสียงสระภาษาไทยโดยใช้โครงข่ายประสาทเทียม (เอกฤทธิ์ มณีน้อย, 2541)	สัมประสิทธิ์ Ceptral	สระ	9	90.34%

สาเหตุที่ทำให้ระบบการรู้จำเสียงพูดภาษาไทยยังไม่สามารถใช้งานได้อย่างกว้างขวางนั้น เนื่องจากรูปแบบทางภาษาที่มีความสลับซับซ้อน โดยเฉพาะอย่างยิ่งเมื่อนำไปใช้กับการพิมพ์เอกสารเนื่องจากภาษาไทยมีรูปแบบของสระ, วรรณยุกต์, พยัญชนะ, ตัวเลข และสัญลักษณ์ต่างๆ มากมายซึ่งมีจำนวนรวมแล้วมากกว่า 100 ตัว เมื่อยังไม่รวมกับภาษาอังกฤษ ที่มีอีกกว่า 40 ตัว ดังนั้นหากเพิ่มจำนวนการรู้จำให้แก่ระบบการรู้จำจะเป็นการเพิ่มภาระแก่การเรียนรู้ เพราะจำนวนคำที่ต้องการให้รู้จำมีจำนวนมากขึ้น เวลาในการเรียนรู้จะสูงมากขึ้นเป็นเท่าตัว ในขณะที่ความถูกต้องของการรู้จำจะมีประสิทธิภาพที่ลดต่ำลงด้วย ซึ่งจากการทดลองระบบการรู้จำโดยใช้การวิเคราะห์ทางสเปกตรัมกับโครงข่ายประสาทเทียมแบบส่งค่าย้อนกลับ (Backpropagation) กับข้อมูลการรู้จำที่มีจำนวนตั้งแต่ 2, 5, 10, 15,...,60 พบว่าค่าเฉลี่ยความถูกต้องของการรู้จำจะอยู่ในช่วงประมาณ 80-100% และมีแนวโน้มที่ความถูกต้องจะมีค่าที่ลดลงเรื่อยๆเมื่อจำนวนคำในการรู้จำมากขึ้น แต่หากสังเกตค่าความถูกต้องของการรู้จำเป็นรายตัวจะพบว่าค่าความถูกต้องของคำบางคำจะต่ำกว่าระดับค่าเฉลี่ยรวมมาก เช่น การรู้จำคำจำนวน 60 คำ บางคำมีความถูกต้องเพียง 10% เท่านั้น แต่เนื่องจากคำอื่นๆ มีค่าความถูกต้องที่สูง ดังนั้นค่าเฉลี่ยของการรู้จำจึงมีค่าสูงด้วย ดังแสดงในรูปที่ 1.2



รูปที่ 1.2 ความถูกต้องของการรู้จำเมื่อจำนวนคำเพิ่มขึ้น

ดังนั้นหากสามารถแบ่งคำออกเป็นกลุ่มๆ โดยจำกัดจำนวนคำในแต่ละกลุ่มให้มีขนาดน้อยๆ ก็จะส่งผลให้การรู้จำมีประสิทธิภาพที่สูงได้เนื่องจากตัวอักษรภาษาไทยมีทั้งคำที่มีขนาด 1, 2, 3 และ 4 พยางค์ ซึ่งหากจัดกลุ่มแล้วก็จะได้กลุ่มโดยเฉลี่ยอยู่ที่ประมาณ 30-40 คำ

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาและพัฒนาระบบการรู้จำเสียงพูดภาษาไทยแบบขึ้นกับผู้พูดโดยใช้โครงข่ายประสาทเทียม

1.2.2 เพื่อศึกษาและพัฒนาระบบการรู้จำเสียงพูดคำไทย 1, 2 และมากกว่า 2 พยางค์

1.2.3 เพื่อศึกษาและพัฒนาโปรแกรมใช้เสียงพูดเพื่องานพิมพ์เอกสาร

1.3 ขอบเขตงานวิจัย

1.3.1 ศึกษาระบบการรู้จำเสียงพูดอักษรภาษาไทยแบบขึ้นต่อบุคคล

1.3.2 พัฒนาโปรแกรมใช้เสียงพูดเพื่องานพิมพ์เอกสารบน Microsoft Word 2000

1.3.3 โปรแกรมจะสามารถรู้จำเสียงที่ได้รับการฝึกไว้เท่านั้น โดยการพูดต้องพูดแบบเป็นพยางค์ไม่ต่อเนื่อง

1.3.4 ระดับสัญญาณรบกวนต้องไม่เกิน 15 dB

1.3.5 การพูดที่ใช้ในการฝึกและใช้งานโปรแกรมจะมีช่วงว่างระหว่างคำไม่น้อยกว่า 1 วินาที

1.3.6 การพูดที่ใช้ในการฝึกและใช้งานโปรแกรมจะต้องมีระดับเสียงผ่านเข้าสู่ Sound Card ไม่น้อยกว่า 55 dB

1.4 แผนการดำเนินงาน

1.4.1 ขั้นตอนการดำเนินการ

1.4.1.1 ศึกษาระบบการรู้จำเสียงพูดในลักษณะต่างๆ ที่ได้มีการศึกษามาแล้ว

1.4.1.2 กำหนดขอบเขตและกลุ่มคำที่ใช้ในงานวิจัย

1.4.1.3 ออกแบบและพัฒนาโปรแกรมเพื่อใช้ในระบบรู้จำเสียง

1.4.1.4 บันทึกข้อมูลเสียงพูดโดยใช้เสียงกลุ่มผู้ทดลอง

1.4.1.5 นำข้อมูลเสียงที่บันทึกไว้ของผู้ทดลองแต่ละคนป้อนสู่กระบวนการฝึกฝนของโครงข่ายประสาทเทียมและทดสอบการรู้จำ

1.4.1.6 สรุปผลการวิจัย ประเมินผล และเสนอแนะแนวทางในการวิจัยต่อไป

1.4.1.7 เสนอรายงานการวิจัยในรูปแบบวิทยานิพนธ์

1.4.2 ระยะเวลาที่ใช้ในแต่ละขั้นตอน

ตารางที่ 1.2 แสดงระยะเวลาที่ใช้ในงานวิจัย

ขั้นตอน	ระยะเวลาดำเนินการ (เดือน)											
	1	2	3	4	5	6	7	8	9	10	11	12
1. ศึกษาระเบียบวิธีการดำเนินงาน	↔											
2. กำหนดคุณสมบัติของผู้ใช้งาน		↔										
3. ออกแบบระบบสารสนเทศ			↔	↔	↔	↔			↔			
4. พัฒนาระบบสารสนเทศ							↔					
5. ทดสอบระบบสารสนเทศ									↔			
6. ติดตั้งระบบสารสนเทศ										↔	↔	
7. จัดทำคู่มือการใช้งานระบบสารสนเทศ												↔

บทที่ 2
การวิเคราะห์สัญญาณเสียงพูด

การวิเคราะห์สัญญาณเสียงพูดเปรียบเสมือนขั้นตอนในการเตรียมข้อมูลเพื่อใช้ในการแปลความหมาย หรือค้นหารายละเอียดความสำคัญที่มีอยู่ภายในรูปแบบของคลื่นสัญญาณทางไฟฟ้า ซึ่งหากเราสามารถรู้จักและเข้าใจในรูปแบบสัญลักษณ์ที่แสดงออกมานี้ได้จะมีประโยชน์อย่างมากต่อการนำมาประยุกต์ใช้งานเพื่อการควบคุมและสั่งการอุปกรณ์เครื่องจักรกล แต่รูปแบบของสัญญาณเสียงพูดนี้เป็นรูปแบบทางสัญญาณไฟฟ้า ทำให้การที่จะศึกษาและทำความเข้าใจนั้นกระทำได้ยาก จึงจำต้องเปลี่ยนรูปแบบสัญญาณนั้นให้อยู่ในรูปของสัญญาณทางดิจิตอล ซึ่งง่ายต่อการคำนวณด้วยคอมพิวเตอร์ เนื่องจากการประมวลผลในรูปแบบของดิจิตอล การสื่อสารและการแสดงผลสามารถกระทำได้ง่ายและมีประสิทธิภาพมากกว่า

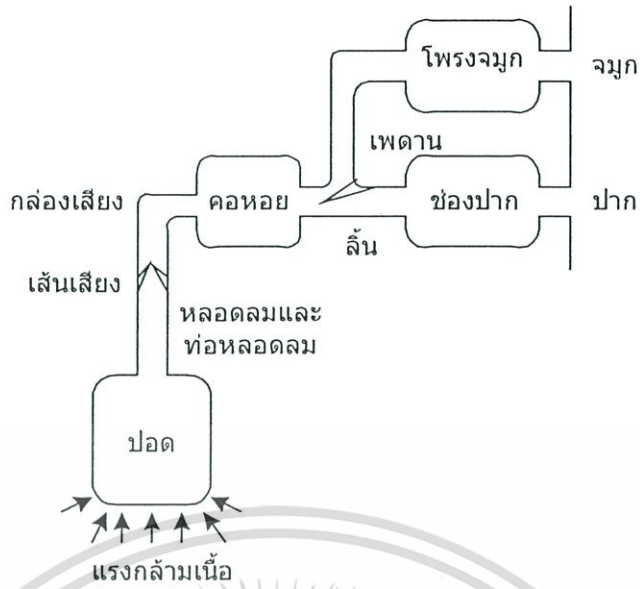
2.1 สัญญาณเสียงพูด (Speech Signal)

การประมวลสัญญาณเสียงพูดมีอยู่มากมายหลายวิธีแต่ก็ล้วนอยู่บนพื้นฐานหลักการของการเปล่งเสียงพูดของมนุษย์ ซึ่งรูปแบบโครงสร้างนั้นตั้งอยู่บนสมมติฐานว่าเสียงพูดสามารถที่จะแสดงให้อยู่ในรูปของเอาต์พุตแบบเชิงเส้นและการแปรผันตามเวลา โดยอวัยวะที่ช่วยในการออกเสียงของมนุษย์สามารถแสดงได้ดังรูปที่ 2.1. และเสียงพูด (Speech Sound) สามารถจัดแบ่งได้ 3 กลุ่มดังนี้คือ [2]

1. เสียงโฆษะ (Voiced Sound) [3] กำเนิดโดยแรงลมผ่านช่องว่างระหว่างเส้นเสียง เมื่อความตึงของเส้นเสียงมีการสั่นสะเทือนจึงเกิดเป็นจังหวะ (pulse) ซึ่งก่อเกิดเป็นเสียงก้องขึ้นมาโดยที่ความแตกต่างของสัญญาณเสียงนั้นจะแตกต่างกันเนื่องจากการทำงานของกล้ามเนื้อที่เปลี่ยนลักษณะรูปร่างของระดับที่เส้นเสียง และความถี่ของเสียงที่เกิดการสะท้อน ซึ่งอัตราการกำเนิดคลื่นการสั่นสะเทือนนี้จะเรียกว่า ความถี่มูลฐาน (Fundamental Frequency) โดยชายจะมีย่านความถี่ของเสียงอยู่ที่ 50 – 250 Hz ส่วนผู้หญิงจะมี 120 – 500 Hz ตัวอย่าง Voiced sound เช่น เสียง *i* ในคำว่า *Lie* , *a* ใน *Baby* และ *ee* ในคำว่า *Beet*

2. เสียงอโฆษะ (Unvoiced Sound) ถูกกำหนดโดยแรงลมที่ผ่านการหดตัวที่ความเร็วสูงที่เพียงพอจนทำให้เกิดการสับสนหรือแปรปรวนทิศทางอยู่ภายในท่อหลอดลม ตัวอย่างเช่นเสียง *f* และ *sh* ในคำว่า *fish* หรือ *s* ในคำว่า *silly*

3. Plosive Sound ถูกสร้างโดยการที่ปิดช่องทางการไหลของแรงดันที่สร้างขึ้นภายหลังที่ได้ปิดกั้นลมไว้แล้วปล่อยโดยฉับพลัน เช่นเสียง *p* ในคำว่า *punch* หรือ *b* ในคำว่า *butcher*

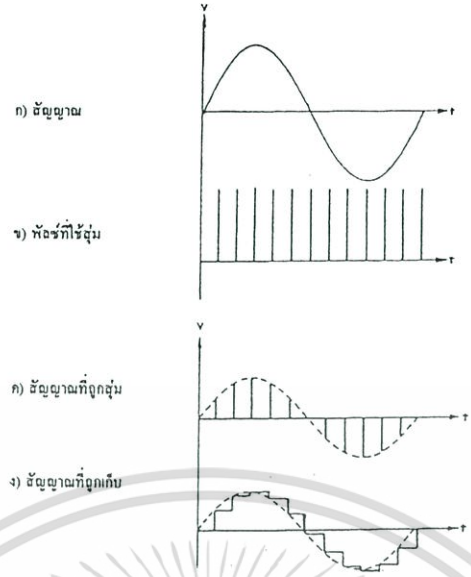


รูปที่ 2.1 แสดงอวัยวะส่วนต่างๆที่ใช้ในช่องทางเดินเสียง

2.2 การแปลงรูปแบบสัญญาณเสียงพูด

รูปแบบสัญญาณไฟฟ้าที่เราพบเห็นและคุ้นเคยในชีวิตประจำวันจะอยู่ในรูปแบบของสัญญาณที่ต่อเนื่อง หรือที่เรียกว่าสัญญาณอนาล็อก ซึ่งแต่เดิมการจะนำเอาสัญญาณไฟฟ้าดังกล่าวมาประมวล (Processed) ในแบบอนาล็อก แต่เมื่อมีการคิดค้นเทคนิคการประมวลผลด้วยดิจิตอล จึงได้มีการเปลี่ยนการประมวลผลมาเป็นแบบดิจิตอล เนื่องจากพบว่าการสื่อสารและแสดงผลในรูปแบบทางดิจิตอลสามารถกระทำได้ง่ายและมีประสิทธิภาพมาก โดยวงจรแปลงสัญญาณอนาล็อกเป็นสัญญาณดิจิตอล (Analog to Digital converters) หรือ ADC และนำมาประมวลด้วยตัวประมวลทางดิจิตอล (Digital processors) เช่น คอมพิวเตอร์ หรือ วงจรดิจิตอล (Digital Circuit)

หลักการพื้นฐานสำหรับแปลงสัญญาณอนาล็อกเป็นสัญญาณดิจิตอล จะใช้ระบบการสุ่มสัญญาณ โดยสัญญาณอนาล็อกจะถูกสุ่มเป็นระยะด้วยค่าคงที่ ดังแสดงในรูปที่ 2.2 กลุ่มของสัญญาณสุ่มจะแทนแบนด์วิธที่ทำงานด้วยความเร็วสูง ซึ่งจะทำการตัดต่อสัญญาณอนาล็อกในช่วงเวลาอันสั้น



รูปที่ 2.2 ภาพแสดงกระบวนการสุ่มสัญญาณ

ผลของการสุ่มสัญญาณด้วยความเร็วจะเสมือนกับการคูณขบวนสัญญาณพัลส์แคบๆ กับสัญญาณอนาล็อก ซึ่งจะได้เป็นสัญญาณที่เกิดการมอดูเลทระหว่างขบวนพัลส์กับสัญญาณอนาล็อกดังแสดงในรูป 2.2 (ค) โดยสัญญาณอนาล็อกจะขี่มาบนขบวนพัลซ์ถ้าหากเอาสวิทซ์และตัวเก็บประจุแทนสวิทซ์แล้วสัญญาณอนาล็อกที่ถูกสุ่มจะถูกเก็บไว้ในตัวเก็บประจุจนกว่าสัญญาณค่าใหม่ถูกสุ่มเข้ามา ซึ่งลักษณะของเอาต์พุตที่แสดงในรูป 2.2 (ง) สำหรับปัญหาที่ว่าอัตราการสุ่มสัญญาณนั้นควรมีขนาดเท่าใดนั้นจึงจะไม่ทำให้ข้อมูลสูญเสียไปเมื่อสัญญาณนั้นถูกเปลี่ยนกลับมาเป็นเช่นเดิมคำตอบก็คือขึ้นอยู่กับความถี่ของสัญญาณอนาล็อก โดยทฤษฎีของการสุ่มกล่าวไว้ว่า "ถ้าสัญญาณต่อเนื่องซึ่งมีความถี่และฮาร์โมนิคไม่เกิน f_c แล้ว สัญญาณดังกล่าวจะสามารถเปลี่ยนกลับมาเป็นอย่างเดิมโดยไม่สูญเสียรายละเอียดหรือผิดเพี้ยนไป ถ้าอัตราการสุ่มไม่น้อยกว่า $2f_c$ ต่อวินาที"

2.3 ฟังก์ชันกรองความถี่ดิจิตอล (Digital Filter Function)

วงจรกรองความถี่ที่มีจำนวนเทอมของสัมประสิทธิ์ที่มากพอ จะให้คุณสมบัติที่ใกล้เคียงกับทฤษฎีมากขึ้น แต่การสร้างวงจรดังกล่าวในทางปฏิบัติเป็นสิ่งที่ทำได้ยากและไม่เหมาะสม เพราะว่าการคำนวณจะต้องใช้เวลานานและเป็นการสิ้นเปลืองอุปกรณ์อีกด้วย วิธีที่ง่ายในการออกแบบวงจรกรองความถี่ก็คือการจำกัดจำนวนอิมพัลส์ การกระทำดังกล่าวย่อมส่งผลให้วงจรที่ได้มีคุณสมบัติที่ไม่สมบูรณ์ภายใต้ข้อแลกเปลี่ยนคือวงจรง่ายขึ้นและมีความสิ้นเปลืองน้อยกว่าการปรับค่า h(n) ที่ได้ โดยการปรับน้ำหนักใหม่ตามลักษณะการคูณเฉพาะช่วง (Window Function, w(n))

สามารถที่จะลดการกระเพื่อม และปรับปรุงช่วงการตอบสนองเปลี่ยนแปลง (Transition) ได้ ดังนั้น ค่าสัมประสิทธิ์ (ค่าการตอบสนองอิมพัลส์) ที่ใช้คือ [4]

$$\hat{h}(n) = h_d(n) \cdot w(n) \tag{2.1}$$

เมื่อ $h_d(n)$ คือ ค่าสัมประสิทธิ์ที่ได้จากการออกแบบ การคูณจุดต่อจุดในโดเมนเวลาก็คือการคูณประสานในโดเมนความถี่ ดังนั้น

$$\hat{H}\left(e^{j\theta}\right) = H\left(e^{j\theta}\right) * W\left(e^{j\theta}\right) \tag{2.2}$$

ซึ่งฟังก์ชั่นหน้าต่างที่นิยมใช้กันมากมีดังนี้

- ฟังก์ชันสี่เหลี่ยม (Rectangular Function)

$$w(n) = \begin{cases} 1; & 0 \le n \le N-1 \\ 0; & elsewhere \end{cases} \tag{2.3}$$

- ฟังก์ชันสามเหลี่ยม (Triangle Function)

$$w(n) = \begin{cases} 2n/(N-1); & 0 \le n \le N-1 \\ 2 - 2n/(N-1); & (N-1)/2 \le n \le N-1 \\ 0; & elsewhere \end{cases} \tag{2.4}$$

- ฟังก์ชันฮานนิ่ง (Hanning Function)

$$w(n) = \begin{cases} \left\{1 - \cos\dfrac{2\pi n}{N-1}\right\}/2; & 0 \le n \le N-1 \\ 0; & elsewhere \end{cases} \tag{2.5}$$

- ฟังก์ชันแฮมมิ่ง (Hamming Function)

$$w(n)=\begin{cases}0.54-0.5\cos\dfrac{2\pi n}{N-1}; 0\le n\le N-1\\ 0; \qquad\qquad elsewhere\end{cases} \tag{2.6}$$

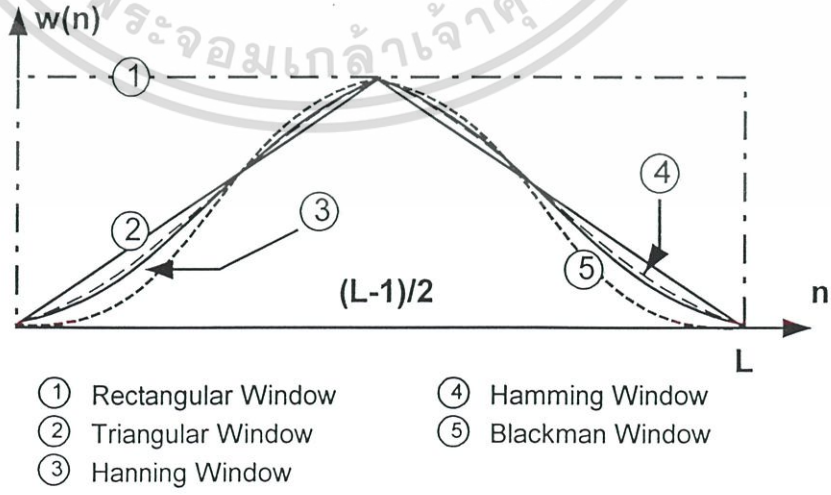
- ฟังก์ชันแบลคแมน (Blackman Function)

$$w(n)=\begin{cases}0.42-0.5\cos\dfrac{2\pi n}{N-1}+0.8\cos\dfrac{4\pi n}{N-1}; 0\le n\le N-1\\ 0; \qquad\qquad\qquad\qquad elsewhere\end{cases} \tag{2.7}$$

- ฟังก์ชันไคเซอร์ (Kaiser Function)

$$w(n)=\begin{cases}\dfrac{I_0\theta_a\sqrt{\left\{\left(\dfrac{N-1}{2}\right)^2-\left(n-\dfrac{N-1}{2}\right)^2\right\}}}{I_0\theta_a\left(\dfrac{N-1}{2}\right)}; 0\le n\le N-1\\ 0; \qquad\qquad\qquad\qquad elsewhere\end{cases} \tag{2.8}$$

โดยที่ $I_0=\int_0^{2\pi}\dfrac{e^{x\cos\theta}}{2\pi}d\theta$



รูปที่ 2.3 คุณสมบัติของหน้าต่างชนิดต่างๆ

2.4 การจัดเตรียมข้อมูลก่อนการประมวลผล (Preprocessing)

2.4.1 กรรมวิธีเน้นล่วงหน้า (Pre-emphasis)

เป็นขั้นตอนแรกที่กระทำกับสัญญาณเสียง x'(n) โดยผ่านวงจรกรองเพื่อปรับสเปกตรัมให้มีความราบเรียบ ดังสมการ

$$H(z) = 1 - a \cdot z^{-1} \quad 0 \le a \le 1 \tag{2.9}$$

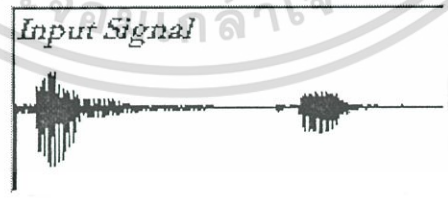
ซึ่งสัมพันธ์กับสัญญาณอินพุตดังสมการ

$$x'(n) = x(n) - ax(n-1) \tag{2.10}$$

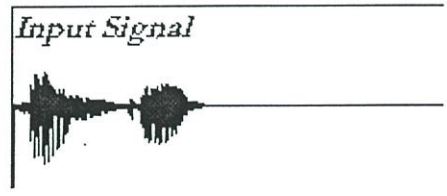
ค่าสัมประสิทธิ์ของวงจรกรอง a จะถูกกำหนดให้เท่ากับ 0.95 [1]

2.4.2 กรรมวิธีหาจุดสิ้นสุดเสียงพูด (Endpoint Detection)

พื้นฐานของสัญญาณในการรับรู้เสียงพูดกลุ่มคำโดดคือ สัญญาณจะประกอบด้วยส่วนของเสียงพูดที่อาจมีสัญญาณรบกวน (Noise Signal) นำหน้าหรือต่อท้ายด้วย ดังนั้นจึงต้องคัดเลือกเฉพาะส่วนที่เป็นข้อมูลจริงๆออกมาและตัดส่วนที่ไม่ใช่เสียงพูดดังกล่าวทิ้งไป ซึ่งเรียกว่าการหาขอบเขตของสัญญาณ (Endpoint Detection) การหาขอบเขตของสัญญาณในการรับรู้เสียงพูดกลุ่มคำโดดมีความจำเป็นอย่างมาก เนื่องจากความถูกต้องของการรับรู้เสียงพูดจะขึ้นตรงกับความถูกต้องของการหาขอบเขตของสัญญาณ และจะทำให้เวลาที่ใช้ในการคำนวณน้อยลง นอกจากนั้นยังทำให้ทราบได้ว่ากลุ่มคำนั้นประกอบด้วยพยางค์มากน้อยเพียงใด [5]



(ก) สัญญาณก่อนผ่านการ Endpoint Detection



(ข) สัญญาณหลังผ่านการ Endpoint Detection

รูปที่ 2.4 ตัวอย่างสัญญาณเสียงพูดก่อนและหลังผ่านการหาจุดสิ้นสุด

จากรูปที่ 2.4(ก) เป็นการรับสัญญาณเสียงพูด "บันทึก" ซึ่งพบว่าจะมีช่วงห่างระหว่างพยางค์ที่เปล่งออกมา แต่เมื่อผ่านขั้นตอนของการตัดแบ่งหัวท้ายพยางค์โดยการหาจุดเริ่มและสิ้นสุดของแต่ละพยางค์ โดยการตั้งระดับค่าพลังงานหรือจุดเปลี่ยนพลังงาน (Energy Level Threshold) เพื่อใช้เป็นเกณฑ์ในการตัดสินช่วงที่เป็นข้อมูลสัญญาณเสียงและปรับระดับเสียงหรือนอร์มอลไลซ์(Normalize) โดยวิธีที่เลือกใช้ในการนอร์มอลไลซ์นี้ก็คือ Max-Min Scale สามารถแสดงได้ดังสมการ

$$e_{\min} = Min[e(m)] \qquad ;m = 0 \ \ to \ \ n \tag{2.11}$$

$$e_{\max} = Max[e(m)] \qquad ;m = 0 \ \ to \ \ n \tag{2.12}$$

เมื่อ e_{min} คือ ค่าพลังงานต่ำสุด

e_{max} คือ ค่าพลังงานสูงสุด

e(m) คือ ค่าพลังงานลำดับที่ 0 ถึง n

โดย e(m) หาได้จากสมการ

$$e(m) = \sum_{n=0}^{N-1} s^2(n) \tag{2.13}$$

จากนั้นจึงทำการปรับค่าสูงสุดและต่ำสุดของพลังงานดังสมการ

$$e(m) = \left[e(m) - e_{\min}\right] \times \frac{100}{e_{\max} - e_{\min}} \qquad ;m = 0 \ \ to \ \ n \tag{2.14}$$

การหารูปคลื่นพลังงาน (Energy Pulse Detection) โดยทั่วไปสามารถทำได้ด้วยการกำหนดระดับค่าพลังงานอ้างอิง (Energy Thresholds) ไว้ 4 ค่าคือ k1, k2, k3 และ k4 ซึ่งในการหารูปคลื่นหรือจำนวนพยางค์นี้สามารถกระทำได้โดยการเปรียบเทียบระดับพลังงานโดยผ่านขั้นตอนการปรับปรุงระดับพลังงาน (Adaptive Level Equalizer) เพื่อเทียบกับค่า Log ของระดับสัญญาณเบื้องหลัง (Background Signal) ซึ่งค่า Equalizer Energy นี้สามารถหาได้ด้วยการคำนวณของสมการ

$$\hat{R}_l(0) = \log\left[R_l(0)\right] - Q \qquad ;l = 1,2,\ldots,L \tag{2.15}$$

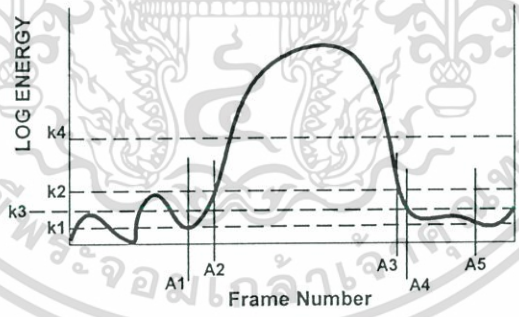
เมื่อ $\hat{R}_l$ แทนค่า Equalizer Energy ลำดับที่ *l*

R_l แทนค่าพลังงานของเฟรมลำดับที่ *l*

Q แทนค่าเฉลี่ยของระดับสัญญาณเบื้องหลัง ซึ่งสามารถหาได้จากค่า E_{min}

$$E_{min} = \min_{1 \le l \le L} \{\log[R_l(0)]\} \tag{2.16}$$

ค่า Q จะได้จากค่าเฉลี่ยของ 3 ค่าสูงสุดของกราฟที่แสดงค่าระดับ log energy ซึ่งต่ำกว่า 10 dB (ค่า log $[R_l(0)]$) เมื่อได้ค่า $\hat{R}_l$ แล้วจะทำการเทียบระดับของสัญญาณด้วยค่าระดับอ้างอิงทั้ง 4 ค่า ดังแสดงในรูปที่ 2.5 โดยหากค่า $\hat{R}_l$ มีสูงกว่า จุดอ้างอิงจุดแรก คือ k1 เฟรมที่ A1 จะถูกบันทึกไว้ และถ้ายังคงมีค่าเกินจุดอ้างอิง k2 ก่อนที่จะตกกลับมาที่จุด k1 จุดเริ่มต้นของคลื่นจะเป็นเฟรมที่ A1 แต่ถ้าระยะเวลาจากจุด A1 ถึง A2 มีค่านานเกินไป จุดเริ่มต้นของรูปคลื่นจะเปลี่ยนเป็น A2 ส่วนการหาจุดสิ้นสุดของรูปคลื่นจะใช้การพิจารณาในลักษณะเดียวกันคือใช้จุด k2 และ k3 แต่หากระยะเวลาระหว่าง k3 และ k4 มีค่าสูงเกินไปจุดสิ้นสุดก็จะเปลี่ยนจาก k4 เป็น k3 และถ้าหากระดับของ $\hat{R}_l$ มีค่าสูงไม่ถึงระดับ k4 ก็จะถือว่าจุดเริ่มต้นของรูปคลื่น (A1 หรือ A2) จะไม่ถูกนำมาคิด (Reject) กล่าวคือเริ่มต้นการวิเคราะห์จุดเริ่มต้นใหม่อีกครั้งหรือหากช่วงเวลาในระดับ k4 นี้สั้นเกินไป (น้อยกว่า 5 เฟรม หรือ 75 ms.)รูปคลื่นนี้ก็จะไม่ถูกนำมาคิดด้วยเช่นกัน [6]



รูปที่ 2.5 การใช้ Energy Threshold หาขอบเขตพยางค์

2.5 การแปลงสัญญาณเสียงให้อยู่ในรูปแบบสเปกตรัม

การวิจัยด้านการสื่อสารทางเสียงจะใช้หลักการของฟูริเยร์ในการแก้ไข เพราะฟูริเยร์จะช่วยในการสร้างรูปแบบสำหรับสัญญาณเสียงของระบบเชิงเส้นที่เป็นคาบเวลา หรือการสุ่มของสัญญาณที่แปรตามเวลา โดยทั่วไปสเปกตรัม(Spectrum) ของสัญญาณที่ออกมาจะอยู่ในรูปของผลตอบสนองทางด้านความถี่ ดังนั้นจึงสามารถคาดเดาได้ว่า สเปกตรัมของเอาต์พุตจะสะท้อนให้เห็นถึงคุณสมบัติจากองค์ประกอบด้านความถี่ของสัญญาณเสียง เนื่องจากระบบการรู้จำเสียง

จำเป็นต้องอาศัยการคำนวณอย่างสูง การแปลงฟูริเยร์อย่างเร็วหรือ Fast Fourier Transform (FFT) จึงถูกเลือกใช้ในงานวิจัยในครั้งนี้ เนื่องจากช่วยลดการคำนวณลงจากการคำนวณแบบ Discrete Fourier Transform (DFT) ถึง $N/\log_2 N$ เท่า (DFT ต้องคำนวณเลขเชิงซ้อน N*N ครั้ง ส่วน FFT ใช้การคูณเชิงซ้อนเพียง $N \log_2 N$ ครั้ง) [7]

2.5.1 การแปลง FFT (Fast Fourier Transform)

คำว่า FFT เป็นชื่อกลางๆ ที่ไม่ได้บ่งบอกว่าเป็นวิธีการคำนวณแบบใด ซึ่งในที่นี้จะเสนอวิธีทำ FFT พื้นฐานวิธีหนึ่งคือ Radix-2 แบบแตกเป็นส่วนย่อยทางฝั่งเวลา (Decimation-in-time) หากพิจารณาสมการที่ใช้แปลง DFT (สมการที่ 2.17) จะได้ว่า ถ้าให้ N เป็นเลขคู่จะสามารถกระจาย X(k) ให้อยู่ในรูปของผลบวกของเทอมที่ n เป็นคู่ และเทอมที่ n เป็นคี่ได้สมการที่ 2.18

$$X(k)=\sum_{n=0}^{N-1}x(n)W_N^{kn} \quad ;W_N=e^{-j2\pi/N},\ k=0,1,\ldots N-1 \tag{2.17}$$

$$X(k)=\underbrace{\sum_{n=0}^{\frac{N}{2}-1}x(2n)*W_N^{2nk}}_{\text{เทอมคู่}}+\underbrace{X(k)=\sum_{n=0}^{\frac{N}{2}-1}x(2n+1)*W_N^{(2n+1)k}}_{\text{เทอมคี่}} \tag{2.18}$$

$$X(k)=\sum_{n=0}^{\frac{N}{2}-1}x(2n)*W_N^{2nk}+\sum_{n=0}^{\frac{N}{2}-1}x(2n)*W_N^{2nk}*W_N^{k} \tag{2.19}$$

ถ้าพิจารณาเทอม W_N^{ab} ที่มี a และ b เป็นจำนวนใดๆ ที่มี a และ b ไม่เท่ากับ 0 จะพบว่าสามารถย้ายเลขยกกำลังของ W ไปเป็นตัวหารของ N ได้ดังนี้

$$W_N^{ab}=e^{\frac{-2\pi}{N}ab}=e^{\frac{-2\pi}{N/b}a}=W_{N/b}^{a} \tag{2.20}$$

เมื่อแทนค่าเทอม W_N^{2nk} ด้วย $W_{N/2}^{nk}$ ในสมการที่ 2.19 จะได้

$$X(k)=\underbrace{\sum_{n=0}^{\frac{N}{2}-1}x(2n)*W_{N/2}^{nk}}_{\text{DFT }\frac{N}{2}\text{ จุด}}+\underbrace{\sum_{n=0}^{\frac{N}{2}-1}x(2n)*W_{N/2}^{nk}}_{\text{DFT }\frac{N}{2}\text{ จุด}}*W_N^{k} \tag{2.21}$$

สัมประสิทธิ์พิเศษ (เทอมคี่)

จะเห็นได้ว่า X(k) ได้กลายเป็นผลบวกของสองเทอม แต่ละเทอมเป็นรูปแบบของการคำนวณ DFT N/2 จุด โดยเทอมแรกกระทำกับสัญญาณ x(0), x(2), ..., x(N-2) และเทอมที่สองกระทำกับสัญญาณ x(1), x(3), ..., x(N-1)

ถ้ายุติการแตกกระจายเทอมย่อยแต่เพียงเท่านี้ และคำนวณ DFT โดยใช้สมการที่ 2.21 จะพบว่าต้องคำนวณ DFT N/2 จุด เป็นจำนวน 2 ชุด ซึ่งแต่ละชุดจะต้องใช้จำนวน CMAC (Complex Multiplication and Accumulation: หน่วยวัดการคำนวณ ซึ่ง 1 CMAC เท่ากับการกระทำทางคณิตศาสตร์ที่ประกอบด้วยการคูณเลขเชิงซ้อน 2 จำนวน เสร็จแล้วนำผลลัพธ์ที่ได้ไปบวกสมทบกับเลขเชิงซ้อนอีกจำนวนหนึ่ง) ประมาณ $2(N/2)^2 = N^2/2$

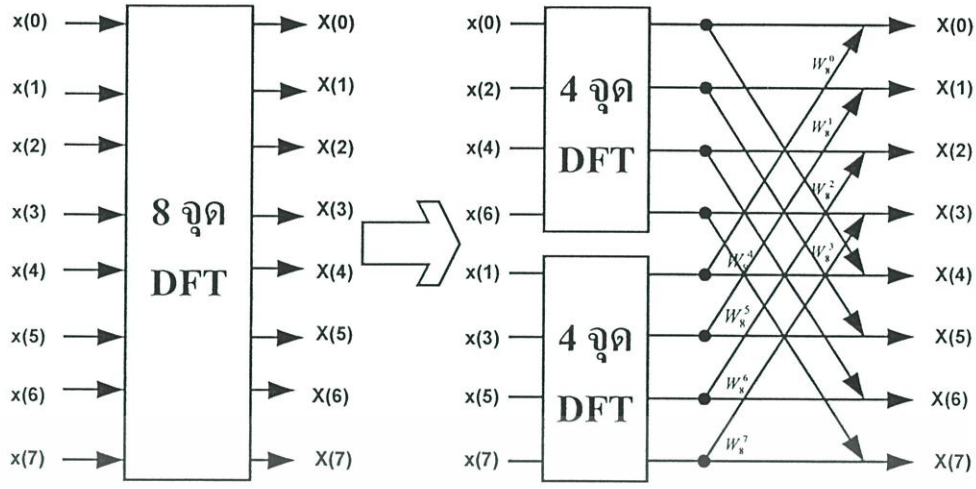
สรุปว่าการหา W(k) ซึ่งเป็น DFT N จุด สามารถกระจายให้อยู่ในเทอมของ DFT N/2 จุด ซึ่งทำให้จำนวน CMAC ที่ต้องใช้ลดลงประมาณครึ่งหนึ่ง เช่นเดียวกันหากทำการแตกเทอม DFT N/2 จุดที่อยู่ในสมการที่ 2.21 ต่อไปแต่ละเทอมจะสามารถกระจายกลายเป็นผลบวกของ DFT N/4 จุดสองเทอม ซึ่งจะทำให้จำนวน CMAC ที่ต้องใช้ลดลงไปอีกประมาณครึ่งหนึ่ง จึงสามารถกระจายการแตกของเทอมออกได้เรื่อยๆ จนกระทั่งทุกจุดอยู่ในรูปของ DFT 2 จุด ซึ่งหากสมมติให้ x(n) ยาว 2 จุดจะได้

$$X(k) = \sum_{n=0}^{1} x(n) * W_2^{kn} \qquad (2.22)$$

อาศัยหลักความจริงที่ $W_2^0 = 1$ และ $W_2^1 = e^{-j\pi} = -1$ จะได้

$$\left.\begin{aligned} X(0) &= x(0) + x(1) \\ X(1) &= x(0) - x(1) \end{aligned}\right\} \qquad (2.23)$$

ขั้นตอนที่กล่าวมาทั้งหมดนี้รวมเรียกว่า การแปลง FFT ซึ่งมักจะเขียนโดยใช้แผนภาพผีเสื้อ (butterfly diagram) ซึ่งการกระจาย DFT 8 จุดให้อยู่ในรูปของ DFT 4 จุด สองเทอมบวกกันสามารถแสดงได้ดังรูปที่ 2.6



รูปที่ 2.6 การกระจาย DFT 8 จุด เป็น DFT 4 จุด

รูปที่ 2.7 สัญลักษณ์ที่ใช้ในแผนภาพผีเสื้อ

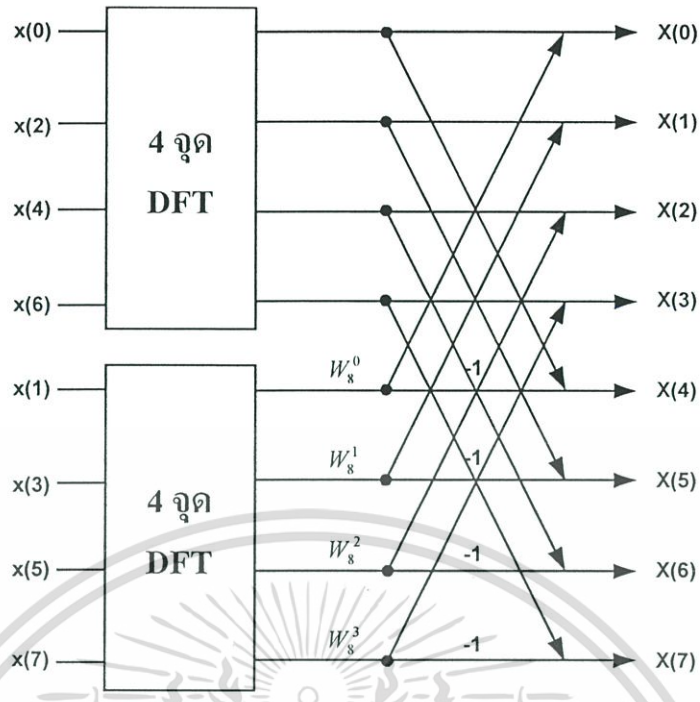
จากรูปที่ 2.12 สามารถจัดรูปแบบสัมประสิทธิ์ให้ง่ายลงได้โดยคุณสมบัติความสามารถของ W_N ดังนี้

$$W_N^{k+N/2} = W_N^k * W_N^{N/2} = W_N^k(-1) = -W_N^k \qquad (2.24)$$

ใช้คุณสมบัติตามสมการที่ 2.24 จะได้ว่า

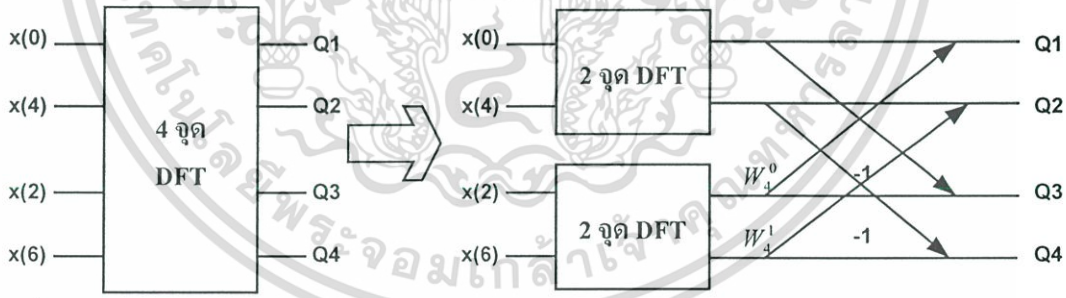
$$W_8^4 = -W_8^0,\ W_8^5 = -W_8^1,\ W_8^6 = -W_8^2,\ W_8^7 = -W_8^3 \qquad (2.25)$$

แทนค่าทั้งหมดลงในแผนภาพรูปที่ 2.12 จะได้แผนภาพดังรูปที่ 2.8

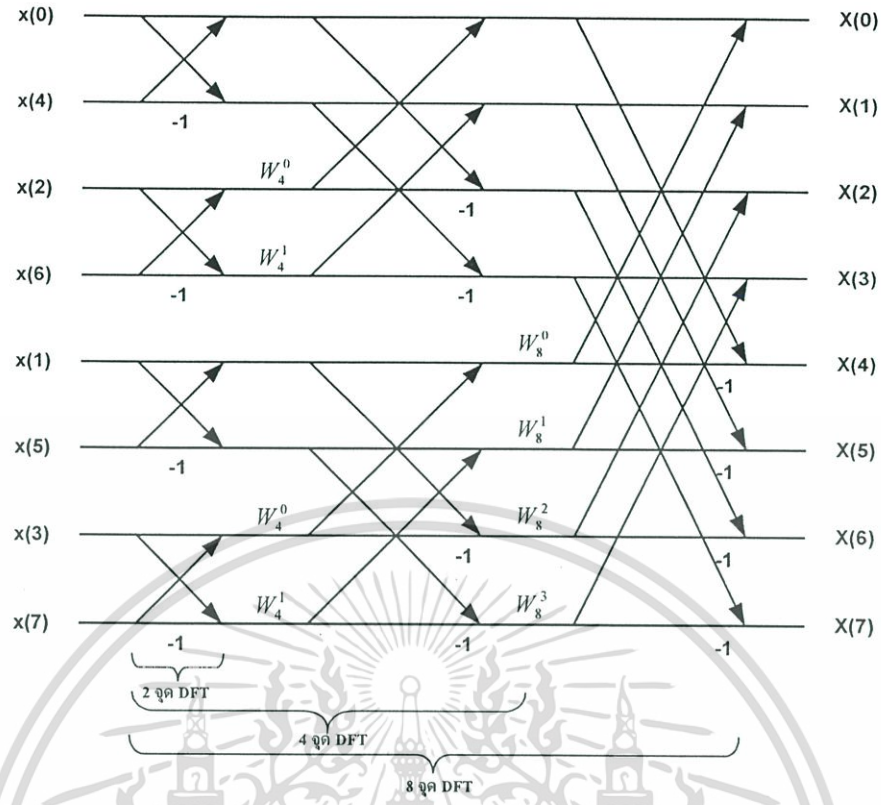


รูปที่ 2.8 การกระจาย DFT 8 จุด เป็น DFT 4 จุด หลังใช้คุณสมบัติสมมาตร

ในลักษณะเดียวกันกับ DFT 4 จุด ยังสามารถกระจายเป็น DFT 2 จุดได้ดังรูปที่ 2.8 และเมื่อรวมผลลัพธ์แต่ละส่วนเข้าเป็นแผนภาพเดียวกัน จะปรากฏดังในรูปที่ 2.9 ซึ่งสามารถใช้เป็นแนวทางในการเขียนแผนภาพและโปรแกรมสำหรับ FFT จำนวน N จุดใดๆ ได้ทันที



รูปที่ 2.9 การกระจาย DFT 4 จุด เป็น DFT 2 จุด



รูปที่ 2.10 แผนภาพรวมของการคำนวณ FFT 8 จุด

จากแผนภาพผีเสื้อของคำนวณ FFT มีสิ่งที่ควรสังเกตดังนี้คือ

1. หากต้องการได้ผลตอบในเชิงความถี่เรียงตามลำดับจาก X(0), X(1), …, X(7) ต้องทำการเรียงลำดับสัญญาณขาเข้าใหม่ดังนี้ x(0), x(4), x(2), x(6) ,x(1), x(5), x(3) และ x(7) เมื่อเขียนลำดับเหล่านี้ในเลขฐานสองจะได้ดังตารางต่อไปนี้

ตารางที่ 2.1 การเรียงลำดับสัญญาณขาเข้าใหม่ของหาคำนวณ FFT

ลำดับใหม่ฐานสิบ	ลำดับใหม่ฐานสอง	ลำดับปกติฐานสิบ	ลำดับปกติฐานสอง
0	000	0	000
4	100	1	001
2	010	2	010
6	110	3	011
1	001	4	100
5	101	5	101
3	011	6	110
7	111	7	111

จะเห็นได้ว่าลำดับใหม่เกิดจากการเรียงลำดับบิตจากหลังไปหน้าของลำดับปกติ (bit-reversed order) ซึ่งจะเป็นจริงสำหรับการคำนวณหา FFT ที่จำนวนจุดใดๆด้วย

2. ค่าคงที่ W ที่ใช้คูณกับส่วนคี่ สามารถเปลี่ยนให้อยู่ในรูปของ W_8 ได้ทั้งหมด โดยคูณตัวห้อยและตัวยกกำลังด้วยค่าเดียวกัน ดังนั้นจึงสามารถเปลี่ยนเทอมต่อไปนี้

$$W_4^0 \rightarrow W_8^0 \quad \text{และ} \quad W_4^1 \rightarrow W_8^2 \tag{2.26}$$

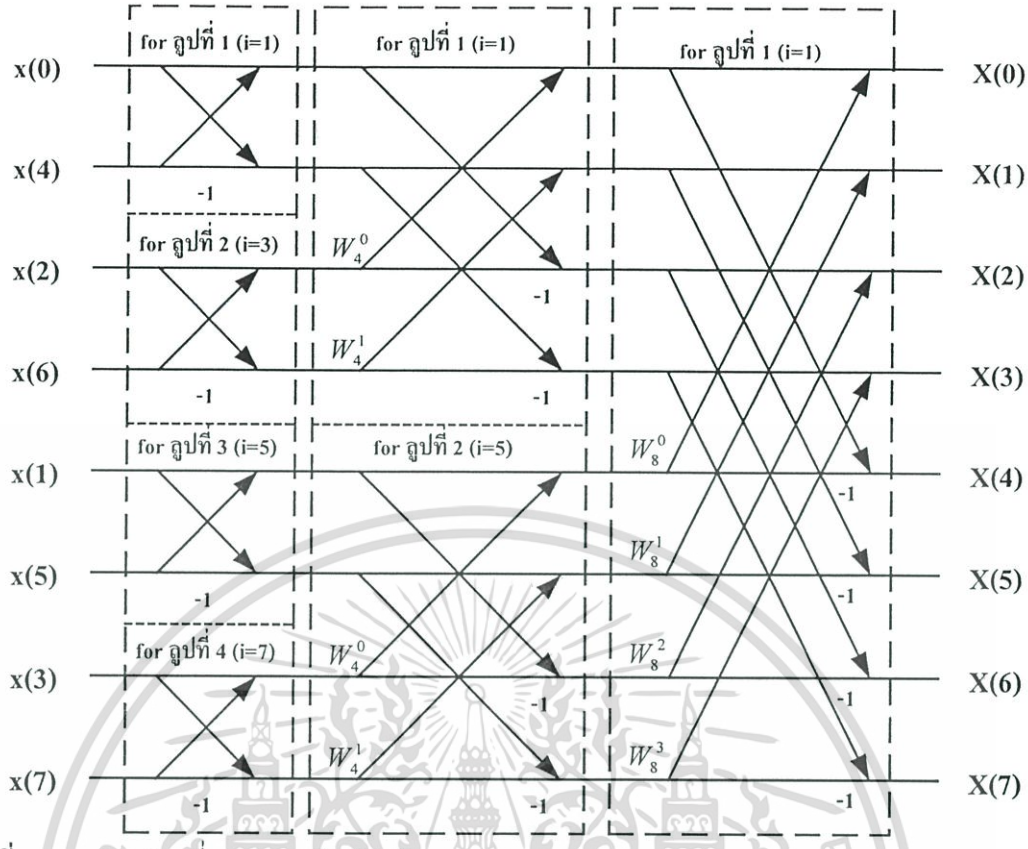
ดังนั้นสามารถใช้ W_8 แทนค่าได้ทั้งหมด ซึ่งสามารถคำนวณ W_8 ที่ k ต่างๆนี้ไว้ล่วงหน้าได้ และใช้เสมือนเป็นค่าคงที่สำหรับ FFT 8 จุด ซึ่งจะเป็นจริงสำหรับ FFT จำนวน N จุดใดๆ

3. พิจารณาโดยรวมแล้วจะได้ว่า การคำนวณ FFT จุด ถูกแบ่งเป็น Log_2 N ขั้นตอนโดยอาจประมาณได้ว่าแต่ละขั้นตอนมีการคำนวณเท่ากับ N CMAC's (มีเส้นทแยงในแผนภาพ N เส้นในทุกๆขั้นตอน) ดังนั้นจะได้ว่า

$$\text{จำนวน CMAC ที่ต้องใช้ในการคำนวณ FFT N จุด} = N \text{ Log}_2 N \tag{2.27}$$

4. วิธี radix-2 นี้ใช้ได้เฉพาะเมื่อค่า N เท่ากับ 2^b โดย b เป็นจำนวนเต็มบวกใดๆ ซึ่งหากจำนวนข้อมูลมีไม่ถึงสามารถแก้ได้โดยการใช้เทคนิคเติมศูนย์ เพื่อให้ได้ความยาวตามต้องการ

ดังนั้นจึงสามารถสร้างโปรแกรมเพื่อช่วยในการคำนวณ FFT แบบ radix-2 สำหรับความยาวใดๆ โดยในส่วนที่วนลูปเพื่อคำนวณ FFT ตามแผนภาพผีเสื้อนั้นจะประกอบด้วยการวนลูป 2 ลูปด้วยกันคือ ลูป For ซ้อนอยู่ในลูป While ตัวอย่างเช่น การแปลง FFT 8 จุดจะมีจำนวนของลูปที่วนดังแสดงในรูปที่ 2.11



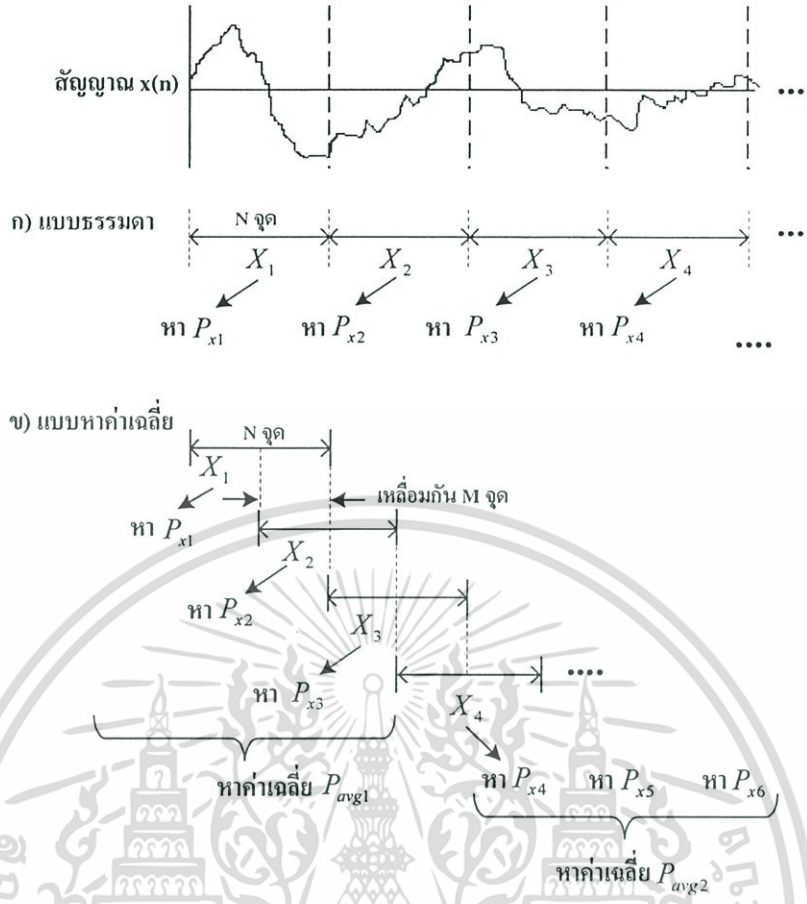
รูปที่ 2.11 ส่วนย่อยที่ลูปต่างๆ สำหรับคำนวณหา FFT 8 จุด

2.5.2 การวิเคราะห์สเปกตรัมและการปรับปรุงผลลัพธ์ที่ได้จาก FFT

ตัววิเคราะห์สเปกตรัม (Spectral analyzer) คือ การที่รับสัญญาณขาเข้าเป็นอะนาล็อก และแสดงสเปตรัม (ของกำลัง) ของสัญญาณออกทางหน้าจอ และสามารถติดตามการเปลี่ยนแปลงสเปตรัมของสัญญาณได้อย่างทันทีทันใด การวิเคราะห์สเปกตรัมอย่างง่ายสามารถสร้างได้โดยใช้ FFT ตามสมการที่ 2.28 เพื่อหาค่าสเปกตรัมของกำลัง

$$P_x(k) = \frac{1}{N}|X(k)|^2 \qquad (2.28)$$

โดยหลังจากที่สัญญาณถูกแปลงเป็นสัญญาณที่ไม่ต่อเนื่องแล้ว จะทำการจัดสัญญาณเป็นเฟรมๆ ละ N จุด แล้วทำการหาสเปกตรัมของสัญญาณทีละเฟรม

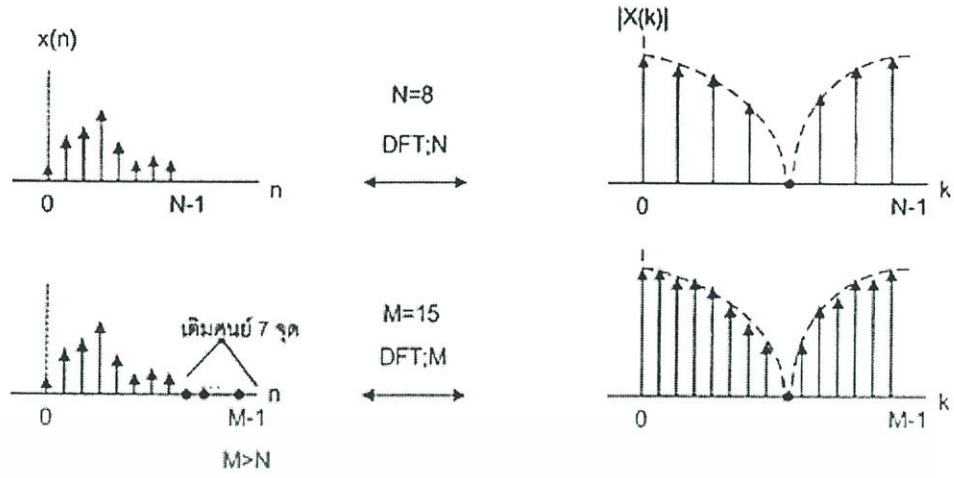


รูปที่ 2.12 การจัดเฟรมของสัญญาณเพื่อหาสเปกตรัม

ในงานวิจัยนี้สัญญาณเสียงในแต่ละเฟรม (หลังตัดหัวท้ายพยางค์แล้ว) จะถูกคูณด้วยฟังก์ชันหน้าต่างแบบแฮมมิ่ง โดยจะลดทอนแอมปลิจูดอย่างช้าๆ ที่บริเวณขอบด้านข้างของกรอบข้อมูลเพื่อหลีกเลี่ยงความไม่ต่อเนื่องที่จุดปลาย ซึ่งข้อมูลจะเหลื่อมซ้อนกันเท่ากับ 512 ข้อมูล

2.5.3 เทคนิคการเติมศูนย์ (Zero Padding)

"การเติมศูนย์" เป็นการเติมจุดที่มีค่าเป็นศูนย์ต่อท้ายเข้าไปในสัญญาณ x(n) ก่อนที่จะทำการแปลง DFT หรือ FFT ซึ่งจะส่งผลให้สเปกตรัมที่ได้มีจำนวนจุดมากขึ้น ซึ่งเสมือนเป็นการสุ่มสเปกตรัมด้วยจำนวนจุดที่มากขึ้น จึงช่วยให้มองเห็นรูปร่างของสเปกตรัมได้ละเอียดและชัดเจนขึ้น แต่ในทางทฤษฎีไม่ได้เป็นการเพิ่มข้อมูลใดๆ ให้แก่สัญญาณเลย [8] ดังแสดงในรูปที่ 2.13



รูปที่ 2.13 ผลการเติมศูนย์กับการแปลง DFT

2.6 การสกัดค่าลักษณะสำคัญ (Feature Extraction)

หลักการพื้นฐานของการสกัดค่าสำคัญคือ การวิเคราะห์หาค่าที่จะใช้แทนสัญญาณเสียงเพื่อนำไปในขั้นตอนการรู้จำ โดยสามารถแบ่งได้ 3 กลุ่มหลักคือ

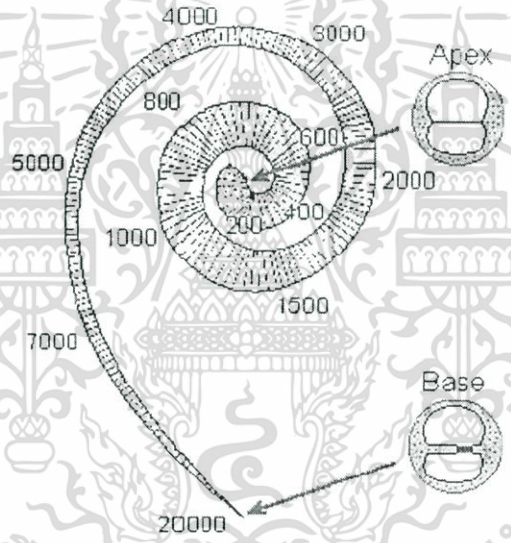
1. การหาค่าลักษณะสำคัญระดับสูง (High level feature) ได้แก่ สำเนียงการพูด, รูปแบบการพูด และความเร็วในการพูด เป็นต้น

2. การหาค่าลักษณะสำคัญทางฉันทลักษณ์ (Prosodic feature) เช่นค่าความถี่มูลฐาน (Fundamental frequency), ความถี่ฟอร์แมนท์ (Formant frequency) และระดับพลังงาน (Energy Profile) เป็นต้น ซึ่งวิธีการเหล่านี้จะมีประสิทธิภาพสูงในการรู้จำ แต่ยากต่อการสกัดจากสัญญาณ

3. การหาค่าลักษณะสำคัญแบบค่าแวดล้อมของสเปกตรัม (Spectral envelop feature) เนื่องจากค่าลักษณะสำคัญส่วนใหญ่สำหรับการรู้จำเสียงจะรวมอยู่ในข้อมูลเชิงสเปกตรัม แต่เนื่องจากข้อมูลทั้งหมดที่ได้จากการคำนวณ FFT นั้นมีขนาดเท่ากับ จำนวนเฟรมทั้งหมด 40 เฟรม คูณด้วยจำนวนค่าความถี่ที่ได้จากการหา FFT (1024 ค่า) ซึ่งถือว่าเป็นข้อมูลที่มีขนาดใหญ่มาก และจากการทดลองพบว่า เนื่องจากข้อมูลส่วนใหญ่ที่ใช้ในการทดสอบเป็นกลุ่มคำที่มีความยาวเสียงสั้นๆ ดังนั้นข้อมูลสัญญาณเสียงหลังจากผ่านกระบวนการตัวหัวท้ายพยางค์แล้วจะเหลือไม่เกิน 6000 จุด (ประมาณ 10 เฟรม) อย่างไรก็ตามหากนำไปใช้งานในระบบรู้จำเสียงจะต้องใช้เวลาในการฝึกสอนมาก ดังนั้นจึงต้องทำการลดปริมาณข้อมูลให้มีขนาดที่เหมาะสมแก่การใช้งาน โดยใช้เทคนิคที่เรียกว่า การหาค่าสัมประสิทธิ์เซปสตรัม (Cepstral Coefficient) [9]

2.7 การหาค่าสัมประสิทธิ์เซปสตรัม

พื้นฐานการหาค่าสัมประสิทธิ์เซปสตรัมคือการแปลงค่าโคซายน์แบบไม่ต่อเนื่อง (Discrete cosine transform) ของค่าลอการิธึม (Logarithm) ของสเปกตรัม (Spectral) ของสัญญาณเสียงส่วนย่อย สเปกตรัมของสัญญาณเสียงสามารถหาได้โดยการแปลงฟูริเยร์แบบไม่ต่อเนื่อง (Discrete Fourier Transformation) หรือการแปลงฟูริเยร์อย่างเร็ว (Fast Fourier Transformation) ขั้นตอนดังกล่าวตั้งอยู่บนพื้นฐานแนวคิดที่ว่า สเปกตรัมของสัญญาณเสียงกำเนิดจากส่วนประกอบ 2 ส่วนคือ ค่าแวดล้อมของสเปกตรัม (Spectral Envelop) และโครงสร้างรายละเอียดของสเปกตรัม (Spectral fine structure) ทั้งสองส่วนนี้สามารถแยกกันได้โดยการใส่ค่าลอการิธึม สัมประสิทธิ์เซปสตรัมเป็นการแทนสัญญาณในส่วนของค่าแวดล้อมของสเปกตรัมเท่านั้น

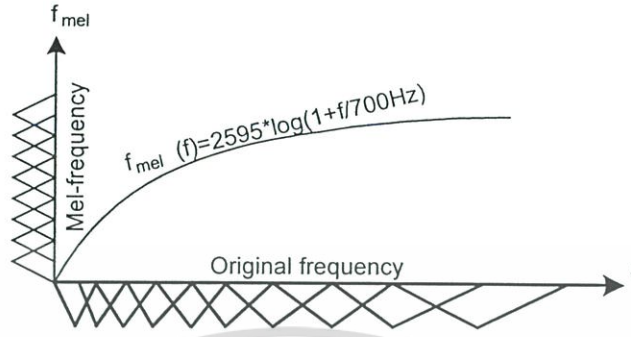


รูปที่ 2.14 การตอบสนองเสียงด้านความถี่ของหูชั้นใน

ในงานวิจัยนี้จะขอนำเสนอการวิเคราะห์สเปกตรัมโดยใช้ค่าสัมประสิทธิ์บนสเกลเมล (Mel Frequency Cepstral Coefficient: MFCC) โดยการผ่านสเปกตรัมของสัญญาณเสียงเข้าไปในกลุ่มของตัวกรอง (Filter bank) ซึ่งกระจายอยู่บนสเกลความถี่แบบไม่สม่ำเสมอตามเมลสเกล (Mel Scale) ซึ่งออกแบบมาให้เหมาะสมกับการรับฟังของหูมนุษย์

การคำนวณหาค่าเซปสตรัมแบบ MFCC จะต้องแปลงความถี่จริง (Hz) ให้อยู่ในรูปแบบความถี่แบบเมล คือ ช่วงความถี่ที่ต่ำกว่า 1 KHz จะมีลักษณะค่าความเป็นเชิงเส้น แต่ความถี่ที่สูงกว่า 1 KHz จะมีความเปลี่ยนแปลงในลักษณะที่เป็นค่าลอการิธึม ซึ่งการแปลงความถี่จริงให้เป็นค่าความถี่เมลสามารถคำนวณหาได้จากสมการ 2.29 [10]

$$F_{mel} = 2595 \times Log_{10}\left(1 + \frac{F_{Hz}}{700}\right) \quad (2.29)$$



รูปที่ 2.15 สเกลเมลและหน้าต่างเมล [11]

หากสมมติให้ค่า y(n) แทนสัญญาณเสียงที่รับเข้ามาจะสามารถคำนวณค่าเซปสตรัมได้ดังนี้

- **ขั้นตอนที่ 1.** แปลงสัญญาณเสียงจากสัญญาณเชิงเวลาให้อยู่ในเชิงความถี่โดยวิธีการแปลงฟูริเยร์อย่างเร็วดังสมการ

$$Y(m) = \sum_{m=0}^{F-1} y(n) * W(n) * e^{-j2n\pi\frac{m}{F}} \quad ;m = 0,1,2...F-1 \quad (2.30)$$

เมื่อ F คือ ขนาดของเฟรม

W(n) คือ ฟังก์ชั่นหน้าต่างแฮมมิ่ง (Hamming window function)

- **ขั้นตอนที่ 2.** หากำลังงานของสเปกตรัม

$$X(m) = |Y(m)|^2 \quad (2.31)$$

- **ขั้นตอนที่ 3.** คำนวณค่ากำลังงานในแต่ละหน้าต่างเมล

$$S[k] = \sum_{j=0}^{K/2-1} W_k(j) * X(j) \quad ;1 \le k \le M \quad (2.32)$$

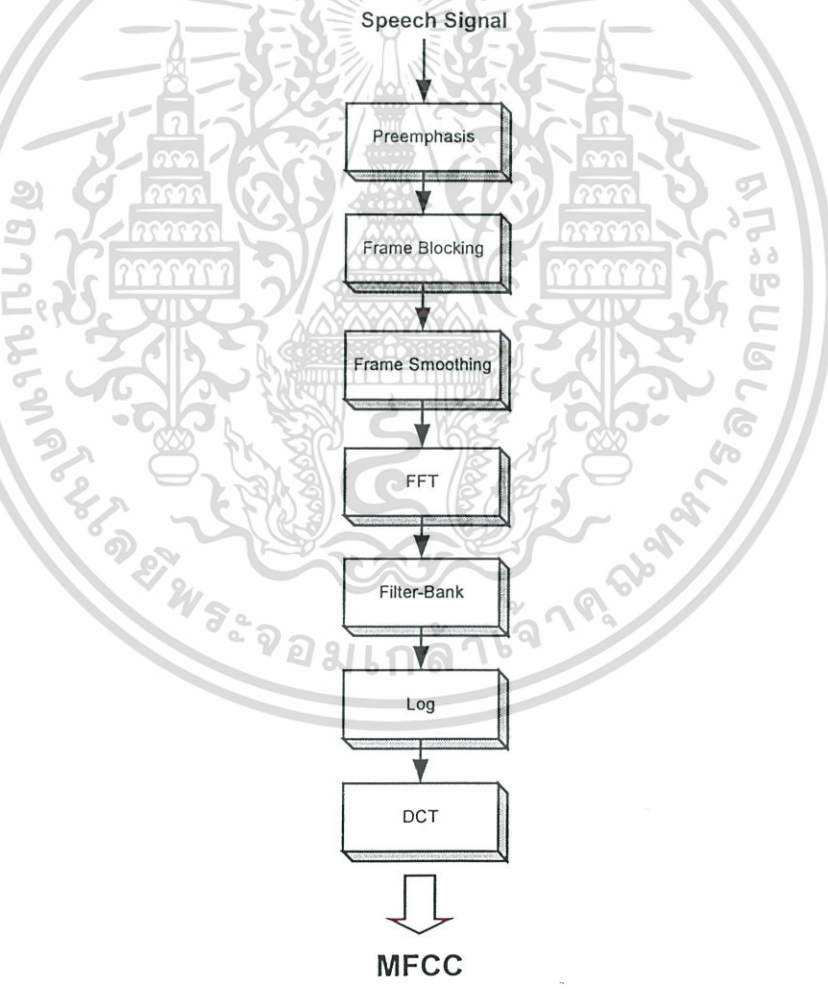
เมื่อ M คือ จำนวนหน้าต่างเมลในเมลสเกล ซึ่งโดยทั่วไปกำหนดให้มี 20-24 หน้าต่าง

$W_k(j)$ คือ ฟังก์ชั่นค่าน้ำหนักของแต่ละตัวกรองสามเหลี่ยม

- **ขั้นตอนที่** 4. คำนวณผลที่ได้กับค่าลอการิธึมและการแปลงโคซายน์เพื่อหาค่าสัมประสิทธิ์เซปสตรัมบนความถี่เมล ด้วยสมการ

$$C[n]=\sum_{k=1}^{M} Log\big(S[k]\big)*\cos\left[n*(k-0.5)*\frac{\pi}{M}\right] \qquad ;1\le n\le L \qquad (2.33)$$

เมื่อ L คือจำนวนลำดับ MFCC ที่ต้องการ



รูปที่ 2.16 ขั้นตอนการคำนวณ MFCC

บทที่ 3
โครงข่ายประสาทเทียม

เมื่อสัญญาณเสียงถูกส่งเข้ามาในกระบวนการประมวลผลการรู้จำเสียงพูด และผ่านกระบวนการในการวิเคราะห์หาคุณลักษณะเฉพาะของสัญญาณเสียง (Feature Analysis) เป็นที่เรียบร้อยแล้วจะถูกส่งผ่านมายัง Pattern classification Pattern Classification ซึ่งเป็นการหากลุ่มของของข้อมูลที่เราไม่ทราบเทียบกับข้อมูลอ้างอิงแต่ละตัว วิธีที่ใช้ในขั้นตอนนี้มีหลายวิธี เช่น DTW, HMM, โครงข่ายประสาทเทียม DTW ใช้เทคนิคการปรับยืดขยายหรือหดรูปคลื่นสัญญาณตามแกนเวลาแบบไดนามิค ซึ่งในงานวิจัยนี้ได้เลือกใช้ Multilayer Percepton หรือ MLP ซึ่งเป็นโครงข่ายประสาทเทียม (Neural Network) อีกแบบหนึ่งที่มีโครงสร้างและลักษณะการทำงานที่มีการจัดเรียงเป็นชั้น (Layer) และมีความยืดหยุ่นในการทำงานสูง จึงสามารถใช้ในการแก้ปัญหากับข้อมูลอินพุตที่มีความซับซ้อนดังเช่น ข้อมูลเสียง ได้ดี

3.1 โครงข่ายประสาทเทียมและการประยุกต์ใช้ในการรู้จำเสียงพูด

การรู้จำเสียงพูด (Speech Recognition) ถือว่าเป็นอีกหนึ่งแหล่งความรู้ที่เป็นประโยชน์ในการพัฒนาด้านปัญญาประดิษฐ์ (Artificial Intelligence) โดยมีแนวคิดหลักๆ อยู่ 2 ประการคือ

- การได้มาซึ่งความรู้ (Knowledge Acquisition) หรือการเรียนรู้ (Learning)
- การประยุกต์ใช้งาน (Adaptation)

โครงข่ายประสาทเทียมเป็นอีกวิธีการหนึ่งที่ใช้ในการปฏิบัติหรือแก้ไขปัญหา โดยมีรูปแบบแนวคิดเหมือนการจำลองปมประสาทรับความรู้สึก (Neural) ของสมองมนุษย์แบบง่ายๆ ซึ่งปมประสาทนี้มีอยู่เป็นจำนวนมากในสมองมนุษย์ โดยแต่ละปมประสาทมีการเชื่อมต่อกันเป็นจำนวนมหาศาล จึงเป็นเหตุให้การทำงานของสมองมนุษย์ทำงานได้อย่างรวดเร็ว เนื่องจากมีความไม่เป็นเชิงเส้นและเป็นการทำงานแบบขนาน จึงอาจเรียกอีกชื่อว่า Artificial Neural Network หรือ ANN ซึ่งสามารถจัดแบ่งได้ 3 ประเภทหลักคือ [12]

1. Single/Multilayer perceptrons
2. Hopfield หรือ Recurrent Network
3. Kohonen หรือ Self-organizing Networks

โดยในเอกสารนี้จะเน้นการใช้งานโครงข่ายประสาทเทียมกับการรู้จำเสียงพูดด้วยวิธีการ Multilayer Perceptrons (MLP)

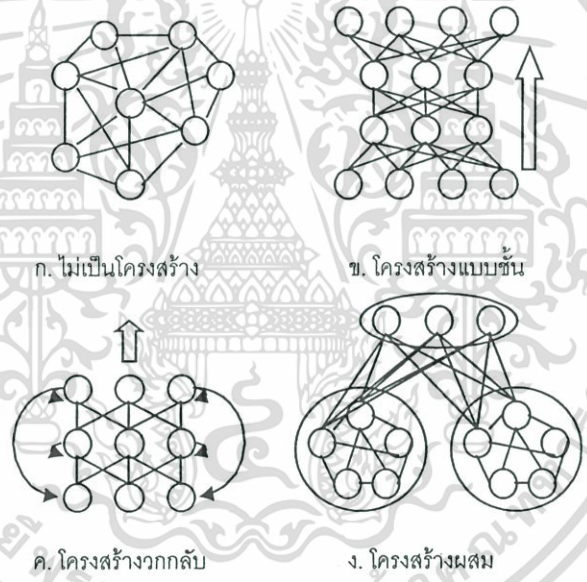
3.2 องค์ประกอบพื้นฐานของโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมประเภทต่างๆจะมีองค์ประกอบที่เหมือนกันอยู่ 4 อย่างคือ [13]

1. หน่วยประมวลผล (Processing Units)
2. การเชื่อมต่อ (Connections)
3. กระบวนการคำนวณ (Computing Procedure)
4. กระบวนการฝึกฝน (Training Procedure)

3.2.1 หน่วยประมวลผล (Processing Units)

หน่วยประมวลผลในโครงข่ายประสาทเทียมจะแบ่งออกเป็น 3 ส่วนคือ หน่วยข้อมูลเข้า (Input unit) ทำหน้าที่รับข้อมูลจากภายนอก หน่วยซ่อนตัว (Hidden unit) ที่แปลงข้อมูลภายในและหน่วยข้อมูลออก (Output unit) ทำหน้าที่ตัดสินหรือควบคุมสัญญาณ



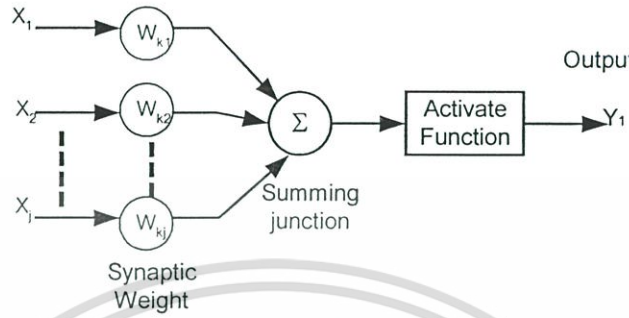
รูปที่ 3.1 แสดงโครงสร้างโครงข่ายประสาทเทียมรูปแบบต่างๆ

3.2.2 การเชื่อมต่อ (Connections)

หน่วยการประมวลผลในโครงข่ายประสาทเทียมจะถูกจัดเรียงเป็นโครงสร้างต่างๆ โดยการเชื่อมต่อซึ่งมีค่ากำกับไว้ ค่าที่กำกับการเชื่อมต่อเรียกว่าน้ำหนักเชื่อมต่อ (Weight) โครงข่ายประสาทเทียมสามารถมีการเชื่อมต่อเป็นลักษณะต่างๆได้ 4 แบบดังแสดงในรูปที่ 3.1 คือไม่เป็นโครงสร้าง (Unstructured) โครงสร้างแบบชั้น (Layered) โครงสร้างวนกลับ (Recurrent) และโครงสร้างแบบผสม (Modular)

3.2.3 **กระบวนการคำนวณ** (Computing Procedure)

โหนดเป็นหน่วยประมวลผลข้อมูลซึ่งเป็นพื้นฐานของโครงข่ายประสาทเทียมมีลักษณะดังรูปที่ 3.2 ซึ่งมีองค์ประกอบดังต่อไปนี้คือ



รูปที่ 3.2 แบบจำลองการทำงานโครงข่ายประสาทเทียม

1. การเชื่อมต่อซึ่งแต่ละการเชื่อมต่อจะมีคุณลักษณะคือ น้ำหนักการเชื่อมต่อสัญญาณเข้า X_j ของการเชื่อมต่อ j กับโหนดที่ k จะถูกคูณด้วยน้ำหนักการเชื่อมต่อ W_{kj}
2. การบวกสำหรับการรวมสัญญาณเข้าที่ถูกคูณด้วยน้ำหนักการเชื่อมต่อ
3. ฟังก์ชันการกระตุ้นสำหรับจำกัดขนาดของสัญญาณที่ออกมาจากโหนด

3.2.4 **กระบวนการฝึกฝน** (Training Procedure)

การฝึกฝนโครงข่ายประสาทเทียมคือ การปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อหรือบางกรณีเป็นการปรับเปลี่ยนโครงสร้างของโครงข่ายประสาทเทียม ซึ่งเป็นการเพิ่มหรือลบการเชื่อมต่อโหนด การปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อจะมีลักษณะทั่วไปมากกว่าการปรับเปลี่ยนโครงสร้าง เพราะการที่ค่าน้ำหนักการเชื่อมต่อเท่ากับศูนย์ถือเป็นการลบการเชื่อมต่อนั้นออกจากโครงข่ายประสาทเทียม อย่างไรก็ตามการเปลี่ยนแปลงโครงสร้างของโครงข่ายประสาทเทียมจะเป็นการเพิ่มความเร็วในการเรียนรู้และเพิ่มความสามารถในการรู้จำรูปแบบทั่วไป

โครงข่ายประสาทเทียมมีโครงสร้างเป็นชั้นและมีความไม่เป็นเชิงเส้น การปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อทำโดยวิธีการวนซ้ำ การปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อแต่ละครั้งจะต้องไม่ทำให้การเรียนรู้ที่ผ่านมาสูญเสียไป ค่าคงที่ที่ใช้ควบคุมขนาดของการปรับเปลี่ยนน้ำหนักการเชื่อมต่อเรียกว่าอัตราการเรียนรู้ (Learning Rate) การกำหนดค่าอัตราการเรียนรู้มีความสำคัญมาก ถ้ากำหนดค่าอัตราการเรียนรู้น้อยเกินไปจะทำให้การเรียนรู้นานมาก และถ้ากำหนดค่าอัตราการเรียนรู้มากจะทำให้สูญเสียการเรียนรู้ที่ผ่านมา

3.3 โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP)

โครงข่ายประสาทเทียมแบบ MLP ประกอบด้วย 3 ส่วนคือ [14]

1. ชั้นข้อมูลสำหรับรับข้อมูล (Input Layer)
2. ชั้นซ่อนตัว (Hidden Layer)
3. ชั้นข้อมูลออกสำหรับการคำนวณ (Output Layer)

สัญญาณเข้าจะผ่านเข้าไปในโครงข่ายประสาทเทียมในทิศทางเดียวจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง โครงข่ายประสาทเทียมแบบ MLP ถูกประยุกต์ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบ Supervise และใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (error-correction) คือผลต่างของผลตอบที่แท้จริง (actual response) กับผลตอบเป้าหมาย (target response) เกิดเป็นสัญญาณผิดพลาด (error signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ ค่าน้ำหนักการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย

3.4 คุณสมบัติของโครงข่ายประสาทเทียมแบบ MLP

1. แบบจำลองของแต่ละโหนดจะมีความไม่เป็นเชิงเส้นและความไม่เป็นเชิงเส้นดังกล่าวจะต้องมีความราบเรียบคือสามารถหาอนุพันธ์ได้ทุกจุด ความไม่เป็นเชิงเส้นดังกล่าวถูกกำหนดจากฟังก์ชันซิกมอยด์

$$y_j = \frac{1}{1+e^{-\lambda v_j}} \tag{3.1}$$

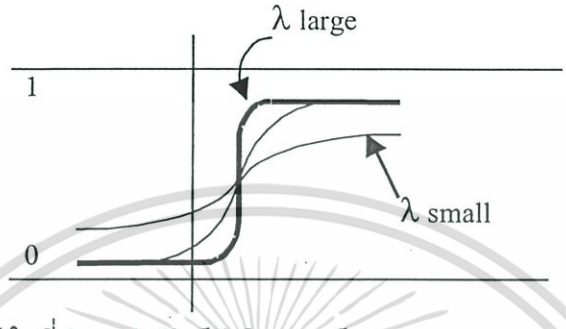
เมื่อ V_j คือผลรวมภายในโหนดที่ j

Y_j คือสัญญาณขาออกของโหนดที่ j

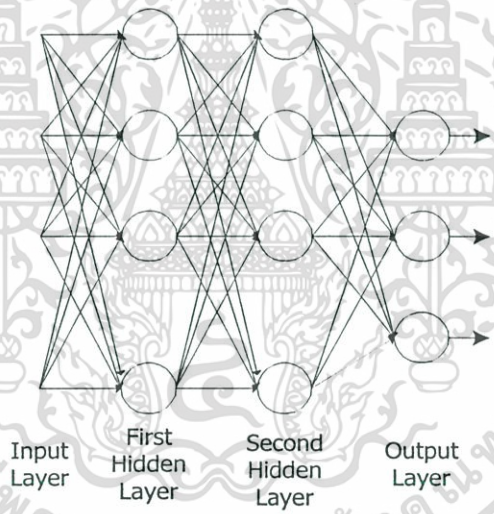
λ คือค่าที่กำหนดความชันของฟังก์ชัน

2. โครงข่ายประสาทเทียมจะมีจำนวนชั้นซ่อนตัวที่มากกว่าหนึ่งชั้นได้ ซึ่งไม่ใช่ชั้นข้อมูลเข้าและชั้นข้อมูลออก โหนดในชั้นซ่อนตัวนี้จะทำให้โครงข่ายประสาทเทียมสามารถเรียนรู้งานที่มีความซับซ้อนได้ดีขึ้น

3. โครงข่ายประสาทเทียมแบบ MLP จะมีการเชื่อมต่อกันมาก



รูปที่ 3.3 ความสัมพันธ์ λ ที่ส่งผลต่อฟังก์ชันซิกมอยด์

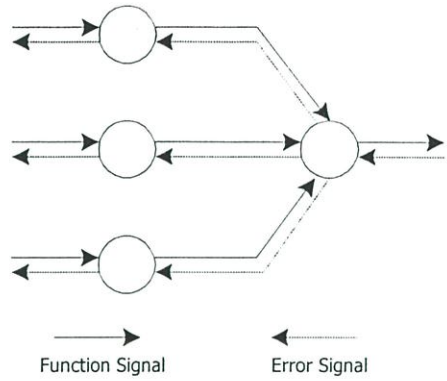


รูปที่ 3.4 โครงสร้างประสาทเทียมแบบ MLP [15]

จากรูปที่ 3.4 เป็นภาพการแสดงตัวอย่างโครงสร้าง โครงข่ายประสาทเทียมแบบ MLP ที่มีจำนวนชั้นซ่อนตัว 2 ชั้น สัญญาณที่มีโครงข่ายประสาทเทียมแบบ MLP มี 2 ประเภทคือ Function Signal และ Error Signal ดังแสดงในรูปที่ 3.5 และมีรายละเอียดดังนี้

1. Function Signal เป็นสัญญาณเข้าที่มาจากโหนดในชั้นก่อนหน้าจะส่งผ่านไปข้างหน้าจากโหนดหนึ่งไปสู่อีกโหนดหนึ่ง

2. Error Signal เป็นสัญญาณที่เกิดขึ้นที่โหนดในชั้นข้อมูลออกของโครงข่ายประสาทเทียมและถูกส่งผ่านย้อนกลับจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง



รูปที่ 3.5 แสดงทิศทาง Function Signal และ Error Signal

3.5 ขั้นตอนวิธีการส่งค่าย้อนกลับ (Backpropagation Algorithm)

การส่งค่าย้อนกลับเรียนรู้โดยกระบวนการวนซ้ำของกลุ่มตัวอย่างที่ใช้ในการฝึกฝนเปรียบเทียบกับค่าที่ได้จากการทำนายของเครือข่ายสำหรับแต่ละค่าตัวอย่างด้วยค่าที่แท้จริง แต่สำหรับค่าตัวอย่างการฝึกฝนนั้น น้ำหนักการเชื่อมต่อจะถูกปรับให้มีค่า Mean Square Error ที่ต่ำอยู่ระหว่างค่าที่เครือข่ายทำนายได้กับค่าจริง ซึ่งการปรับเปลี่ยนนี้เป็นการส่งค่าย้อนกลับ (Backward Pass) มาจากชั้นข้อมูลซ่อนตัวไปยังชั้นรับข้อมูล (จากนี้ไปจะเรียกว่า แบคพรอพาเกชัน) แม้จะไม่มีการรับรองในค่าน้ำหนักการเชื่อมต่อว่าในที่สุดจะทำให้ได้ค่าที่แท้จริงตรงตามที่ต้องการ และเป็นการทำให้ขั้นตอนการเรียนรู้นั้นสิ้นสุดลงได้ก็ตาม ซึ่งขั้นตอนและวิธีการส่งค่าย้อนกลับมีดังนี้ [16]

- ***ขั้นตอนที่*** *0.* Initialize the weights

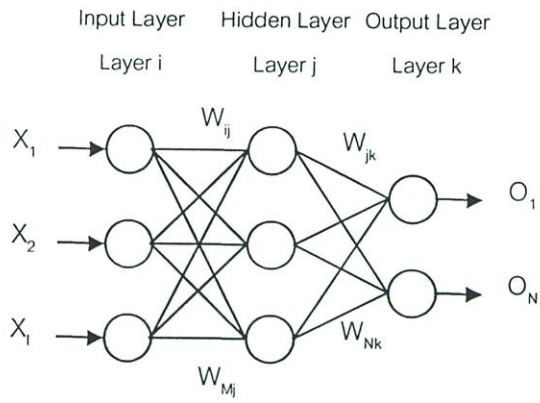
เป็นการกำหนดค่าน้ำหนักเชื่อมต่อในเครือข่ายให้มีค่าน้อยๆ หรืออาจเกิดจากการสุ่มเป็นค่าตัวเลขที่มีค่าต่ำๆ (-1.0 ถึง +1.0 หรือ –0.5 ถึง +0.5)

- ***ขั้นตอนที่*** *1.*ลูปวนการทำงาน (ขั้นตอน ที่ 2-9)

เป็นขั้นตอนที่จะตรวจสอบเงื่อนไขการหยุดการทำงานหากเงื่อนไขไม่ถูกต้องจะดำเนินตามขั้นตอนที่ 2-9 ต่อไป

- ***ขั้นตอนที่*** *2.* การฝึกฝน (ขั้นตอน ที่ 3-9)

เป็นขั้นตอนกระบวนการฝึกฝนเพื่อให้โครงข่ายประสาทเทียมได้เรียนรู้ ซึ่งประกอบด้วย 3 ส่วนหลักคือ Feedforward (ขั้นตอนที่ 3-5), Backpropagation of Error (ขั้นตอนที่ 6-7) และ Update weight and biase (ขั้นตอนที่ 8)



รูปที่ 3.6 โครงสร้างประสาทเทียมแบบมี 1 ชั้นซ่อนตัว

- ***ขั้นตอนที่ 3.*** Receives input signal

เป็นการรับสัญญาณอินพุตเข้ามาในแต่ละ Input node และแพร่กระจายไปสู่ทุกๆโหนดในชั้นถัดไป (ชั้นซ่อนตัว)

- ***ขั้นตอนที่ 4.*** คำนวณผลรวมน้ำหนักและค่าอินพุตที่รับเข้ามาของชั้นซ่อนตัว

ในการคำนวณหาสัญญาณอินพุตที่นับเข้ามาในชั้นซ่อนตัวนั้นสามารถหาได้จากสมการ โดย Input Lj คือ ค่าผลรวมซึ่งเป็นค่าอินพุตที่โหนดลำดับที่ j ในชั้นซ่อนตัว

$$Input\,Lj = \sum_{i=1}^{M} X_i W_{ij} \qquad (3.2)$$

X_i คือ ค่าสัญญาณอินพุตที่ถูกส่งมายังโหนดที่ i ในชั้นรับข้อมูล

W_{ij} คือ ค่าน้ำหนักประจำตัวของอินพุตโหนดที่ j ในชั้นซ่อนตัวซึ่งเชื่อมต่อมาจากโหนดที่ i ในชั้นรับข้อมูล

จากนั้นทำการปรับสัญญาณออกจากโหนดในชั้นซ่อนตัวด้วย Activation Function หรือฟังก์ชันซิกมอยด์แล้วทำการส่งออกเพื่อใช้เป็นสัญญาณอินพุตของโหนดในชั้นต่อไป (ชั้นข้อมูลออก) ดังสมการ

$$Y_j = f(Input\,L_j) \qquad (3.3)$$

- ***ขั้นตอนที่ 5.*** คำนวณผลรวมน้ำหนักและค่าอินพุตที่รับเข้ามาของชั้นข้อมูลออก

$$Input\ Lk = \sum_{j=1}^{N} Y_j W_{jk} \tag{3.4}$$

ในขั้นตอนนี้จะมีการคำนวณคล้ายกับในขั้นตอนที่ 4 ทุกประการ โดยทำการเปลี่ยนค่าสัญญาณและน้ำหนัก จาก X_i เป็น Y_j และเปลี่ยนค่า W_{ij} เป็น W_{jk} ซึ่งจะได้สมการใหม่ดังนี้

จากนั้นทำการปรับสัญญาณออกจากโหนดในชั้นซ่อนตัวด้วย Activation Function หรือฟังก์ชันซิกมอยด์แล้วทำการส่งออกเพื่อใช้เป็นสัญญาณอินพุตของโหนดในชั้นต่อไป (ชั้นข้อมูลออก) ดังสมการ

$$Z_k = f(Input\ L_k) \tag{3.5}$$

- ***ขั้นตอนที่ 6.*** หาค่า Error Signal term (Output Layer)

เป็นการหาค่าความคลาดเคลื่อน (Error Signal Term) ในชั้นข้อมูลออกระหว่างค่าที่ได้จริง (actual output values) กับเป้าหมายหรือค่าที่ต้องการ (desired output values หรือ target) เพื่อใช้ในการปรับน้ำหนักโดยเป็นสัดส่วนกับความคลาดเคลื่อนคูณกับค่าอินพุต ดังสมการ

$$\delta_k = -\frac{1}{2}(t_k - Z_k) f'(Input\ Lk) \tag{3.6}$$

คำนวณ Weight correction term เพื่อใช้ในการปรับปรุงน้ำหนัก W_{jk} ต่อไปภายหลังด้วยสมการ

โดย α คือ อัตราการเรียนรู้ (Learning Rate)

$$\Delta w_{jk} = \alpha \delta_k Y_j \tag{3.7}$$

- ***ขั้นตอนที่ 7.*** หาผลรวม Delta input (Hidden Layer)

ในขั้นตอนนี้มีหลักในการคำนวณเหมือนกับในขั้นตอนที่ 6 เพียงแต่ชั้นซ่อนตัวจะไม่มีค่าเป้าหมาย (Output target) จึงแทนด้วยสมการ

$$\delta_j = \sum_{k=1}^{N} \delta_k W_{jk} f'(Input\ Lj) \tag{3.8}$$

คำนวณ Weight correction term เพื่อใช้ในการปรับปรุงน้ำหนัก W_{ij} ต่อไปภายหลังด้วยสมการ

$$\Delta w_{ij} = \alpha \delta_j x_i \tag{3.9}$$

- ***ขั้นตอนที่ 8.*** ปรับปรุงค่าน้ำหนัก

สามารถปรับค่าน้ำหนักใหม่แต่ละตัวในชั้นข้อมูลออกได้ด้วยสมการ

$$W_{jk}^{t} = W_{jk}^{t-1} + \Delta w_{jk} \tag{3.10}$$

เมื่อ W^t คือ ค่าน้ำหนักหลังปรับ (ค่าใหม่)

W^{t-1} คือ ค่าน้ำหนักก่อนปรับ

จากนั้นปรับค่าน้ำหนักใหม่แต่ละตัวในชั้นซ่อนตัวด้วยสมการ

$$W_{ij}^{t} = W_{ij}^{t-1} + \Delta w_{ij} \tag{3.11}$$

- ***ขั้นตอนที่ 9.*** Testing stopping condition

ในขั้นตอนนี้เป็นการทดสอบการทำงานของโครงข่ายประสาทเทียมหลังจากที่ผ่านกระบวนการฝึกฝนเสร็จสิ้นแล้ว โดยจะใช้การคำนวณเพื่อให้ได้เอาต์พุต ในลักษณะเช่นเดียวกับการฝึกฝนในส่วนของ Forward ทุกประการ เพียงแต่ป้อนข้อมูลที่ต้องการใช้ในการทดสอบ ส่วนเอาต์พุตที่ได้จะเกิดจากการตรวจสอบ ซึ่งอาจใช้เวลาน้อยกว่ากระบวนการฝึกฝนมาก เพราะเป็นเพียง forward pass เพียงรอบเดียวโดยใช้ค่าน้ำหนักที่ได้จากการฝึกฝนเป็นตัวถอดรหัสความสัมพันธ์

3.6 เงื่อนไขการหยุดฝึกฝน

การกำหนดเงื่อนไขในการหยุดการฝึกฝนนั้นสามารถทำได้ 2 กรณีคือ [17]

1. เมื่อผลรวมของค่าผิดพลาดเฉลี่ยระหว่างเอาต์พุตกับเป้าหมายทั้งหมด (SSQERR) มีค่าลดลงน้อยกว่าค่าที่กำหนด โดยกำหนดสมการหาค่าเฉลี่ยของเอาต์พุตทั้งหมดดังนี้

$$SSQERR = \frac{1}{2}\sum (TARGET - OUTPUT)^2 \tag{3.17}$$

SSQERR คือค่า Sum square error ที่ได้จากเอาต์พุตของโครงข่ายประสาทเทียมเทียบกับค่าเป้าหมายที่ใช้สำหรับบ่งชี้ผลของการฝึกฝนเพียงพอกับการใช้งานหรือไม่

2. จำนวนรอบของการฝึกฝนดำเนินไปจนถึงค่าที่ตั้งไว้ ซึ่งจำนวนที่เหมาะสมได้จากการทดลอง

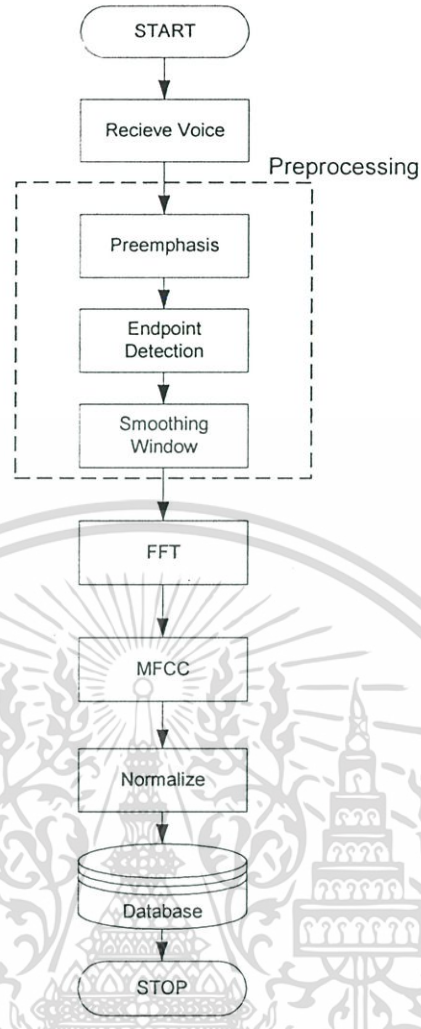
บทที่ 4
ระบบรู้จำเสียงพูดแบบแยกวิเคราะห์ตามจำนวนพยางค์

ระบบรู้จำเสียงโดยทั่วไปหากมีจำนวนของข้อมูลที่ใช้ในการรู้จำมาก ประสิทธิภาพในการรู้จำจะต่ำเมื่อเทียบกับระบบที่มีจำนวนคำในการรู้จำที่น้อยกว่า งานวิจัยนี้จึงนำเสนอการจัดแบ่งกลุ่มคำของการรู้จำโดยแยกชุดในการฝึกฝนออกเป็น 3 กลุ่มตามจำนวนพยางค์ เพื่อเพิ่มประสิทธิภาพในการรู้จำทำให้การรู้จำถูกต้องมากขึ้น และใช้เวลาในการฝึกฝนน้อยลง โดยใช้หลักการวิเคราะห์ค่าพลังงานจากสเปกตรัมในการแบ่งกลุ่มของพยางค์แทนการใช้ค่าระดับพลังงานแบบเดิม และสกัดค่าลักษณะสำคัญด้วยการหาเซปสตรัมจากสเปกตรัมของข้อมูลเสียงพูด

4.1 ขั้นตอนการรู้จำเสียง

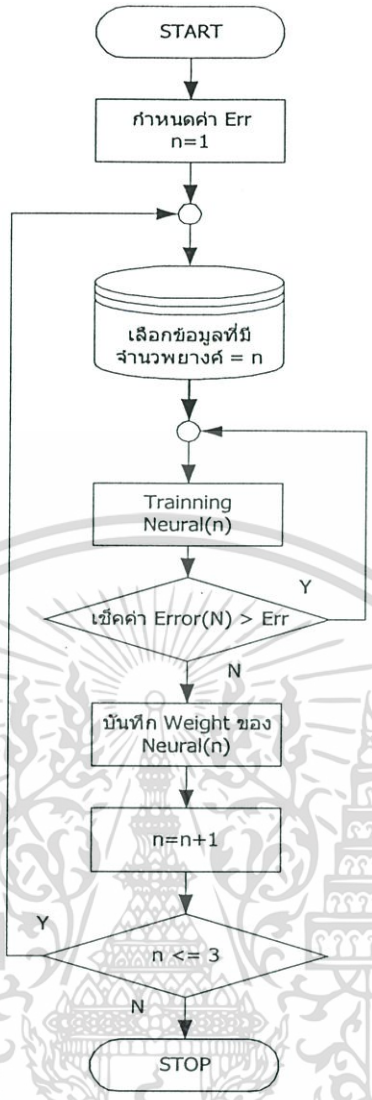
สัญญาณเสียงที่ได้รับจากไมโครโฟนจะถูกส่งผ่านมายังการ์ดเสียง (Sound Card) ที่มีการสุ่มสัญญาณที่ความถี่ 11 KHz ซึ่งการรับสัญญาณเสียงเข้ามานั้นจะจัดเก็บข้อมูลเสียงไว้ประมาณ 0.1 วินาที และในแต่ละครั้งของการรับสัญญาณเสียงจะตรวจสอบผลรวมของระดับสัญญาณ หากมีค่ามากพอจึงจะทำการรับสัญญาณต่อไปอีกเป็นระยะเวลาประมาณ 2 วินาที แต่หากระดับสัญญาณมีค่าต่ำกว่าค่าที่ตั้งไว้จะหยุดและวนรับสัญญาณเพื่อตรวจสอบต่อไปเรื่อยๆ ทุกช่วง 0.1 วินาที ด้วยเหตุนี้จึงทำให้ไม่เกิดการคำนวณที่สูงมากจนเกินไปในขณะที่ไม่มีสัญญาณเกิดขึ้น

สัญญาณเสียงพูดที่ได้จะถูกส่งเข้าสู่กระบวนการจัดเตรียมข้อมูลก่อนประมวลผล และจัดแบ่งข้อมูลออกเป็นเฟรม ซึ่งการวิเคราะห์สัญญาณเสียงขนาดใหญ่อาจจะใช้เฟรมที่มีขนาดใหญ่ 100 ms ได้ [6] ดังนั้นจึงเลือกใช้กรอบวิเคราะห์ขนาดเฟรมละ 1024 ข้อมูล หลังจากจัดแบ่งข้อมูลสัญญาณเสียงออกเป็นเฟรมย่อยๆแล้ว จะนำข้อมูลที่ได้แต่ละเฟรมผ่านเข้าฟังก์ชันหน้าต่าง (Smoothing Window) ซึ่งในที่นี้จะใช้ฟังก์ชันหน้าต่างแฮมมิ่ง (Hamming Window) ก่อนทำการคำนวณ FFT โดยผลลัพธ์ที่ได้จากการคำนวณ FFT ของแต่ละเฟรมนั้นจะสามารถแยกได้เป็นความถี่สเปกตรัมได้ 1024 ข้อมูล (จุด) ครอบคลุมความถี่ตั้งแต่ 0-5.5 KHz จากนั้นจะทำหาเซปสตรัม MFCC ทั้งหมดทีละเฟรม แล้วจัดเก็บลงฐานข้อมูลพร้อมกับจำนวนพยางค์ของคำนั้นลงในฐานข้อมูล เพื่อใช้เป็นข้อมูลในการฝึกสอนโครงข่ายประสาทเทียมดังแสดงในรูปที่ 4.1



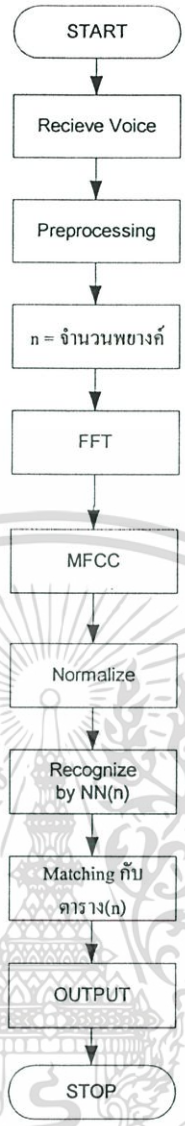
รูปที่ 4.1 ขั้นตอนการบันทึกเสียง

ในงานวิจัยนี้จะใช้โครงข่ายประสาทเทียม 3 ชุด เพื่อฝึกฝนและทดสอบการรู้จำ โดยโครงข่ายประสาทเทียมชุดที่1, 2 และ 3 จะใช้สำหรับการรู้จำคำที่มี 1, 2 และมากกว่า 2 พยางค์ตามลำดับ โดยขั้นตอนการฝึกฝนจะกำหนดค่า Err เพื่อเป็นเงื่อนไขในการหยุดการฝึกฝน โดยจะทำการเลือกข้อมูลทั้ง 3 ชุดพยางค์ แล้วทำการฝึกฝนทีละชุดพยางค์จนกว่าค่า SSQERR จะน้อยกว่าหรือเท่ากับค่า Err ที่กำหนดไว้ ดังแสดงในรูปที่ 4.2



รูปที่ 4.2 ขั้นตอนการฝึกสอนของโครงข่ายประสาทเทียมแยกตามจำนวนพยางค์

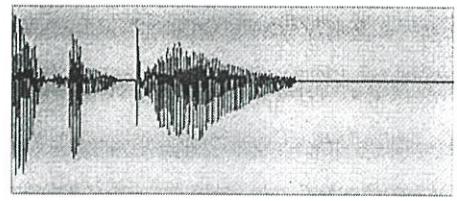
หลังจากที่ได้ทำการฝึกสอนโครงข่ายประสาทเทียมแล้ว สามารถทำการทดสอบการรู้จำได้ โดยการรับสัญญาณเสียงเข้ามา เพื่อคำนวณค่า FFT และเซปสตรัม MFCC ตามลำดับ แล้วป้อนเข้าสู่โครงข่ายประสาทเทียม เพื่อทำการหาคำตอบว่าค่าของผลลัพธ์ที่ได้เป็นอะไร แล้วทำการเทียบค่า (Matching) กับตาราง (Index Table) ของแต่ละชุดพยางค์ว่าตรงกับคำใด ซึ่งจะมีรหัสแอสกี (ASCII) ของอักษรนั้นๆ กำกับไว้เพื่อใช้ในพิมพ์เอกสาร โดยวิธีการทดสอบการรู้จำนั้นสามารถแสดงได้ดังรูปที่ 4.3



รูปที่ 4.3 ขั้นตอนทดสอบการรู้จำของโครงข่ายประสาทเทียมแยกตามจำนวนพยางค์

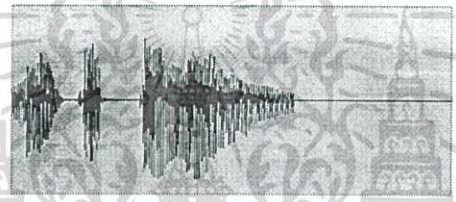
4.2 การหาขอบเขตสัญญาณด้วยการวิเคราะห์สเปกตรัม

การหาขอบเขตของสัญญาณด้วยวิธีการหารูปคลื่นพลังงาน (Energy Pulse Detection) เป็นเทคนิคที่นิยมใช้กันโดยทั่วไปสำหรับการรู้จำแบบเป็นคำโดด แต่อย่างไรก็ตามวิธีการนี้จะไม่สามารถแก้ไขปัญหาของสัญญาณเสียงที่เกิดจากการเปล่งเสียงประเภท Plosive Sound ได้ดีนัก ทั้งนี้เนื่องจากคำในกลุ่มนี้จะมีเสียงลมเกิดขึ้นก่อน หรือระหว่างการเปล่งเสียงพยางค์เช่น “ท” และ “พ” หรือแม้กระทั้งเสียงประเภท Fricative Sound เช่น “ฟ” ดังนั้นหากสังเกตภาพสัญญาณอนาล็อกของเสียงดังกล่าวนี้ จะพบพลังงานในระดับที่สูงชั่วขณะดังแสดงในรูปที่ 4.4



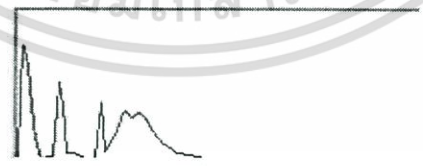
รูปที่ 4.4 สัญญาณเสียงพูด “พอ-สำ-เพา”

หากพิจารณารูปที่ 4.4 จะพบว่ามีเสียงลมเกิดขึ้นหน้าของเสียงของพยางค์สุดท้าย แต่ด้วยเทคนิคการกำหนดค่าการยกเลิก (Reject) ของการหารูปคลื่นพลังงาน (Energy Pulse Detection) ด้วยระดับค่าอ้างอิง k4 อาจสามารถแก้ปัญหาดังกล่าวนี้ได้ บ่อยครั้งที่ระดับของแรงลมนี้มีค่าสูงมากพอๆ กับค่าพลังงานของเสียงพยางค์ปกติ และหากกำหนดช่วงระยะเวลาของ k4 ไว้มากเกินไปก็อาจทำให้ไม่สามารถตรวจพบพยางค์เสียงขนาดเล็กๆ ได้

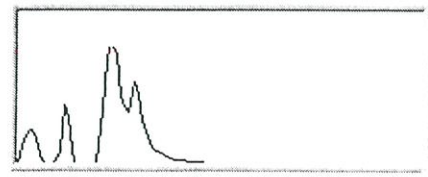


รูปที่ 4.5 สัญญาณเสียงพูด “ไม้-จัด-ตะ-วา”

ปัญหาอีกประการที่อาจพบในการหาขอบเขตของสัญญาณก็คือ เสียงพยางค์อยู่ติดกันมากจนเกินไปอาจทำให้ไม่สามารถแยกพยางค์ทั้งสองออกจากกันได้ ดังแสดงในรูปที่ 4.5 หากพิจารณาด้วยสายตาอาจไม่สามารถแยก 2 พยางค์สุดท้ายออกจากกัน แม้ว่าจะใช้การหารูปคลื่นจากพลังงานดังแสดงในรูปที่ 4.6 (ข)



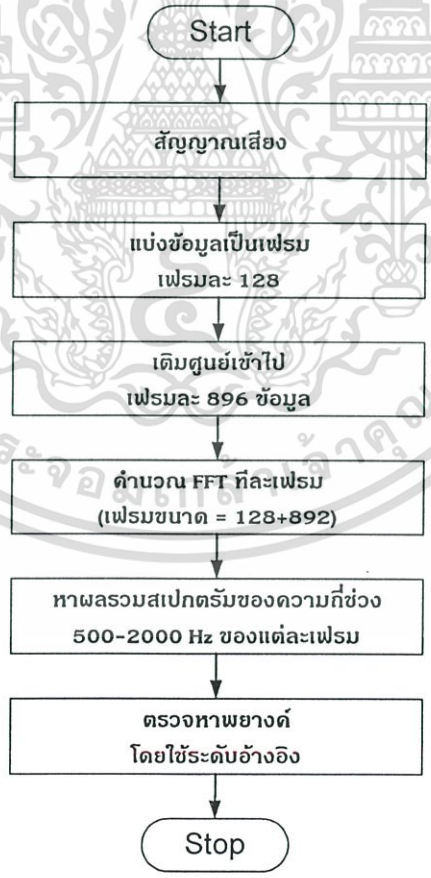
(ก) พลังงานของเสียงพูด “พอ-สำ-เพา”



(ข) พลังงานของเสียงพูด “ไม้-จัด-ตะ-วา”

รูปที่ 4.6 พลังงานเสียงพูดจาก Energy Pulse Detection

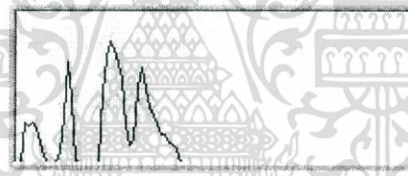
จากปัญหาดังกล่าวข้างต้น งานวิจัยนี้จึงขอนำเสนอการค้นหาขอบเขตพยางค์โดยใช้หลักการหาค่าพลังงานในรูปแบบองค์ประกอบเชิงความถี่ด้วยการแบ่งสัญญาณเสียง ซึ่งอยู่ในรูปของคาบเวลาออกเป็นบล็อกซึ่งเรียกว่าเฟรม (Frame) และนำข้อมูลในแต่ละเฟรมส่งเข้าประมวลผลเพื่อหาสเปกตรัมของกำลังด้วยการแปลงฟูริเยร์อย่างเร็ว (Fast Fourier Transform หรือ FFT) โดยหลังจากสัญญาณเสียงทุกเฟรมผ่านกระบวนการคำนวณด้วย FFT แล้วจะทำการหาผลรวมของพลังงานในแต่ละเฟรมเฉพาะช่วงย่านความถี่ 500-2000Hz ซึ่งเป็นช่วงความถี่เสียงพูดทั่วไป โดยสัญญาณเสียงที่รับเข้ามาจะถูกคูณด้วยอัตราการสุ่ม (Sampling Rate) เท่ากับ 11 KHz และสัญญาณเสียงที่รับเข้ามาจะมีขนาด 1024 ข้อมูล แต่ละจุดของสเปกตรัมจะมีค่าระยะห่างเท่ากับ f_s/N = 10.77 Hz ซึ่งสิ่งที่เราจะต้องพิจารณาคือจำนวนข้อมูลของแต่ละเฟรมหากมีค่าที่มากเกินไปจะลดความสามารถของการติดตามการเปลี่ยนแปลงของสเปกตรัม ในขณะที่หากค่า N มีค่าน้อยไปก็จะทำให้สเปกตรัมไม่มีความละเอียด โดยทั่วไปแล้วการติดตามความเปลี่ยนแปลงของพลังงานนี้จะเกิดอย่างรวดเร็วมาก ดังนั้นจะใช้หน้าต่าง (Window) หรือกรอบวิเคราะห์ประมาณ 5-10 ms แต่ในที่นี้จะใช้ข้อมูล 128 จุด นั่นคือเป็นช่วงเวลาประมาณ 11.6 ms



รูปที่ 4.7 ขั้นตอนการหาขอบเขตพยางค์ด้วย FFT

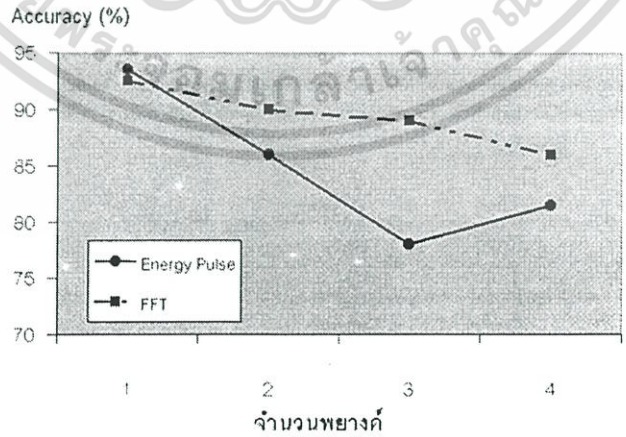
การส่งข้อมูลเพียง 128 จุดเข้าประมวลผลด้วย FFT จะทำให้ได้ช่วงห่างของความถี่เท่ากับ 86.13 Hz เพื่อให้ค่าความละเอียดของสเปกตรัมมีสูงขึ้นจึงส่งข้อมูลเข้าวิเคราะห์ FFT ด้วยเทคนิคการเติมศูนย์ (Zero padding) เข้าไปอีก 896 จุด รวมเป็นข้อมูลทั้งหมด 1024 จุด เพื่อให้ค่าของความถี่มีความห่างเพียง 10.77 Hz เมื่อนำผลรวมของพลังงานจากแต่ละเฟรมในช่วงความถี่ 500-2000 Hz วาดเป็นเส้นกราฟแล้วใช้เส้นอ้างอิงเพียงระดับเดียวเพื่อหาจุดเริ่มต้นและสิ้นสุดของพยางค์ ซึ่งสามารถแสดงขั้นตอนการหาขอบเขตพยางค์ได้ดังรูปที่ 4.7 และเมื่อนำสัญญาณเสียง “พอ-สำ-เพา” และ “ไม้-จัด-ตะ-วา” ที่เคยทดสอบด้วย Energy Pulse Detection มาทดสอบด้วยเทคนิค FFT จะได้รูปคลื่นพลังงานที่มีความชัดเจนของพยางค์มากขึ้นดังรูปที่ 4.8 และจากการทดลองเปรียบเทียบความถูกต้องของการหาขอบเขตพยางค์ขนาด 1-4 พยางค์ระหว่างวิธี Energy pulse detection และการใช้ FFT สามารถแสดงได้ดังรูปที่ 4.9

(ก) พลังงานของเสียงพูด “พอ-สำ-เพา”



(ข) พลังงานของเสียงพูด “ไม้-จัด-ตะ-วา”

รูปที่ 4.8 พลังงานเสียงพูดจากเทคนิค FFT



รูปที่ 4.9 เปรียบเทียบความถูกต้องของการหาขอบเขตพยางค์ด้วยวิธี Energy Pulse Detection และวิธีการวิเคราะห์สเปกตรัม

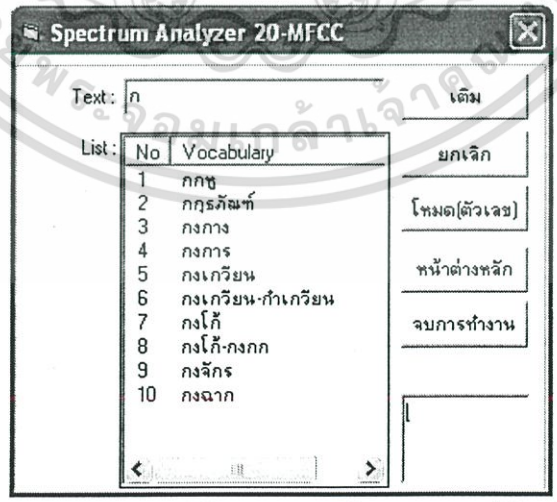
4.3 การประยุกต์ใช้พจนานุกรมในระบบการรู้จำเสียงพูด

ระบบการรู้จำที่ได้กล่าวมาในข้างต้นนี้สามารถนำมาประยุกต์ใช้ร่วมกับพจนานุกรมภาษาไทยได้ ซึ่งประโยชน์ของการเลือกใช้พจนานุกรมร่วมกับระบบการรู้จำคือ ลดข้อจำกัดในการพิมพ์ข้อความ และลดเวลาที่ใช้ในการพิมพ์ด้วยเสียง

เนื่องจากงานวิจัยนี้ใช้คำสำหรับการรู้จำเป็นข้อมูลตัวอักขระต่างๆ ในภาษาไทย การพิมพ์โดยใช้เสียงจึงเป็นการสะกดคำทีละอักษร ดังนั้นจึงสามารถพิมพ์คำเฉพาะต่างๆ เช่น ชื่อบุคคล หรือสถานที่ได้ จึงแตกต่างจากวิธีการรู้จำแบบอื่น เช่น การฝึกสอนเป็นคำศัพท์ ผลการรู้จำจะสามารถพิมพ์ได้เท่ากับจำนวนคำที่ได้รับการฝึกสอนเท่านั้น หรือหากเลือกใช้วิธีการรู้จำแบบวิเคราะห์เสียงพยางค์ก็จะมีความจำกัดในด้านการพิมพ์เนื่องจากไม่สามารถพิมพ์คำเฉพาะได้หากกลุ่มคำดังกล่าวไม่มีในพจนานุกรม

พจนานุกรมเป็นตัวช่วยในการพิมพ์ทำให้สามารถเลือกรายการคำศัพท์ที่มีอยู่ได้ โดยการออกเสียงเลือกลำดับเลขที่ของรายการ (List) ที่มีอยู่ จึงไม่จำเป็นต้องออกเสียงสะกดทุกตัวอักษรของคำที่จะพิมพ์

การพิมพ์เอกสารด้วยเสียงพูดสำหรับงานวิจัยนี้ สามารถทำได้โดยการออกเสียงพูดอักขระที่ต้องการ โดยข้อความดังกล่าวจะไปปรากฏในช่องรับข้อความ (Text) ซึ่งจะมีช่องรายการคำศัพท์แสดงคำต่างๆ จากฐานข้อมูล ผู้ใช้จึงสามารถเลือกรายการดังกล่าวได้โดยไม่ต้องออกเสียงสะกดข้อความทั้งหมด และในกรณีที่ต้องการพิมพ์คำที่อาจไม่มีอยู่ในพจนานุกรม (อาจเป็นชื่อเฉพาะต่างๆ) สามารถออกเสียงคำสั่ง "เติม" เพื่อส่งคำดังกล่าวในช่องรับข้อความไปยังโปรแกรมไมโครซอฟต์เวิร์ด ซึ่งการใช้งานโปรแกรมรู้จำเสียงได้แสดงไว้ในในภาคผนวก ก



รูปที่ 4.10 หน้าจอโปรแกรมการประยุกต์ใช้พจนานุกรมในระบบการรู้จำเสียงพูด

บทที่ 5
ผลการทดลอง

งานวิจัยนี้พัฒนาโปรแกรมโดยใช้ Microsoft Visual Basic Version 6.0 และได้ทำการออกแบบการทดลองดังนี้

5.1 การทดลองหาค่าระดับพลังงานสำหรับการตัดหัวท้ายพยางค์

จากการทดลองกับผู้ทดสอบชาย 5 คน และหญิง 5 คน โดยกำหนดใช้คำในทดสอบคือ 1, 2 และมากกว่า 2 พยางค์ กลุ่มละ 20 คำ (รวมทั้งหมดเป็น 60 คำ) ผู้ทดสอบแต่ละคนจะพูด 20 ครั้งต่อ 1 คำทดสอบ นั่นคือในการทดสอบชายและหญิงจะต้องพูดออกเสียงคำละ 100 ครั้ง (20*5)

ตารางที่ 5.1 ผลการทดลองหาค่าระดับพลังงานสำหรับการแบ่งพยางค์

Energy Level Threshold (%)	ความถูกต้อง(คำ)						เปอร์เซ็นต์ความถูกต้อง		
	1 พยางค์		2 พยางค์		มากกว่า 2 พยางค์				
	ชาย	หญิง	ชาย	หญิง	ชาย	หญิง	ชาย	หญิง	Avg.
5	97	98	86	89	85	86	89.33	91.00	90.17
10	98	98	88	91	90	92	92.00	93.67	92.83
15	98	99	95	92	88	89	93.67	93.33	93.50
20	97	99	92	94	86	79	91.67	90.67	91.17
25	98	98	86	90	75	80	86.33	89.33	87.83
30	99	97	88	84	78	75	88.33	85.33	86.83
35	98	98	84	86	76	68	86.00	84.00	85.00
40	95	95	85	87	79	64	86.33	82.00	84.17
50	94	96	76	88	68	67	79.33	83.67	81.50
60	89	92	79	85	62	58	76.67	78.33	77.50
70	91	94	78	75	70	61	79.67	76.67	78.17
80	87	93	74	79	57	63	72.67	78.33	75.50
90	86	92	68	75	64	55	72.67	74.00	73.33

จากการทดลองจะได้ค่าระดับพลังงาน Energy level threshold ที่ 15 % ให้ค่าความถูกต้องเฉลี่ยสูงสุดคือ 93.50% งานวิจัยนี้จึงเลือกใช้ค่าระดับอ้างอิงเท่ากับ 15% สำหรับการค้นหาขอบเขตและจำนวนพยางค์

5.2 การทดลองหาจำนวนลำดับของเซปสตรัมสำหรับงานวิจัย

ในการทดลองระบบรู้จำเสียงพูดแบบแยกวิเคราะห์ตามจำนวนพยางค์ เพื่อใช้สำหรับประยุกต์ใช้งานพิมพ์เอกสาร ได้กำหนดเงื่อนไขในการทดลอง และกลุ่มคำที่ใช้ในการรู้จำดังนี้

5.2.1 เงื่อนไขการทดลอง

1. ผู้ทดลองเป็นกลุ่มชาย-หญิงอายุ 20-30 ปี โดยแบ่งเป็นชาย 4 คน หญิง 6 คน และแยกทดลองทีละบุคคล

2. การทดลองใช้เครื่องไมโครคอมพิวเตอร์ขนาดประมวลผล Atlon XP +1600, การ์ดเสียง SiS 7018 โดยใช้คำทั้งหมด 92 คำ เพื่อทดสอบการรู้จำ

3. วิเคราะห์หาค่าพารามิเตอร์ที่เหมาะสมสำหรับการหาขอบเขตพยางค์และจำนวนลำดับของเซปสตรัมแบบเมล

4. เก็บตัวอย่างเสียงสำหรับฝึกฝนของโครงข่ายประสาทเทียม โดยใช้จำนวนตัวอย่างเสียง 10 เสียงต่อ 1 คำ รวมเป็นตัวอย่างเสียงทั้งหมด 920 ตัวอย่างเสียง เพื่อเพิ่มความแตกต่างของข้อมูลได้แบ่งการบันทึกเสียงเป็นครั้งละ 5 ตัวอย่างเสียงต่อ 1 คำ (ครั้ง ละ 460 ตัวอย่างเสียง)

5. ฝึกฝนระบบรู้จำเสียงพูดโดยกำหนดค่า SSQERR = 0.01, อัตราการเรียนรู้ (Learning Rate) = 0.9

5.2.2 ข้อมูลที่ใช้ในการทดลอง

ในการทดสอบการรู้จำเสียงพูดนี้จะแบ่งกลุ่มเสียงพูดออกเป็น 3 กลุ่ม ตามจำนวนพยางค์ของคำอ่านอักษรไทย [18] ได้ดังนี้

1. คำที่มี 1 พยางค์ ประกอบด้วยคำทั้งหมด 24 คำ
2. คำที่มี 2 พยางค์ ประกอบด้วยคำทั้งหมด 36 คำ
3. คำที่มีมากกว่า 2 พยางค์ ประกอบด้วยคำทั้งหมด 32 คำ

5.2.3 คำที่ใช้ในการทดสอบแยกตามจำนวนพยางค์

รายละเอียดคำที่ใช้ในแต่ละกลุ่มและคำอ่านแยกตามจำนวนพยางค์สามารถแสดงได้ดังตารางที่ 5.2

ตารางที่ 5.2 แสดงคำและคำอ่านที่ใช้ในการทดลองแต่ละกลุ่ม

	คำ 1 พยางค์	คำอ่าน	คำ 2 พยางค์	คำอ่าน	คำมากกว่า 2 พยางค์	คำอ่าน
1	หนึ่ง	*หนึ่ง*	ก	*กอ-ไก่*	ฆ	*คอ-ระ-คัง*
2	สอง	*สอง*	ข	*ขอ-ไข่*	ฏ	*ดอ-ชะ-ดา*
3	สาม	*สาม*	ค	*คอ-ควาย*	ฏ	*ตอ-ปะ-ตัก*
4	สี่	*สี่*	ฅ	*คอ-คน*	ฐ	*ถอ-สัน-ถาน*
5	ห้า	*ห้า*	ง	*งอ-งู*	ฑ	*ทอ-มน-โท*
6	หก	*หก*	จ	*จอ-จาน*	ฒ	*ทอ-ผู้-เท่า*
7	เจ็ด	*เจ็ด*	ฉ	*ฉอ-ฉิ่ง*	ท	*ทอ-ทะ-หาน*
8	แปด	*แปด*	ช	*ชอ-ช้าง*	บ	*บอ-ใบ-ไม้*
9	เก้า	*เก้า*	ซ	*ซอ-โซ่*	ภ	*พอ-สำ-เพา*
10	สิบ	*สิบ*	ฌ	*ชอ-เชอ*	ศ	*สอ-สา-ลา*
11	ศูนย์	*ศูน*	ญ	*ยอ-หญิง*	ษ	*สอ-รือ-สี*
12	ฤ	*รึ*	ฌ	*นอ-เนน*	ฬ	*ลอ-จุ-ลา*
13	ลบ	*ลบ*	ด	*ดอ-เด็ก*	ฮ	*ฮอ-นก-ฮูก*
14	ซ้าย	*ซ้าย*	ต	*ตอ-เต่า*	ะ	*สะ-หระ-อะ*
15	ขวา	*ขวา*	ถ	*ถอ-ถุง*	า	*สะ-หระ-อา*
16	บน	*บน*	ธ	*ทอ-ทง*	ิ	*สะ-หระ-อิ*
17	ล่าง	*ล่าง*	น	*นอ-หนู*	ี	*สะ-หระ-อี*
18	วรรค	*วัก*	ป	*ปอ-ปลา*	ึ	*สะ-หระ-อึ*
19	เติม	*เติม*	ผ	*ผอ-ผึ้ง*	ื	*สะ-หระ-อือ*
20	หน้า	*หน้า*	ฝ	*ฝอ-ฝา*	ุ	*สะ-หระ-อุ*
21	หลัง	*หลัง*	พ	*พอ-พาน*	ู	*สะ-หระ-อู*
22	จุด	*จุด*	ฟ	*ฟอ-ฟัน*	เ	*สะ-หระ-เอ*
23	ทับ	*ทับ*	ม	*มอ-ม้า*	แ	*สะ-หระ-แอ*
24	โหมด	*โหมด*	ย	*ยอ-ยักษ์*	ไ	*ไม้-มะ-ลาย*
25			ร	*รอ-เรือ*	์	*ไม้-จัด-ตะ-วา*
26			ล	*ลอ-ลิง*	็	*ไม้-ไต่-คู้*
27			ว	*วอ-แหวน*	ั	*ไม้-หัน-อา-กาศ*
28			ส	*สอ-เสือ*	โ	*สะ-หระ-โอ*
29			ห	*หอ-หีบ*	ำ	*สะ-หระ-อำ*
30			อ	*ออ-อ่าง*	ๆ	*ไม้-ยะ-มก*
31			ไ	*ไม้-ม้วน*	ฯ	*ไป-ยาน-น้อย*
32			๋	*ไม้-เอก*	บรรทัดใหม่	*บัน-ทัด-ใหม่*
33			้	*ไม้-โท*		
34			๊	*ไม้-ตรี*		
35			์	*กา-รัน*		
36			ยกเลิก	*ยก-เลิก*		

5.2.4 การแบ่งกลุ่มคำสำหรับการพิมพ์เอกสารด้วยเสียงพูด

นอกจากคำจะถูกแบ่งตามจำนวนพยางค์เพื่อใช้ในการฝึกฝนและรู้จำเสียงแล้วยังสามารถแบ่งตามการใช้งาน (การพิมพ์เอกสาร) ได้ ซึ่งบางคำอาจจัดอยู่ได้หลายประเภท ทั้งนี้ขึ้นอยู่กับสภาวะการกำหนดโหมดระหว่างการใช้งาน เช่น "หนึ่ง" อาจเป็นได้ทั้งคำสั่ง (สำหรับเลือกรายการคำศัพท์ของพจนานุกรม) และเป็นได้ทั้งตัวอักษร (คืออักขระ "1") เป็นต้น

5.2.4.1 กลุ่มอักขระและพยัญชนะ ได้แก่ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ก, ข, ค, ฅ, ฆ, ง, จ, ฉ, ช, ซ, ฌ, ญ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, ผ, ฝ, พ, ฟ, ภ, ม, ย, ร, ล, ว, ศ, ษ, ส, ห, ฬ, อ, ฮ, ฤ, ฯ, ๆ, จุด, ทับ และ วรรค

5.2.4.2 กลุ่มสระและวรรณยุกต์ ได้แก่ ิ, ี, ึ, ื, ุ, ู, เ-, แ-, ไ-, ่, ้, ๊, ๋, โ-, ำ, ์, ็, ั และ ็

5.2.4.3 กลุ่มคำสั่ง ได้แก่ หนึ่ง, สอง, สาม, สี่, ห้า, หก, เจ็ด, แปด, เก้า, สิบ, เติม, หน้า, หลัง,บน, ล่าง, โหมด, ขวา, ซ้าย, ลบ, บรรทัดใหม่ และ ยกเลิก

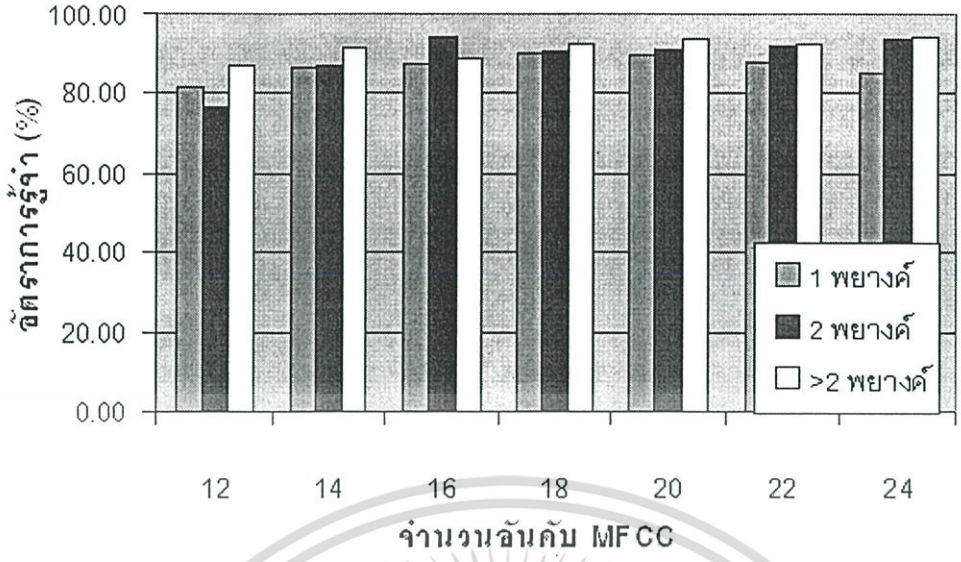
5.2.5 ข้อความสำหรับการทดสอบพิมพ์เอกสารด้วยเสียงพูด

ข้อความที่ใช้ในการทดสอบการพิมพ์เอกสารด้วยเสียงพูดของงานวิจัย คือประวัติสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง [19] ประกอบด้วย 2 ย่อหน้า และมีจำนวนอักขระทั้งสิ้น 631 ตัว (รวมสระ และวรรณยุกต์) ซึ่งมีรายละเอียดดังนี้

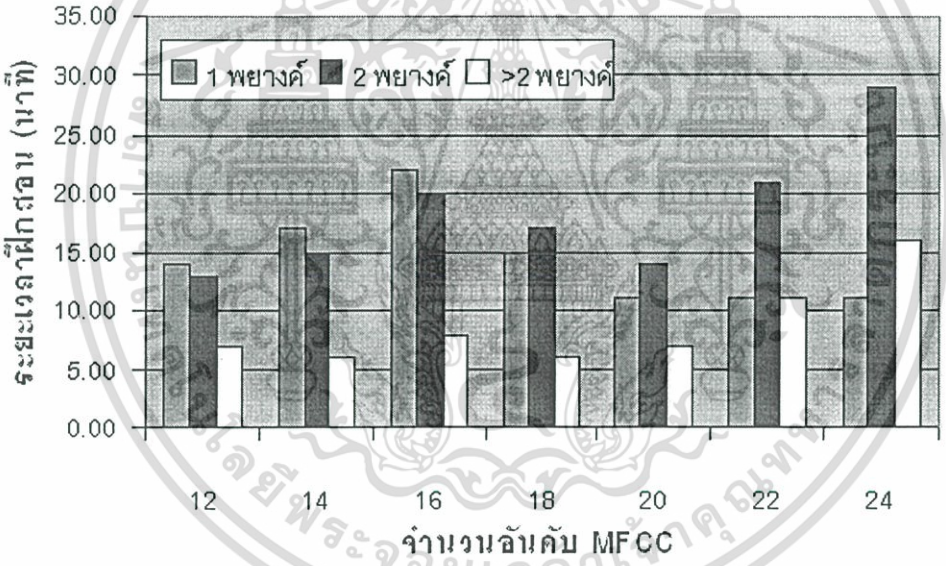
"สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เป็นมหาวิทยาลัยตามพระราชบัญญัติสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง พ.ศ. 2528 โดยมีวัตถุประสงค์เพื่อให้การศึกษา การค้นคว้าและพัฒนาด้านวิทยาศาสตร์ และเทคโนโลยีเพื่อความก้าวหน้าทางอุตสาหกรรม และเศรษฐกิจของประเทศ

เดิมที สถาบันเทคโนโลยีพระจอมเกล้าตั้งขึ้นตามพระราชบัญญัติสถาบันเทคโนโลยีพระจอมเกล้า พ.ศ. 2514 ด้วยการรวมวิทยาลัยโทรคมนาคมนนทบุรี วิทยาลัยเทคนิคพรนครเหนือและวิทยาลัยเทคนิคธนบุรีเข้าด้วยกัน โดยแต่ละแห่งมีฐานะเป็นวิทยาเขต วิทยาลัยโทรคมนาคมนนทบุรี เป็นสถาบันเทคโนโลยีพระจอมเกล้าวิทยาเขตนนทบุรี และในปีเดียวกันนั้นได้ย้ายไปที่อำเภอลาดกระบัง เป็นวิทยาเขตเจ้าคุณทหารลาดกระบัง"

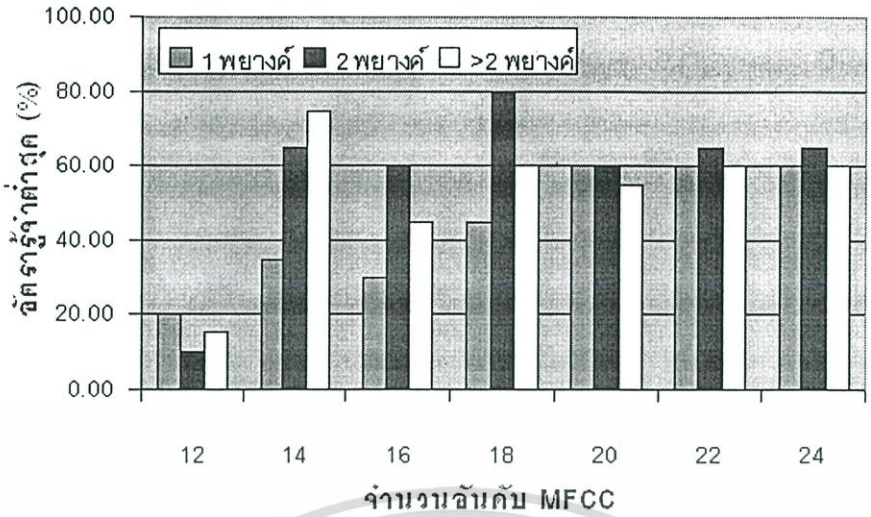
จากการทดลองหาจำนวนลำดับของเซปสตรัม MFCC ที่เหมาะสม โดยการทดสอบจากอัตราการรู้จำด้วยค่าลำดับตั้งแต่ 12-24 สามารถสรุปผลการทดลองได้ดังรูปที่ 5.1, 5.2 และ 5.3



รูปที่ 5.1 อัตราการรู้จำจากการทดลอง MFCC 12-24 ลำดับ



รูปที่ 5.2 ระยะเวลาที่ใช้ในการฝึกฝนของการทดลอง MFCC 12-24 ลำดับ



รูปที่ 5.3 อัตราการรู้จำต่ำสุดของการทดลอง MFCC 12-24 ลำดับ

หากพิจารณารูปที่ 5.1, 5.2 และ 5.3 จะพบว่าอัตราการรู้จำของ MFCC ในเซปสตรัมในแต่ละลำดับไม่แตกต่างกันมากนัก แต่เมื่อพิจารณาระยะเวลาในการฝึกฝนจะพบว่าเซปสตรัมที่มีลำดับ 12 และ 20 จะใช้ระยะเวลาในการรู้จำที่น้อยกว่า และเมื่อพิจารณาในเรื่องของอัตราการรู้จำโดยรวม (วิเคราะห์ทั้งกลุ่มคำ 1, 2 และ มากกว่า 2 พยางค์) พบว่าเซปสตรัมขนาด 20, 22 และ 24 ลำดับจะมีค่าอัตราการรู้จำต่ำสุดที่สูง ดังนั้นจึงเลือกใช้เซปสตรัมขนาด 20 ลำดับสำหรับการทดลองในงานวิจัยนี้

5.3 การทดสอบระบบรู้จำเสียงพูดแบบแยกวิเคราะห์ตามจำนวนพยางค์

ตารางที่ 5.3 อัตราการรู้จำของกลุ่มผู้ทดสอบแบบแยกวิเคราะห์ตามจำนวนพยางค์

ลำดับผู้ทดสอบ	อัตราการรู้จำ (%)			
	คำ 1 พยางค์	คำ 2 พยางค์	คำ >2 พยางค์	ค่าเฉลี่ย
1	89.58	90.97	93.91	91.49
2	85.63	93.61	94.06	91.10
3	95.21	91.67	94.22	93.70
4	84.38	84.17	85.47	84.67
5	92.71	93.06	93.13	92.97
6	90.42	91.94	87.97	90.11
7	86.46	86.67	87.34	86.82
8	89.58	89.44	87.57	88.86
9	94.01	88.75	91.56	91.44
10	89.79	87.64	87.19	88.21

5.4 การทดสอบพิมพ์เอกสารด้วยเสียงพูด

การทดสอบพิมพ์เอกสารจะเป็นการประยุกต์ใช้ฐานข้อมูลพจนานุกรมภาษาไทย เพื่อช่วยให้การพิมพ์เอกสารเป็นไปได้อย่างสะดวกรวดเร็วมากยิ่งขึ้น โดยใช้ข้อความชุดเดียวกันดังแสดงในตารางที่ 5.2 ซึ่งสามารถแสดงอัตราการพิมพ์และเวลาที่ใช้ในการพิมพ์ดังตารางที่ 5.4

ตารางที่ 5.4 ผลการทดสอบพิมพ์เอกสารของกลุ่มผู้ทดสอบโดยใช้เสียงสั่งงาน

ลำดับผู้ทดสอบ	จำนวนอักษรที่ใช้ทดสอบ (ตัว)	เวลาที่ใช้ในการพิมพ์ (นาที)	อัตราการพิมพ์ (ตัว/นาที)
1	631	56	11.27
2	631	51	12.37
3	631	46	13.71
4	631	72	8.76
5	631	58	10.87
6	631	67	9.47
7	631	109	5.78
8	631	92	6.89
9	631	75	8.41
10	631	88	7.17
เฉลี่ย	631	71.4	8.83

ตารางที่ 5.5 ผลการทดสอบพิมพ์เอกสารของกลุ่มผู้ทดสอบโดยใช้เสียงสั่งงานร่วมกับเมาส์

ลำดับผู้ทดสอบ	จำนวนอักษรที่ใช้ทดสอบ (ตัว)	เวลาที่ใช้ในการพิมพ์ (นาที)	อัตราการพิมพ์ (ตัว/นาที)
1	631	42	15.20
2	631	48	13.15
3	631	41	15.39
4	631	61	10.34
5	631	57	11.07
6	631	56	11.26
7	631	90	7.01
8	631	82	7.70
9	631	64	9.85
10	631	69	9.14
เฉลี่ย	631	61	10.34

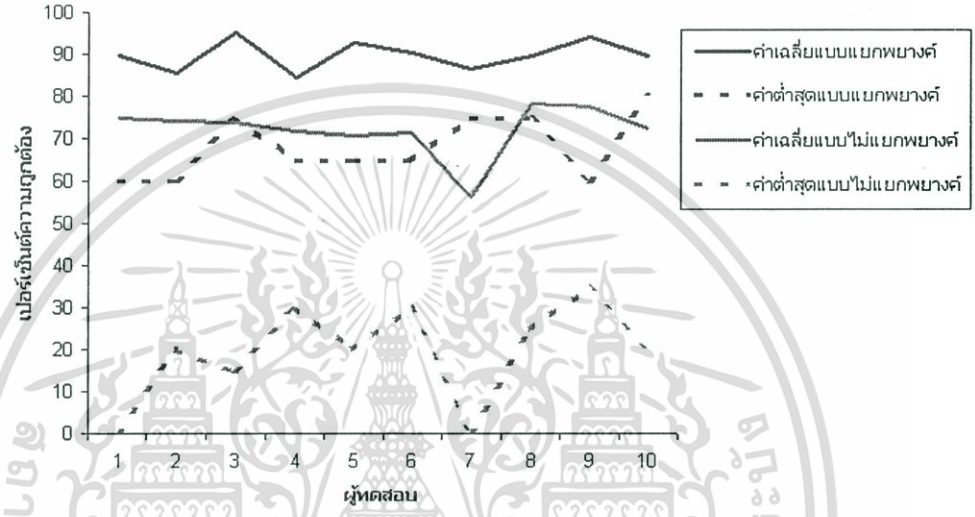
5.5 การทดสอบระบบรู้จำเสียงพูดแบบไม่แยกวิเคราะห์ตามจำนวนพยางค์

การทดสอบระบบรู้จำเสียงพูดแบบไม่วิเคราะห์ตามจำนวนพยางค์จะใช้ข้อมูลชุดเดียวกันกับที่ใช้ในการทดสอบด้วยการแยกวิเคราะห์ตามจำนวนพยางค์ หากแต่ใช้โครงข่ายประสาทเทียมเพียงชุดเดียวในการวิเคราะห์คำทั้ง 92 คำ เมื่อนำมาจัดแบ่งความสามารถในการรู้จำตามจำนวนพยางค์จะได้ผลดังแสดงในตารางที่ 5.5

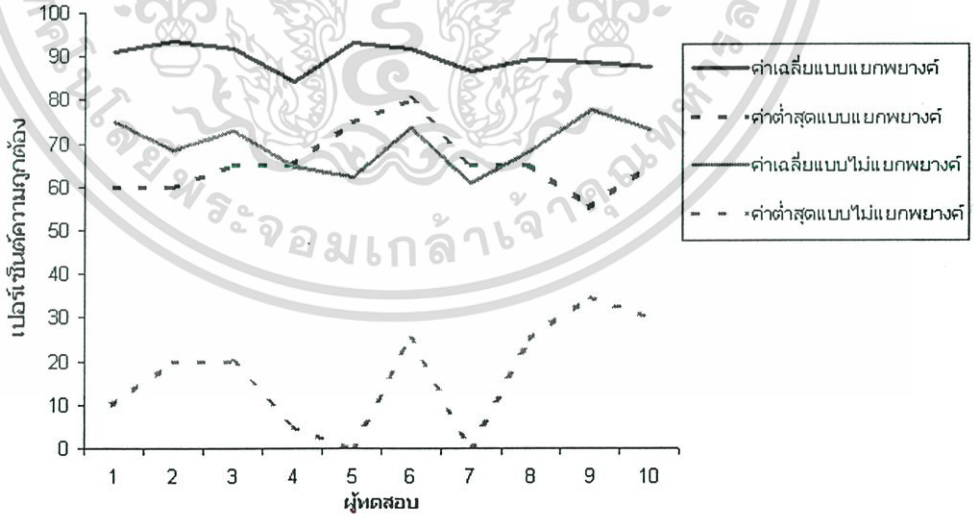
ตารางที่ 5.6 อัตราการรู้จำของกลุ่มผู้ทดสอบแบบไม่แยกวิเคราะห์ตามจำนวนพยางค์

ลำดับผู้ทดสอบ	อัตราการรู้จำ (%)			
	คำ 1 พยางค์	คำ 2 พยางค์	คำ >2 พยางค์	ค่าเฉลี่ย
1	74.79	75.00	79.69	76.49
2	74.17	68.33	81.56	74.69
3	73.96	72.92	76.56	74.48
4	71.67	64.86	77.66	71.39
5	70.63	62.50	81.25	71.46
6	71.25	73.47	70.63	71.78
7	56.04	60.83	74.38	63.75
8	78.13	68.33	78.59	75.02
9	77.50	77.78	84.06	79.78
10	72.29	73.19	76.41	73.96

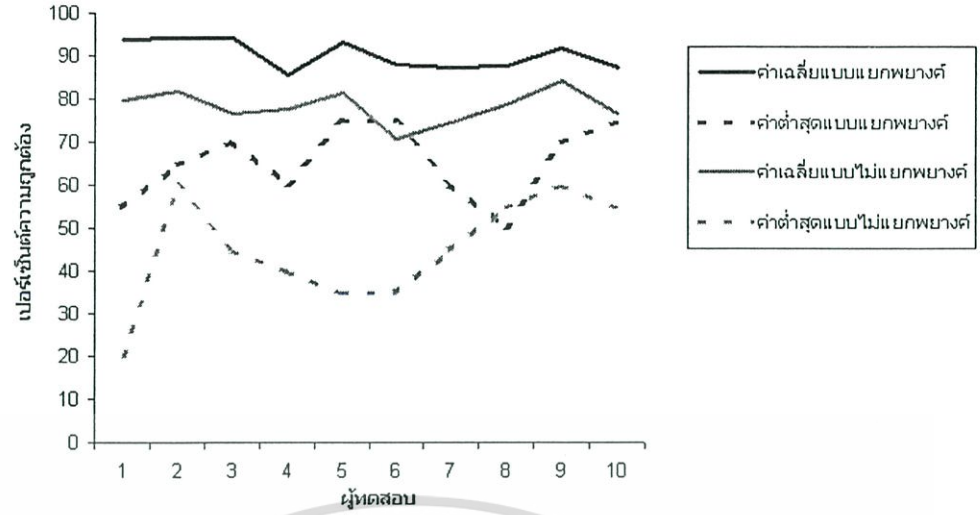
2. เพิ่มความถูกต้องของระบบการรู้จำเสียง การจัดกลุ่มในการรู้จำเสียงพูดจะทำให้จำนวนคำที่ใช้ในการรู้จำในแต่ละกลุ่มมีน้อยลง จึงช่วยลดความยุ่งยากในการจัดแบ่งกลุ่ม (Classification) ซึ่งจากรูปที่ 6.2-6.4 จะพบว่าผลของอัตราการรู้จำเฉลี่ยของแบบแยกวิเคราะห์ตามจำนวนพยางค์มีค่าสูงกว่าแบบไม่แยกวิเคราะห์ตามจำนวนพยางค์อยู่ในช่วงประมาณ 10-30% และเมื่อวิเคราะห์ในส่วนของอัตราการรู้จำต่ำสุดจะพบว่าผลการรู้จำแบบแยกวิเคราะห์ตามจำนวนพยางค์มีค่าการรู้จำสูงกว่าแบบไม่แยกพยางค์มาก



รูปที่ 6.2 เปรียบเทียบอัตรารู้จำคำ 1 พยางค์

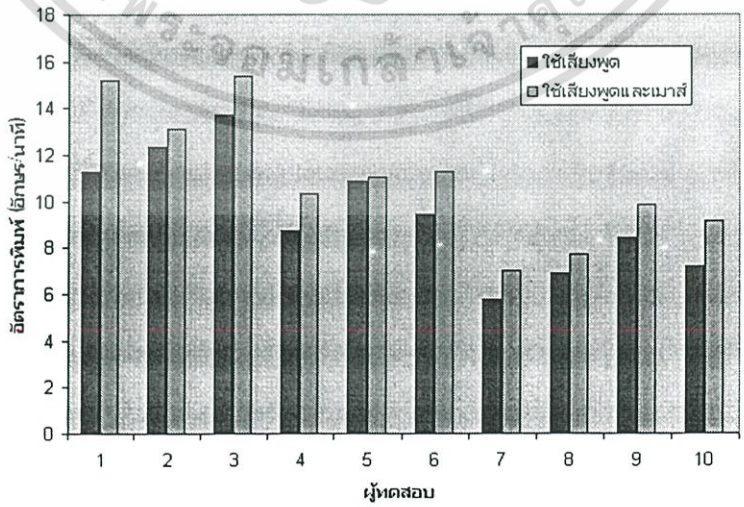


รูปที่ 6.3 เปรียบเทียบอัตรารู้จำคำ 2 พยางค์



รูปที่ 6.4 เปรียบเทียบอัตรารู้จำคำมากกว่า 2 พยางค์

การประยุกต์ใช้งานระบบการรู้จำเสียงพูดกับงานพิมพ์เอกสารนั้นพบว่ามีอัตราการพิมพ์สูงสุดอยู่ที่ประมาณ 14 อักษรต่อนาที โดยหากใช้อุปกรณ์เมาส์เลือกรายการในช่องแสดงคำศัพท์ของพจนานุกรมจะเพิ่มความสามารถในการพิมพ์ได้สูงสุดที่ประมาณ 16 อักษรต่อนาที ดังแสดงในรูปที่ 6.5 เนื่องจากคำศัพท์ในฐานข้อมูลที่ใช้นั้นนำมาจากพจนานุกรมฉบับราชบัณฑิตยสถาน ดังนั้นคำศัพท์ที่มีอยู่จึงอาจไม่เหมาะสมกับการใช้ในงานพิมพ์มากเท่าที่ควร จึงอาจเพิ่มความสามารถในการพิมพ์โดยเพิ่มคำศัพท์ที่นิยมใช้เข้าไปในฐานข้อมูล และลบคำศัพท์ที่สั้นๆ ทั้งนี้เพราะคำศัพท์ที่สั้นๆ เช่น 2-3 อักษรนั้น สามารถที่จะพิมพ์ลงไปได้โดยไม่มีความจำเป็นต้องเลือกจากช่องรายการคำศัพท์ ซึ่งเมื่อตัดคำศัพท์เหล่านี้ออกไปจะทำให้รายการคำศัพท์ของพจนานุกรมสามารถแสดงคำศัพท์ที่ยาวกว่าได้มากขึ้น จึงอาจช่วยเพิ่มอัตราการพิมพ์ได้



รูปที่ 6.5 เปรียบเทียบอัตราการพิมพ์ข้อความด้วยเสียงพูดและใช้เสียงพูด+เมาส์

เอกสารอ้างอิง

[1] Claudio Becchetti and Lucio P. Ricotti. SPEECH RECOGNITION. New York : John Wiley & Sons. 1999.

[2] Britton C. Rorabaugh. DSP Primer. New York : McGraw-Hill, Inc. 1999.

[3] เสาวลักษณ์ อารีย์พงศา. **การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธี ฮิดเดน มาร์คอฟ โมเดล และเวกเตอร์ควอนไตซ์เซชั่น.** วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย. 2538.

[4] สมศักดิ์ ชุ่มช่วย. **การประมวลสัญญาณเชิงเลขเบื้องต้น.** กรุงเทพมหานคร : ภาควิชาอิเล็กทรอนิกส์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2545.

[5] เอกฤทธิ์ มณีน้อย. **การรู้จำเสียงสระภาษาไทยโดยใช้โครงข่ายประสาทเทียม.** วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย. 2541.

[6] Douglas O'Shaughnessy. SPEECH COMMUNICATION - Human and Machine. The Institute of Electrical and Electronics Engineers, Inc., 2000. Pp.387-389

[7] Ken Steiglitz. A Digital Signal Processing Primer. New York : Addison-Wesley. 1996.

[8] John G. Proakis. Digital Signal Processing. New York : Prentice-Hall, Inc. 1996.

[9] ชัย วุฒิวิวัฒน์ชัย, สุทัศน์ แซ่ตั้ง และวารินทร์ อัจฉริยะกุลพร. **ความก้าวหน้าของการพัฒนาระบบระบุผู้พูดภาษาไทย.** NECTEC Technical Journal, Vol. II, No. 7, 2543. หน้า 496-510.

[10] Jia-Ching Wang, Jhing-Fa Wang and Yu-Sheng Weng. Chip Design of Mel Frequency Cepstral Coefficients For Speech Recognition. The Institute of Electrical and Electronics Engineers, Inc., 2000, pp 3658-3661

[11] Sirko Molau, Michael Pitz, Ralf Schluter and Hermann Ney. Computing Mel-Frequency Cepstral Coefficients On The Power Spectrum. The Institute of Electrical and Electronics Engineers, Inc., 2001, pp 73-76.

[12] Lawrence and Biing-hwang Juang. Fundamentals Of Speech Recognition. New York : Prentice-Hall, Inc. 1993.

[13] Tebelskins J. **Speech Recognition using Neural Networks.** Carnegie Mellon University. 1995.

[14] Jiawei Han and Micheline Kamber. **Data Mining: Concepts and Techniques.** Mogan Kaufmann Publishers. 2001.

[15] Haykin S. **Neural Networks.** Macmillan College Publishing Company. 1994.

[16] Jack M Zurade. **Introduction to Artificial Neural System.** West Publishing Company. 1992.

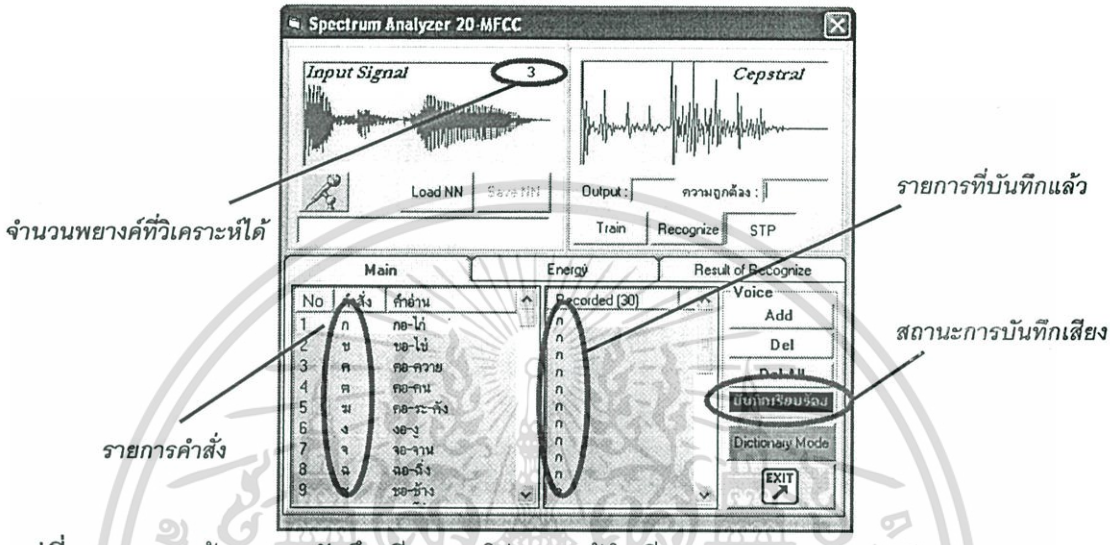
[17] เสรี ปานซาง. **การรู้จำเสียงพูดคำไทยแบบไม่ขึ้นกับผู้พูดด้วยนิวรัลเน็ตเวอร์ค.** วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2540.

[18] สมพงศ์ วิทยศักดิ์พันธุ์, รังรอง นิลประภัสสร, เยาวลักษณ์ กระแสร์สินธุ์ และ อัมพร แก้วสุวรรณ. **แบบเรียนภาษาไทยเบื้องต้นในบริบทไทยศึกษาสำหรับชาวต่างชาติ.** กรุงเทพมหานคร : โครงการพัฒนาความร่วมมือด้านการเรียนการสอนภาษาไทยบนฐานของไทยคดีศึกษา, ทบวงมหาวิทยาลัย. 2545.

[19] บัณฑิตวิทยาลัย. **คู่มือบัณฑิตศึกษา สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.** กรุงเทพมหานคร : สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. ม.ป.ป.

ภาคผนวก ก
การใช้งานโปรแกรมรู้จำเสียงพูด

1. การบันทึกข้อมูลเสียง



รูปที่ ก.1 ภาพหน้าจอการบันทึกเสียงของโปรแกรมรู้จำเสียงพูด

การบันทึกข้อมูลเสียงพูดจะถูกจัดแบ่งหมวดตามจำนวนพยางค์ ซึ่งมีทั้งหมด 3 กลุ่มคือ 1 พยางค์, 2 พยางค์ และมากกว่า 2 พยางค์ ซึ่งสามารถแสดงหน้าจอการทำงานได้ดังรูปที่ ก.1 โดยมีขั้นตอนการบันทึกเสียงดังนี้

1. เลือกคำที่ต้องการบันทึกเสียงโดยการใช้เมาส์คลิกเลือกคำสั่ง (คำศัพท์) ที่ช่อง ***รายการคำสั่ง*** ซึ่งจะมีคำอ่านกำกับไว้ท้ายคำสั่งเสียงเหล่านั้น

2. คลิกปุ่ม Add เพื่อเริ่มบันทึกเสียง และเมื่อบันทึกเสียงเสร็จในแต่ละคำจะมีสถานะบอกให้ทราบว่าผลการบันทึกเป็นอย่างไร เช่น บันทึกเรียบร้อย หรือ ไม่ได้บันทึก (เนื่องจากจำนวนพยางค์ของเสียงที่พูดไม่ตรงกับหมวดพยางค์ของคำนั้นๆ)

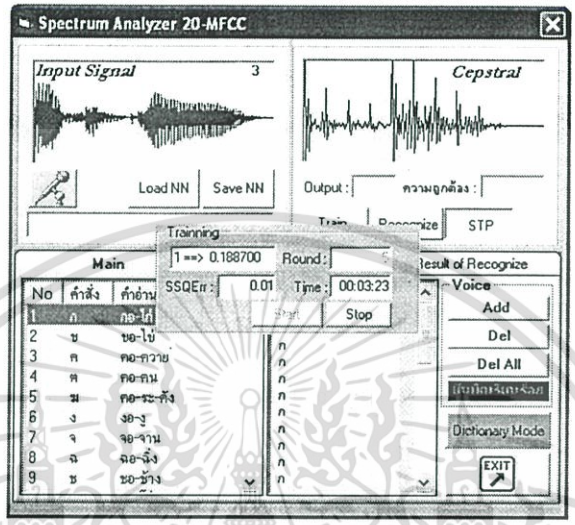
3. เมื่อต้องการบันทึกคำศัพท์อื่นๆ ทำได้โดยการเลือกรายการในช่อง ***รายการคำสั่ง*** ตามที่ต้องการ

4. เมื่อต้องการลบรายการเสียงที่ได้บันทึกไปแล้ว ให้เลือกรายการเสียงที่ต้องการลบจากในช่อง *รายการที่บันทึกแล้ว* แล้วคลิกปุ่ม *Del* เมื่อลบเฉพาะตัวอย่างเสียงนั้นเพียงเสียงเดียว หรือคลิกปุ่ม *Dell All* เพื่อลบเสียงทั้งหมด

2. การฝึกฝนโครงข่ายประสาทเทียม

หลังจากบันทึกตัวอย่างเสียงที่จะใช้ในการฝึกฝนเสร็จเรียบร้อยแล้ว สามารถทำการฝึกฝนโครงข่ายประสาทเทียมได้ดังขั้นตอนต่อไปนี้

1. คลิกปุ่ม *Train* จะปรากฏกรอบเล็กดังแสดงในรูปที่ ก.2



รูปที่ ก.2 การฝึกฝนโครงข่ายประสาทเทียม

2. กำหนดค่า SSQErr แล้วคลิกที่ปุ่ม *Start* จากนั้นจะเริ่มการฝึกฝนโดยจากโครงข่ายประสาทเทียมของชุด 1, 2 และมากกว่า 2 พยางค์ตามลำดับ

3. ระหว่างการฝึกฝนจะแสดงจำนวนรอบและเวลาที่ใช้ในช่อง Round และ Time ตามลำดับ ส่วน Textbox ซึ่งอยู่ด้านซ้ายของ Round จะแสดงค่าผิดพลาดของการรู้จำของการฝึกฝนในแต่ละรอบ

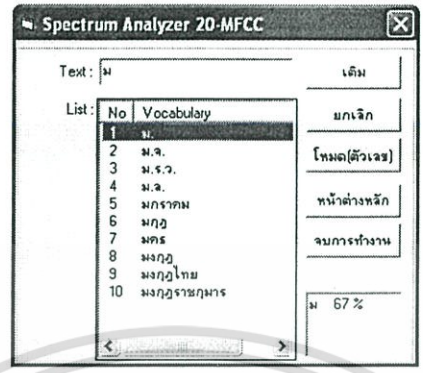
4. เมื่อค่าผิดพลาดของการรู้จำมีค่าต่ำกว่าหรือเท่ากับ ค่า SSQErr โครงข่ายประสาทเทียมจะสิ้นสุดการฝึกฝนในชุดนั้นๆ และจะฝึกฝนโครงข่ายประสาทเทียมชุดต่อไปจนกว่าจะครบทั้ง 3 ชุด

5. เมื่อฝึกฝนเสร็จสามารถบันทึกค่าน้ำหนัก (Weight) ด้วยการคลิกที่ปุ่ม *Save NN*

3. การทดสอบรู้จำเสียง

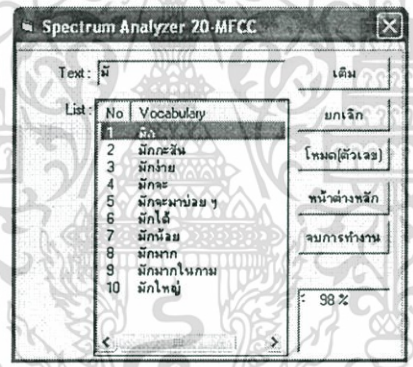
1. คลิกที่ปุ่ม *Recognize* โปรแกรมจะทำการรอรับข้อมูลเสียง หากต้องการใช้งานร่วมกับพจนานุกรมสามารถทำได้โดยคลิกที่ปุ่ม *Dictionary Mode* โปรแกรมจะเปลี่ยนหน้าจอเป็นโหมดพจนานุกรม

2. สมมติต้องพิมพ์คำว่า "มั่นคั่ง" สามารถทำได้โดยการพูดออกเสียง "มอ-ม้า" ช่อง *List* จะเปลี่ยนไปดังรูปที่ ก.3



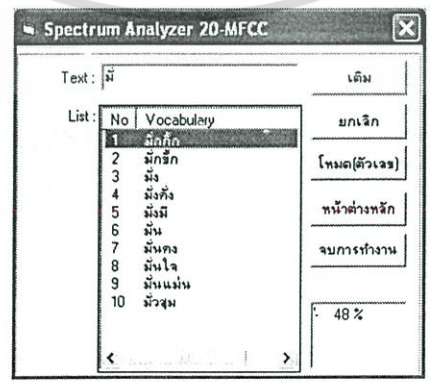
รูปที่ ก.3 ภาพการรู้จำ "มอ-ม้า" ในโหมดพจนานุกรม

3. พูดออกเสียง "ไม้-หัน-อา-กาด" ช่อง *List* จะเปลี่ยนไปดังรูปที่ ก.4



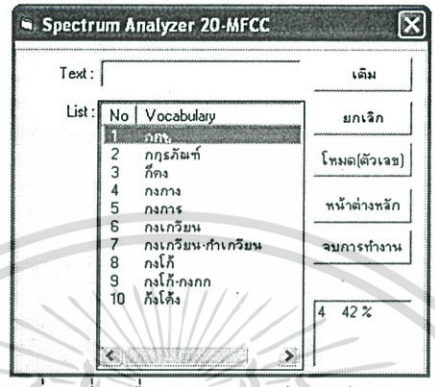
รูปที่ ก.4 ภาพการรู้จำ "ไม้-หัน-อา-กาด"

4. พูดออกเสียง "ไม้-เอก" ช่อง *List* จะเปลี่ยนไปดังรูปที่ ก.5



รูปที่ ก.5 ภาพการรู้จำ "ไม้-เอก"

5. สามารถพูดออกเสียงสะกดไปเรื่อยจนครบแล้ว ออกเสียง "เติม" เพื่อส่งข้อความไปยัง MS-Word หรือสามารถออกเสียง "สี่" เพื่อเลือกรายการที่ 4 นั่นคือคำว่า "มั่นคั่ง" จะถูกส่งไปยัง MS-Word ได้เช่นกัน จากนั้นช่องข้อความ Text จะว่างเพื่อรอรับคำสั่งเสียงเพื่อพิมพ์คำใหม่ดังแสดงในรูปที่ ก.6



รูปที่ ก.6 ภาพหน้าจอหลังใช้คำสั่ง "สี่" เพื่อส่งข้อความไปยังไมโครซอฟต์เวิร์ด

ภาคผนวก ข
สเกลความถี่หน้าต่างเมล

ตารางที่ ข.1 สเกลความถี่หน้าต่างสามเหลี่ยมของเมลแบบ 24 หน้าต่าง

Triangle No.	Original Frequency			Mel Frequency
	Min	Center	Max	Center
1	0	65	140	100
2	65	140	215	200
3	140	215	300	300
4	215	300	390	400
5	300	390	490	500
6	390	490	605	600
7	490	605	725	700
8	605	725	855	800
9	725	855	1000	900
10	855	1000	1160	1000
11	1000	1160	1330	1100
12	1160	1330	1520	1200
13	1330	1520	1725	1300
14	1520	1725	1950	1400
15	1725	1950	2195	1500
16	1950	2195	2465	1600
17	2195	2465	2760	1700
18	2465	2760	3080	1800
19	2760	3080	3430	1900
20	3080	3430	3815	2000
21	3430	3815	4230	2100
22	3815	4230	4690	2200
23	4230	4690	5190	2300
24	4690	5190	5735	2400

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายเอกรินทร์ แซ่เฮ้ง
วัน เดือน ปีเกิด	21 เมษายน 2520 ที่จังหวัดอุบลราชธานี
ที่อยู่	2100/1014 หมู่บ้านพนาสิน2 ถ.รามคำแหง แขวงหัวหมาก เขตบางกะปิ กรุงเทพฯ 10240 โทร. 0-2718-9819, 0-2718-9657
ประวัติการศึกษา	2540 ประกาศนียบัตรวิชาชีพชั้นสูง สาขาวิชาอิเล็กทรอนิกส์ (คอมพิวเตอร์) วิทยาลัยเทคนิคอุบลราชธานี
	2543 วิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีอุตสาหกรรม-อิเล็กทรอนิกส์ (เกียรตินิยมอันดับ 2) สถาบันราชภัฏอุบลราชธานี
ประสบการณ์การทำงานและผลงานวิจัย	
พ.ศ. 2541-2543	อาจารย์พิเศษสถาบันอบรมคอมพิวเตอร์ของเอกชน
พ.ศ. 2543-ปัจจุบัน	ตำแหน่งนักวิเคราะห์และพัฒนาระบบสารสนเทศภายในองค์กรของ บริษัท โกลเด้นไทย อินดรัสทรียส์ จำกัด
ปัจจุบัน	โปรแกรมเมอร์อิสระ