การรู้จำเสียงพูดภาษาไทยระยะที่หนึ่ง : การรู้จำเสียงพูดคำไทยโดด ๆ โดยไม่ขึ้นกับผู้พูด

โดย

สมชาย จิตะพันธ์กุล

โครงการวิจัยเลขที่ 45G-EE-2538

ทุนงบประมาณแผ่นดิน

ปี 2538

สถาบันวิจัยและพัฒนาของคณะวิศวกรรมศาสตร์

คณะวิศวกรรมศาสตร์

จุฬาลงกรณ์มหาวิทยาลัย

กรุงเทพฯ

กันยายน 2540

สถาบันวิจัยและพัฒนาของคณะวิศวกรรมศาสตร์ไม่รับผิดชอบต่อผลเสียใด ๆ อันอาจเกิดจากการนำความคิดเห็นในเอกสารฉบับนี้ไปใช้ ความคิดเห็นที่ปรากฏในเอกสารเป็นความคิดเห็นของผู้เขียนซึ่งไม่จำเป็นต้องเป็นความคิดเห็นของสถาบันฯ

การรู้จำเสียงพูดภาษาไทย ระยะที่หนึ่ง : การรู้จำเสียงพูดคำไทยโดดๆโดยไม่ขึ้นกับผู้พูด

โดย

รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล

วศ.บ., วศ.ม. (วิศวกรรมไฟฟ้า จุฬาลงกรณ์มหาวิทยาลัย)

D.E.A., Dr. Ing. SIGNAUX ET SYSTEMES SPATIO TEMPORELS

มหาวิทยาลัย AIX-MARSEILLE ประเทศฝรั่งเศส

โครงการวิจัยเลขที่ 45G-EE-2538

ทุนงบประมาณแผ่นดิน

ปี 2538

สถาบันวิจัยและพัฒนาของคณะวิศวกรรมศาสตร์

คณะวิศวกรรมศาสตร์

จุฬาลงกรณ์มหาวิทยาลัย

กรุงเทพฯ

กันยายน 2540

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ในการศึกษาและเลือกกรรมวิธีที่มีประสิทธิภาพสูงสุดในการรู้จำเสียงพูดตัวเลขไทย ระหว่าง กรรมวิธีไดนามิก ไทม์วาร์ปิง (DTW) กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ (HMM) และกรรมวิธีนิวรอลเน็ตเวิร์ก (NN) ทั้งสามกรรมวิธีประกอบด้วยกระบวนการ 4 ขั้นตอนคือ การประมวลเบื้องต้น การวัดหาค่าลักษณะสำคัญ การจำแนกรูปแบบ และการตัดสินใจ

ในการประมวลเบื้องต้น กรรมวิธีย่อยในการหาจุดสิ้นสุดเสียงพูดเป็นเรื่องหลักที่ได้รับการพิจารณา ทั้งนี้รายละเอียดของกรรมวิธีย่อยนี้จะแตกต่างกันไปในแต่ละกรรมวิธีหลักข้างต้น แต่ทั้งหมดใช้หลักการพิจารณาระดับพลังงานของสัญญาณ

ในการวัดหาค่าลักษณะสำคัญ DTW ใช้ผลการแปลงฮาร์ตเลย์คำนวณหาค่าพารามิเตอร์ ในขณะที่ HMM ใช้การหาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น ลำดับ 10 ร่วมกับการควอนไตซ์เวกเตอร์ของรหัสขนาด 64 เพื่อคำนวณค่าพารามิเตอร์ สำหรับ NN ใช้การหาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น ลำดับ 10 เท่านั้นในการกำหนดค่าพารามิเตอร์

ในขั้นตอนของการจำแนกรูปแบบ DTW ใช้วิธีการไดนามิก ไทม์วาร์ปิง เพื่อกำหนดรูปแบบ ส่วน HMM ใช้แบบจำลองฮิดเดน มาร์คอฟ จำนวน 3 สถานะ เพื่อคำนวณหารูปแบบ โดยที่ NN ใช้อัลกอริทึมแบบแบคพรอพาเกชันเพื่อหารูปแบบที่เหมาะสม

ในขั้นตอนการตัดสินใจ DTW ใช้เงื่อนไข Nearest Neighbor กับค่าความคลาดเคลื่อนที่ได้จากการเปรียบเทียบรูปแบบทดสอบกับรูปแบบอ้างอิง ในขณะที่ HMM ใช้อัลกอริทึม Viterbi ในการตัดสินใจซึ่งเป็นกระบวนการที่ซับซ้อนที่สุด และ NN ใช้กรรมวิธีที่ธรรมดาที่สุด กล่าวคือใช้เงื่อนไขความคลาดเคลื่อนต่ำสุดในการเปรียบเทียบ

ในการทดสอบและเปรียบเทียบการทำงานของระบบการรู้จำ มีการจัดเตรียมข้อสนเทศ 3 ชุด ชุดแรกเป็นชุดฝึกฝน ชุดที่ 2 และ 3 เป็นชุดทดสอบ 1 และ 2 ตามลำดับ แต่ละชุดเป็นข้อมูลเสียงที่บันทึกจากผู้พูดทั้งเพศหญิงและชายที่มีอายุอยู่ในช่วง 18 ถึง 25 ปี ข้อสนเทศชุดฝึกฝนและชุดทดสอบ 1 เป็นข้อมูลที่บันทึกจากผู้พูดกลุ่มเดียวกัน แต่บันทึกข้อมูลไว้คนละชุด ส่วนข้อสนเทศชุดทดสอบ 2 เป็นข้อมูลที่บันทึกจากผู้พูดต่างกลุ่มออกไป จำนวนตัวอย่างในแต่ละกลุ่มเรียงตามลำดับคือ 20, 20, และ 20 สำหรับ DTW 45, 45, และ 10 สำหรับ HMM และ 30, 30, และ 12 สำหรับ NN อัตราการรู้จำเฉลี่ยของแต่ละกลุ่มที่ได้เรียงตามลำดับคือ ร้อยละ 90.50, 86.50, และ 79.25 สำหรับ DTW ที่ใช้รูปแบบอ้างอิงของ 20 ตัวอย่าง; ร้อยละ 95.30, 89.70, และ 84.00 สำหรับ HMM ที่ใช้รูปแบบอ้างอิงของ 45 ตัวอย่าง; และร้อยละ 98.20, 84.30, และ 89.40 สำหรับ NN ที่ใช้รูปแบบอ้างอิงของ 30 ตัวอย่าง

ABSTRACT

This research has the objective to study and select an efficient algorithm for speaker independent Thai numeral word recognition among the Dynamic Time Warping (DTW), Hidden Markov Model (HMM), and Neural Network (NN). All three methods are composed of 4 steps : Preprocessing, Feature Measurement, Pattern Classification, and Decision Making. The first main consideration is the endpoint detection techniques in preprocessing step that used different details among those three methods but all of them were based on energy level measurement. For feature measurement step, DTW used the discrete Hartley transform to extract required parameters, but HMM used LPC of order 10 in accordance with the vector quantization (VQ) of 64 codebooks to compute its essential features, and NN used also LPC of order 10 to measure its necessary parameters. In pattern classification step, DTW used its time warping algorithm to create pattern, and 3 states of hidden Markov model was used to construct pattern in HMM, but the backpropagation algorithm was executed to form the pattern. The Nearest Neighbor condition was set for DTW in decision making step. For HMM, this step is more complicate than another by using the Viterbi algorithm. The most simple criteria for decision should certainly is that of NN by using the minimum error distance.

To test and compare those three methods, the separated speech training set and testing set 1 and 2 were composed of both male and female speakers within the range of 18 to 25 years of age. The training set and testing set 1 were the same speaker group but different data. The testing set 2 was another speaker group. The number of each set was varied between those methods : 20, 20, and 20 for DTW; 45, 45, and 10 for HMM; and 30, 30, and 12 for NN, respectively. The average recognition rates of each set were : 90.50%, 86.50%, and 79.25% for DTW with 20 reference 20 samples; 95.30%, 89.70%, and 84.00% for HMM with 45 reference samples; and 98.20%, 84.30%, and 89.40% for NN with 30 reference samples, respectively.

กิตติกรรมประกาศ

งานวิจัยชิ้นนี้สำเร็จลุล่วงไปได้ตามเป้าหมายที่ผู้วิจัยวางไว้ และเกินกว่าที่กำหนดไว้ในโครงการวิจัยที่เสนอขอรับทุนอุดหนุน ก็เนื่องมาจากการมีส่วนร่วมของนิสิตสังกัดห้องปฏิบัติการวิจัยกรรมวิธีสัญญาณดิจิทัลจำนวนมาก ที่สมควรระบุถึงเป็นพิเศษในที่นี้ ได้แก่ ความช่วยเหลือที่ได้รับจาก คุณระพีพัฒน์ เพ็ญศิริ และคุณธีระ ภัทราพรนันท์ ในการศึกษาวิจัยกรรมวิธีไดนามิก ไทม์วาร์ปิง จาก คุณเสาวลักษณ์ อารีย์พงศา และคุณวิศรุต อายุบุตร ในการศึกษาวิจัยกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ จาก คุณวุฒิพงษ์ พรสุขจันทรา และคุณกิตติพงษ์ เจนวีถีสุข ในการศึกษาวิจัยกรรมวิธีนิวรอลเน็ตเวิร์ก คุณอรินทรา เลาหะราชพฤกษ์ คุณนวรัตน์ วรพันธ์ตระกูล และคุณธวัช ปานเสถียรกุล ในการรวบรวมและข้อมูลและเปรียบเทียบผลการทดสอบบางส่วน ซึ่งผู้วิจัยต้องขอขอบคุณเป็นอย่างยิ่ง รวมถึงนิสิตที่สละเวลาเพื่อทำการบันทึกเสียง

คำขอบพระคุณอย่างสูงที่ผู้วิจัยต้องขอมอบให้กับ ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน อันเนื่องมาจากคำแนะนำต่าง ๆ ที่งานวิจัยนี้ได้รับผ่านมาทางวิทยานิพนธ์ของนิสิตที่ระบุนามข้างต้นบางคน ซึ่งเป็นประโยชน์อย่างยิ่งต่อการดำเนินการวิจัยชิ้นนี้

งานวิจัยนี้คงจะไม่สามารถเริ่มต้นและดำเนินการไปจนบรรลุผลได้ ถ้าไม่ได้รับความอนุเคราะห์เรื่องเงินทุนจากงบประมาณแผ่นดิน ปี 2538 ที่ผู้วิจัยได้รับจัดสรรจากทางคณะวิศวกรรมศาสตร์ ซึ่งผู้วิจัยขอขอบคุณไว้ ณ ที่นี้เช่นกัน

สุดท้ายนี้ ผู้มีส่วนให้กำลังใจกับผู้วิจัยตลอดการทำงานวิจัยนี้คือ ครอบครัวของผู้วิจัยเอง ได้แก่ สุมาลี จิตะพันธ์กุล ภรรยา ติลยา จิตะพันธ์กุล และพฤศญา จิตะพันธ์กุล บุตรสาวทั้งสอง ของผู้วิจัยเอง ที่ทำให้งานวิจัยนี้สำเร็จลุล่วงไปด้วยดี

สมชาย จิตะพันธ์กุล

สารบัญ

สารบัญตาราง

ตารางที่	ชื่อตาราง	หน้าที่

สารบัญรูปภาพ

สารบัญคำศัพท์

ขั้นตอนวิธีการ	algorithm
ไม่แปรเปลี่ยนตามเวลาโดยค่าเฉลี่ยเชิงเส้นกำกับ	asymtotically mean stationary
อัตสหสัมพันธ์	autocorrelation
การประมาณพันธะแบบถดถอย	backward prediction
การค้นหาแบบทวิภาค	binary search
เงื่อนไขขอบเขต	boundary condition
จุดศูนย์ถ่วง	centriod
การแบ่งกลุ่ม	clustering
เวกเตอร์รหัส	code vector
ชุดรหัส	codebook
คำต่อเนื่อง	connected word
น้ำหนักการเชื่อมจ่อ	connection weight
เงื่อนไขความต่อเนื่อง	continuity condition
เสียงพูดต่อเนื่อง	continuous speech
กรรมวิธีความแปรปรวนร่วม	covariance method
การบีบอัดข้อมูล	data compression
ดีสครีต	discrete
ความเพี้ยน	distortion
ไดนามิก ไทม์วาร์ปิง	dynamic time warping, DTW
การตรวจวัดหัวท้ายคำ, การตัดหัวท้ายคำ, กรรมวิธีหาจุดสิ้นสุด	end point detection
ลักษณะสำคัญ	feature
การประมาณพันธะแบบก้าวหน้า	forward prediction
การค้นหาทั่วทั้งหมด	full search
ค่าที่เหมาะสมที่สุดที่ครอบคลุมทั้งหมด	global optimum
ผลการแปลงฮาร์ตเลย์	Hartley transform
อุปนัย	induction
ผลคูณภายใน	inner product
รูปแบบอินพุตเอาต์พุต	input output pattern
คำโดด	isolated word
วนซ้ำ	iterative
ผลรวมเชิงเส้น	linear combination
การประมาณค่าเชิงเส้น	linear interpolation
การประมาณพันธะเชิงเส้น	linear prediction
สัมประสิทธิ์ของการประมาณพันธะเชิงเส้น	linear prediction coefficient

การเข้ารหัสโดยการประมาณพันธะเชิงเส้น	linear predictive coding (LPC)
ค่าที่เหมาะสมที่สุดเฉพาะแห่ง	local optimum
ความน่าจะเป็นจริงสูงสุด	maximum likelihood
ค่าความผิดพลาดกำลังสองเฉลี่ย	mean-square error
เงื่อนไขโมโนโทนิก	monotony condition
กฎการเลือกบริเวณที่ใกล้เคียงที่สุด	nearest neighbour rule
นิวรอลเน็ตเวิร์ก, เครือข่ายประสาท	neural network
ทำให้เป็นบรรทัดฐานเดียวกัน, นอร์แมลไลซ์	normalize
ความสัมพันธ์เชิงพิกัดฉาก	orthogonal relationship
ค่าปรากฏ, ค่าสังเกต	observation
พารามิเตอร์	parameter
อัตสหสัมพันธ์ส่วนย่อย	PARCOR (Partial Autocorrelation)
อนุพันธ์ส่วนย่อย	partial derivation
รูปแบบ	pattern
ความคล้ายคลึงกันของรูปแบบ	pattern similarity
ฟังก์ชันพหุคูณ	polynomial function
กรรมวิธีประมวลผลเบื้องต้น	preprocessing
ฟังก์ชันความหนาแน่นความน่าจะเป็น	probability density function
การรู้จำ	recognition
สัมประสิทธิ์การสะท้อน	reflection coefficient
ปริภูมิ	space
เชิงสเปกตรัม	spectral
เสียงพูด	speech
การเข้ารหัสเสียงพูด	speech coding
การบีบอัดเสียงพูด	speech compression
ค่าความคลาดเคลื่อนกำลังสอง, ค่าความผิดพลาดกำลังสอง	square error
สถานะ	state
จุดเริ่มเปลี่ยน	threshold
การฝึกฝน	training
ชุดฝึก, ชุดฝึกฝน	training set
การควอนไทซ์แบบเวกเตอร์	vector quantization

บทที่ 1

บทนำ

1.1. ความสำคัญและที่มาของปัญหาที่ทำการวิจัย

การวิจัยทางด้านกรรมวิธีสัญญาณเสียงพูดในปัจจุบันนี้ มีความก้าวหน้าไปมาก การรู้จำเสียงพูดจะปรากฏให้เห็นเป็นรูปเป็นร่างมากยิ่งขึ้นในการประยุกต์ใช้งาน ในปัจจุบันการรู้จำเสียงพูด (Speech Recognition) ถูกนำมาประยุกต์ใช้งานกันอย่างกว้างขวาง เช่น การให้บริการของธุรกิจการธนาคาร การป้อนข้อมูล สันทนาการ โดยเฉพาะในระบบสื่อสารโทรคมนาคม จุดมุ่งหมายหลักของการรู้จำเสียงพูดก็คือการเพิ่มพูนความสามารถของอุปกรณ์ต่างๆ ในการรับรู้และสื่อสารโต้ตอบกับมนุษย์ได้ เพื่อเพิ่มทางเลือกในการควบคุมสั่งการอุปกรณ์เครื่องมือต่างๆ โดยเฉพาะเครื่องคอมพิวเตอร์ ซึ่งการใช้เสียงพูดควบคุมสั่งการนี้ถือได้ว่าเป็นวิธีการที่เป็นธรรมชาติมากที่สุดของมนุษย์ ศาสตร์ทางด้านการรู้จำเสียงพูดเกิดขึ้นมากว่า 40 ปีแล้ว แต่เพิ่งจะเริ่มมีการศึกษากันอย่างจริงจังในช่วงสองทศวรรษที่ผ่านมานี้เอง

อย่างไรก็ดี เทคนิคต่าง ๆ ที่พัฒนาขึ้นใช้เหล่านั้นมีพื้นฐานการพัฒนาอยู่บนหลักการออกเสียงที่ไม่ใช่ภาษาไทยของเรา ถ้ามีการนำมาปรับใช้กับเสียงพูดภาษาไทยโดยตรง ย่อมมีโอกาสที่จะทำให้ได้ผลลัพธ์ที่ด้อยลงไปไม่มากก็น้อย เนื่องจากรูปแบบการออกเสียงและโครงสร้างของภาษาไทยเรา แตกต่างจากภาษาของประเทศตะวันตก หรือแม้แต่เมื่อเทียบกับภาษาอื่น ๆ ในประเทศทางเอเชีย การวิจัยนี้จึงเน้นการศึกษา ค้นคว้า ดัดแปลง และพัฒนากรรมวิธีรู้จำเสียงพูดที่เหมาะสมกับเสียงพูดภาษาไทย เพื่อให้มีอัตราการรู้จำสูงเพียงพอที่จะนำไปประยุกต์ใช้งานได้ต่าง ๆ เช่นเดียวกับของภาษาอื่น ๆ

การรู้จำเสียงพูดนั้นมีการศึกษาวิจัยกันอย่างแพร่หลาย แนวทางในการศึกษาวิจัยก็มีหลากหลายว่าจะเน้นในด้านใด เนื่องจากกระบวนการรู้จำเสียงพูด ต้องการขั้นตอนและกรรมวิธีที่แตกต่างกันไป ทั้งนี้อาจแบ่งออกเป็นแนวทางหลัก ๆ ได้เป็น การศึกษาวิจัยที่เน้น กลุ่มผู้พูด ลักษณะการพูด กรรมวิธีหลัก กรรมวิธีและเทคนิคเสริมอื่น ๆ เป็นต้น

ก) การศึกษาวิจัยที่เน้นตามกลุ่มผู้พูด แบ่งย่อยได้เป็น

ก.1) แบบขึ้นกับผู้พูด (Speaker-Dependent)

ก.2) แบบไม่ขึ้นกับผู้พูด (Speaker-Independent)

ก.3) แบบตรวจรู้ผู้พูด (Speaker Identification)

งานวิจัยส่วนใหญ่เน้นที่แบบไม่ขึ้นกับผู้พูด เพราะสามารถนำไปประยุกต์ใช้งานได้กว้างขวางกว่าแบบอื่น อีกทั้งลักษณะสำคัญที่ต้องใช้จะไม่มีข้อกำหนดที่เข้มงวด ต่างจาก 2 แบบหลัง อย่างไรก็ตาม การศึกษาวิจัยในระยะแรก เริ่มจากแบบขึ้นกับผู้พูด เพราะสามารถควบคุมความแปรปรวนของกลุ่มตัวอย่างได้ดีกว่า อีกทั้งผลิตภัณฑ์ทางด้านการรู้จำเสียงพูดที่เผยแพร่ในระยะแรก ๆ จนถึงปัจจุบัน ก็เป็นแบบขึ้นกับผู้พูดหรืออยู่ในแนวทางดังกล่าว เพราะต้องมีการฝึกฝนผู้ที่จะนำผลิตภัณฑ์เหล่านั้นไปใช้งานเป็นระยะเวลาหนึ่ง ในขณะที่แทบไม่ปรากฏงานวิจัยแบบตรวจรู้ผู้พูด เพราะเรายังไม่สามารถสร้างแบบจำลองกลไกการพูดของมนุษย์ที่สมบูรณ์ ที่สำคัญการหาลักษณะสำคัญเฉพาะในเสียงพูดของแต่ละบุคคล เป็นเรื่องที่ยังทำไม่ได้ด้วยความรู้ที่มีในปัจจุบัน

ข) การศึกษาวิจัยที่เน้นตามลักษณะการพูด สามารถจำแนกตามเสียงพูดได้เป็น

ข.1) แบบคำโดด (Isolated Word) เป็นการศึกษาวิจัยที่เริ่มก่อนแบบอื่น เพราะรูปแบบของสัญญาณง่ายต่อการวิเคราะห์และวัดหาลักษณะสำคัญโดยรวมที่สุด มีงานวิจัยจำนวนมากที่เดินไปในแนวทางนี้แม้จนถึงปัจจุบัน เป็นต้นว่า งาน

วิจัยของ Rabiner, 1978; Rabiner et al., 1978; Rabiner et al., 1979; Furui, 1980; Myers et al., 1980; Kuhn, and Tomaschewski, 1983; Rabiner et al., 1983; Furui, 1986; Euler, and Wolf, 1987; Rashwan, and Fahmy, 1988; Kammerer, and Kupper, 1989; McInnes et al., 1989; Lleida et al., 1990; Dermatas et al., 1991; Huang, and Tseng, 1991; Peinado et al., 1991; Reynolds, and Tarassenko, 1991; Haiyan, and Chengyi, 1992; Ying, and Jamieson, 1993

ข.2) แบบคำติดกัน (Connected Word) หรือคำหลายพยางค์ (Polysyllabic Word) เป็นแบบที่มีการศึกษาวิจัยในลำดับต่อจากแบบแรก เพื่อพยายามก้าวไปสู่แบบที่สาม ดังตัวอย่างงานวิจัยของ Rabiner, and Sambur, 1976; Rabiner, and Schmidt, 1980; Rabiner, and Leevinson, 1985; Bocchieri, and Doddington, 1987; Hunt, 1988; Rabiner et al., 1989; Silverman, and Morgan, 1990; Ukita et al., 1992

ข.3) แบบเสียงพูดต่อเนื่อง (Continuous Speech) เป็นแบบที่ให้ความสนใจทำการศึกษาวิจัยกันมากในปัจจุบัน ดังเช่นงานวิจัยของ Zelinski, and Class, 1983; Mattia, and Giachin, 1988; Picone, 1990; Bourlard, and Morgan, 1993; Gopalakrishnan, and Nahamoo, 1993; Dugast et al., 1994; Zhao, 1994; Zavaliagkos et al., 1994; Chen, and Wang, 1995; Morgan, and Bourlard, 1995

ค) การศึกษาวิจัยที่เน้นตามลักษณะของเสียงพูด เป็นการวิจัยที่มีการกำหนดขอบเขตหรือเป้าหมายของงานวิจัยหรือประโยชน์ที่จะนำไปประยุกต์ แบ่งเป็น

ค.1) แบบจำกัดเฉพาะเสียงตัวเลข

ค.2) แบบเสียงคำสั่งเฉพาะงาน ที่เป็นคำ ๆ หรือคำหลายพยางค์ มีจำนวนคำไม่มาก

ค.3) แบบเสียงคำทั่วไปที่มีคำศัพท์จำนวนมากพอประมาณ

ค.4) แบบเสียงคำทั่วไปที่มีคำศัพท์จำนวนมาก

ง) การศึกษาวิจัยที่เน้นตามกรรมวิธีหลัก ที่แบ่งได้เป็น 4 กรรมวิธี (Roe and Wilpon, 1993) คือ

ง.1) การเข้าคู่ต้นแบบ (Template Matching)

เทคนิคที่อยู่ในแนวทางกรรมวิธีนี้และเป็นที่นิยมกันมากคือ Dynamic Time Warping (DTW) ซึ่งมีกรรมวิธีรุ่นแรก ๆ ที่นำมาวิจัยกัน มีความง่ายต่อการพัฒนา การฝึกฝนใช้เวลาน้อย แต่เวลาในการรู้จำขึ้นกับจำนวนแบบอ้างอิง และสมรรถนะของการรู้จำจะต่ำกว่ากรรมวิธีอื่น ๆ และลดลง ถ้าเพิ่มจำนวนคำศัพท์มากขึ้น ตัวอย่างเช่น Rabiner, 1978; Rabiner et al., 1978; Myers et al., 1980; Rabiner, and Schmidt, 1980; Furui, 1986; McInnes, 1989

ง.2) ระบบตามกฎเกณฑ์ (Rule-Based System)

ระบบตามกฎเกณฑ์นั้นจะอาศัยเงื่อนไขในการตัดสิน ซึ่งเมื่อระบบมีขนาดใหญ่และซับซ้อนมากขึ้นจะทำให้การตัดสินใจผิดพลาดได้ง่ายตั้งแต่ขั้นตอนแรก Fissore et al., 1988; Reynolds, and Tarassenko, 1991; Carlson, and Clements, 1994; Sheikhzadeh, and Denng, 1994; Chen, and Soong, 1994; Zhao, 1994

ง.3) แบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Model, HMM)

กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ เป็นที่นิยมกันมากที่สุดในปัจจุบัน เพราะข้อดีหลายประการ เป็นต้นว่า ความยืดหยุ่นในการปรับให้ใช้กับการรู้จำเสียงพูดที่มีลักษณะการพูดต่าง ๆ กัน สมรรถนะในการรู้จำที่สูง ความสามารถรองรับคำศัพท์จำนวนมาก อย่างไรก็ดีข้อเสียของมันคือ ต้องการเวลาในการฝึกฝนและรู้จำ โดยเฉพาะเมื่อมีคำศัพท์ใหม่ ๆ เพิ่มขึ้น จะต้องเริ่มกระบวนการฝึกฝนใหม่ทุกครั้งไป ตัวอย่างงานวิจัยทางด้านนี้ ได้แก่ Rabiner et al., 1983; Levinson et al., 1983; Rabiner, and Levinson, 1985; Rabiner, and Juang, 1986; Euler, and Wolf, 1987; Bahl et al., 1988; Deng et al., 1988; Lee, and Hon, 1988; Leevinson et al., 1988; Lee, and Hon, 1989; Rabiner, 1989; Rabiner et al., 1989; Picone, 1990; Peinado et al., 1991; Huang, 1992; Bahl et al., 1993; Jianing et al., 1994; Su, and Lee, 1994; Levinson, and Ljolje, 1994; Renals et al., 1994; Matsui, and Furui, 1994

ง.4) เครือข่ายนิวรอล (Neural Network)

เครือข่ายนิวรอล หรือนิวรอลเน็ตเวิร์ก เป็นกรรมวิธีที่เริ่มมีผู้สนใจทำวิจัยกัน โดยการเลียนแบบเครือข่ายประสาทของมนุษย์ ที่มีความเร็วในการรู้จำเหนือกว่ากรรมวิธีอื่น ข้อเสียของกรรมวิธีนี้จะเหมือนกับของฮิดเดน มาร์คอฟ ที่ต้องต้องการเวลาในการฝึกฝน และต้องเริ่มการฝึกฝนใหม่ทุกครั้งที่มีการเพิ่มคำศัพท์ ตัวอย่างของงานวิจัยในแนวทางนี้ ได้แก่ Lebensky, 1991; Elvira, and Carrasco, 1992; Haiyan, and Chengyi, 1992; Chen, and Wang, 1995

จ) การศึกษาวิจัยเกี่ยวกับกรรมวิธีและเทคนิคเสริมอื่น ๆ

เนื่องจากกระบวนการรู้จำประกอบด้วยขั้นตอนการประมวลผลหลายขั้นตอน ดังนั้น แต่ละขั้นตอนที่เสริมเข้าไปในกระบวนการอาจได้รับการพัฒนา การปรับแต่ง การเปลี่ยนแปลง ให้ระบบทำงานได้ดีขึ้น ตัวอย่างที่น่าสนใจ เช่น

จ.1) กรรมวิธี Hybrid HMM-NN ที่นำเอาส่วนดีของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ผสมกับส่วนดีของนิวรอลเน็ตเวิร์ก ดังเช่นงานวิจัยของ Jin, and Chung, 1993; Dugast et al., 1994; Rigoll, 1994; Cerf et al., 1994; Zavaliagkos et al, 1994

จ.2) เทคนิค End-Point Detection ที่ให้ความสนใจในเรื่องการหาจุดสิ้นสุดเสียงพูดและการตัดคำ เพราะปรากฏว่ามีโอกาสมากที่ข้อสนเทศหรือลักษณะสำคัญของเสียงพูดไปอยู่ที่ส่วนหัวและท้ายของคำ งานวิจัยที่น่าสนใจ เช่น Lamel et al, 1981; Dermatas et al., 1991; Evangelos et al., 1991; Huang, and Tseng, 1991; Ying, and Jamieson, 1993

จ.3) เทคนิค Vector Quantization ที่เน้นการลดจำนวนข้อมูลที่ต้องใช้ในการรู้จำเพื่อเพิ่มความเร็วในการรู้จำ ตัวอย่างของงานวิจัย ได้แก่ Rabiner att al., 1983; Gray, 1984; Makhoul et al., 1985; Pan et al., 1985; Matsui, and Sadaoki Furui, 1994

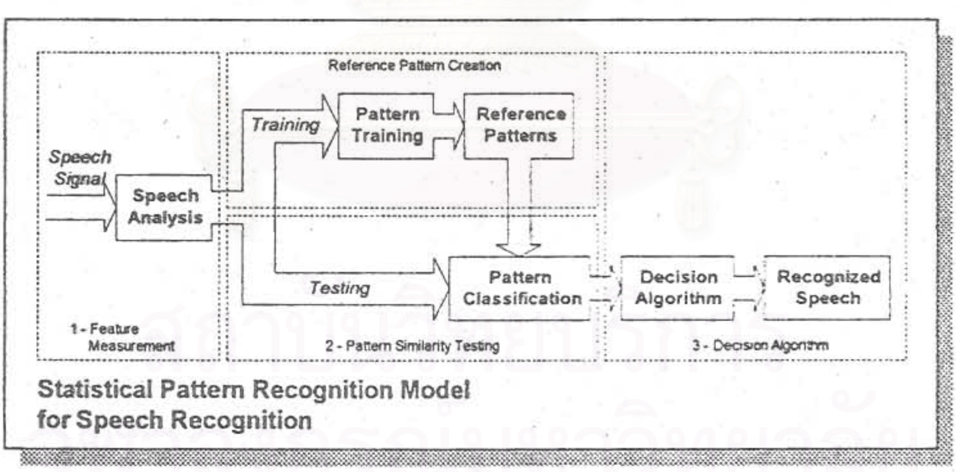
จ.4) เทคนิค Fuzzy-HMM ที่นำเสนอตรรกในการตัดสินใจแบบฟัซซีเข้าช่วยในกระบวนการรู้จำ ตัวอย่างงานวิจัยในแนวทางนี้ คือ Koo, and Un, 1990; Kim, and Lee, 1991

การค้นคว้าวิจัยในด้านการรู้จำเสียงพูดภาษาไทยที่มีในประเทศไทยนั้น เท่าที่สามารถสืบค้นเอกสารได้ ปรากฏว่ามีไม่มากนักและเพิ่งเริ่มมีการศึกษาค้นคว้าวิจัยอย่างจริงจังตั้งแต่ประมาณปี พ.ศ. 2525 ดังจะเห็นได้จากผลงานวิจัยต่าง ๆ ได้แก่ การประมาณพันธะเชิงเส้นเสียงพูด (สุเธียร เกียรติสุนทร, 2525) การตรวจรู้เสียงพูดภาษาไทยโดยใช้หน่วยพยางค์ (ทวิ ประทุมทาน, 2530) ระบบการรับรู้เสียงพูดแบบต่างบุคคล (ไพศาล ธรรมโพธิทอง, 2533) การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดน มาร์คอฟ โมเดลและเวกเตอร์ควอนไตซ์เซชั่น (เสาวลักษณ์ อารีพงศา, 2538; Areeponsa, and Jitapunkul, 1995) การรู้จำเสียงพูดสระภาษาไทยโดดๆ ไม่ขึ้นกับผู้พูดโดยการวัดสเปกตรัมดิสแตนส์และใช้ไดนามิกไทม์วาร์ปปิ้ง (ธีระ ภัทราพรนันท์, 2538; Phatrapornnant, and Jitapunkul, 1995) การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นกับผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปิ้ง (ระพีพัฒน์ เพ็ญศิริ, 2538; Pensiri, and Jittapunkul, 1995) และการรู้จำคำพูดภาษาไทยโดยใช้ลักษณะบ่งความต่างของหน่วยเสียง (ณัฐกร ทับทอง, 2538) ระบบการรู้จำคำไทยหลายพยางค์แบบไม่ขึ้นกับผู้พูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟ (วิศรุต อาชุบุตร, 2539; Ahkuputra et al., 1997) การรู้จำเสียงตัวเลขภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยใช้แอลพีซี และนิวรอลเน็ตเวิร์กแบบแบ็กพรอพาเกชัน (วุฒิพงษ์ พรสุขจันทรา, 2539; Pornsukjantra, and Jitapunkul, 1996) การตรวจหาจุดเริ่มต้นและจุดสิ้นสุดของคำโดด ๆ (ชัยศรี เอี่ยมอำไพ) โดยงานวิจัยเหล่านี้ล้วนเป็นการวิจัยเกี่ยวกับคำโดดภาษาไทย แต่ไม่เคยมีการศึกษาเชิงเปรียบเทียบถึงกรรมวิธีต่าง ๆ เพื่อพิจารณากรรมวิธีที่ควรจะปรับใช้กับเสียงพูดภาษาไทยเราได้อย่างมีประสิทธิภาพ

ระบบการรู้จำเสียงพูดจะอาศัยการฝึกฝน (Training) เพื่อจดจำรูปแบบอ้างอิง (Reference Patterns) ไว้ใช้ในการเปรียบเทียบกับเสียงพูดที่ยังไม่ทราบรูปแบบ และใช้รูปแบบอ้างอิงเหล่านั้นในการตัดสินใจเลือกรูปแบบที่ใกล้เคียงกับเสียงพูด ที่ถูกนำมาเปรียบเทียบมากที่สุด ขั้นตอนในการรู้จำเสียงพูดโดยทั่วไปแบ่งได้เป็น 3 ขั้นตอน (Rabiner and Levinson, 1981) ได้แก่

1. การวิเคราะห์และวัดค่าลักษณะสำคัญ (Feature Measurement)
2. การจำแนกรูปแบบ (Pattern Classification) หรือ
   การทดสอบความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing)
3. ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)

แบบจำลองการรู้จำเสียงพูดคำต่อเนื่องบนพื้นฐานของคำศัพท์นั้น ดังแสดงในรูปที่ 1.1 เริ่มต้นจากการสร้างและเก็บลักษณะสำคัญ (Feature) จากเสียงพูดที่ยังไม่ทราบรูปแบบ แล้วนำไปเปรียบเทียบกับรูปแบบของคำศัพท์แต่ละคำที่ได้เก็บไว้แล้ว เพื่อหารูปแบบที่ใกล้เคียงกันมากที่สุดกับเสียงพูดที่เข้ามา ถ้ารูปแบบทั้งสองใกล้เคียงกันมากพอ ระบบจะตัดสินให้เป็นคำดังกล่าวทันที แต่ถ้ารูปแบบทั้งสองไม่ใกล้เคียงกัน ระบบจะไม่ตัดสินว่าเป็นคำใด แต่จะให้ผู้พูดพูดซ้ำอีกครั้งหนึ่ง วิธีการหาลักษณะสำคัญของสัญญาณเสียงเพื่อใช้ในการเปรียบเทียบนั้นมีหลายวิธีการด้วยกัน (Rabiner and Wilpon, 1979; Rabiner and Levinson, 1981; Rabiner, 1994) เช่น Digital Filter Bank, Fourier Transform, Cepstral Coefficient, Linear Prediction Coefficient, LP-Derived Filter Bank, LP-Derived Cepstral Coefficient เป็นต้น โดยการวิเคราะห์จะสามารถแบ่งได้เป็น 2 ประเภท (Rabiner, 1994) ได้แก่ การวิเคราะห์ในเชิงเวลา (Time Domain Analysis) และการวิเคราะห์ในเชิงความถี่ (Frequency Domain Analysis) ซึ่งในงานวิจัยนี้จะใช้วิธีการวิเคราะห์ในเชิงเวลาเป็นหลัก



รูปที่ 1.1 แบบจำลองรูปแบบการรู้จำทางสถิติที่ใช้ในการรู้จำเสียงพูด (Rabiner and Levinson, 1981)

ขั้นตอนในการเปรียบเทียบรูปแบบระหว่างรูปแบบที่ยังไม่ทราบ กับรูปแบบของคำศัพท์ที่ได้จัดเก็บไว้แล้วนั้น จัดอยู่ในขั้นตอนการจำแนกประเภทรูปแบบ (Pattern Classification) (Rabiner and Levinson, 1981; Levinson and Roe, 1990; Roe and Wilpon, 1993) ดังแสดงในรูปที่ 1.1 วิธีการจำแนกประเภทรูปแบบในการรู้จำเสียงพูดสามารถแบ่งได้เป็น 4 ดังได้บรรยายไว้ข้างต้นแล้วในหัวข้อ ง.1 ถึง ง.4

1.2. วัตถุประสงค์

2.1. เพื่อศึกษาการรู้จำเสียงพูดด้วยกรรมวิธี ไดนามิก ไทม์วาร์ปิง, แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

2.2 เพื่อศึกษาเชิงเปรียบเทียบกรรมวิธีกรรมวิธีทั้งสามในการรู้จำเสียงพูดภาษาไทยที่เป็นคำโดด และไม่ขึ้นกับผู้พูด

2.3 เพื่อพัฒนากรรมวิธีที่เหมาะสมในการรู้จำเสียงพูดตัวเลขไทย

1.3. เป้าหมายและขอบเขตของงานวิจัย

สามารถพัฒนากรรมวิธีรู้จำเสียงพูดตัวเลขไทย 0 - 9 แบบไม่ขึ้นกับผู้พูดได้ด้วยอัตราการรู้จำร้อยละ 80

1.4. ขั้นตอนและวิธีการดำเนินการ

1.4.1. ศึกษาคุณลักษณะของเสียงพูด และแบบจำลองการเปล่งเสียงพูด

1.4.2. ค้นคว้าและเก็บรวบรวมข้อมูลรายละเอียดที่เกี่ยวข้องกับวิธีการดังต่อไปนี้

1.4.2.1. การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

1.4.2.2. การวิเคราะห์และวัดค่าลักษณะสำคัญ (Feature Measurement)

1.4.2.3 การทดสอบหรือวัดหาความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing or Determination)

1.4.2.4. ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)

1.4.3. เก็บข้อมูลเสียงพูดตัวเลขภาษาไทยของกลุ่มตัวอย่าง ได้แก่

1.4.3.1. กลุ่มตัวอย่างเสียงพูดเพื่อฝึกฝน (Training Group) สำหรับใช้ในการฝึกฝนระบบเพื่อสร้างรูปแบบอ้างอิงของฐานข้อมูลคำศัพท์

1.4.3.2. กลุ่มตัวอย่างเสียงพูดเพื่อทดสอบ (Testing Group) สำหรับใช้ในการทดสอบรูปแบบอ้างอิงที่สร้างขึ้นมาจากกลุ่มตัวอย่างเสียงพูดเพื่อฝึกฝน

1.4.4. วิเคราะห์และพัฒนาโปรแกรมในแต่ละส่วน

1.4.5. ทำการฝึกฝนพร้อมทั้งจำแนกกลุ่มคำเพื่อสร้างรูปแบบคำพูดอ้างอิงจากเสียงพูดของกลุ่มตัวอย่าง

1.4.6. ทำการทดสอบอัตราความแม่นยำในการรู้จำเสียงพูด โดยใช้เสียงพูดของทั้งสามกลุ่ม

1.4.7. ปรับปรุงแก้ไขโปรแกรมหรือเพิ่มจำนวนการฝึกฝนเพื่อเพิ่มอัตราความถูกต้องในการรู้จำ

1.4.8. สรุปรวบรวมผลการวิจัยทั้งหมด พร้อมทั้งจัดทำรายงานฉบับสมบูรณ์

1.5. ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

5.1. ทราบถึงวิธีการรู้จำคำพูดภาษาไทยและลักษณะทางภาษาศาสตร์ของคำพูดภาษาไทย

5.2. ทราบถึงกรรมวิธีที่เหมาะสมในการรู้จำเสียงพูดภาษาไทย

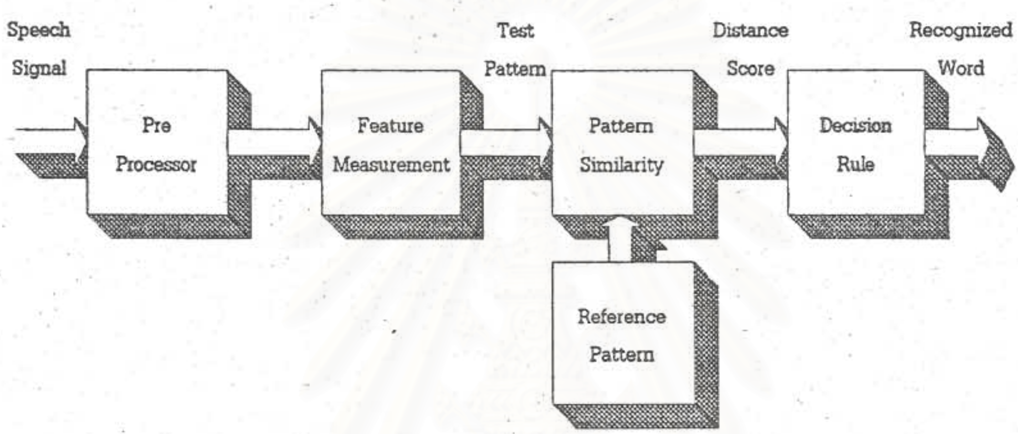
5.3. เป็นการเพิ่มเติมความสามารถให้แก่เครื่องคอมพิวเตอร์ในการรู้จำคำพูดภาษาไทย

5.4. เพื่อเป็นประโยชน์แก่ผู้พิการทางสายตาและผู้ที่ไม่สามารถช่วยตนเองได้ ให้สามารถใช้งานเครื่องมือและอุปกรณ์ต่างๆ ด้วยการสั่งงานเป็นเสียงพูดได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.0. คำนำ

แบบจำลองของการรู้จำเสียงพูด ที่ใช้ในงานวิจัยนี้เป็นแบบจำลองดังแสดงอยู่ในรูปที่ 2.1 (Rabiner and Levinson ,1981) จะประกอบด้วยขั้นตอนดำเนินการหลัก 4 ขั้นตอน คือ



รูปที่ 2.1 แบบจำลองของการรู้จำเสียงพูด

2.0.1. Preprocessing เป็นขั้นตอนการประมวลผลเบื้องต้น เพื่อจัดเตรียมข้อมูลให้เหมาะสมต่อการประมวลผลในขั้นตอนต่อ ๆ ไป

2.0.2. Feature measurement เป็นขั้นตอนของการสกัดเพื่อนำลักษณะเด่นของข้อมูลออกมา และจะเพิ่มกระบวนการลดจำนวนข้อมูลเพื่อลดเวลาในการประมวลผลในขั้นตอนต่อไป

2.0.3. Pattern similarity determination เป็นขั้นตอนในการหาค่ารูปแบบใกล้เคียงของคำที่ยังไม่ทราบเมื่อเทียบกับคำอ้างอิงที่มีอยู่

2.0.4. Decision rule เป็นกฎเกณฑ์ในการเลือกว่าคำที่ต้องการรู้จำ ว่าใกล้เคียก ฐ ปบบ อ้างอิงของคำใดมากที่สุด

2.1. การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

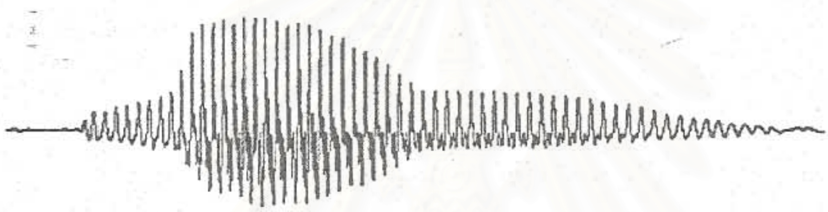
การประมวลผลสัญญาณเบื้องต้นเป็นขั้นตอนกรรมวิธีในการจัดเตรียมข้อมูล จากข้อมูลดิบของเสียงพูดที่ได้จากการบันทึกเสียงนำมาผ่านกรรมวิธีประมวลผลสัญญาณเชิงเลข เพื่อใช้ในการประมวลผลในขั้นตอนต่อไป เนื่องจากสัญญาณเสียงพูดโดยรวมจะมีสัญญาณรบกวน มีการแปรเปลี่ยนตามเวลาและไม่เสถียร (Nonstationary) ดังรูปที่ 2.2 ที่อาจแบ่งลักษณะของเสียงได้ออกเป็น 3 ส่วน ดังรูปที่ 2.3 (ที่เป็นการขยายสัดส่วนทางเวลาของรูปที่ 2.2 ออกและตัดแบ่งเป็นเฟรมหรือกรอบ เพื่อเห็นรายละเอียดได้ชัดเจนขึ้น) ได้ดังนี้คือ

ก. ช่วงที่ยังไม่มีการเปล่งเสียงหรือสภาวะเงียบ (silence) เสียงในช่วงนี้จะค่อนข้างเรียบถ้าไม่มีสัญญาณรบกวนจากภายนอก ดูรูปที่ 2.3ก และ 2.3ข

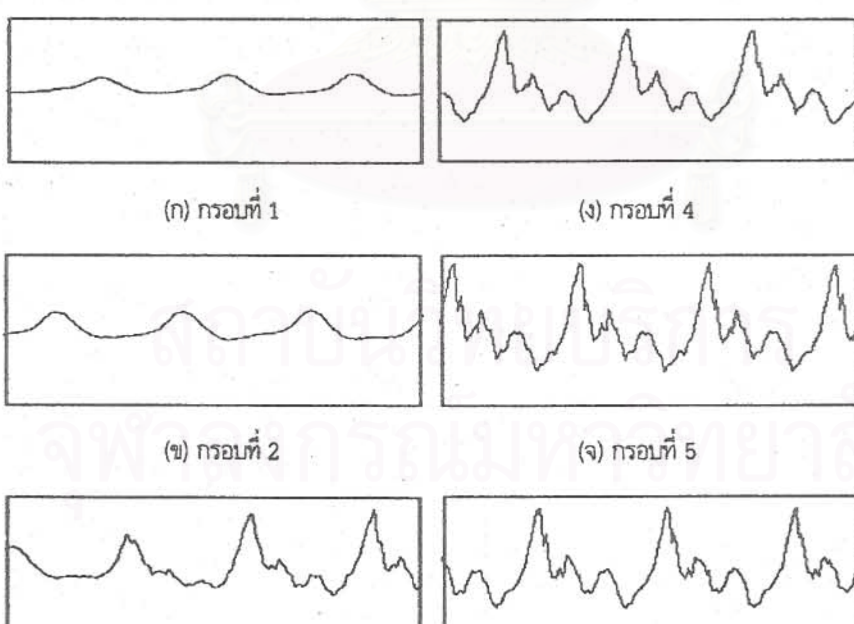
ข. ช่วงก่อนที่จะเปล่งเสียงออกมาหรือที่เรียกว่า เสียงอโฆษะ (unvoice speech) ในช่วงนี้แอมพลิจูดของเสียงจะต่ำและจะไม่มีความเป็นคาบ ดูรูปที่ 2.3ค

ค. ช่วงที่เป็นคำพูดหรือที่เรียกว่าเสียงโฆษะ (voice speech) ในช่วงนี้เสียงพูดจะมีลักษณะเป็นคาบจะมีแอมพลิจูดสูง ดูรูปที่ 2.3ง ถึง 2.3ฉ

ดังนั้นในการประยุกต์ใช้งานกรรมวิธีประมวลผลสัญญาณเชิงเลขกับสัญญาณเสียงพูด จึงต้องทำกรรมวิธีประมวลสัญญาณเบื้องต้นก่อน เช่น การกำจัดสัญญาณรบกวน การแบ่งสัญญาณเสียงพูดออกเป็นส่วนย่อย (Rabiner and Levinson, 1981; Furui, 1985) ที่เรียกว่า "กรอบเสียงพูด" (Speech Frame) โดยแต่ละกรอบเสียงจะมีความยาวประมาณ 10 - 40 มิลลิวินาที ซึ่งถือได้ว่าสัญญาณเสียงพูดในแต่ละกรอบเสียงพูดที่มีข้อสนเทศ (information) จริง ๆ จะมีความเสถียรและไม่แปรเปลี่ยนตามเวลา (Stationary) ดังรูปที่ 2.3ง ถึง 2.3ฉ จากนั้นจึงสามารถทำการประมวลผลสัญญาณเชิงเลขกับสัญญาณเสียงพูดในแต่ละกรอบเสียงพูดด้วยกรรมวิธีที่ง่ายสะดวกและรวดเร็วได้



รูปที่ 2.2 ตัวอย่างสัญญาณเสียงพูด "หนึ่ง"



(ก) กรอบที่ 1

(ง) กรอบที่ 4

(ข) กรอบที่ 2

(จ) กรอบที่ 5

(ค) กรอบที่ 3

(ฉ) กรอบที่ 6

รูปที่ 2.3 แสดงรูปคลื่นในแต่ละกรอบของเสียงพูด "หนึ่ง" ขนาดของกรอบเท่ากับ 25 มิลลิวินาที

2.1.1 กรรมวิธีเน้นล่วงหน้า (Preemphasis)

ขั้นตอนกรรมวิธีเน้นล่วงหน้านี้เป็นการบีบอัดช่วงพิสัยพลวัต (Dynamic Range) ของสัญญาณเสียงพูด โดยการลดความแปรปรวนของสัญญาณด้วยการพิจารณาความเปลี่ยนแปลงของสัญญาณในแต่ละช่วงเวลาแทนการใช้ตัวของสัญญาณ ทำให้ความลาดเอียงในเชิงความถี่แบนราบลง ดังสมการที่ (2.1) (Furui, 1985) เมื่อ a เป็นสัมประสิทธิ์ของตัวกรอง $\widetilde{s}(n)$ เป็นค่าของสัญญาณเสียงพูดขาออกที่ผ่านกรรมวิธีเน้นล่วงหน้าที่ n $s(n)$ เป็นค่าของสัญญาณเสียงพูดขาเข้าที่ n และ $s(n-1)$ เป็นค่าของสัญญาณเสียงพูดขาเข้าค่าก่อนหน้าที่ n-1 ดังนี้

$$\widetilde{s}(n) = s(n) - as(n-1) \qquad (2.1)$$

ซึ่งจะส่งผลให้ค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าสูงขึ้น ในทางปฏิบัติแล้วจะนำสัญญาณผ่านตัวกรองเชิงเลขลำดับหนึ่ง (First-Order Digital Filter) ที่มีฟังก์ชันถ่ายโอนดังแสดงในสมการที่ (2.2) (Furui, 1985)

$$H(z) = 1 - az^{-1} \qquad (2.2)$$

โดยกำหนดให้ค่าสัมประสิทธิ์ของตัวกรอง a มีค่าเข้าใกล้ 1 เมื่อใช้ร่วมกับการวิเคราะห์หาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นจะกำหนดให้ค่า $a = 0.95$ เนื่องจากเป็นค่าที่ให้ผลดีที่สุดสำหรับการวิเคราะห์ (Rabiner, Levinson, Rosenberg, and Wilpon, 1979)

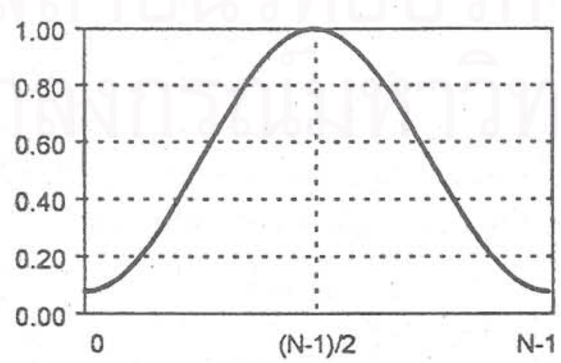
2.1.2 กรรมวิธีวางกรอบขนาดสัญญาณ (Smoothing Window)

ขั้นตอนกรรมวิธีวางกรอบขนาดสัญญาณ เป็นขั้นตอนในการเตรียมข้อมูลในแต่ละกรอบข้อมูลเสียงพูดเพื่อการวิเคราะห์อัตสหสัมพันธ์ (autocorrelation) โดยการคูณแต่ละค่าของสัญญาณในกรอบข้อมูลเสียงพูดด้วยค่าฟังก์ชันกรอบ (Window Function) ซึ่งมีหลายประเภทได้แก่ Rectangular Window, Hamming Window, Hanning Window, Blackman Window, Kaiser Window เป็นต้น (Oppenheim and Schafer, 1989) ผลของการวางกรอบขนาดสัญญาณมี 2 ประการ ประการแรก เป็นการลดทอนแอมปลิจูดอย่างช้าๆ ที่บริเวณปลายแต่ละข้างของกรอบข้อมูลเสียงพูดเพื่อป้องกันการเปลี่ยนแปลงอย่างกระทันหันที่จุดปลาย ประการที่สอง เป็นการสร้างค่าการประสานสำหรับการผลแปลงฟูริเยร์ของฟังก์ชันกรอบและแถบสเปกตรัมของเสียงพูด สำหรับการวิเคราะห์เสียงพูดในงานวิจัยนี้จะใช้ฟังก์ชันกรอบชนิด Hamming Window ดังแสดงในรูปที่ 2.4 ซึ่งใช้ในการวิเคราะห์เสียงพูดโดยเฉพาะดังสมการที่ (2.3) และ (2.4) เมื่อ L เป็นจำนวนกรอบข้อมูลเสียงพูดทั้งหมด N เป็นจำนวนข้อมูลในแต่ละกรอบข้อมูลเสียงพูด l เป็นกรอบที่ l ของกรอบ L ทั้งหมด และ n เป็นข้อมูลที่ n ของข้อมูลทั้งหมด N ค่าซึ่งอยู่ภายในกรอบที่ l (Furui, 1985; Oppenheim and Schafer, 1989)

$$\widetilde{x}_l(n) = x_l(n) \cdot w(n) \qquad (2.3)$$

$$w(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \qquad (2.4)$$

เมื่อ $l = 0, 1, \ldots, L-1 \qquad n = 0, 1, \ldots, N-1$



รูปที่ 2.4 ฟังก์ชันกรอบชนิด Hamming Window

2.1.3 กรรมวิธีหาจุดสิ้นสุดเสียงพูด (Endpoint Detection)

ขั้นตอนกรรมวิธีหาจุดสิ้นสุดเสียงพูดนี้เป็นขั้นตอนที่สำคัญที่สุดขั้นตอนหนึ่งในกระบวนการรู้จำเสียงพูด (Rabiner and Levinson,1981) เพราะ

ก) ความผิดพลาดในการหาจุดสิ้นสุดเสียงพูด จะทำให้ความน่าจะเป็นของความผิดพลาดในการรู้จำเสียงพูดเพิ่มขึ้น

ข) การหาจุดสิ้นสุดเสียงพูดที่ถูกต้อง ช่วยให้การคำนวณทั้งหมดของระบบต่ำสุด

การหาจุดสิ้นสุดเสียงพูดทั้งทางส่วนหน้าและส่วนท้ายหรือการหาหัวท้ายของเสียงพูดเป็นกระบวนการค้นหาช่วงที่เป็นข้อสนเทศจริงของเสียงพูดที่ได้จากการบันทึก มีกรรมวิธีหลัก ๆ ดังนี้

2.1.3.1. กรรมวิธีหาจุดสิ้นสุดเสียงพูดโดยใช้ค่าแอมพลิจูด

เมื่อสัญญาณมีค่าแอมพลิจูดมากกว่าค่าที่กำหนดไว้เท่ากับจำนวนครั้งที่กำหนด จะให้จุดนั้นเป็นจุดเริ่มต้นของเสียงพูด และทำเช่นเดียวกันในส่วนท้ายของเสียงที่บันทึกมาเพื่อหาจุดสิ้นสุด ข้อดีของวิธีนี้คือ ใช้การคำนวณง่าย ๆ และใช้เวลาในการคำนวณน้อยมาก ข้อเสียของวิธีนี้คือความผิดพลาดเมื่อมีสัญญาณรบกวนที่มีแอมพลิจูดสูง ๆ ในบริเวณส่วนหัวหรือท้ายคำ เช่น เสียงหายใจ

2.1.3.2. กรรมวิธีหาจุดสิ้นสุดเสียงพูดโดยใช้ค่าพลังงาน (Rabiner and Levinson,1981)

วิธีนี้ใช้คอนทัวร์ (contour) ของพลังงานในแต่ละส่วนย่อย เพื่อหาจุดที่มีพลังงานมากกว่าระดับที่กำหนดไว้ ติดต่อกันนานกว่าคาบเวลาที่กำหนด จุดเริ่มต้นของเสียงพูดจะอยู่ก่อนจุดที่ตรวจพบด้วยระดับพลังงานเท่ากับคาบเวลาคงที่ค่าหนึ่ง ข้อดีของวิธีนี้คือ สามารถลดการตัดคำผิดพลาดเมื่อมีสัญญาณรบกวนที่มีแอมพลิจูดสูง ข้อเสียของวิธีนี้คือจุดเริ่มต้นที่คำนวณได้อาจคลาดเคลื่อนจากจุดเริ่มต้นที่แท้จริงของเสียงพูด

2.1.3.3. กรรมวิธีหาจุดสิ้นสุดเสียงพูดโดยใช้ค่าพลังงานและอัตราการตัดค่าศูนย์ (zero-crossing rate) (Furui, 1989)

เหมือนกับกรรมวิธีที่ใช้ค่าพลังงาน แต่มีการปรับปรุงการหาจุดเริ่มต้นของเสียงพูด โดยใช้อัตราการตัดค่าศูนย์แทนการใช้คาบเวลาคงที่ ทำให้สามารถหาจุดเริ่มต้นได้ถูกต้องมากขึ้น ซึ่งก็ต้องแลกเปลี่ยนด้วยเวลาที่ใช้ในการคำนวณอัตราการตัดค่าศูนย์

อย่างไรก็ตาม มีงานวิจัยจำนวนมากที่นำเสนอกรรมวิธีหาจุดสิ้นสุดเสียงพูดด้วยวิธีการที่ดัดแปลงไปจากกรรมวิธีพื้นฐานทั้งสามข้างต้นเพื่อให้บรรลุวัตถุประสงค์เฉพาะที่ตั้งไว้ (Huang and Tseng, 1991; Dermatas et al., 1991; Ying, Mitchell, and Jamiesen, 1993)

งานวิจัยนี้ที่ศึกษาและพัฒนากรรมวิธีรู้จำ 3 วิธี ทำให้ต้องมีการพัฒนากรรมวิธีหาจุดสิ้นสุดเสียงพูดที่เหมาะสมสำหรับแต่ละกรรมวิธี แต่ทั้งหมดจะอยู่บนพื้นฐานของการใช้พลังงานเป็นตัวกำหนดการตัดหัวท้ายคำ เพราะกรรมวิธีประเภทนี้สามารถแก้ปัญหาการตัดคำผิดพลาด โดยใช้เวลาในการคำนวณไม่มากเกินไป ถึงแม้ว่าจุดเริ่มต้นของเสียงพูดที่คำนวณได้อาจคลาดเคลื่อนไปบ้าง แต่ก็สามารถแก้ไขได้โดยใช้การประมาณค่าคาบเวลาที่เหมาะสม กับกลุ่มคำที่ต้องการรู้จำ

เนื่องจากเป็นการกำหนดจุดเริ่มต้นและจุดสิ้นสุดของคำทั้งหมด หรือของแต่ละพยางค์สำหรับการแยกแยะเสียงพูด เพื่อนำแต่เฉพาะส่วนที่เป็นข้อสนเทศจริงของเสียงพูด ไปดำเนินการวิเคราะห์ในขั้นตอนการรู้จำเสียงพูดของคำศัพท์แต่ละคำ สำหรับงานวิจัยนี้จะอาศัยแผนภูมิเส้นระดับพลังงาน (Energy Level Contour) ของเสียงพูดโดยการวิเคราะห์ค่าระดับพลังงานของแต่ละกรอบข้อมูลเสียงพูด ดังแสดงในสมการที่ (2.5) เมื่อ $E(m)$ เป็นค่าระดับพลังงานของกรอบข้อมูลเสียงพูดที่ m และ $s(n)$ แทนค่าของสัญญาณค่าที่ n ในกรอบข้อมูลเสียงพูด (Rabiner and Levinson, 1981)

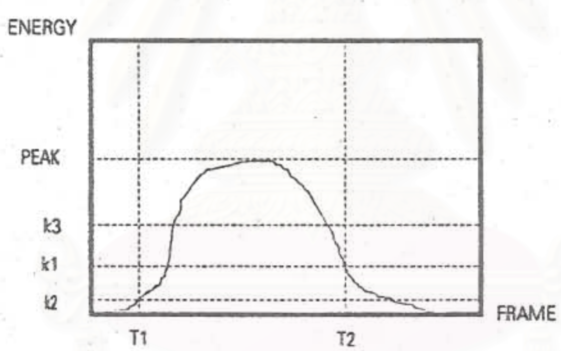
$$E(m) = \sum_{n=0}^{N-1} |s(n)| \qquad (2.5)$$

ในการวิเคราะห์เพื่อกำหนดจุดเริ่มต้นและจุดสิ้นสุดของคำที่ใช้ในการรู้จำด้วย กรรมวิธีไดนามิก ไทม์วาร์ปิง จะเหมือนกับของกรรมวิธีฮิดเดน มาร์คอฟ กล่าวคือ จะใช้ค่าระดับพลังงานของกรอบข้อมูลเสียงพูดตามสมการที่ (2.5) โดยจะกำหนดระดับพลังงานอ้างอิง (Energy Thresholds) 2 ค่า คือ k1, k2 และมีพารามิเตอร์ช่วยอีก 2 ตัว คือ PEAK และ ค่า T1, T2

เนื่องจากการบันทึกเสียงในสภาวะแวดล้อมปกติ บางครั้งมีเสียงรบกวนค่อนข้างมาก อันเนื่องจากสภาวะภายนอกรวมทั้งเกิดจากอุปกรณ์ที่ใช้ในการบันทึกเสียง โดยที่ พารามิเตอร์ PEAK จะแทน ค่าสูงสุด (peak) ของสัญญาณเสียงที่วิเคราะห์ ซึ่งจะนำค่านี้ไปกำหนดค่าของระดับพลังงานอ้างอิงดังนี้คือ

$$k1 = a.PEAK$$
$$k2 = b.PEAK \quad \text{.........(2.6)}$$
$$k3 = c.PEAK$$

ค่าของ a, b, และ c จะเป็นตัวกำหนดค่าของระดับพลังงาน k1, k2, และ k3 ตามลำดับ ค่าของระดับพลังงานอ้างอิงจะเปลี่ยนแปลงตามค่าของพารามิเตอร์ PEAK ที่ได้จากการคำนวณก่อนหน้าที่จะกำหนดระดับพลังงานอ้างอิงดังในรูปที่ 2.5 โดยค่าระดับพลังงาน k1 จะเป็นตำแหน่งเริ่มต้นของเสียง, k2 จะเป็นระดับพลังงานสิ้นสุดของเสียงที่ตัดคำได้ เมื่อทำการตรวจสอบได้ค่าเริ่มต้นและสิ้นสุดสัญญาณเสียงแล้วจะทำการตรวจสอบว่า ระดับพลังงานอ้างอิง k3 อยู่ภายในช่วงดังกล่าวหรือไม่ ถ้าไม่อยู่ภายในช่วงดังกล่าวจะทำการหาจุดเริ่มต้นของเสียงใหม่ แต่ถ้าค่าระดับพลังงานอ้างอิง k3 นี้อยู่ภายใน จะทำการตรวจสอบอีกครั้งว่าช่วงดังกล่าวนี้มีจำนวนเฟรมของสัญญาณน้อยกว่า ค่า T2 หรือไม่ ถ้ามีค่าน้อยกว่าจะไม่ถือว่าเป็นส่วนของคำ แต่ถ้าจำนวนกรอบมีค่าไม่น้อยกว่า T2 แล้ว จะถือว่าเป็นส่วนของคำที่ตรวจสอบได้



รูปที่ 2.5 แสดงจุดอ้างอิงเพื่อหาจุดเริ่มต้นและจุดสิ้นสุดของรูปคลื่นพลังงาน

ในการวิเคราะห์เพื่อกำหนดจุดเริ่มต้นและจุดสิ้นสุดของคำที่ใช้ในการรู้จำด้วย กรรมวิธีนิวรอล เน็ตเวิร์กได้พัฒนารายละเอียดแตกต่างไปจากที่ใช้ในสองกรรมวิธีข้างต้น เนื่องจากกรรมวิธีนี้มีความจำเป็นจะต้องผ่านกรรมวิธีปรับบรรทัดฐานเชิงเวลา เพื่อให้ได้จำนวนข้อมูลที่จะป้อนให้กับนิวรอล เน็ตเวิร์ก คงที่ ดังนั้นการตัดคำ จึงต้องคำนึงถึงข้อสนเทศในส่วนหน้าและหลังของสัญญาณเสียงพูดมากขึ้น เพราะจะส่งผลให้การรู้จำด้วยกรรมวิธีนี้ดีขึ้นได้ ซึ่งกระบวนการกำหนดจุดเริ่มต้นและสิ้นสุด มีดังนี้

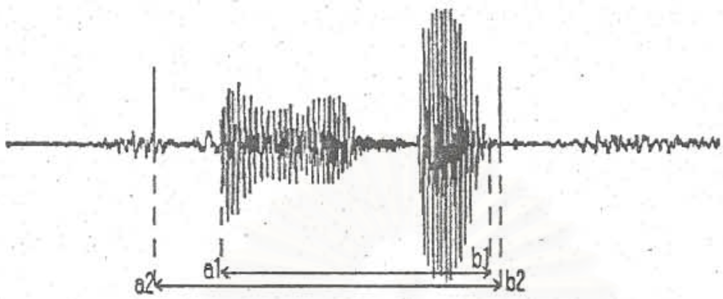
กำหนดให้ $E_{\max}$ เป็นค่าพลังงานในส่วนย่อยที่มีค่าพลังงานสูงสุดในเสียงพูดที่กำลังพิจารณา จากนั้นกำหนดระดับพลังงานอ้างอิง 2 ระดับเพื่อใช้นับจำนวนพยางค์โดยข้อมูลที่นับว่าเป็นพยางค์จะต้องมีคุณสมบัติทั้ง 2 ข้อ ดังนี้

(1) มีส่วนย่อยที่มีพลังงานมากกว่า a เท่าของ $E_{\max}$ เป็นจำนวน m ส่วนย่อยติดกันขึ้นไป

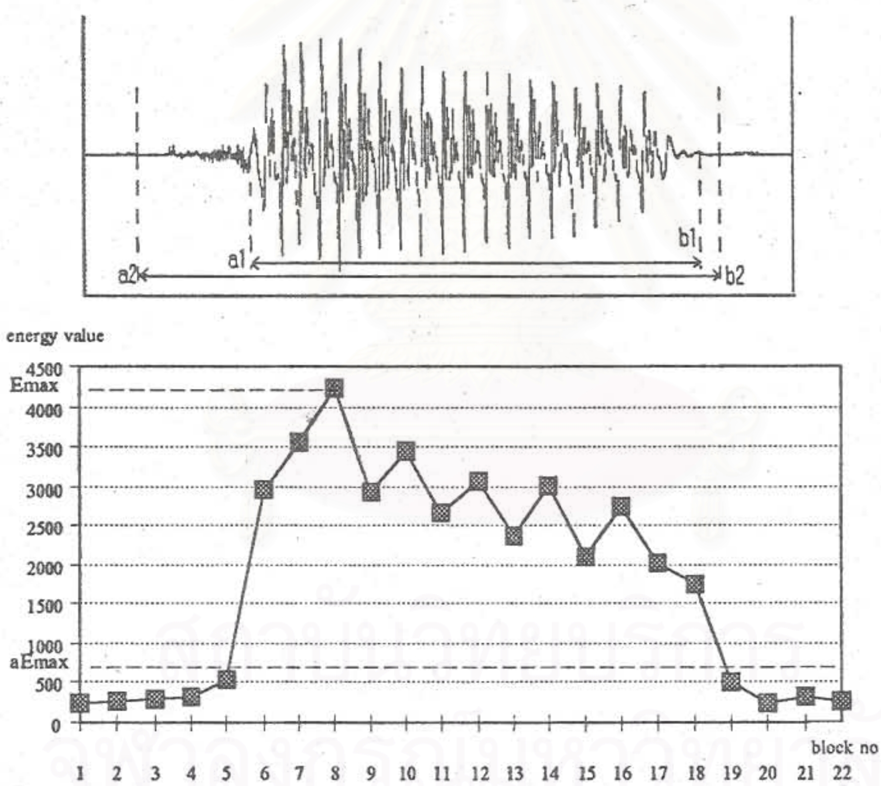
(2) มีส่วนย่อยที่มีพลังงานมากกว่า b เท่าของ $E_{\max}$ เป็นจำนวน n ส่วนย่อยติดกันขึ้นไป

เมื่อ $b > a$ และ $m \geq n$

ค่าระดับพลังงานในคุณสมบัติข้อ (1) เป็นระดับพลังงานเริ่มต้นที่นับว่าเป็นพยางค์ ใช้เพื่อแยกส่วนที่เป็นเสียงพูดออกจากเสียงสภาพแวดล้อมในขณะบันทึกเสียง ส่วนค่าระดับพลังงานในคุณสมบัติข้อ (2) เป็นระดับพลังงานที่ยอมรับว่าเป็นพยางค์จริง ๆ ใช้เพื่อแยกเสียงพูดจากเสียงรบกวนที่มีระดับความดังสูง เช่น เสียงหายใจหรือเสียงเคาะไมโครโฟน ตัวอย่างของเสียงรบกวนประเภทนี้แสดงในรูปที่ 2.6 การใช้ระดับพลังงานเป็นจำนวนเท่าของพลังงานในส่วนย่อยที่มีค่ามากสุด ทำให้โปรแกรมนี้สามารถทำงานได้ถูกต้องแม้ว่าเสียงพูด มีระดับความดังต่างกัน



รูปที่ 2.6 รูปคลื่นของคำสองพยางค์ที่มีเสียงรบกวน



รูปที่ 2.7 รูปคลื่นและพลังงานของคำพยางค์เดียว

จากรูปที่ 2.7 ซึ่งเป็นรูปคลื่นของคำพยางค์เดียว ส่วนของคำจะเริ่มจากจุดเริ่มต้นของพยางค์แรกจนถึงจุดสิ้นสุดของพยางค์สุดท้ายที่ตรวจพบ นั่นคือระยะจากจุด a_1 ถึง b_1 จากนั้นทำการเลื่อนส่วนหัวและส่วนท้ายคำออกไปเท่ากับคาบเวลาคงที่ค่าหนึ่ง เพื่อให้ได้รายละเอียดของเสียงพูดส่วนต้นและท้ายคำ โดยที่คาบเวลาในการเลื่อนส่วนหัวอาจไม่เท่ากับคาบเวลาในการเลื่อนส่วนท้าย เสียงพูดจากส่วนหัวจนถึงส่วนท้ายคำคือระยะจากจุด a_2 ถึง b_2 ถูกใช้เป็นสัญญาณเสียง

พูดที่จะทำการวิเคราะห์ต่อไป ส่วนสัญญาณเสียงพูดที่อยู่นอกช่วงนี้จะถือว่าเป็นเสียงของสภาพแวดล้อมและสัญญาณรบกวนซึ่งถูกตัดทิ้ง

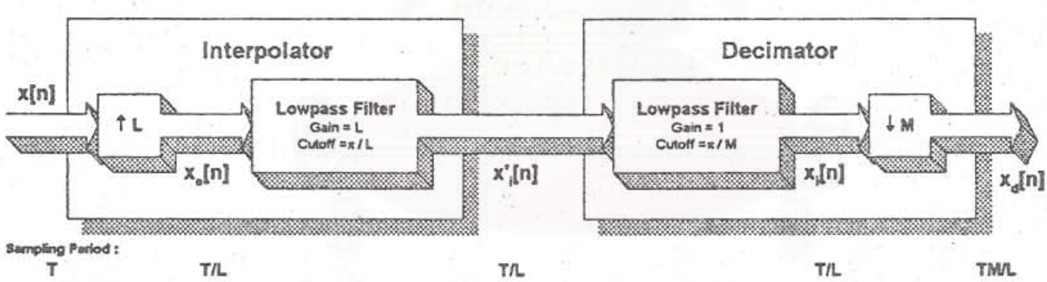
2.1.4 กรรมวิธีปรับบรรทัดฐานเชิงเวลา (Time Normalization)

ขั้นตอนกรรมวิธีปรับบรรทัดฐานเชิงเวลา เป็นขั้นตอนในการเพิ่มหรือลดขนาดความยาวของสัญญาณในเชิงเวลา เพื่อปรับแต่งขนาดความยาวของสัญญาณให้เหมาะสม เนื่องจากสัญญาณเสียงพูดที่ได้จากเสียงพูดของแต่ละบุคคลมีความยาวไม่เท่ากัน จึงจำเป็นต้องมีการปรับขนาดความยาวของสัญญาณในเชิงเวลาให้มีขนาดเท่าที่กำหนดเพื่อใช้ในการหาค่าลักษณะสำคัญและเปรียบเทียบสัญญาณเสียงต่อไป

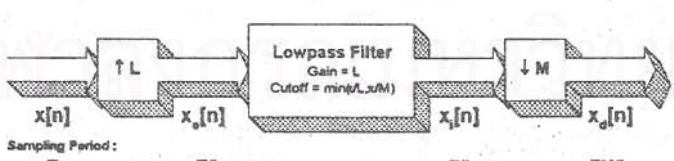
กรรมวิธีในการปรับบรรทัดฐานเชิงเวลา จะอาศัยขั้นตอนในการลดหรือเพิ่มอัตราการสุ่มตัวอย่าง (Sampling Rate) หรือความถี่ในการสุ่มตัวอย่าง (Sampling Frequency) หรือคาบเวลาในการสุ่มตัวอย่าง (Sampling Period) เพื่อปรับขนาดความยาวของสัญญาณในเชิงเวลาให้เป็นไปตามที่ต้องการ เนื่องจากการเปลี่ยนแปลงอัตราการสุ่มตัวอย่างของเสียงพูดนั้นเป็นการเปลี่ยนแปลงแบบไม่เป็นจำนวนเต็ม ดังนั้นการเปลี่ยนแปลงอัตราการสุ่มจึงต้องเพิ่มอัตราการสุ่มให้สูงขึ้นเป็นจำนวน L เท่า จากนั้นจึงจะลดอัตราการสุ่มลงเป็นจำนวน M เท่าตามต้องการ และเพื่อเป็นการป้องกันไม่ให้ข้อมูลเกิดการสูญหายเมื่อเปรียบเทียบกับการลดอัตราสุ่มลงเพียงอย่างเดียว โดยจำนวนเท่าของการเปลี่ยนแปลงจะเป็นอัตราส่วนระหว่างจำนวนเท่าของข้อมูลเสียงพูดที่ลดลงกับจำนวนเท่าของข้อมูลเสียงพูดที่เพิ่มขึ้นดังนี้ (Oppenheim and Schafer, 1989)

$$T' = \frac{M}{L} T \qquad (2.7)$$

เมื่อ T เป็นคาบเวลาในการสุ่มตัวอย่างของสัญญาณที่ต้องการเปลี่ยนแปลงความถี่ในการสุ่มตัวอย่าง T' เป็นคาบเวลาในการสุ่มตัวอย่างของสัญญาณที่ได้รับการเปลี่ยนแปลง M เป็นจำนวนเท่าของอัตราการสุ่มตัวอย่างที่เพิ่มขึ้น และ L เป็นจำนวนเท่าของอัตราการสุ่มตัวอย่างที่ลดลง โดยมีขั้นตอนกรรมวิธีดังแสดงในรูปที่ 2.8 และ 2.9 ดังนี้ (Oppenheim and Schafer, 1989)



รูปที่ 2.8 ขั้นตอนกรรมวิธีการเปลี่ยนแปลงอัตราการสุ่มตัวอย่างแบบไม่เป็นจำนวนเต็ม



รูปที่ 2.9 ขั้นตอนกรรมวิธีการเปลี่ยนแปลงอัตราการสุ่มตัวอย่างเมื่อรวมตัวกรองแบบผ่านต่ำไว้ด้วยกัน

ขั้นตอนกรรมวิธีการเปลี่ยนแปลงอัตราการสุ่มตัวอย่างแบบไม่เป็นจำนวนเต็มดังแสดงในรูปที่ 2.8 และ 2.9 นั้น เป็นการเพิ่มอัตราการสุ่มตัวอย่างให้สูงขึ้น L เท่าซึ่งเรียกว่า Upsamping หรือ Interpolation จากนั้นจึงทำการลด

อัตราการสุ่มตัวอย่างลง M เท่าซึ่งเรียกว่า Downsampling หรือ Decimation โดยมีรายละเอียดดังนี้ (Oppenheim and Schafer, 1989)

2.1.4.1. การเพิ่มอัตราการสุ่มตัวอย่าง (Upsampling)

การเพิ่มอัตราการสุ่มตัวอย่าง เริ่มจากการสุ่มตัวอย่างสัญญาณที่ต่อเนื่องทางเวลา (Continuous - time Signal) $x_c(nT)$ เมื่อ $x[n]$ เป็นสัญญาณที่สุ่มได้ ดังนี้

$$x[n] = x_c(nT) \qquad (2.8)$$

เมื่อทำการเพิ่มอัตราการสุ่มตัวอย่างขึ้น L เท่าด้วยการลดขนาดของคาบเวลาในการสุ่มตัวอย่าง T ลง L เท่า โดยที่ $T' = T/L$ จะได้ว่า

$$x_i[n] = x_c(nT') \qquad (2.9)$$

ดังนั้น

$$x_i[n] = x[n/L] = x_c(nT/L), \qquad n = 0, \pm L, \pm 2L, K \qquad (2.10)$$

เมื่อ $x_i[n]$ เป็นสัญญาณที่ได้รับการเพิ่มอัตราการสุ่มตัวอย่าง และเพื่อเป็นการลดผลของการเคลือบแฝง (Aliasing) ของสัญญาณอันเนื่องมาจากจำนวนตัวอย่างที่เพิ่มขึ้น จึงจำเป็นต้องทำการกรองสัญญาณด้วยตัวกรองแบบผ่านต่ำ (Lowpass Filter) ที่มีความถี่ตัด (Cutoff Frequency) ที่ตำแหน่ง π/L ภายหลังการเพิ่มอัตราการสุ่มตัวอย่างดังแสดงในรูปที่ 2.8 ดังนั้นระบบการเพิ่มอัตราการสุ่มตัวอย่างจึงประกอบไปด้วยตัวเพิ่มอัตราการสุ่มตัวอย่างและตัวกรองแบบผ่านต่ำ โดยเรียกระบบนี้ว่า Interpolator

2.1.4.2. การลดอัตราการสุ่มตัวอย่าง (Downsampling)

การลดอัตราการสุ่มตัวอย่าง จะเป็นการเปลี่ยนแปลงลำดับการสุ่มตัวอย่างของทั้งสัญญาณที่สุ่มได้ $x[n]$ หรือสัญญาณที่ต่อเนื่องทางเวลา $x_c(nT)$ เพื่อให้เกิดลำดับใหม่ที่ต้องการขึ้นมาด้วยการเพิ่มขนาดของคาบเวลาในการสุ่มตัวอย่าง T เพิ่มขึ้น M เท่า โดยที่ $T' = MT$ และเมื่อ $x_d[n]$ เป็นสัญญาณที่สุ่มได้ดังนี้

$$x_d[n] = x[nM] = x_c(nMT) \qquad (2.11)$$

เพื่อเป็นการลดผลของการเคลือบแฝง (Aliasing) ของสัญญาณอันเนื่องมาจากแถบความถี่ของสัญญาณที่เพิ่มขึ้น จึงจำเป็นต้องทำการลดขนาดแถบความถี่ของสัญญาณด้วยตัวกรองแบบผ่านต่ำ (Lowpass Filter) ที่มีความถี่ตัด (Cutoff Frequency) ที่ตำแหน่ง π/M ก่อนการลดอัตราการสุ่มตัวอย่างดังแสดงในรูปที่ 2.8 ดังนั้นระบบการเพิ่มอัตราการสุ่มตัวอย่างจึงประกอบไปด้วย ตัวกรองแบบผ่านต่ำและตัวลดอัตราการสุ่มตัวอย่าง โดยเรียกระบบนี้ว่า Decimator

ดังนั้นในทางปฏิบัติ การเพิ่มหรือลดอัตราการสุ่มตัวอย่างแบบไม่เป็นจำนวนเต็ม จะใช้ระบบดังแสดงในรูปที่ 2.9 โดยรวมตัวกรองแบบผ่านต่ำของระบบ Interpolator และระบบ Decimator เข้าด้วยกันเป็นตัวเดียว โดยมีเงื่อนไขในการเลือกความถี่ตัดตามขนาดของ M และ L ที่มีค่าน้อยที่สุด

2.2. การวิเคราะห์และวัดค่าลักษณะสำคัญ (Feature Measurement)

เป็นเทคนิคการลดจำนวนข้อมูล โดยที่ข้อมูลจำนวนมากจะถูกแปลงเป็นชุดของข้อมูลที่มีจำนวนน้อยลง และยังคงแสดงคุณสมบัติสำคัญของรูปคลื่นสัญญาณเสียงได้อย่างถูกต้อง โดยทั่วไปสัญญาณเสียงถูกวิเคราะห์โดยใช้ลักษณะเด่นเชิงสเปกตรัม (spectral feature) เพราะลักษณะเด่นส่วนใหญ่สำหรับการรับรู้เสียงพูดโดยหูของมนุษย์ รวมอยู่ในข้อมูลเชิงสเปกตรัม วิธีการสกัดเอนเวโลปเชิงสเปกตรัม (spectral envelope) แบ่งออกเป็นการวิเคราะห์โดยใช้พารามิเตอร์ (parametric analysis) และการวิเคราะห์โดยไม่ใช้พารามิเตอร์ (nonparametric analysis) การวิเคราะห์โดยใช้พารามิเตอร์จะเลือกแบบจำลองที่เหมาะสมกับสัญญาณ และปรับแต่งพารามิเตอร์ลักษณะเด่นที่ใช้แทนแบบจำลองนั้น ในขณะที่การวิเคราะห์โดยไม่ใช้พารามิเตอร์สามารถประยุกต์ใช้กับสัญญาณหลายชนิดได้เพราะการวิเคราะห์วิธีนี้ไม่ได้สร้างแบบจำลองสัญญาณ ถ้าแบบจำลองที่ใช้มีความเหมาะสมกับสัญญาณ การวิเคราะห์โดยใช้พารามิเตอร์จะสามารถแสดงลักษณะเด่นของสัญญาณได้ดีกว่า

ก) การวิเคราะห์โดยไม่ใช้พารามิเตอร์ มีวิธีการหลัก ๆ ดังนี้

1. ชุดวงจรกรองผ่านแถบ (band-pass filter bank) วิธีนี้นำสัญญาณเสียงมาผ่านวงจร กรองผ่านแถบหลายวงจร ที่มีช่วงความถี่ผ่านแตกต่างกัน วงจรกรองผ่านแถบแต่ละวงจรจะให้สัญญาณเอาต์พุตที่สัมพันธ์กับพลังงานของสัญญาณในช่วงความถี่ผ่านของวงจรกรองนั้น วิธีนี้มีข้อดีคือสร้างเป็นฮาร์ดแวร์ได้ง่ายและเหมาะสมสำหรับการประมวลผลเวลาจริง (real-time processing)

2. การวิเคราะห์การตัดค่าศูนย์ (zero-crossing analysis) จะนับจำนวนการเปลี่ยนเครื่องหมายของสัญญาณ ซึ่งเป็นการประมาณค่าความถี่ฟอร์แมนท์ (formant frequency) คือ ความถี่ที่มีพลังงานสูงสุด การวิเคราะห์วิธีนี้มักใช้ร่วมกับวิธีชุดวงจรกรองผ่านแถบ

3. การวิเคราะห์โดยใช้เซ็ปสตรัม (cepstrum) การวิเคราะห์วิธีนี้มีข้อดีคือ สามารถแยกเอนเวโลปเชิงสเปกตรัมและโครงสร้างย่อยเชิงสเปกตรัม (spectral fine structure) ออกจากกันได้ในโดเมนคิวเฟร็นซี่ (quefrency domain) ซึ่งเป็นพารามิเตอร์ในโดเมนเวลา แต่มีข้อเสียคือต้องคำนวณผลการแปลงฟูริเยร์แบบเร็ว (fast fourier transform) 2 ครั้ง และคำนวณค่าลอการิทึม ซึ่งต้องใช้เวลาในการคำนวณมาก Chantawekul, and Jitapunkul, 1993 ได้แสดงให้เห็นว่าสามารถนำผลการแปลงฮาร์ตเลย์ใช้งานแทนผลการแปลงฟูริเยร์ ในการวิเคราะห์สัญญาณเสียงพูดในเชิงความถี่ได้เป็นอย่างดี และผลการแปลงฮาร์ตเลย์มีอัลกอริทึมอย่างเร็วเช่นเดียวกับผลการแปลงฟูริเยร์ แต่ใช้เวลาในการคำนวณน้อยกว่าประมาณครึ่งหนึ่ง

ข) การวิเคราะห์โดยใช้พารามิเตอร์ มีวิธีการหลัก ๆ ดังนี้

1. การวิเคราะห์โดยการสังเคราะห์ (analysis-by-synthesis) วิธีนี้สามารถสร้างแบบจำลองที่ถูกต้องแม่นยำได้ โดยการใช้พารามิเตอร์หลายค่าเช่น ค่าความถี่ฟอร์แมนท์,ความกว้างแถบ(bandwidth), เอนเวโลปเชิงสเปกตรัม และอื่น ๆ แต่มีข้อเสียคือต้องใช้เวลาคำนวณในการวนซ้ำมากเพราะค่าพารามิเตอร์หลายค่ามีผลกระทบต่อกัน

2. การประมาณพันธะเชิงเส้น (linear predictive coding) ซึ่งเป็นวิธีที่ใช้กันแพร่หลายในการหา feature ของเสียง เป็นการสร้างแบบจำลองของสเปกตรัมอย่างง่ายโดยใช้โพล (all-pole spectrum modeling) พารามิเตอร์สามารถประมาณค่าได้จากค่าความแปรปรวนร่วมหรือค่าอัตสหสัมพันธ์ โดยไม่ใช้การวนซ้ำ วิธีนี้มีข้อดีคือ สามารถแทนสัญญาณเสียงได้อย่างมีประสิทธิภาพโดยใช้พารามิเตอร์จำนวนน้อยและใช้การคำนวณที่ค่อนข้างง่าย

หลักการประมาณพันธะเชิงเส้นมีวิธีการหลัก 2 วิธีคือวิธีการหาค่าความแปรปรวนร่วมและวิธีอัตสหสัมพันธ์ เนื่องจากวิธีอัตสหสัมพันธ์ใช้การคำนวณน้อยกว่าวิธีความแปรปรวนร่วม และมีความแน่นอนด้านเสถียรภาพ (Furui,1989) ดังนั้นงานวิจัยนี้จึงเลือกใช้การประมาณพันธะเชิงเส้นโดยวิธีอัตสหสัมพันธ์ การประมาณพันธะเชิงเส้นสามารถแสดงคุณสมบัติได้ใกล้เคียงกับพื้นฐานรูปแบบการกำเนิดเสียงของมนุษย์ (Rabiner and Levinson , 1981)

การวิเคราะห์และวัดค่าลักษณะสำคัญเป็นการวิเคราะห์สัญญาณเสียงพูด เพื่อเก็บรวบรวมลักษณะสำคัญของเสียงพูดแต่ละเสียง สำหรับการฝึกฝนระบบให้รับรู้ถึงความแตกต่างของเสียงพูดแต่ละเสียงและเพื่อใช้ในการเปรียบเทียบแบ่งแยกความแตกต่างของเสียงพูดแต่ละเสียงออกจากกัน ในการวัดค่าลักษณะสำคัญของกรรมวิธีไดนามิก ไทม์วาร์ปิง จะใช้ผลการแปลงฮาร์ตเลย์วิเคราะห์สัญญาณเสียงพูดในเชิงความถี่เพื่อหาค่าพารามิเตอร์หรือลักษณะสำคัญ ในขณะที่กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ จะใช้วิธีการหาสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น ส่วนการวิเคราะห์ลักษณะสำคัญจะอาศัยขั้นตอนการควอนไทซ์แบบเวกเตอร์เป็นหลัก ส่วยกรรมวิธีนิวรอลเน็ตเวิร์ก จะใช้เพียงวิธีการหาสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นเท่านั้น

2.2.1. ผลการแปลงฮาร์ตเลย์ (Hartley Transform)

Chantawekul, and Jitapunkul (1993) และ สุนิสา จันทวีกูล (2536) ได้แสดงให้เห็นว่าสามารถนำผลการแปลงฮาร์ตเลย์ ไปใช้วิเคราะห์สเปกตรัมและสเปกตรัมกำลังของสัญญาณที่แปรเปลี่ยนตามเวลาและไม่เสถียรได้เป็น

อย่างดี โดยใช้เวลาในการประมวลผลน้อยกว่าผลการแปลงฟูริเยร์ประมาณครึ่งหนึ่ง ในงานวิจัยนี้มีการนำผลการแปลงฮาร์ตเลย์แบบดีสครีต ดังสมการที่ (2.11) มาวิเคราะห์หาพารามิเตอร์สำหรับกรรมวิธีรู้จำแบบไดนามิก ไทม์วาร์ปิง

$$H[k] = \sum_{n=0}^{N-1} x[n]cas(2\pi nk / N) \quad , \quad 0 \leq n \leq N-1 \quad \text{.........} (2.12)$$

โดยที่ $cas(\theta) \underset{=}{\Delta} cos(\theta) + sin(\theta)$

$H[k]$ เป็นผลการแปลงฮาร์ตเลย์แบบดีสครีตของ x[n]

$x[n]$ เป็นค่าตัวอย่างของข้อมูลที่จะหาผลการแปลง

N แทนจำนวนตัวอย่างของข้อมูลที่จะนำมาวิเคราะห์

k เป็นลำดับของผลการแปลงฮาร์ตเลย์แบบดีสครีต

โดยการแบ่งข้อมูลเสียง x[n] ออกเป็นกรอบ ๆ แต่ละกรอบมีจำนวนค่าสุ่ม N ค่า และกรอบที่อยู่ติดกันจะมีค่าสุ่มเหลื่อมกัน N/2 ค่า ค่าในแต่ละกรอบจะถูกวางกรอบด้วยฟังก์ชันกรอบชนิด Hamming Window จากนั้นจึงหาค่าผลการแปลงฮาร์ตเลย์แบบดีสครีต เพื่อนำไปใช้เป็นพารามิเตอร์ในการรู้จำต่อไป ดังสมการที่ (2.12)

$$P_i[k] = |H_i[k]| \quad \text{.........} (2.13)$$

เมื่อ i แทนหมายเลขเฟรม $i = 0,1,2,\ldots,I-1$

k แทนลำดับจะอยู่ในช่วง 0 ถึง N-1

ค่าของ $P_i[k]$ นี้จะเป็นค่าสมบูรณ์ (abslute value) ของผลการแปลงฮาร์ตเลย์แบบดีสครีต ซึ่งจะแทนข้อมูลของความถี่เสียงพูดในช่วง 0-4 kHz ข้อมูลเสียงที่ได้นี้ จะทำการปรับบรรทัดฐาน ด้วยค่าของค่าเฉลี่ยของสัญญาณดังในสมการที่ (2.13) เพื่อที่จะปรับสัญญาณให้เหมาะสมต่อการนำมาเปรียบเทียบ

$$P_{av} = \frac{\sum_{i=0}^{I-1}\sum_{k=0}^{N_p} P_i[k]}{I.N_p} \quad \text{.........} (2.14)$$

ค่าของ P_{av} จะเป็นค่าเฉลี่ยของสัญญาณที่ผ่านการแปลงข้อมูล ซึ่งจำนำค่าของ P_{av} ไปใช้ในการหาค่าที่ปรับบรรทัดฐานของสัญญาณดังแสดงในสมการที่ (2.14)

$$\widetilde{P}_i[k] = \frac{P_i[k]}{P_{av}} \quad \text{.........} (2.15)$$

การวัด distance ของพารามิเตอร์ที่ได้จะเป็นดังสมการที่ (2.15)

$$d(i,j) = \sum_{n=0}^{K-1}(a_{in} - b_{jn})^2 \quad \text{.........} (2.16)$$

โดยที่ $d(i,j)$ เป็น distance ของกรอบที่ i และ j

a_{in} เป็นพารามิเตอร์ของเสียงทดสอบที่กรอบที่ i

b_{jn} เป็นพารามิเตอร์ของเสียงอ้างอิงที่กรอบที่ j

K เป็นจำนวนพารามิเตอร์ที่ใช้ในการเปรียบเทียบ

2.2.2. สัมประสิทธิ์ของการประมาณพันธะเชิงเส้น (Linear Prediction Coefficient)

คำว่า "การประมาณพันธะเชิงเส้น" หรือ Linear Prediction ถูกนำเสนอเป็นครั้งแรกโดย N. Weiner ในปี ค.ศ. 1966 โดยเทคนิคนี้ถูกนำมาใช้เป็นครั้งแรกกับการวิเคราะห์และการสังเคราะห์เสียงโดย Itakura กับ Saito และ Atal กับ Schroeder ในปี ค.ศ. 1968 (Furui, 1991) ความสำคัญของเทคนิคการประมาณพันธะเชิงเส้นนี้ก็คือ การที่รูปคลื่นและลักษณะสมบัติทางความถี่ของเสียงพูดสามารถแสดงด้วย ค่าพารามิเตอร์เพียงไม่กี่ค่าได้อย่างแม่นยำ และมีประสิทธิภาพ นอกจากนี้ค่าพารามิเตอร์ดังกล่าวยังสามารถคำนวณได้ง่ายอีกด้วย

กำหนดให้สัญญาณเสียงที่ถูกสุ่มอย่างไม่ต่อเนื่องทุกเวลา ΔT วินาที แทนได้ด้วย $\{x(t)\}$ เมื่อ t เป็นจำนวนเต็ม และ $\Delta T \leq 1/2W$ วินาทีเมื่อความถี่ของสัญญาณเสียงอยู่ในช่วง 0 - W เฮิรตซ์ ดังนั้นความสัมพันธ์ระหว่างสัญญาณสุ่ม $x(t)$ กับสัญญาณสุ่มก่อนหน้านี้ p ค่าแสดงได้เป็น (Furui, 1991)

$$x_t + \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} = \varepsilon_t \qquad (2.17)$$

เมื่อ $\{\varepsilon_t\}$ เป็นตัวแปรทางสถิติชนิดอสหสัมพันธ์ ที่มีค่าเฉลี่ยเป็นศูนย์และมีค่าความแปรปรวนเป็น σ^2 ซึ่งจะสามารถประมาณสัญญาณสุ่มค่าถัดไปเมื่อทราบค่าก่อนหน้านี้ p ค่าได้เป็น

$$\hat{x}_t = -\sum_{i=1}^{p} \alpha_i x_{t-i} \qquad (2.18)$$

จากสมการที่ (2.17) และ (2.18) จะได้ว่า

$$x_t - \hat{x}_t = \varepsilon_t \qquad (2.19)$$

ดังนั้นจากสมการที่ (2.17) จะได้ว่า $\{\alpha_i\}$ เป็นสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นและ ε_t เป็นค่าความผิดพลาดตกค้างดังสมการที่ (2.19)

กำหนดคุณสมบัติของตัวกรองการประมาณพันธะเชิงเส้นให้เป็น

$$F(z) = -\sum_{i=1}^{p} \alpha_i z^{-i} \qquad (2.20)$$

กำหนดให้ $\hat{X}(z) \leftrightarrow \hat{x}_t$ และ $X(z) \leftrightarrow x_t$ เป็นคู่การแปลง Z จะได้ว่า

$$\hat{X}(z) = F(z)\,X(z) \qquad (2.21)$$

ดังนั้นจากสมการที่ (2.18) และ (2.19) รูปแบบของการประมาณพันธะเชิงเส้นในรูปของการแปลง Z จะเขียนได้เป็น

$$X(z)\big(1 - F(z)\big) = E(z) \qquad (2.22ก)$$

หรือ

$$X(z)\,A(z) = E(z) \qquad (2.22ข)$$

เมื่อ

$$A(z) = 1 + \sum_{i=1}^{p} \alpha_i z^{-i} = 1 - F(z) \qquad (2.23)$$

วิธีการประมาณค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น $\{\alpha_i\}$ สามารถกระทำได้โดยใช้วิธีการค่าความผิดพลาดกำลังสองเฉลี่ยน้อยที่สุด (Least Mean Square Error) กับสมการ (2.19) เมื่อพิจารณาในช่วง $[t_0, t_1]$ จะได้ค่าความเพี้ยนกำลังสองรวม β มีค่าเป็น

$$\begin{aligned}\beta &= \sum_{t=t_0}^{t_1} \varepsilon_t^2 = \sum_{t=t_0}^{t_1} \left(\sum_{i=0}^{p} \alpha_i x_{t-i} \right)^2 \\ &= \sum_{t=t_0}^{t_1} \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i \alpha_j x_{t-i} x_{t-j}\end{aligned} \qquad (2.24)$$

เมื่อ $\alpha_0 = 1$ กำหนดให้

$$c_{ij} = \sum_{t=t_0}^{t_1} x_{t-i} x_{t-j} \qquad (2.25)$$

ดังนั้น β สามารถเขียนได้เป็น

$$\beta = \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i c_{ij} \alpha_j \qquad (2.26)$$

ทำการหาค่าต่ำสุดของ β ด้วยการหาอนุพันธ์ย่อยของ β เทียบกับ α_j $(j = 1,2,\ldots,p)$ โดยกำหนดให้เป็นศูนย์ ดังนั้นจากสมการ (2.26) จะได้เป็น

$$\frac{\partial \beta}{\partial \alpha_j} = 2\sum_{i=0}^{p} \alpha_i c_{ij} = 0 \quad \left(j = 1,2,\mathrm{K}\ ,p \right) \qquad (2.27)$$

ดังนั้นการหาค่าสัมประสิทธิ์ของการประมาณ $\{\alpha_i\}$ ก็คือการหาคำตอบของชุดสมการเชิงเส้น p สมการ เมื่อทราบค่า c_{ij} $\left(i = 0,1,2,\mathrm{K}\ ,p; j = 1,2,\mathrm{K}\ ,p\right)$ ได้โดยหาจากสมการ (2.25) ซึ่งแสดงว่าค่า x_t จาก $t_0 - p$ ถึง t_1 มีนัยสำคัญ

ในการหาคำตอบของลำดับเสียงพูด N ค่า สามารถหาได้ 2 วิธีคือ วิธีการแปร ปรวนร่วม (Covariance Method) และวิธีการอัตสหสัมพันธ์ (Autocorrelation Method) แต่ในที่นี้จะกล่าวถึงเพียงวิธีการอัตสัมพันธ์เท่านั้น โดยวิธีการอัตสหสัมพันธ์จะกำหนดให้เวลาอยู่ในช่วง $t_0 = -\infty$ และ $t_1 = \infty$ โดยกำหนดให้ $x_t = 0$ เมื่อ $t < 0$ และ $t \geq N$ จะได้ว่า

$$\begin{aligned} c_{ij} &= \sum_{t=-\infty}^{\infty} x_{t-i} x_{t-j} = \sum_{t=-\infty}^{\infty} x_t x_{t+|i-j|} \\ &= \sum_{t=0}^{N-1-|i-j|} x_t x_{t+|i-j|} = r_{|i-j|} \end{aligned} \qquad (2.28)$$

ดังนั้นค่า α_i จึงสามารถหาได้จาก

$$\sum_{i=0}^{p} \alpha_i r_{|i-j|} = 0 \quad \left(j = 1,2,\mathrm{K}\ ,p \right) \qquad (2.29)$$

เมื่อ

$$r_\tau = \sum_{t=0}^{N-1-\tau} x_t x_{t+\tau} \quad (\tau \geq 0) \qquad (2.30)$$

โดยที่สมการ (2.30) สามารถเขียนในรูปสมการเมตริกซ์ได้ดังนี้

$$\begin{bmatrix} r_0 & r_1 & \Lambda & r_{p-1} \\ r_1 & r_0 & & \mathrm{M} \\ \mathrm{M} & & \mathrm{O} & r_1 \\ r_{p-1} & \Lambda & r_1 & r_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \mathrm{M} \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ \mathrm{M} \\ r_p \end{bmatrix} \qquad (2.31)$$

ก) การหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น

จากการหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นโดยอาศัยอัตสหสัมพันธ์นั้น เริ่มต้นจากกำหนดให้ E เป็นค่าพลังงานของความผิดพลาด เมื่อ $e(n)$ เป็นความผิดพลาดตกค้างที่สัมพันธ์กับสัญญาณ $x(n)$

$$\begin{aligned} E &= \sum_{-\infty}^{\infty} e^2(n) \\ &= \sum_{-\infty}^{\infty} \left[x(n) - \sum_{k=1}^{p} a_k x(n-k) \right]^2 \end{aligned} \qquad (2.32)$$

การหาค่า a_k ที่ทำให้ E มีค่าต่ำที่สุดสามารถกระทำได้โดยการกำหนดให้ $\partial E / \partial a_k = 0$ สำหรับแต่ละ $k = 1,2,3,\mathrm{K}\ ,p$ จะได้สมการเชิงเส้น p สมการ ซึ่งก็คือ a_k ที่ไม่ทราบค่า p ตัวดังนี้

$$\sum_{-\infty}^{\infty} x(n-i) x(n) = \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} x(n-i) \cdot x(n-k) \qquad (2.33)$$

เนื่องจากพจน์แรกเป็นอัตสหสัมพันธ์ $R(i)$ ของ $x(n)$ และ $x(n)$ มีความยาวจำกัดจะได้

$$\sum_{k=1}^{p} a_k R(i-k) = R(i), \quad 1 \leq i \leq p \qquad (2.34)$$

เมื่อ

$$R(i) = \sum_{n=i}^{N-1} x(n) x(n-i) \qquad (2.35)$$

เมื่อเขียนในรูปของสมการเมตริกซ์จะได้เป็น $\mathbf{Ra} = \mathbf{r}$ ตามสมการที่ (2.31) และเนื่องจากเมตริกซ์ $\mathbf{R}$ เป็นชนิด Toepliz เมตริกซ์ซึ่งมีค่าในแนวทแยงเท่ากันทั้งหมดทุกแนว ดังนั้นการคำนวณหาค่า a_k ในเมตริกซ์ $\mathbf{a}$ จะอาศัยขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin (O'Shaughnessy, 1988) เข้ามาช่วย

ข) ขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin

ขั้นตอนวิธีการนี้เป็นเทคนิคที่ช่วยในการคำนวณหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น a_m ในเมตริกซ์ $\mathbf{a}$ เมื่อ $m = 1,2,K\ ,p$ และ p เป็นลำดับ (Order) ของการวิเคราะห์หาสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น ซึ่งขั้นตอนวิธีการนี้จะให้คำตอบที่มีความเสถียรและแน่นอน โดยมีขั้นตอนแสดงในตารางที่ 2.1 ดังนี้

ตารางที่ 2.1 รายละเอียดขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin

---

ขั้นตอนที่ 1 **เงื่อนไขเริ่มต้น**

$$E_0 = R(0), \quad a_0 = 0 \qquad (2.36)$$

ขั้นตอนที่ 2 **สัมประสิทธิ์การสะท้อน**

$$k_m = \frac{R(m) - \sum_{i=1}^{m-1} a_{m-1}(i) R(m-i)}{E_{m-1}}, \quad |k_m| \le 1 \qquad (2.37)$$

ขั้นตอนที่ 3 **สัมประสิทธิ์ของการประมาณพันธะเชิงเส้นในแต่ละรอบของการคำนวณ**

$$\begin{aligned} a_m(m) &= k_m \\ a_m(i) &= a_{m-1}(i) - k_m a_{m-1}(m-i), \quad (1 \le i < m) \\ E_m &= \left(1 - k_m^2\right) E_{m-1} \end{aligned} \qquad (2.38)$$

ขั้นตอนที่ 4 **สมการอัตสหสัมพันธ์**

$$R(i) = \sum_{n=i}^{N-1} x(n) x(n-i), \quad i = 1,2,K\ ,p \qquad (2.39)$$

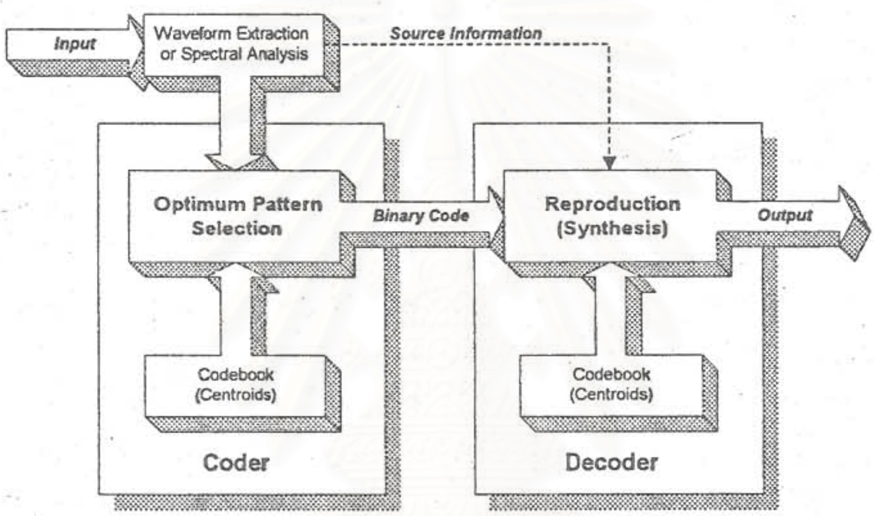
ขั้นตอนที่ 5 **สัมประสิทธิ์ของของการประมาณพันธะเชิงเส้นเมื่อสิ้นสุดการคำนวณ**

$$a_m = a_m(p), \quad 1 \le m \le p \qquad (2.40)$$

2.2.3. การควอนไทซ์แบบเวกเตอร์ (Vector Quantization)

การควอนไทซ์แบ่งได้เป็น 2 ประเภท (Makhoul, Roucos, and Gish, 1985) ได้แก่การควอนไทซ์แบบสเกลาร์และการควอนไทซ์แบบเวกเตอร์ การควอนไทซ์แบบสเกลาร์นั้นชุดของพารามิเตอร์หรือลำดับของสัญญาณจะถูกควอนไทซ์แยกจากกัน ส่วนการควอนไทซ์แบบเวกเตอร์นั้นชุดของพารามิเตอร์จะถูกควอนไทซ์รวมกันเป็นเวกเตอร์เดียว จุดประสงค์ของการควอนไทซ์แบบเวกเตอร์ก็เพื่อการลดขนาดจำนวนข้อมูลลง ซึ่งถือได้ว่าเป็นการบีบอัดข้อมูลวิธีการหนึ่ง เมื่อนำมาประยุกต์ใช้งานกับเสียงพูดจึงเรียกว่าเป็นการบีบอัดเสียงพูดหรือการเข้ารหัสเสียงพูดนั่นเอง

ในงานวิจัยด้วยกรรมวิธีฮิดเดน มาร์คอฟ นี้ได้นำเทคนิคการควอนไทซ์แบบเวกเตอร์มาใช้ในการควอนไทซ์ข้อมูลสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นของเสียงพูด เพื่อลดขนาดของข้อมูลให้เหลือน้อยลงโดยมีความเพี้ยนน้อยที่สุด ด้วยการแทนที่เวกเตอร์ของสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นจากสัญญาณเสียงพูด ซึ่งตัดออกมาพิจารณาด้วยเวกเตอร์ค่าหนึ่งที่ใกล้เคียงมากที่สุด โดยอาศัยการวัดความเพี้ยนระหว่างเวกเตอร์ทั้งสองให้มีค่าน้อยที่สุด



รูปที่ 2.10 รายละเอียดขั้นตอนการควอนไทซ์แบบเวกเตอร์ (Sadaoki Furui, 1989)

รายละเอียดขั้นตอนการควอนไทซ์แบบเวกเตอร์ดังแสดงในรูปที่ 2.10 เริ่มต้นจากการที่สัญญาณเสียงได้รับการเปรียบเทียบกับชุดรหัสที่มีอยู่ โดยการคำนวณหาค่าความเพี้ยนแล้วพิจารณาชุดรหัสที่ให้ค่าความเพี้ยนน้อยที่สุด สัญญาณเสียงดังกล่าวจะถูกแทนที่ด้วยชุดรหัสนั้น

ทฤษฎีของขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์ เริ่มจากสมมติให้เวกเตอร์ขนาด N มิติ $\mathbf{x} = [x_1 \quad x_2 \quad \Lambda \quad x_N]^T$ ซึ่งมีส่วนประกอบ $\{x_k,\ 1 \le k \le N\}$ เป็นตัวแปรสุ่มซึ่งมีแอมปลิจูดต่อเนื่องและเป็นจำนวนจริง ในการควอนไทซ์แบบเวกเตอร์นั้นเวกเตอร์ $\mathbf{x}$ จะถูกจับคู่ให้ตรงกับเวกเตอร์ $\mathbf{y}$ ซึ่งมีแอมปลิจูดไม่ต่อเนื่องและเป็นจำนวนจริง หรือเรียกได้ว่า $\mathbf{x}$ ถูกควอนไทซ์เป็น $\mathbf{y}$ และ $\mathbf{y}$ เป็นค่าควอนไตซ์ของ $\mathbf{x}$ นั่นคือ

$$\mathbf{y} = q(\mathbf{x}) \qquad (2.41)$$

เมื่อ $q(\cdot)$ เป็นตัวดำเนินการควอนไทซ์ นอกจากนี้ $\mathbf{y}$ ยังอาจเรียกได้ว่า "เวกเตอร์สร้างใหม่" หรือ "เวกเตอร์ขาออก" ที่สัมพันธ์กับ $\mathbf{x}$ โดยทั่วไปแล้ว $\mathbf{y}$ เป็นเพียงหนึ่งในชุดของ $\mathbf{Y} = \{\mathbf{y}_i,\ 1 \le i \le L\}$ เมื่อ $\mathbf{y}_i = [y_{i1} \quad y_{i2} \quad \Lambda \quad y_{iN}]^T$ ซึ่งชุด $\mathbf{Y}$ เป็นชุดรหัสที่สร้างใหม่หรือชุดรหัส L เป็นขนาดของชุดรหัส และ $\mathbf{y}_i$ เป็นชุดของเวกเตอร์รหัส ในด้านการรู้จำรูปแบบนั้น $\mathbf{y}_i$ ถือว่าเป็นรูปแบบอ้างอิงหรือชุดรูปร่องต้นแบบ ขนาด L ของชุดรหัสเรียกว่า "จำนวนชั้น" ดังนั้นการออกแบบชุดรหัสจะใช้การแบ่งปริภูมิ N มิติของเวกเตอร์สุ่ม $\mathbf{x}$ ออกเป็น L ภาคหรือ "เซลล์" $\{C_i,\ 1 \le i \le L\}$ และกำหนดความ

สัมพันธ์ระหว่างแต่ละเซลล์ C_i กับเวกเตอร์ $\mathbf{y}_i$ ดังนั้นการควอนไทซ์จะเป็นการกำหนดเวกเตอร์รหัส $\mathbf{y}_i$ ถ้า $\mathbf{x}$ อยู่ใน C_i ตามเงื่อนไขในสมการที่ (2.42)

$$q(\mathbf{x}) = \mathbf{y}_i, \qquad \text{if } \mathbf{x} \in C_i \qquad (2.42)$$

ขั้นตอนการออกแบบชุดรหัสตามเงื่อนไขนี้เรียกว่า "การฝึกฝน" หรือ "การสร้างชุดรหัส" เมื่อ $\mathbf{x}$ ถูกควอนไทซ์เป็น $\mathbf{y}$ จะเกิดความเพี้ยนจากการควอนไทซ์ขึ้น โดยการวัดค่าความเพี้ยน $d(\mathbf{x},\mathbf{y})$ ระหว่าง $\mathbf{x}$ และ $\mathbf{y}$ ซึ่ง $d(\mathbf{x},\mathbf{y})$ ก็คือการวัดความไม่คล้ายคลึงกันหรือการวัดระยะห่าง ดังนั้นความเพี้ยนเฉลี่ยโดยรวมแสดงได้ดังนี้

$$D = \lim_{M\to\infty} \frac{1}{M}\sum_{n=1}^{M} d[\mathbf{x}(n),\mathbf{y}(n)] \qquad (2.43)$$

ถ้า $\mathbf{x}(n)$ ไม่แปรเปลี่ยนตามเวลาและมีความเป็นเออร์กอดิก ดังนั้นความเพี้ยนเฉลี่ยโดยรวมจะถูกจำกัดไว้ดังนี้

$$\begin{aligned} D &= E[d(\mathbf{x},\mathbf{y})] \\ &= \sum_{i=1}^{L} P(\mathbf{x}\in C_i) E[d(\mathbf{x},\mathbf{y}_i) | \mathbf{x}\in C_i] \\ &= \sum_{i=1}^{L} P(\mathbf{x}\in C_i) \int_{\mathbf{x}\in C_i} d(\mathbf{x},\mathbf{y}_i) p(\mathbf{x}) d\mathbf{x} \end{aligned} \qquad (2.44)$$

เมื่อ $P(\mathbf{x}\in C_i)$ เป็นความน่าจะเป็นแบบไม่ต่อเนื่องที่ $\mathbf{x}$ อยู่ใน C_i ส่วน $p(\mathbf{x})$ เป็นฟังก์ชันความหนาแน่นความน่าจะเป็นแบบหลายมิติของ $\mathbf{x}$ และอินทิกรัลจะครอบคลุมส่วนประกอบของเวกเตอร์ $\mathbf{x}$ ทั้งหมด

2.2.3.1. การวัดค่าความเพี้ยน

การวัดค่าความเพี้ยนนั้นเป็นส่วนที่มีความสำคัญมากที่สุดส่วนหนึ่งในระบบการรู้จำคำพูด โดยถูกนำมาใช้ในขั้นตอนการสร้างและฝึกฝนชุดรหัสรวมทั้งในขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์ เนื่องจากความแตกต่างของค่าความเพี้ยนจะเป็นตัวบ่งชี้ถึงความเหมือนหรือความแตกต่างในคุณภาพของเสียงพูด วิธีการวัดค่าความเพี้ยนมีหลายวิธีได้แก่ ค่าความผิดพลาดกำลังสองเฉลี่ย (Mean-Square Error, MSE) ค่าความผิดพลาดกำลังสองเฉลี่ยถ่วงน้ำหนัก (Weighted Mean-Square Error) การวัดค่าความเพี้ยนโดยอาศัยการประมาณพันธะเชิงเส้น (Linear Prediction Distortion Measures) เป็นต้น (Makhoul, 1985) ดังมีรายละเอียดของแต่ละวิธีการดังนี้

ก) ค่าความผิดพลาดกำลังสองเฉลี่ย (Mean-Square Error, MSE)

การวัดค่าความเพี้ยนโดยอาศัยค่าความผิดพลาดกำลังสองเฉลี่ย ถือว่าเป็นวิธีการวัดค่าที่ใช้กันมากที่สุด เนื่องจากความง่ายทั้งในการประยุกต์ใช้งานและในทางคณิตศาสตร์ โดยมีรูปแบบดังนี้

$$d_2(\mathbf{x},\mathbf{y}) = \frac{1}{N}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) = \frac{1}{N}\sum_{k=1}^{N}(x_k - y_k)^2 \qquad (2.45)$$

หน่วยของการวัดค่าความเพี้ยนกำหนดให้มีหน่วยต่อมิติ กรณีทั่วไปของการวัดค่าความเพี้ยนในกรณีนี้บนพื้นฐานของ L_r Norm ดังนี้

$$d_r(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{k=1}^{N}|x_k - y_k|^r \qquad (2.46)$$

จากกรณีทั่วไปของการวัดค่าความเพี้ยนตามสมการที่ (2.36) นั้นจะสมมูลกับสมการที่ (2.45) ได้ในกรณีที่ $r=2$ นอกจากนี้ยังมีค่าที่นิยมใช้กันมากอีก 2 ค่าได้แก่ค่า $r=1$ และ $r=\infty$ โดยที่ d_1 แทนค่าความผิดพลาดสัมบูรณ์เฉลี่ย (Average Absolute Error) และ d_∞ มีแนวโน้มที่จะเข้าสู่ค่าความผิดพลาดสูงสุด ซึ่งแสดงว่า

$$\lim_{r\to\infty} [d_r(\mathbf{x},\mathbf{y})]^{1/r} = \max \{|x_k - y_k|, 1 \le k \le N\} \qquad (2.47)$$

การทำให้ D ที่ $r=\infty$ มีค่าน้อยที่สุด จะเทียบเท่ากับการทำให้ค่าความผิดพลาดที่มากที่สุดในการควอนไทซ์มีค่าน้อยที่สุด ในการเข้ารหัสเสียงพูดนั้น d_2 เป็นวิธีการวัดค่าความเพี้ยนที่ถูกนำมาใช้มากที่สุด ส่วน d_1 และ d_∞ ก็ถูกนำมาใช้บ้างเป็นครั้งคราว

ข) ค่าความผิดพลาดกำลังสองเฉลี่ยถ่วงน้ำหนัก

(Weighted Mean Square Error)

จากค่าความผิดพลาดกำลังสองเฉลี่ย d_2 นั้นกำหนดให้ค่าความเพี้ยนที่เกิดขึ้นโดยการควอนไทซ์ค่าพารามิเตอร์ $\{x_k\}$ ต่างกันมีน้ำหนักสมดุลย์ โดยทั่วไปแล้วน้ำหนักที่ไม่สมดุลย์สามารถกำหนดขึ้นได้โดยการให้ ความสำคัญกับค่าความเพี้ยนค่าหนึ่งมากกว่าค่าอื่น ๆ กรณีทั่วไปของค่าความผิดพลาดกำลังสองเฉลี่ยถ่วงน้ำหนักเป็นดังนี้

$$d_w(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \mathbf{W}(\mathbf{x}-\mathbf{y}) \quad \text{.......... (2.48)}$$

เมื่อ $\mathbf{W}$ เป็นเมตริกซ์ถ่วงน้ำหนักที่เป็นบวกทั้งหมดเสมอ และถ้า $\mathbf{W} = N^{-1}\mathbf{I}$ เมื่อ $\mathbf{I}$ เป็นเมตริกซ์เอกลักษณ์ ผลลัพธ์ที่ได้ก็คือ $d_w = d_2$ ตัวเลือกหนึ่งของ $\mathbf{W}$ ที่นำมาประยุกต์ใช้กับการจำแนกรูปแบบก็คือ $\mathbf{W} = \Gamma^{-1}$ เมื่อ Γ เป็นเมตริกซ์ของความแปรปรวนร่วมของเวกเตอร์สุ่ม $\mathbf{x}$

$$\Gamma = E\left[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T\right], \qquad \mathbf{x} = E[\overline{\mathbf{x}}] \quad \text{.......... (2.49)}$$

ในกรณีนี้ทำให้ d_w ลดรูปลงอยู่ในรูปของระยะทาง Mahalanobis ดังนี้

$$d_w(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Gamma^{-1}(\mathbf{x}-\mathbf{y}) \quad \text{.......... (2.50)}$$

ถ้าเมตริกซ์ $\mathbf{W}$ มีความสมมาตรนอกเหลือจากการเป็นบวกทั้งหมดเสมอ ดังนั้น

$$\mathbf{W} = \mathbf{P}^T\mathbf{P} \quad \text{.......... (2.51)}$$

เวกเตอร์ $\mathbf{x}$ และ $\mathbf{y}$ สามารถแปลงไปเป็นชุดเวกเตอร์ $\tilde{\mathbf{x}}$ และ $\tilde{\mathbf{y}}$ ดังนี้

$$\tilde{\mathbf{x}} = \mathbf{Px} \qquad \tilde{\mathbf{y}} = \mathbf{Py} \quad \text{.......... (2.52)}$$

$$\begin{aligned} d_w(\mathbf{x},\mathbf{y}) &= (\mathbf{Px}-\mathbf{Py})^T(\mathbf{Px}-\mathbf{Py}) \\ &= (\tilde{\mathbf{x}}-\tilde{\mathbf{y}})^T(\tilde{\mathbf{x}}-\tilde{\mathbf{y}}) \quad \text{.......... (2.53)} \\ &= d_2(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \end{aligned}$$

ซึ่งแสดงว่าค่าความผิดพลาดกำลังสองเฉลี่ยถ่วงน้ำหนักระหว่างค่าเวกเตอร์ต้นฉบับจะมีค่าเท่ากับค่าความผิดพลาดกำลังสองเฉลี่ยระหว่างเวกเตอร์ที่เปลี่ยนรูปแล้ว ดังนั้นจึงควรเปลี่ยนรูปเวกเตอร์ตามสมการที่ (2.42) กับข้อมูลทั้งหมดก่อนทำการควอนไทซ์แบบเวกเตอร์

ค) การวัดค่าความเพี้ยนโดยอาศัยการประมาณพันธะเชิงเส้น

(Linear Prediction Distortion Measures)

ในการวิเคราะห์หาสัมประสิทธิ์ ของการประมาณพันธะเชิงเส้นนั้น สัมประสิทธิ์ของการประมาณ $\{a(k)\}$ หาได้จากผลลัพธ์ของการทำให้ค่าพลังงานจากค่าคงเหลือของการประมาณมีค่าต่ำที่สุด ซึ่งก็คือผลลัพธ์ของชุดสมการเชิงเส้นดังนี้

$$\sum_{k=1}^{N} a(k)\phi(i-k) = -\phi(i), \qquad 1 \le i \le N \quad \text{.......... (2.54)}$$

เมื่อ $\{\phi(i),\ 0 \le i \le N\}$ เป็นสัมประสิทธิ์ของอัตสหสัมพันธ์ช่วงสั้นของสัญญาณเสียงในกรอบเดียว ดังนั้นอัตราขยาย G ของตัวกรอง $H(z)$ จะถูกกำหนดขึ้นโดยให้พลังงานงานขาออกมีค่าเท่ากับ $\phi(0)$ เมื่อถูกกระตุ้นด้วยแหล่งที่มีความแปรปรวนหนึ่งหน่วยดังนี้

$$G^2 = \phi(0) + \sum_{k=1}^{N} a(k)\phi(k) \qquad (2.55)$$

ซึ่งมีค่าเท่ากับพลังงานคงเหลือที่มีค่าต่ำที่สุด และเนื่องจากความไม่เสถียรของตัวกรองที่เป็นโพลทั้งหมด จึงต้องมีการแปลงไปเป็นพารามิเตอร์ชุดใหม่เรียกว่า "สัมประสิทธิ์การสะท้อน" $\{K_k,\ 1 \le k \le N\}$ หรือ "สัมประสิทธิ์ของอัตสหสัมพันธ์ส่วนย่อย" ซึ่งสามารถหาได้จากการหาผลลัพธ์ของสมการที่ (2.54) หรือหาจากค่าสัมประสิทธิ์ของการประมาณ สำหรับ $H(z)$ ที่เสถียรนั้นสัมประสิทธิ์การสะท้อนจะต้องมีคุณสมบัติดังนี้

$$|K_k| < 1, \qquad 1 \le k \le N \qquad (2.56)$$

เนื่องจากเมื่อค่าของ $|K_k|$ มีค่าเข้าใกล้ 1 จะทำให้โพลเข้าใกล้วงกลมหนึ่งหน่วยมากขึ้น ดังนั้นเมื่อ K_k เปลี่ยนไปเพียงเล็กน้อยจะส่งผลให้สเปกตรัมเปลี่ยนแปลงไปมาก ดังนั้นจึงต้องแปลงไปเป็นพารามิเตอร์ชุดใหม่ดังนี้

$$S_k = \frac{2}{\pi}\sin^{-1} K_k, \qquad 1 \le k \le N \qquad (2.57)$$

$$G_k = \frac{1}{2}\log\frac{1-K_k}{1+K_k} = \tanh^{-1} K_k, \qquad 1 \le k \le N \qquad (2.58)$$

เมื่อ G_k เป็นอัตราส่วนพื้นที่ล็อก (Log-Area-Ratios, LARs) ซึ่งมีสัดส่วนสัมพันธ์กับการเปลี่ยนแปลงของล็อกสเปกตรัมของ $H(z)$ โดยใช้ค่าความผิดพลาดกำลังสองเฉลี่ย d_2 และค่าความผิดพลาดสูงสุดที่ต่ำที่สุด d_∞ ในการควอนไทซ์ S_k และ G_k

ง) การวัดค่าความเพี้ยนของ Itakura-Saito

(Itakura-Saito Distortion Measures)

เทคนิคการวัดค่าความเพี้ยน ในการควอนไทซ์สัมประสิทธิ์การประมาณพันธะของ Itakura-Saito นี้อยู่บนพื้นฐานของหลักการความน่าจะเป็นจริงสูงสุด (Maximum-Likelihood Principles) โดยอาศัยหลักการของการวัดค่าความเพี้ยนระหว่างเวกเตอร์ของสัมประสิทธิ์การประมาณพันธะ

$$\mathbf{x} = [a(1)\ a(2)\ \Lambda\ a(N)]^T$$

กับเวกเตอร์ของสัมประสิทธิ์การประมาณพันธะ $\mathbf{y}$ ดังนี้

$$d_I(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Phi_x (\mathbf{x}-\mathbf{y}) \qquad (2.59)$$

$$\Phi_x = \{\phi(i-k)/\phi(0), 0 \le i,k \le N-1\} \qquad (2.60)$$

เมื่อ Φ_x เป็นเมตริกซ์อัตสหสัมพันธ์ที่ได้รับการปรับให้เป็นบรรทัดฐานเดียวกัน โดยมีสัมประสิทธิ์ $\phi(i-k)$ ที่ใช้ในการคำนวณหาค่าเวกเตอร์ของสัมประสิทธิ์การประมาณพันธะ $\mathbf{x}$ ตามสมการที่ (2.54) เนื่องจากค่าสัมประสิทธิ์ของอัตสหสัมพันธ์ในสมการที่ (2.60) ได้รับการปรับให้เป็นบรรทัดฐานเดียวกันด้วยค่า $\phi(0)$ จึงแสดงว่าเมตริกซ์ Φ_x และเวกเตอร์ $\mathbf{x}$ ใช้ในการหาค่าซึ่งกันและกัน นอกจากนี้ Φ_x ในสมการที่ (2.59) เป็นเมตริกซ์ถ่วงน้ำหนักที่มีการเปลี่ยนแปลงค่าเมื่อค่าของ $\mathbf{x}$ เปลี่ยนแปลงไป แต่ในสมการที่ (2.48) นั้นค่า $\mathbf{W}$ มีค่าคงที่เสมอ ดังนั้นการหาค่าความเพี้ยนด้วยเทคนิคนี้ไม่มีความสมมาตรอันเนื่องมาจากค่าเมตริกซ์ $\Phi_x \ne \Phi_y$ เมื่อ $\mathbf{x} \ne \mathbf{y}$ เช่น $d_I(\mathbf{x},\mathbf{y}) \ne d_I(\mathbf{y},\mathbf{x})$ เป็นต้น

2.2.3.2. การออกแบบชุดรหัส

การออกแบบชุดรหัส L ขั้นนั้น อาศัยวิธีการแบ่งปริภูมิ N มิติออกเป็น L เซลล์ $\{C_i, 1 \le i \le L\}$ และกำหนดให้แต่ละเซลล์ C_i สัมพันธ์กับเวกเตอร์ $\mathbf{y}_i$ โดยตัวควอนไทซ์จะกำหนดเวกเตอร์รหัส $\mathbf{y}_i$ ถ้า $\mathbf{x}$ อยู่ใน C_i ซึ่งจะมีความเหมาะสมที่สุดหรือมีความเพี้ยนน้อยที่สุดเมื่อค่าความเพี้ยนตามสมการที่ (2.44) มีค่าน้อยที่สุดจากตัวควอนไทซ์ L ขั้นทั้งหมด โดยมีเงื่อนไขของความเหมาะสม 2 ประการดังนี้ เงื่อนไขแรก ตัวควอนไทซ์ที่เหมาะสมที่สุดต้องเป็นไปตามกฎความเพี้ยนน้อยที่สุดหรือกฎการเลือกบริเวณที่ใกล้เคียงที่สุด (Nearest Neighbour Rule) โดยตัวควอนไทซ์จะเลือกเวกเตอร์รหัสที่มีความเพี้ยนน้อยที่สุดเมื่อเทียบกับ $\mathbf{x}$ ดังนี้

$$q(\mathbf{x}) = \mathbf{y}_i, \;\; iff \; d(\mathbf{x}, \mathbf{y}_i) \le d(\mathbf{x}, \mathbf{y}_j), \; j \ne i, \;\; 1 \le j \le L \quad (2.61)$$

เงื่อนไขที่สอง การเลือกเวกเตอร์รหัส $\mathbf{y}_i$ จะต้องให้ความเพี้ยนเฉลี่ยในเซลล์ C_i มีค่าน้อยที่สุด นั่นคือ $\mathbf{y}_i$ เป็นเวกเตอร์ $\mathbf{y}$ ที่ทำให้สมการที่ (2.62) มีค่าน้อยที่สุด

$$D_i = E\left[d(\mathbf{x}, \mathbf{y}) \middle| \mathbf{x} \in C_i\right] = \int_{\mathbf{x} \in C_i} d(\mathbf{x}, \mathbf{y})\, p(\mathbf{x})\, d\mathbf{x} \quad (2.62)$$

เวกเตอร์ดังกล่าวนี้เรียกว่า "จุดศูนย์ถ่วง" (Centroid) ของเซลล์ C_i ซึ่งเขียนได้เป็น

$$\mathbf{y}_i = cent(C_i) \quad (2.63)$$

การหาค่าจุดศูนย์ถ่วงของแต่ละบริเวณนั้น จะขึ้นอยู่กับนิยามของการวัดค่าความเพี้ยน ในทางปฏิบัติจะกำหนดชุดของเวกเตอร์ฝึกฝน $\{\mathbf{x}(n), 1 \le n \le M\}$ โดยที่ชุดย่อย M_i ของเวกเตอร์นี้อยู่ในเซลล์ C_i ซึ่งมีความเพี้ยนเฉลี่ย D_i ดังนี้

$$D_i = \frac{1}{M} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \mathbf{y}_i) \quad (2.64)$$

เมื่อพิจารณาในกรณีของค่าความผิดพลาดกำลังสองเฉลี่ยและค่าความผิดพลาดกำลังสองเฉลี่ยถ่วงน้ำหนัก โดยที่ $\mathbf{y}_i$ เป็นเพียงค่าเฉลี่ยสุ่มของเวกเตอร์ฝึกฝนทั้งหมดที่อยู่ภายใน C_i ดังนั้นความเพี้ยนเฉลี่ย D_i จะลดรูปลงเหลือเพียง

$$\mathbf{y}_i = \frac{1}{M_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}(n) \quad (2.65)$$

เมื่อพิจารณาในกรณีค่าความเพี้ยนของ Itakura-Saito d_I นั้น การหาค่า $\mathbf{y}_i$ จะอาศัยการเฉลี่ยค่าอัตสหสัมพันธ์ที่ถูกทำให้เป็นบรรทัดฐานเดียวกันและสัมพันธ์กับเวกเตอร์สุ่ม โดยที่ $\phi_x(k)$ ได้รับการทำให้เป็นบรรทัดฐานเดียวกันทำให้ $\phi_x(0) = 1$ ดังนั้นเวกเตอร์สามารถหาได้จากคำตอบของสมการที่ (2.44) โดยใช้ $\phi_{\mathbf{y}_i}(k)$ เป็นสัมประสิทธิ์ของอัตสหสัมพันธ์

$$\phi_{\mathbf{y}_i}(k) = \frac{1}{M_i} \sum_{\mathbf{x} \in C_i} \phi_x(k), \qquad 0 \le k \le N \quad (2.66)$$

นอกจากนี้ ยังมีอีกวิธีการหนึ่งในการออกแบบชุดรหัสก็คือขั้นตอนวิธีการแบ่งกลุ่มแบบวนซ้ำ (Iterative Clustering Algorithm) หรือขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วน (K-Means Algorithm) เมื่อกำหนดให้ K = L จะเป็นการแบ่งชุดของเวกเตอร์ฝึกฝน $\{\mathbf{x}(n)\}$ ออกเป็น L กลุ่ม C_i โดยเป็นไปตามเงื่อนไขของความเหมาะสมที่จำเป็นทั้งสองประการ ขั้นตอนวิธีการนี้กำหนดให้ m เป็นดัชนีของการวนซ้ำและ $C_i(m)$ เป็นกลุ่มที่ i ในรอบที่ m โดยมี $\mathbf{y}_i(m)$ เป็นจุดศูนย์ถ่วง ดังแสดงในตารางที่ 2.2

เนื่องจากขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนนี้จะลู่เข้าสู่ค่าเหมาะสมที่สุดเฉพาะแห่ง (Local Optimum) จึงทำให้ผลลัพธ์ที่ได้ไม่เป็นหนึ่งเดียว ดังนั้นการหาค่าเหมาะสมที่สุดที่ครอบคลุมทั้งหมด (Global Optimum) จะกระทำได้โดยการใช้ค่าเริ่มต้นของเวกเตอร์รหัสหลายค่าที่แตกต่างกันไป จากนั้นจึงทำซ้ำขั้นตอนวิธีการดังกล่าวนี้กับชุดของค่าเริ่มต้นหลายชุดที่แตกต่างกัน แล้วจึงเลือกชุดรหัสที่ให้ค่าความเพี้ยนทั้งหมดน้อยที่สุด

ตารางที่ 2.2 รายละเอียดขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วน

ขั้นตอนที่ 1 **กระบวนการเริ่มต้น**	กำหนดให้ $m=0$ และเลือกชุดของเวกเตอร์รหัสเริ่มต้น $y_i(0),\ 1 \le i \le L$ ด้วยวิธีการที่เหมาะสม
ขั้นตอนที่ 2 **กระบวนการจำแนก**	ทำการจำแนกชุดของเวกเตอร์ฝึกฝน $\{\mathbf{x}(n),\ 1 \le n \le M\}$ ไปตามกลุ่มด้วยกฎการเลือกเพื่อนบ้านใกล้เคียงที่สุดดังนี้ $\mathbf{x} \in C_i(m),\ \textit{iff}\ d[\mathbf{x}, \mathbf{y}_i(m)] \le d[\mathbf{x}, \mathbf{y}_j(m)],\ \text{all}\ j \ne i$
ขั้นตอนที่ 3 **กระบวนการปรับค่า** **เวกเตอร์รหัส**	กำหนดให้ $m \leftarrow m+1$ เพื่อปรับค่าของเวกเตอร์รหัสของทุกกลุ่มโดยการคำนวณจุดศูนย์ถ่วงของเวกเตอร์รหัสในแต่ละกลุ่มใหม่ดังนี้ $\mathbf{y}_i(m) = cent(C_i(m)),\ \ 1 \le i \le L$
ขั้นตอนที่ 4 **กระบวนการสิ้นสุด**	ถ้าค่าความเพี้ยนโดยรวม $D(m)$ ของการวนซ้ำรอบที่ m เมื่อเทียบกับ $D(m-1)$ มีค่าลดลงต่ำกว่าจุดเริ่มเปลี่ยนให้ถือว่าสิ้นสุดกระบวนการ ถ้ามีค่าสูงกว่าให้ทำซ้ำขั้นตอนที่ 2 ใหม่

การควอนไทซ์แบบเวกเตอร์นั้น มีข้อได้เปรียบทางด้านประสิทธิภาพเหนือการควอนไทซ์แบบสเกลาร์โดยเฉพาะกับแหล่งข้อมูลที่ขึ้นแก่กันอย่างไม่เป็นเชิงเส้น ดังนั้นในการออกแบบสร้างชุดรหัสที่มีประสิทธิภาพจะต้องอาศัยวิธีการที่เหมาะสม ในการเข้ารหัสโดยการค้นหาทั่วทั้งหมดนั้นค่าใช้จ่ายจะเพิ่มขึ้นแบบเอกซ์โพเนนเชียลตามจำนวนบิตต่อเวกเตอร์ ค่าใช้จ่ายในการคำนวณและการจัดเก็บจะเพิ่มขึ้นเป็นสองเท่าต่ออัตราเร็วที่เพิ่มขึ้นแต่ละบิต

ขั้นตอนวิธีการค้นหาอย่างรวดเร็ว ที่ได้รับการนำเสนอในการรู้จำรูปแบบนั้นมีด้วยกันหลายวิธี ซึ่งได้รับการนำมาประยุกต์ใช้กับการควอนไทซ์แบบเวกเตอร์เพื่อลดการคำนวณที่เกิดจากการค้นหาหมดทั้งชุดรหัส ขั้นตอนวิธีการส่วนใหญ่จะอยู่บนพื้นฐานของหลักการเชิงเรขาคณิตในปริภูมิชนิด Euclidean ซึ่งต้องการกรรมวิธีประมวลผลเบื้องต้นกับชุดรหัส และเป็นการประนีประนอมกันระหว่างการคูณกับการเปรียบเทียบและความต้องการเนื้อที่จัดเก็บที่เพิ่มสูงขึ้น โดยจำนวนการคูณสามารถลดลงได้เป็นจำนวนเท่าของจำนวนเดิม

ขั้นตอนวิธีการออกแบบสร้างและฝึกฝนชุดรหัส ซึ่งเป็นขั้นตอนหนึ่งในการสร้างชุดรูปร่องต้นแบบของคำศัพท์แต่ละคำ จัดเป็นขั้นตอนการฝึกฝนชุดรหัสเพื่อใช้ในการควอนไทซ์แบบเวกเตอร์ ขั้นตอนวิธีการออกแบบสร้างชุดรหัสมีหลายวิธี (Makhoul, Roucos, and Gish, 1985) ได้แก่ การค้นหาแบบทวิภาค (Binary Search) การควอนไทซ์แบบต่อเรียงกัน (Cascade Quantization) รหัสผลคูณ (Product Codes) และชุดรหัสสุ่ม (Random Codebooks) สำหรับงานวิจัยนี้จะใช้วิธีการชุดรหัสสุ่มในการสร้างชุดรหัสเริ่มต้นเพื่อการฝึกฝนชุดรหัส เนื่องจากมีความรวด

เร็วและมีประสิทธิภาพใกล้เคียงกับวิธีการค้นหาแบบทวิภาค ดังนั้นในที่นี้จะกล่าวถึงเพียงรายละเอียดของวิธีการค้นหาแบบทวิภาคและชุดรหัสสุ่มเท่านั้น

ก) การค้นหาแบบทวิภาค (Binary Search)

เนื่องจากขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนโดยมี $K = L$ ขั้นนั้น จำเป็นต้องอาศัยการค้นหาทั่วทั้งหมดใน L เวกเตอร์รหัสเพื่อควอนไทซ์แต่ละเวกเตอร์ที่รับเข้ามา ดังนั้นการค้นหาแบบทวิภาคหรือการแบ่งกลุ่มเป็นลำดับชั้น (Hierarchical Clustering) จึงเป็นวิธีการสำหรับแบ่งปริภูมิที่ทำให้การค้นหาเวกเตอร์รหัสที่มีค่าความเพี้ยนน้อยที่สุดเป็นสัดส่วนกับ $\log_2 L$ แทนที่จะเป็น L โดยมีขั้นตอนคือ ในขั้นแรกปริภูมิ N มิติจะได้รับการแบ่งออกเป็น 2 บริเวณซึ่งใช้ขั้นตอนวิธีการแบ่งเฉลี่ย K เมื่อ $K = 2$ จากนั้นทั้งสองบริเวณก็จะถูกแบ่งออกเป็นสองบริเวณต่อไปเรื่อยๆ จนกระทั่งปริภูมิถูกแบ่งออกเป็น L บริเวณหรือเซลล์ จึงทำให้ค่าของ L จะต้องเป็นกำลังของ 2 เท่านั้น นั่นคือ $L = 2^B$ เมื่อ B เป็นจำนวนบิตทั้งหมด โดยมีจุดศูนย์ถ่วงที่สัมพันธ์กับแต่ละบริเวณในการแบ่งแบบทวิภาคแต่ละครั้ง

แผนภูมิการแบ่งแบบทวิภาคของปริภูมิออกเป็น L = 8 เซลล์ดังแสดงในรูปที่ 2.7 ในการแบ่งแบบทวิภาคครั้งแรกจะได้ $\mathbf{v}_1$ และ $\mathbf{v}_2$ เป็นจุดศูนย์ถ่วง ในการแบ่งแบบทวิภาคครั้งที่สองประกอบด้วย 4 บริเวณซึ่งมีจุดศูนย์ถ่วงเป็น $\mathbf{v}_3$ ถึง $\mathbf{v}_6$ จุดศูนย์ถ่วงของบริเวณภายหลังการแบ่งแบบทวิภาคครั้งที่สามจะเป็นเวกเตอร์รหัส $\mathbf{y}_i$ ดังนั้นเวกเตอร์ขาเข้า $\mathbf{v}_1$ จะได้รับการควอนไทซ์โดยการไล่ค้นหาตามแผนภูมิต้นไม้ดังรูปที่ 2.2 ไปตามแต่ละปมบนเส้นทาง โดยจะเปรียบเทียบ $\mathbf{x}$ กับ $\mathbf{v}_1$ และ $\mathbf{v}_2$ ถ้า $d(\mathbf{x},\mathbf{v}_2) < d(\mathbf{x},\mathbf{v}_1)$ จึงเลือกเส้นทางไปยัง $\mathbf{v}_2$ จากนั้นจะเปรียบเทียบ $\mathbf{x}$ กับ $\mathbf{v}_5$ และ $\mathbf{v}_6$ ถ้า $d(\mathbf{x},\mathbf{v}_5) < d(\mathbf{x},\mathbf{v}_6)$ จึงเลือกเส้นทางไปยัง $\mathbf{v}_5$ และในขั้นตอนสุดท้ายจะเปรียบเทียบ $\mathbf{x}$ กับ $\mathbf{y}_5$ และ $\mathbf{y}_6$ ถ้า $d(\mathbf{x},\mathbf{y}_6) < d(\mathbf{x},\mathbf{y}_5)$ ดังนั้น $\mathbf{y}_6$ จึงเป็นค่าควอนไทซ์ของ $\mathbf{x}$

จากขั้นตอนวิธีการแบ่งแบบทวิภาค จำนวนการคำนวณค่าความเพี้ยนจะมีค่าเท่ากับ $2\log_2 L$ เมื่อกำหนดให้การคำนวณค่าความเพี้ยนแต่ละครั้งมีการคูณและบวก N ครั้งจะมีค่าใช้จ่ายรวมทั้งหมดในการคำนวณเป็น

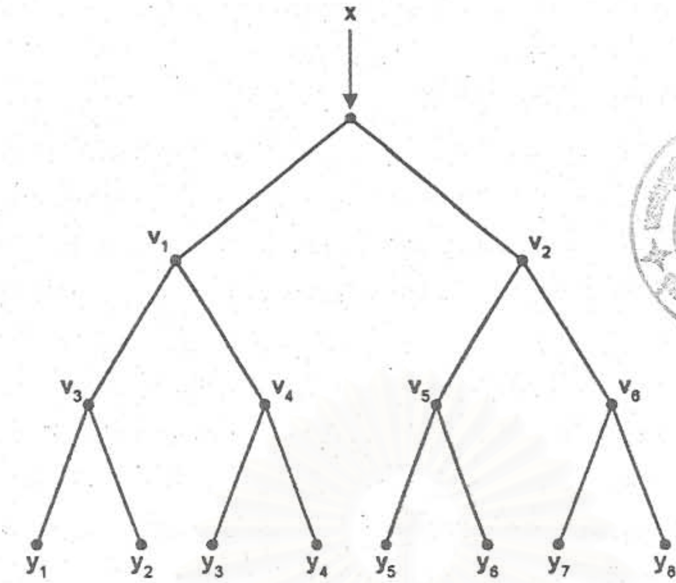
$$C = 2N\log 2L = 2NB \qquad (2.67)$$

โดยมีความสัมพันธ์เป็นเชิงเส้นกับจำนวนบิต ซึ่งเป็นผลให้มีการลดจำนวนการคำนวณไปอย่างมหาศาล แต่ค่าใช้จ่ายในการจัดเก็บกลับเพิ่มสูงขึ้น นอกจากการจัดเก็บเวกเตอร์รหัส $\mathbf{y}_i$ ไว้แล้ว ยังต้องจัดเก็บเวกเตอร์ระหว่างกลาง (Intermediate Vector) ทั้งหมดไว้ด้วย ดังนั้นค่าใช้จ่ายทั้งหมดในการจัดเก็บจะเพิ่มขึ้นเป็นสองเท่าดังนี้

$$M = 2N(L-2) \qquad (2.68)$$

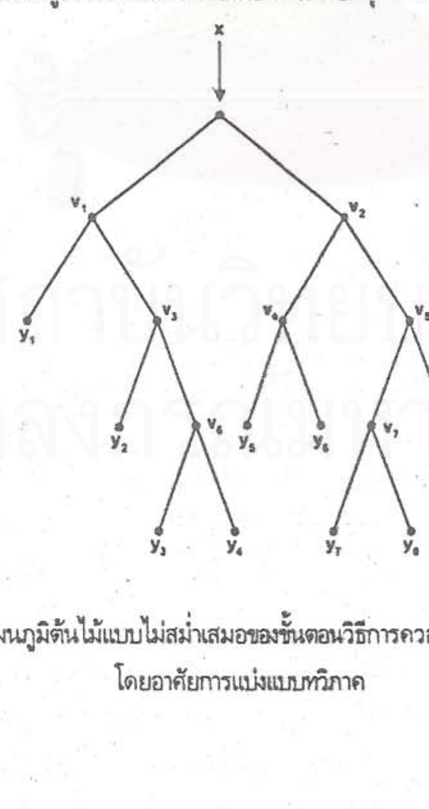
ค่าใช้จ่ายที่จำเป็นข้างต้นสามารถลดลงครึ่งหนึ่งได้โดยการใช้ค่าความผิดพลาดกำลังสองเฉลี่ยในการวัดค่าความเพี้ยน ในกรณีนี้จะอาศัยการเปรียบเทียบบริเวณที่ $\mathbf{x}$ อยู่เมื่อเทียบกับ ระนาบซึ่งแบ่งเป็นสองบริเวณ แทนที่การเปรียบเทียบ $\mathbf{x}$ กับเวกเตอร์ทั้งสอง โดยการเปรียบเทียบนี้จะเป็นการคำนวณผลคูณเชิงสเกล่าร์ของเวกเตอร์ทั้งสองเพียงครั้งเดียวเท่านั้น

จากรูปที่ 2.7 เป็นแผนภูมิต้นไม้เพื่อการควอนไทซ์ชนิดสม่ำเสมอ (Uniform Quantization Tree) โดยบริเวณทั้งหมดในแต่ละขั้นตอนจะถูกแบ่งออกเป็นสองบริเวณย่อย เมื่อเริ่มทำการฝึกฝนเพื่อสร้างแผนภูมิต้นไม้เพื่อการควอนไทซ์นั้น อาจแบ่งออกได้มากกว่าหนึ่งหรือสองกลุ่มย่อยที่แต่ละจุดในขั้นตอนการแบ่งย่อยซึ่งมีข้อมูลฝึกฝนน้อยเกินไป กลุ่มย่อยประเภทนี้จะเป็นการสูญเสียบิตข้อมูลไปโดยเปล่าประโยชน์เนื่องจากการแบ่งย่อยลงไปอีกไม่ช่วยให้ค่าความเพี้ยนลดลงไปได้อีก ดังนั้นเพื่อให้ได้ค่าความเพี้ยนเฉลี่ยที่ต่ำลงและเป็นการใช้ประโยชน์บิตข้อมูลให้ได้มากที่สุดจึงไม่ควรแบ่งแขนงของแผนภูมิต้นไม้ให้สม่ำเสมอ โดยในระหว่างการฝึกฝนที่แต่ละขั้นตอนการแบ่งย่อยนั้นจะต้องทำการตรวจสอบค่าความเพี้ยนรวมที่เกิดจากแต่ละกลุ่มเสมอ กลุ่มที่มีค่าความเพี้ยนมากที่สุดจะถูกแบ่งย่อยต่อไปและกระทำกระบวนการซ้ำอีกครั้ง ผลที่ได้จะเป็นแผนภูมิต้นไม้ชนิดไม่สม่ำเสมอ (Nonuniform Quantization Tree) ในกรณีเช่นนี้จำนวนขั้นจะเป็นจำนวนเต็มค่าใดก็ได้ โดยไม่จำกัดอยู่แต่เพียงกำลังของ 2 เท่านั้น ดังแสดงในรูปที่ 2.8 ด้วยค่า L = 9



รูปที่ 2.11 แผนภูมิต้นไม้แบบสม่ำเสมอของขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์ด้วยการแบ่งแบบทวิภาค

การค้นหาแบบทวิภาคนี้ เป็นกรณีพิเศษของวิธีการควอนไทซ์แบบเวกเตอร์ที่เรียกว่า "การควอนไทซ์แบบเวกเตอร์โดยอาศัยแผนภูมิต้นไม้ในการค้นหา" (Tree-searched Vector Quantization) ซึ่งถือว่าการค้นหาแบบทวิภาคเป็นวิธีการที่ง่ายที่สุด โดยทั่วไปแล้วจะสามารถแบ่งปริภูมิที่แต่ละปมในแผนภูมิต้นไม้ได้มากกว่าสองบริเวณย่อยด้วยขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนที่มี $K > 2$ วิธีการดังกล่าวนี้จะเพิ่มปริมาณการคำนวณมากกว่าการค้นหาแบบทวิภาค แต่ประสิทธิภาพของการค้นหาแบบทวิภาคก็ใกล้เคียงกับการค้นหาทั่วทั้งหมดในการประยุกต์ใช้งานทั่วไป ดังนั้นจึงควรใช้การค้นหาแบบทวิภาคแบบแผนภูมิต้นไม้ไม่สม่ำเสมอเพื่อการใช้จ่ายที่คุ้มค่ามากที่สุด



รูปที่ 2.12 แผนภูมิต้นไม้แบบไม่สม่ำเสมอของขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์
โดยอาศัยการแบ่งแบบทวิภาค

ข) ชุดรหัสสุ่ม (Random Codebooks)

นอกจากความต้องการลดค่าใช้จ่ายที่เกิดขึ้น ในกระบวนการควอนไทซ์แบบเวกเตอร์แล้ว ยังต้องคำนึงถึงระยะเวลาที่ใช้ในกระบวนการฝึกฝนอีกด้วย วิธีการในการออกแบบสร้างชุดรหัสโดยไม่เสียค่าใช้จ่ายในการคำนวณระหว่างกระบวนการฝึกฝนก็คือ การสุ่มเลือกเวกเตอร์รหัสเริ่มต้นจากชุดข้อมูลฝึกฝน โดยชุดรหัสถูกออกแบบสร้างด้วยวิธีการนี้เรียกว่า "ชุดรหัสสุ่ม" (Random Codebook) การนำชุดรหัสสุ่มมาใช้งานนั้นเป็นทางเลือกที่เหมาะสมในกรณีที่ค่า L และ N มีค่ามาก แต่ก็ไม่เหมาะสมในทางปฏิบัติเมื่อไม่ตรงตามเงื่อนไข อย่างไรก็ตามถึงแม้ว่าชุดรหัสสุ่มจะง่ายต่อการออกแบบสร้าง แต่ก็เป็นชุดรหัสแบบค้นหาทั่วทั้งหมด ซึ่งต้องใช้เวลาในการคำนวณมาก และต้องการเนื้อที่เพื่อจัดเก็บชุดรหัสทั้งหมดไว้ด้วยเช่นกัน

2.2.3.3. การฝึกฝนและการทดสอบชุดรหัส (Codebook Training and Testing)

ในการออกแบบสร้างชุดรหัสใดก็ตาม สิ่งที่สำคัญก็คือกระบวนการฝึกฝนเพื่อเสริมสร้างชุดรหัสขึ้นมา โดยขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนจะเป็นกระบวนการฝึกฝน ยกเว้นขั้นตอนกระบวนการเริ่มต้นซึ่งเป็นการเลือกชุดเวกเตอร์รหัสเริ่มต้น ในหัวข้อนี้จะกล่าวถึงกระบวนการเริ่มต้นที่ใช้ในการควอนไทซ์แบบเวกเตอร์กับสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นในการเข้ารหัสเสียงพูด และจะกล่าวถึงกระบวนการทดสอบและความเสถียรของชุดรหัสที่ได้ต่อไป

2.2.3.3.1. การฝึกฝนชุดรหัส (Codebook Training)

2.2.3.3.1.1. การแบ่งกลุ่มแบบทวิภาค (Binary Clustering)

การค้นหาแบบทวิภาคนั้น ในแต่ละตอนจำเป็นต้องมีกระบวนการเริ่มต้นเพื่อการแบ่งปริภูมิออกเป็นสองบริเวณ โดยปรกติแล้วจะอาศัยสร้างระนาบที่วิ่งผ่านค่าเฉลี่ยของชุดข้อมูลฝึกฝนและตั้งฉากกับเวกเตอร์เจาะจงที่มีค่ามากที่สุด โดยใช้ค่าความเพี้ยนกำลังสองเฉลี่ยเป็นหลัก ระนาบดังกล่าวนี้จะแบ่งปริภูมิออกเป็นสองบริเวณเริ่มต้นเพื่อเริ่มกระบวนการฝึกฝนแบบวนซ้ำ สำหรับข้อมูลที่มิติใดมิติหนึ่งมีค่าความแปรปรวนมากกว่ามิติอื่นจะประมาณทิศทางของเวกเตอร์เจาะจงด้วยมิติที่มีค่าความแปรปรวนมากที่สุด และนำจุดศูนย์ถ่วงของกลุ่มทั้งสองที่ถูกแบ่งด้วยระนาบตั้งฉากมาใช้เป็นเวกเตอร์รหัสเริ่มต้นสองค่าแรก จากนั้นจึงดำเนินขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนด้วยค่า K = 2 เพื่อหาเวกเตอร์รหัสสุดท้าย ซึ่งโดยทั่วไปจะลู่เข้าเมื่อทำซ้ำกระบวนการไปประมาณ 5 ถึง 10 รอบ ภายหลังการแบ่งครั้งแรกสิ้นสุดลงแต่ละบริเวณย่อยจะถูกแบ่งออกเป็นสองบริเวณด้วยกระบวนการเดียวกัน

2.2.3.3.1.2. กระบวนการเริ่มต้นของขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วน

เนื่องจากผลของขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนไม่สามารถให้ชุดรหัสที่เหมาะสมที่สุดที่ครอบคลุมทั้งหมดได้ ดังนั้นจึงต้องทำซ้ำขั้นตอนวิธีการด้วยชุดของเวกเตอร์รหัสเริ่มต้นที่แตกต่างกันไป แล้วเลือกผลของชุดรหัสที่มีค่าความเพี้ยนต่ำที่สุดมาใช้งานจริง

เนื่องจากการค้นหาแบบทวิภาคมีประสิทธิภาพใกล้เคียงกับขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนที่เหมาะสมที่สุดที่ครอบคลุมทั้งหมด ดังนั้นชุดรหัสที่ถูกสร้างขึ้นโดยการค้นหาแบบทวิภาคที่ไม่สม่ำเสมอจะเป็นชุดรหัสเริ่มต้นที่ดีสำหรับเริ่มต้นขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วน การเริ่มต้นด้วยวิธีการนี้จะให้ผลลัพธ์ที่ดีกว่าการสุ่มชุดรหัสเริ่มต้น

2.2.3.3.2. การทดสอบชุดรหัส (Codebook Testing)

ภายหลังจากการออกแบบสร้างชุดรหัสด้วยชุดข้อมูลฝึกฝนแล้ว จำเป็นต้องมีการทดสอบประสิทธิภาพของชุดรหัสด้วยข้อมูล ที่ไม่ขึ้นกับข้อมูลอื่นและไม่อยู่ในชุดข้อมูลฝึกฝน การทดสอบกับเพียงแค่ชุดข้อมูลฝึกฝน เป็นเพียงการแสดงวิสัยทัศน์ของชุดรหัสที่ปฏิบัติต่อข้อมูลเชิงดำเนินการเท่านั้น จากการทดสอบชุดข้อมูลที่ไม่ขึ้นกับข้อมูลอื่นและชุดข้อมูลฝึกฝนนั้น เมื่อเพิ่มจำนวนการฝึกฝนจะมีผลให้ค่าความเพี้ยนของชุดข้อมูลทั้งสองเริ่มลู่เข้าหากันมากขึ้น แต่การลู่เข้าจะเริ่มลดลงเมื่อจำนวนของเวกเตอร์ฝึกฝนต่อชุดรหัสมีจำนวนเพิ่มขึ้นมากกว่า 20 เวกเตอร์ จึงสามารถยึดถือเป็นกฎเกณฑ์ได้ว่าการใช้ชุดข้อมูลฝึกฝน 50 ชุดต่อ เวกเตอร์รหัสก็เพียงพอสำหรับการประยุกต์ใช้งานโดยทั่วไป ถ้า

จำนวนข้อมูลไม่เพียงพอหรือถ้าการคำนวณหรือการจัดเก็บไม่เพียงพอ ดังนั้นการใช้ข้อมูลเพียง 10 ข้อมูลฝึกฝนต่อเวกเตอร์รหัสก็อาจจะเพียงพอแล้ว

ความแตกต่างระหว่างชุดข้อมูลที่ไม่ขึ้นกับข้อมูลอื่นและชุดข้อมูลฝึกฝนเกิดขึ้นเนื่องจากเสียงพูดที่ใช้ทดสอบแตกต่างจากเสียงพูดที่ใช้ฝึกฝน ซึ่งเกี่ยวข้องกับความทนทานของชุดรหัส โดยทั่วไปแล้วการลู่เข้าของการฝึกฝนและประสิทธิภาพที่ได้จากการทดสอบของแหล่งข้อมูลใดๆ อาจไม่เพิ่มขึ้นตามจำนวนการฝึกฝน ผลลัพธ์ที่ได้ในทางทฤษฎีเมื่อใช้การวัดค่าความเพี้ยนที่เหมาะสมจะทำให้เกิดการลู่เข้าเฉพาะกับแหล่งข้อมูลประเภทที่ไม่แปรเปลี่ยนตามเวลาโดยค่าเฉลี่ยเชิงเส้นกำกับ (Asymtotically Mean Stationary) ซึ่งไม่จำเป็นต้องเป็นไปตามเงื่อนไขการไม่แปรเปลี่ยนตามเวลา ดังนั้นเสียงพูดจากผู้พูดคนเดียวจะต้องอยู่ภายใต้ชุดเงื่อนไขสภาวะแวดล้อมที่ตายตัว ตัวอย่างเช่น มีคุณสมบัติการไม่แปรเปลี่ยนตามเวลาทั้งช่วงเวลาสั้นและช่วงเวลายาว รวมทั้งถูกจำลองตามแหล่งข้อมูลที่ไม่แปรเปลี่ยนตามเวลาโดยค่าเฉลี่ยเชิงเส้นกำกับ เป็นต้น อย่างไรก็ตามแบบจำลองอาจจะไม่สามารถใช้งานได้ภายใต้เงื่อนไขที่แปรเปลี่ยนตามเวลาและมีความแปรเปลี่ยนมากขึ้น

2.2.3.3.3. ความทนทานของชุดรหัส (Codebook Robustness)

ความทนทานของชุดรหัสหมายความถึง ความต้านทานของชุดรหัสที่มีต่อประสิทธิภาพที่ลดลงเมื่อทดสอบกับข้อมูลที่มีการกระจายแตกต่างไปจากข้อมูลฝึกฝน เมื่ออยู่ภายใต้เงื่อนไขเชิงดำเนินการจะไม่สามารถประมาณสถานการณ์ที่ตัวควอนไทซ์นำไปใช้งานได้ ซึ่งโดยทั่วไปแล้วการกระจายของข้อมูลเชิงดำเนินการจะแตกต่างไปจากข้อมูลฝึกฝน ดังนั้นจึงสามารถแบ่งแยกความแปรเปลี่ยนได้เป็น 2 ประเภทที่มีผลกระทบต่อการออกแบบสร้างและประสิทธิภาพเชิงดำเนินการของชุดรหัสได้แก่ (Makhoul, Roucos, and Gish, 1985) ความแปรเปลี่ยนของสัญญาณขาเข้า (Input Signal Variability) และความผิดพลาดของช่องสื่อสารเชิงเลข (Digital Transmission Channel Errors)

สำหรับเสียงพูดนั้น ความแปรเปลี่ยนของสัญญาณสามารถแบ่งย่อยออกได้เป็นความแปรเปลี่ยนของผู้พูด (Speaker Variability) และความแปรเปลี่ยนของสภาวะแวดล้อม (Environmental Variability) ความแปรเปลี่ยนระหว่างผู้พูด (Interspeaker Variability) เกิดขึ้นเนื่องจากการเปลี่ยนแปลงเสียงพูดของผู้พูดแต่ละคน ได้แก่การเปลี่ยนแปลงปรกติธรรมดาทุกวัน การเปลี่ยนแปลงอันเนื่องมาจากสุขภาพ และการเปลี่ยนแปลงที่เกิดจากอารมณ์ เป็นต้น ความแปรเปลี่ยนภายในตัวผู้พูด (Intraspeaker Variability) หมายความถึงความแตกต่างของเสียงพูดระหว่างผู้พูดด้วยกันเอง ความแปรเปลี่ยนของสภาวะแวดล้อม หมายความถึงระดับความดังและประเภทของเสียงรบกวนเบื้องหลังที่อยู่รอบตัวผู้พูด และลักษณะสมบัติในการรับสัญญาณซึ่งรวมไปถึงชนิดของไมโครโฟนและสิ่งอำนวยความสะดวกในการส่งผ่านสัญญาณ ประสิทธิภาพของชุดรหัสจะถูกลดทอนลงเมื่อนำมาใช้กับสัญญาณที่ไม่ได้รับการออกแบบมา ถ้าชุดรหัสที่ได้รับการออกแบบหรือฝึกฝนมาสำหรับเสียงของผู้พูดคนหนึ่ง อาจจะมีประสิทธิภาพไม่ดีสำหรับผู้พูดคนอื่นๆ ดังนั้นเพื่อให้ได้ประสิทธิภาพที่เหมาะสมที่สุดจึงควรฝึกฝนชุดรหัสโดยใช้ข้อมูลที่เทียบเท่ากับการใช้งานจริงในการฝึกฝนเท่านั้น

การลู่เข้าของค่าความเพี้ยนของชุดข้อมูลที่ไม่ขึ้นกับข้อมูลอื่น และชุดข้อมูลฝึกฝนนั้นจะลู่เข้าหากันมากขึ้นเมื่อข้อมูลที่นำมาทดสอบมาจากผู้พูดคนเดียวกันกับที่ใช้ในการฝึกฝน ค่าความเพี้ยนระหว่างข้อมูลฝึกฝนและข้อมูลทดสอบที่แตกต่างกันประมาณ 1 dB ก็ถือว่ามีนัยสำคัญ การเพิ่มจำนวนข้อมูลฝึกฝนจากผู้พูดเพศชาย 15 คนไม่อาจทำให้ความแตกต่างลดลงได้ เพียงแต่ทำให้ชุดรหัสที่ใช้ในการควอนไทซ์นั้นมีประสิทธิภาพดีขึ้นเมื่อใช้กับผู้พูดเพียงคนเดียวกัน ดังนั้นถ้าต้องการให้ประสิทธิภาพดีที่สุดบนระบบที่ไม่ขึ้นกับผู้พูดจะต้องลดช่องว่างระหว่างชุดฝึกฝนและชุดทดสอบได้โดยการเพิ่มจำนวนผู้พูดในการฝึกฝนมากกว่าการเพิ่มจำนวนเสียงของผู้พูดแต่ละคน โดยถ้าต้องการใช้งานกับเสียงพูดผู้หญิงก็ต้องรวมเสียงพูดผู้หญิงเข้าไปในการฝึกฝนด้วยเช่นกัน

วิธีการที่เป็นไปได้ ในการเพิ่มประสิทธิภาพของการควอนไทซ์แบบเวกเตอร์ให้ได้มากที่สุดนั้นก็คือ การออกแบบชุดรหัสที่ไม่ขึ้นกับผู้พูดตั้งแต่เริ่มต้นซึ่งภายหลังเมื่อใช้งานระบบจะปรับให้เข้ากับเสียงพูดของผู้พูดคนใหม่ ระบบเช่นนี้มีข้อได้เปรียบในด้านการปรับตัวโดยอัตโนมัติให้เข้ากับสภาพแวดล้อมของเสียงพูดของผู้พูดได้เอง การ

ปรับตัวของชุดรหัสหมายถึงการที่เวกเตอร์รหัสเปลี่ยนไปตามเวลาซึ่งจำเป็นต่อการรับส่งเวกเตอร์รหัสชุดใหม่ไปยังตัวรับด้วยเช่นกัน ตัวอย่างเช่น ระบบแรกอาศัยชุดรหัสแบบค้นหาทั่วทั้งหมดที่ลดความเพี้ยนสูงสุดระหว่างเวกเตอร์ขาเข้าและเวกเตอร์รหัสให้มีค่าต่ำที่สุด เมื่อค่าความเพี้ยนของการควอนไทซ์สูงเกินกว่าจุดเริ่มเปลี่ยน ทำให้เวกเตอร์ขาเข้ากลายเป็นเวกเตอร์รหัสใหม่และทิ้งเวกเตอร์รหัส ที่ถูกใช้งานน้อยที่สุดไป จากนั้นเวกเตอร์รหัสใหม่ที่ได้ก็จะถูกส่งไปยังตัวรับ เป็นต้น

2.3. การทดสอบความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing)

การทดสอบความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing) หรือการจำแนกรูปแบบ (Pattern Classification) เกี่ยวข้องกับการเปรียบเทียบรูปแบบเสียงพูดระหว่างคำพูด หรือวลีที่ไม่ทราบรูปแบบกับรูปแบบที่ได้จัดเก็บไว้แล้ว ซึ่งก็คือชุดรูปร่องต้นแบบ (Templates) หรือแบบจำลอง (Models) ของเสียงพูด ภายหลังการเปรียบเทียบแต่ละครั้งจะได้ค่าความไม่คล้ายคลึงกัน (Dissimilarity Scores) หรือค่าระยะทาง (Distance Scores) เพื่อใช้ในขั้นตอนวิธีการตัดสินใจในการเลือกรูปแบบที่เหมาะสมใกล้เคียงที่สุดต่อไป

ขั้นตอนวิธีการในการจำแนกรูปแบบที่ถูกนำมาใช้มากที่สุด สามารถแบ่งออกได้เป็น 4 วิธีการ (Roe and Wilpon, 1993) ได้แก่ การเข้าคู่ต้นแบบ (Template Matching) ระบบตามกฎเกณฑ์ (Rule-Based System) ระบบแบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Model, HMM) และเครือข่ายประสาท (Neural Network) แต่เนื่องจากงานวิจัยนี้ใช้เทคนิคระบบแบบจำลองฮิดเดน มาร์คอฟ จึงจะกล่าวถึงแต่เฉพาะรายละเอียดและทฤษฎีของวิธีการจำแนกรูปแบบวิธีนี้เท่านั้น

เพราะงานวิจัยนี้เลือกใช้ 3 กรรมวิธีที่มีหลักการแตกต่างกัน ได้แก่ ไดนามิก ไทม์วาร์ปิง ซึ่งเป็นการเข้าคู่ต้นแบบ ๆ หนึ่ง แบบจำลองฮิดเดน มาร์คอฟ และ นิวรอลเน็ตเวิร์ก เป็นเหตุให้การพิจารณาความเป็นไปได้และความเหมาะสมของแต่ละกรรมวิธี ที่จะนำไปใช้ในการรู้จำเสียงพูดภาษาไทยมีเงื่อนไขแตกต่างกันมากบ้างน้อยบ้าง เช่นกัน

ก) ไดนามิก ไทม์วาร์ปิง เป็นกรรมวิธีที่ใช้เทคนิคการปรับยืดขยายหรือหดรูปคลื่นสัญญาณตามแกนเวลาแบบไดนามิค มีการนำกรรมวิธีนี้ไปใช้ในการรู้จำเสียงตัวเลขไทย (Pensiri and Jitapunkul, 1995) และเสียงสระภาษาไทย (Phatrapornnant and Jitapunkul, 1995) กรรมวิธีไดนามิก ไทม์วาร์ปปิง จะใช้เวลาในการสร้างแบบอ้างอิงแต่ละแบบไม่มาก แต่มีขีดจำกัดในเรื่องแบบอ้างอิงของคำ ที่ต้องเป็นคำโดดและมีจำนวนแบบไม่มากนัก

ข) กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ เป็นกรรมวิธีที่กำหนดให้เสียงพูดเป็นสัญญาณประเภทสโตคาสติก (stochastic signal) และถูกนำไปใช้ในการรู้จำเสียงตัวเลขไทย (Areepongsa and Jitapunkul, 1995) และเสียงคำไทยหลายพยางค์ (Ahkuputra, Jitapunkul, Laksaneeyanawin, and Pornsukchandra, 1997) ที่กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ มีความยืดหยุ่นสูงสามารถพัฒนาใช้กับเสียงพูดทั้งที่เป็นคำโดด คำหลายพยางค์ และคำพูดต่อเนื่องได้ แต่ใช้เวลาในการสร้างแบบอ้างอิงสูง ทั้งสองวิธีการแรกนี้ มีข้อด้อยในเรื่องเวลาของการรู้จำที่แปรตามจำนวนคำที่ต้องการรู้จำ เพราะต้องทดสอบกับแบบอ้างอิงของเสียงทุกคำ

ค) กรรมวิธีนิวร็อลเน็ตเวิร์ก จะนำเครือข่ายประสาทของมนุษย์มาจำลองการทำงาน โดยการกำหนดให้รูปแบบที่จะรู้จำมีค่าพารามิเตอร์ต่าง ๆ ที่จะถูกใส่น้ำหนักมากบ้างน้อยบ้างตามความสำคัญของพารามิเตอร์เหล่านั้น และเปรียบเทียบผลลัพธ์ว่าใกล้เคียงกับรูปแบบอ้างอิงใดสูงสุด กรรมวิธีนี้ถูกนำไปใช้ในการรู้จำเสียงตัวเลขไทย (Pornsukjantra and Jitapunkul, 1996) กรรมวิธีนิวรอลเน็ตเวิร์ก มีความยืดหยุ่นเช่นเดียวกับกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ แต่ยังคงใช้เวลาในการรู้จำเท่าเดิมถ้าโครงสร้างของเครือข่ายประสาทไม่เปลี่ยนแปลง ทั้งนี้เป็นเพราะกรรมวิธีนิวรอลเน็ตเวิร์กเก็บความรู้เกี่ยวกับลักษณะของเสียงทุก ๆ คำรวมอยู่ในน้ำหนักการเชื่อมต่อ ไม่ได้แยกเก็บเป็นแบบอ้างอิงสำหรับแต่ละคำ อย่างไรก็ตามถ้าเพิ่มจำนวนคำขึ้นมาก ๆ ต้องเพิ่มขนาดของนิวรอลเน็ตเวิร์กขึ้นด้วย เพื่อให้นิวรอลเน็ตเวิร์กมีความสามารถเพียงพอในการรู้จำเสียง เป็นเหตุให้ยุ่งยากในการดำเนินการเพราะต้องใช้ทรัพยากรมากทั้งหน่วยความจำและเวลาในการฝึกหัด

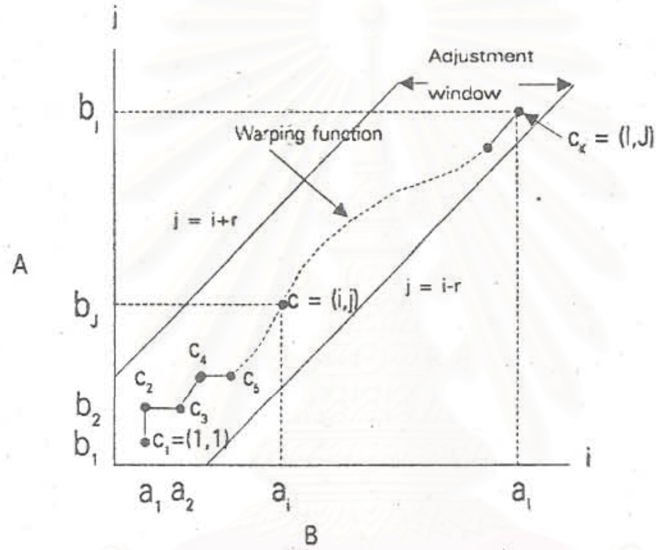
2.3.1. ไดนามิก ไทม์วาร์ปิง (Dynamic Time Warping) (Furui, 1989)

เทคนิคของ dynamic programming ที่นำมาใช้กับ time registration ของแบบทดสอบและแบบอ้างอิงอย่างกว้างขวาง ใน isolated word recognition จากที่ผ่านมาพื้นฐานของ time warping algorithm ถูกนำเสนอโดย Sakoe และ Chiba และ Rabiner , Rosenberg และ Levinson ซึ่ง algorithms เหล่านี้จะใช้ time input ในรูปของ time pattern of feature vector ซึ่งได้จาก isolated word ซึ่งรู้จุดสิ้นสุดที่แน่นอน ซึ่งในการวิเคราะห์เราจะกำหนดลำดับของข้อมูลตามแกนเวลาจำนวน 2 ชุด ซึ่งในแต่ละชุดจะประกอบด้วยเวกเตอร์ตัวแปร (feature vectors) คือ

$$\begin{aligned} A &= a_1, a_2, \ldots, a_i, \ldots, a_I \\ B &= b_1, b_2, \ldots, b_j, \ldots, b_J \end{aligned} \qquad (2.69)$$

โดยที่ A และ B จะถูกแสดงอยู่บนระนาบ i-j ดังแสดงในรูปที่ 2.13 ซึ่ง time warping function จะแทนตำแหน่งของจุดต่าง ๆ ในระนาบ i-j เมื่อ c = (i,j) แทนจุดต่าง ๆ ในระนาบ i-j จะสามารถเขียนลำดับได้เป็น

$$F = c_1, c_2, \ldots, c_k, \ldots c_K \qquad (2.70)$$



รูปที่ 2.13 ไดนาร์มิก ไทม์วาร์ปิง ระหว่าง A และ B (Furui, 1989)

โดยที่ d(c) = d(i,j) จะแทน spectral distance ระหว่าง feature vectors ทั้งสอง a_i และ b_j และผลรวมของ distance ตาม F จะสามารถหาได้จาก

$$D(F) = \frac{\sum_{k=1}^{K} d(c_k) w_k}{\sum_{k=1}^{K} w_k} \qquad (2.71)$$

ค่า D(F) ที่คำนวณได้ยิ่งมีค่าน้อยจะถือว่า feature vectors ระหว่าง A และ B เป็น feature vectors ที่ใกล้เคียงกันที่สุด โดยที่ w_k เป็นสัมประสิทธิ์น้ำหนัก (weight cofficient) ซึ่งจะทำให้การวัดมีความยืดหยุ่นขึ้น โดยที่ตัวหาร $\sum w_k$ จะเป็นตัวชดเชยค่าของ k โดยที่สมการที่ (2.71) จะสามารถเปลี่ยนให้อยู่ในรูปที่ง่ายตาม function บน F ภายใต้เงื่อนไขดังต่อไปนี้

ก. เงื่อนไขโมโนโทนิคและความต่อเนื่อง (Monotony and continuity condition)

$$\begin{aligned} 0 \le i_k - i_{k-1} \le 1 \\ 0 \le j_k - j_{k-1} \le 1 \end{aligned} \qquad (2.72)$$

ตัวอย่างรูปแบบแสดงในรูปที่ 2.14 โดยที่สามารถเขียนเส้นทางการเดินทางไปยังจุด (n,m) ได้ 3 เส้นทางตาม local constraints รูปแบบที่ 1 ได้คือ

$$P \rightarrow (1,0)\ (1,1)$$
$$P \rightarrow (1,1) \quad \text{...(2.73)}$$
$$P \rightarrow (0,1)\ (1,1)$$

ข. เงื่อนไขขอบเขต (Boundary condition)

$$\begin{aligned} i(1) &= 1\ ,\ j(1) = 1 \\ i(K) &= I\ ,\ j(K) = J \end{aligned} \quad \text{..(2.74)}$$

ค. เงื่อนไขหน้าต่างการปรับตัว (Adjustment window condition)

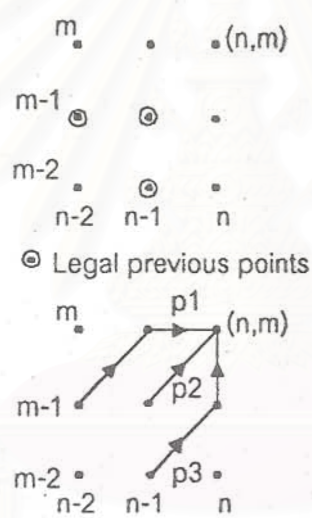
$$|i_k - j_k| \le r \quad ,\ r = \text{ ค่าคงที่} \quad \text{..(2.75)}$$

ถ้า $w_k = (i_k - i_{k-1}) + (j_k - j_{k-1}), (i_0 = j_0 = 0)$ จะได้ว่า

$$\sum_{k=1}^{K} w_k = I + J \quad \text{...(2.76)}$$

จากสมการที่ (2.71) จะได้ว่า

$$D(F) = \frac{1}{I+J}\sum_{k=1}^{K} d(c_k) w_k \quad \text{..(2.77)}$$



Type 1 local constraints

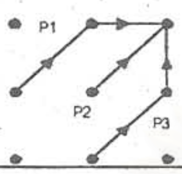
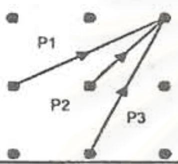
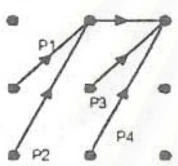
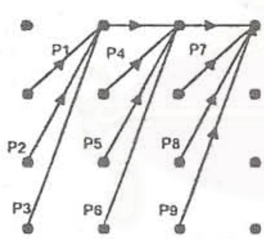
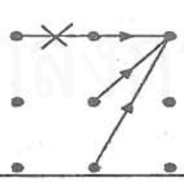
รูป 2.14 แสดง local path constraints ที่ไปยังจุด (n,m) (Myers,Rabiner, and Ronsenberg, 1980)

$$\begin{aligned} w_k &= \min(i(k)-i(k-1), j(k)-j(k-1)) \\ w_k &= \max(i(k)-i(k-1), j(k)-j(k-1)) \\ w_k &= i(k)-i(k-1) \\ w_k &= i(k)-i(k-1)+j(k)-j(k-1) \end{aligned} \quad \text{.........................(2.78)}$$

จากในรูปที่ 2.15 จะแสดงการ ให้น้ำหนัก (weighting) 4 แบบที่ใช้กับtype 2 ในตารางที่ 2.3 โดยที่ i(0) = j(0) = 0 การให้น้ำหนักจะเท่ากันหมดในแบบ (a) ส่วนในแบบ (b) นั้นค่าของความชัน (slope) เป็น 1/2 และ 2 จะมี การให้น้ำหนักมากกว่าที่ความชัน 1 ส่วนแบบ (c) การให้น้ำหนักจะขึ้นกับ distance ที่เคลื่อนที่ไปตามแกน x สำหรับในแบบ (d) นั้นการให้น้ำหนักจะเป็นไปตาม distance ที่เคลื่อนไปตามแกน x และแกน y ส่วนในรูปที่ 2.12 จะแสดงการ

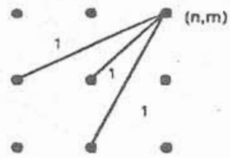
weight ที่ประยุกต์ใช้กับ Type 1 constraints ดังในรูป ทางด้านซ้ายมือ ส่วนทางด้านขวามือจะใช้ smoothing function กับการ weight ซึ่งเสนอโดย Sakoe และ Chiba สำหรับ Type (c) และ Type (d) ผลการ normalization จะได้ว่า

ตารางที่ 2.3 ตัวอย่างของชนิดของ local constraints (Myers et al., 1980)

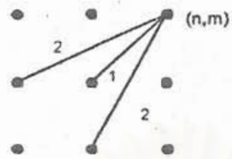
Type	pictorial	productions	E_{max}	E_{min}
1		P1 → (1,0)(1,1) P2 → (1,1) P3 → (0,1)(1,1)	2	1/2
2		P1 → (2,1) P2 → (1,1) P3 → (1,2)	2	1/2
3		P1 → (1,0)(1,1) P2 → (1,0)(1,2) P3 → (1,1) P4 → (1,2)	2	1/2
4		P1 → (1,0)(1,0)(1,1) P2 → (1,0)(1,0)(1,2) P3 → (1,0)(1,0)(1,3) P4 → (1,0)(1,1) P5 → (1,0)(1,2) P6 → (1,0)(1,3) P7 → (1,1) P8 → (1,2) P9 → (1,3)	3	1/3
itakura		no production rule characterization	2	1/2

$$
\begin{aligned}
N(w_c) &= \sum_{k=1}^{K} i(k)-i(k-1) = i(K)-i(0) = I \\
N(w_d) &= \sum_{k=1}^{K} i(k)-i(k-1)+j(k)-j(k-1)) \\
&= i(K)-i(0)+j(K)-j(0) = I+J
\end{aligned}
\qquad \text{........(2.79)}
$$

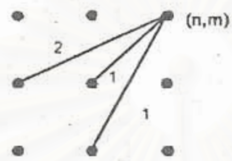
แบบ (a) $w_k = \min(i(k)-i(k-1), j(k)-j(k-1))$



แบบ (b) $w_k = \max(i(k)-i(k-1), j(k)-j(k-1))$



แบบ (c) $w_k = i(k)-i(k-1)$



แบบ (d) $w_k = i(k)-i(k-1)+j(k)-j(k-1)$

รูปที่ 2.15 ตัวอย่างของ weighting function ของ Type 2 constraints (Myers et al., 1980)

จากสมการที่ (2.77) ในส่วนของ $\sum_{k=1}^{K} d(c_k)w_k$ จะเป็นการหาผลบวกที่มีค่าน้อยที่สุดภายใต้เส้นทางเดิน F ตามสมการที่ (2.70) ซึ่งจะสามารถหาผลรวมของ distance ลำดับ $c_1, c_2, \ldots, ck$ ($ck = (i,j)$) ได้ดังนี้

$$
\begin{aligned}
g(c_k) = g(i,j) &= \min_{c_1,\ldots,c_{k-1}} \left[\sum_{m=1}^{k} d(c_m)w_m \right] \\
&= \min_{c_1,\ldots,c_{k-1}} \left[\sum_{m=1}^{k-1} d(c_m)w_m + d(c_k)w_k \right] \\
&= \min_{c_{k-1}} \left[\min_{c_1,\ldots,c_{k-2}} \left\{ \sum_{m=1}^{k-1} d(c_m)w_m \right\} + d(c_k)wk \right] \\
&= \min_{c_{k-1}} \left[g(c_k - 1) + d(c_k)w_k \right] \quad \ldots\ldots\ldots\ldots(2.80)
\end{aligned}
$$

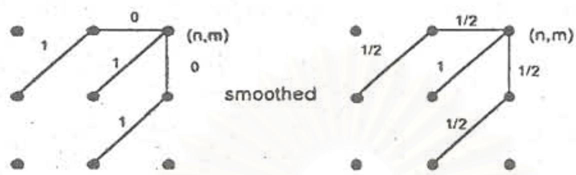
จากสมการที่ (2.80) สามารถเขียนสมการได้เป็น

$$g(i,j) = \min\begin{pmatrix} g(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-1,j)+d(i,j) \end{pmatrix} \quad (2.81)$$

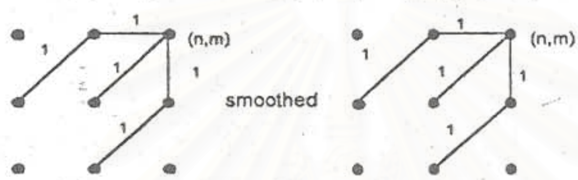
โดยกำหนดเงื่อนไขเริ่มต้นเป็น

$$g(1,1) = 2d(1,1) \quad (2.82)$$

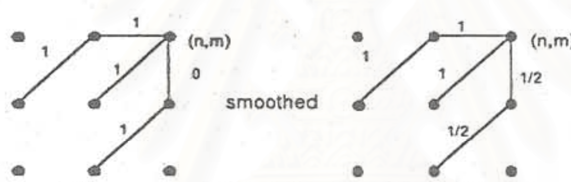
แบบ (a) $w_k = \min(i(k)-i(k-1), j(k)-j(k-1))$



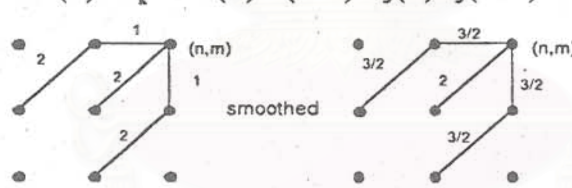
แบบ (b) $w_k = \max(i(k)-i(k-1), j(k)-j(k-1))$



แบบ (c) $w_k = i(k)-i(k-1)$



แบบ (d) $w_k = i(k)-i(k-1)+j(k)-j(k-1)$



รูปที่ 2.16 ตัวอย่างการทำ smoothed weighting function ของ Type 1 constraints (Myers et al., 1980)

การหาระยะทางของการวัดจะกำหนดให้ j = 1 จากนั้นทำการคำนวณตามค่าของ i ตาม adjustment window ที่กำหนด ทำการคำนวณโดยการเปลี่ยนค่าของ j ไปจนกระทั่ง j มีค่าเท่ากับ J ซึ่งจะให้ค่าของการวัดเป็น

$$D(F) = \frac{1}{I+J} G(I,J) \quad (2.83)$$

ในตารางที่ 2.4 แสดงตัวอย่างสมการของรูปแบบเส้นทางของ dynamic time warping แบบต่าง ๆ ส่วนของการทำไดนาร์มิก ไทม์วาร์ปิง จะเป็นแบบสมมาตร ของ P ค่าเท่ากับ 0 นั้น สามารถแสดงได้ดังในรูปที่ 2.17

ในรูปที่ 2.18 จะเป็นการทำ normalize/warp DTW algorithm เพื่อปรับความยาวของแบบทดสอบและแบบอ้างอิง โดยให้อัตราส่วนของจำนวนเฟรมของแบบทดสอบและแบบอ้างอิง ที่ผ่านการปรับ ($\widetilde{N} / \widetilde{M}$) มีค่าเท่ากับ 1 โดยที่การ normalize แบบ อ้างอิง $\widetilde{R}(n)$ จะหาได้จาก

$$\widetilde{R}(\widetilde{n}) = (1-s)R(n)+s(R(n+1)), \quad \widetilde{n} = 1,2,\ldots,\widetilde{N} \quad (2.84)$$

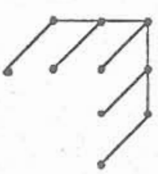
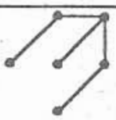
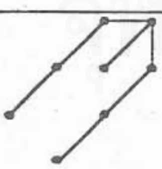
โดยที่ R(n) แทน parameter vector ที่แทนรูปแบบอ้างอิงเฟรมที่ n

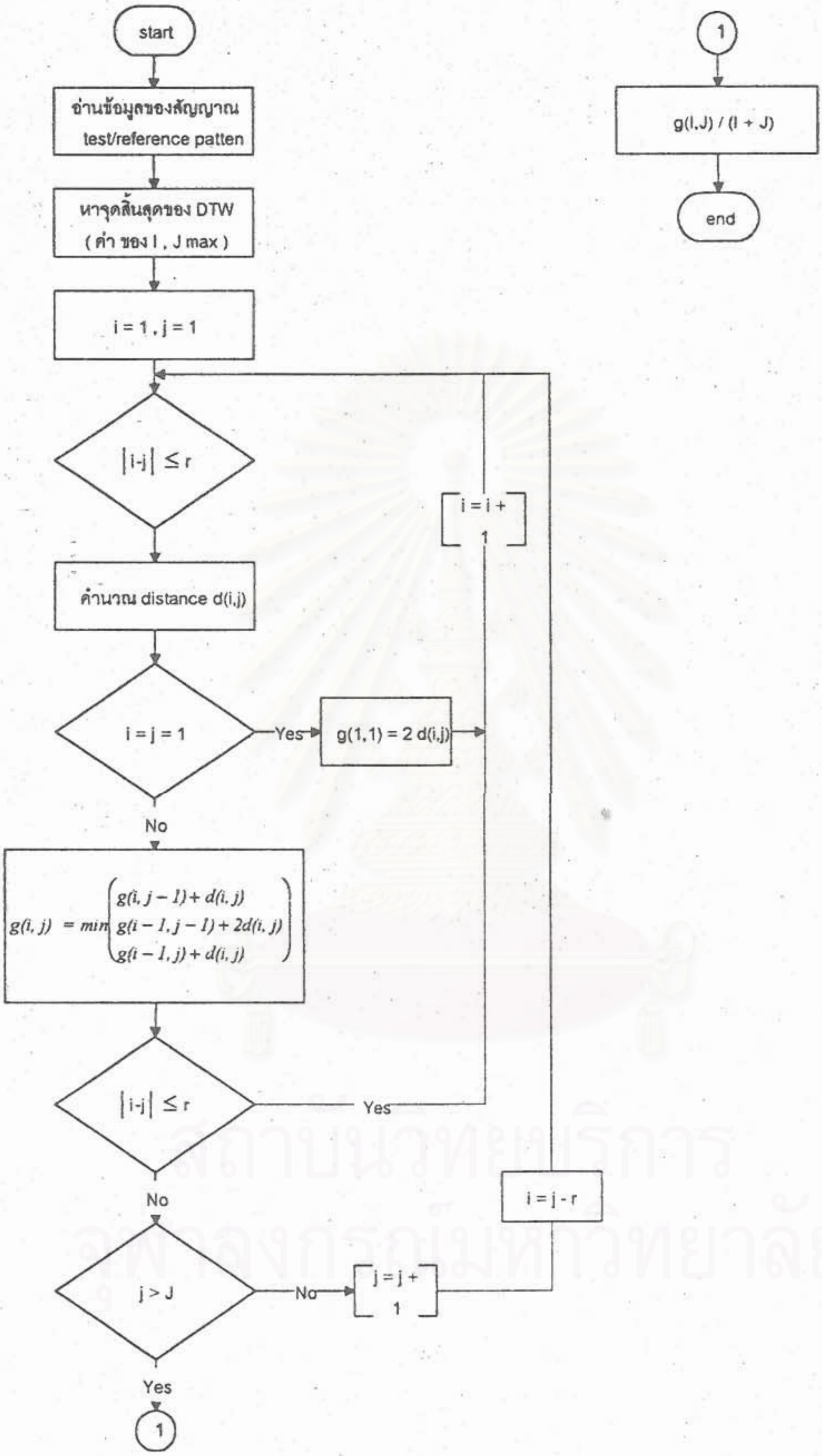
N แทนจำนวนเฟรมของแบบอ้างอิง

$\tilde{R}(n)$ แทน parameter vector ของแบบอ้างอิงที่ normalized

$\tilde{N}$ แทนขนาดของเฟรมของแบบอ้างอิงที่ mormalized

ตารางที่ 2.4 แสดงสมการไดนามิกโปรแกรมมิ่งต่าง ๆ (ไพศาล ธรรมโพธิทองม, 2533 อ้างถึงใน Sakoe, 1978)

P	แผนภาพ แสดงทางเดิน	สมมาตร / ไม่สมมาตร	สมการไดนามิกโปรแกรมมิ่ง g(i,j) =
0		สมมาตร	$\min \left[\begin{array}{l} g(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-1, j) + d(i, j) \end{array} \right.$
		ไม่สมมาตร	$\min \left[\begin{array}{l} g(i, j-1) \\ g(i-1, j-1) + d(i, j) \\ g(i-1, j) + d(i, j) \end{array} \right.$
1/2		สมมาตร	$\min \left[\begin{array}{l} g(i-1, j-3) + 2d(i, j-2) + d(i, j-1) + d(i, j) \\ g(i-1, j-2) + 2d(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-2, j-1) + 2d(i-1, j) + d(i, j) \\ g(i-3, j-1) + 2d(i-2, j) + d(i-1, j) + d(i, j) \end{array} \right.$
		ไม่สมมาตร	$\min \left[\begin{array}{l} g(i-1, j-3) + (d(i, j-2) + d(i, j-1) + d(i, j)) / 3 \\ g(i-1, j-2) + (d(i, j-1) + d(i, j)) / 2 \\ g(i-1, j-1) + d(i, j) \\ g(i-2, j-1) + d(i-1, j) + d(i, j) \\ g(i-3, j-1) + d(i-2, j) + d(i-1, j) + d(i, j) \end{array} \right.$
1		สมมาตร	$\min \left[\begin{array}{l} g(i-1, j-2) + 2d(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-2, j-1) + 2d(i-1, j) + d(i, j) \end{array} \right.$
		ไม่สมมาตร	$\min \left[\begin{array}{l} g(i-1, j-2) + (d(i, j-1) + d(i, j)) / 2 \\ g(i-1, j-1) + d(i, j) \\ g(i-2, j-1) + d(i-1, j) + d(i, j) \end{array} \right.$
2		สมมาตร	$\min \left[\begin{array}{l} g(i-2, j-3) + 2d(i-1, j-1) + 2d(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-3, j-2) + 2d(i-2, j-1) + 2d(i-1, j) + d(i, j) \end{array} \right.$
		ไม่สมมาตร	$\min \left[\begin{array}{l} g(i-2, j-3) + 2(d(i-1, j-1) + 2d(i, j-1) + d(i, j)) / 3 \\ g(i-1, j-1) + d(i, j) \\ g(i-3, j-2) + d(i-2, j-1) + d(i-1, j) + d(i, j) \end{array} \right.$



รูปที่ 2.17 ขั้นตอนการทำไดนาร์มิก ไทม์วาร์ปิง

และ จะหาค่าของ n และ s ในสมการที่ (2.84) ได้จาก

$$n = \left\lfloor (\tilde{n}-1)\frac{(N-1)}{(\tilde{N}-1)} + 1 \right\rfloor \quad \text{(2.85ก)}$$

$$s = (\tilde{n}-1)\frac{(N-1)}{(\tilde{N}-1)} + 1 - n \quad \text{(2.85ข)}$$

ในลักษณะเดียวกัน จะสามารถทำการ normalize แบบทดสอบได้จาก

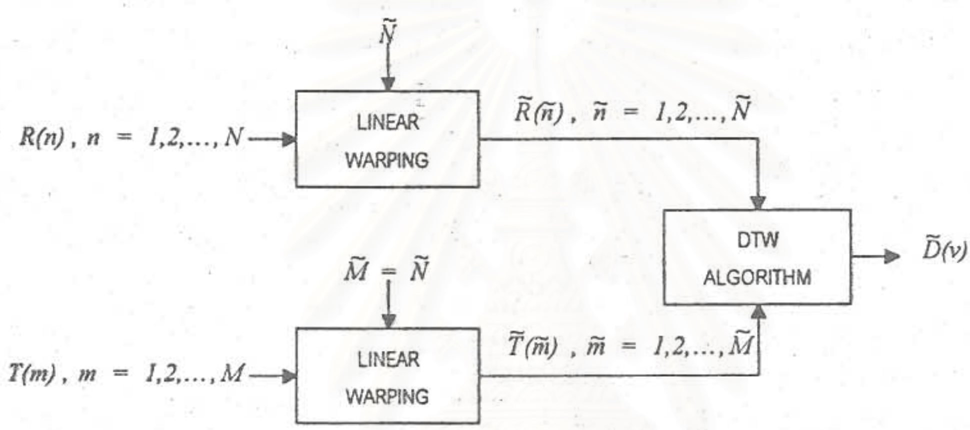
$$\tilde{T}(\tilde{m}) = (1-s)T(m) + s(T(m+1)), \quad \tilde{m} = 1,2,\ldots,\tilde{M} \quad \text{(2.86)}$$

โดยที่ T(m) แทน parameter vector ที่แทนรูปแบบทดสอบกรอบที่ n

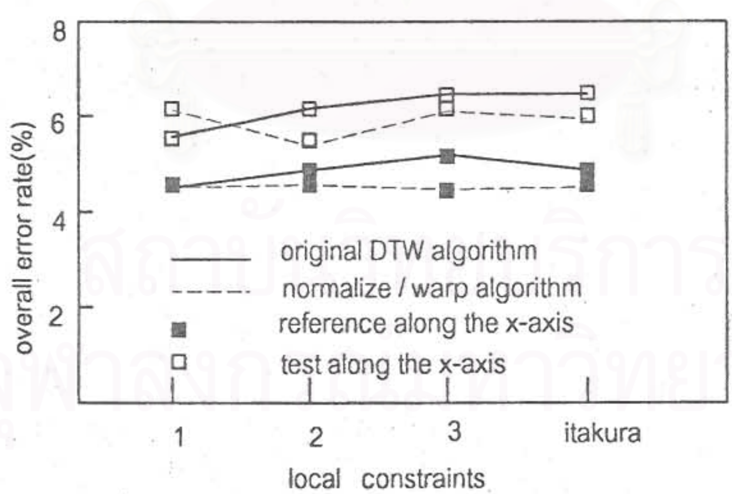
M แทนจำนวนกรอบของแบบอ้างอิง

$\tilde{T}(\tilde{m})$ แทน parameter vector ของแบบทดสอบที่ได้รับการนอร์แมลไลซ์

$\tilde{M}$ แทนขนาดของเฟรมของแบบทดสอบที่ได้รับการนอร์แมลไลซ์



รูปที่ 2.18 แสดงการใช้ normalize/warp DTW algorithm (Myers et al., 1980)



รูปที่ 2.19 แสดงการเปรียบเทียบระหว่าง standard DTW และ normalize/warp DTW (Myers et al., 1980)

และ จะหาค่าของ n และ s ในสมการที่ (2.84) ได้จาก

$$m = \left\lfloor (\tilde{m}-1)\frac{(M-1)}{(\tilde{M}-1)} + 1 \right\rfloor \quad \text{(2.87ก)}$$

$$s = (\widetilde{m}-1)\frac{(M-1)}{(\widetilde{M}-1)}+1-m \quad \text{.......................................(2.87ข)}$$

ส่วน $\widetilde{D}(v)$ จะเป็นผลลัพธ์จากการทำไดนาร์มิก ไทม์วาร์ปิง ระหว่าง $\widetilde{R}(n)$ และ $\widetilde{T}(\widetilde{m})$

จากสมการที่ (2.85ก) และ (2.87ก) ในเทอมของ [x] จะเป็นค่าของเลขจำนวนเต็มที่มีค่าไม่เกิน x ในส่วนของแบบอ้างอิงและแบบทดสอบตามลำดับ ซึ่งผลการทดสอบที่ผ่านมาแสดงดังในรูปที่ 2.19 โดยใช้ค่าของ $\widetilde{N}$ เท่ากับ 40

2.3.2. แบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Model)

แบบจำลองฮิดเดน มาร์คอฟ นี้ถือเป็นขั้นตอนวิธีการจำแนกรูปแบบที่ดีที่สุดวิธีการหนึ่งที่มีอยู่ในขณะนี้โดยอาศัยวิธีการทางสถิติ (Roe and Wilpon, 1993) ขั้นตอนวิธีการนี้มีข้อได้เปรียบที่สำคัญเหนือวิธีการเข้าคู่ต้นแบบก็คือ สามารถเก็บข้อมูลรายละเอียดในทางสถิติเกี่ยวกับเสียงพูดไว้ได้มากกว่าวิธีการเข้าคู่ต้นแบบ โดยเก็บข้อมูลการกระจายที่สมบูรณ์ของลักษณะสำคัญที่มีอยู่ในข้อมูลฝึกฝน จึงสามารถจำแนกความแตกต่างระหว่างเสียงพูดได้ดีมากยิ่งขึ้น อีกทั้งขั้นตอนวิธีการนี้ยังอาศัยการโปรแกรมแบบพลวัต (Dynamic Programming) ทำให้มีความรวดเร็วในการประมวลผลมากยิ่งขึ้น

เหตุผลที่ระบบแบบจำลองฮิดเดน มาร์คอฟ เป็นที่นิยมมีด้วยกัน 2 ประการ (Rabiner, 1989) ประการแรก แบบจำลองนี้อาศัยโครงสร้างทางคณิตศาสตร์และสามารถเปลี่ยนแปลงทฤษฎีพื้นฐานเพื่อประยุกต์ใช้งานได้อย่างกว้างขวาง ประการที่สอง แบบจำลองนี้สามารถทำงานได้เป็นอย่างดีเมื่อประยุกต์ใช้งานอย่างเหมาะสม

สำหรับกระบวนการโดยทั่วไปจะสร้างผลลัพธ์ที่สังเกตได้ซึ่งสามารถแสดงได้เป็นสัญญาณ สัญญาณอาจมีความไม่ต่อเนื่องโดยธรรมชาติหรือมีความต่อเนื่องโดยธรรมชาติก็ได้ แหล่งของสัญญาณอาจเป็นได้ทั้งที่ไม่แปรเปลี่ยนตามเวลาหรือแปรเปลี่ยนตามเวลาก็ได้ ปัญหาที่สำคัญก็คือการสร้างแบบจำลองสัญญาณขึ้นมา โดยมีเหตุผล 3 ประการ (Rabiner, 1989) ดังนี้ ประการแรก แบบจำลองของสัญญาณเป็นพื้นฐานของรายละเอียดในทางทฤษฎีของระบบประมวลผลสัญญาณ ซึ่งใช้ในการประมวลผลสัญญาณเพื่อให้ได้ผลลัพธ์ตามต้องการ ประการที่สอง ช่วยในการศึกษาแหล่งของสัญญาณโดยปราศจากแหล่งของสัญญาณ ประการที่สาม สามารถใช้งานได้เป็นอย่างดีในทางปฏิบัติ

ประเภทของแบบจำลองสัญญาณสามารถแบ่งได้เป็น 2 ประเภท ได้แก่ แบบจำลองที่กำหนดถาวร (Deterministic Models) และแบบจำลองทางสถิติ (Statistical Models) แบบจำลองที่กำหนดถาวรจะบอกถึงคุณสมบัติเฉพาะของสัญญาณ โดยอาศัยเพียงการประมาณค่าพารามิเตอร์ที่จำเป็นให้แก่แบบจำลองสัญญาณเท่านั้น ส่วนแบบจำลองทางสถิติจะอาศัยคุณสมบัติทางสถิติของสัญญาณในการบอกคุณสมบัติของสัญญาณ โดยอาศัยสมมติฐานที่ว่าสัญญาณสามารถแสดงได้ด้วยกระบวนการสุ่มแบบพาราเมตริก และค่าพารามิเตอร์ของกระบวนการเฟ้นสุ่มสามารถประมาณค่าได้อย่างแม่นยำ

ส่วนแบบจำลองฮิดเดน มาร์คอฟ จัดอยู่ในประเภทหนึ่งของแบบจำลองสัญญาณเฟ้นสุ่ม โดยการออกแบบเป็นการแก้ไขปัญหาพื้นฐานสำคัญ 3 ประการของแบบจำลองฮิดเดน มาร์คอฟ (Rabiner, 1989) ได้แก่ การประเมินค่าความน่าจะเป็นของลำดับค่าสังเกตสำหรับเฉพาะแบบจำลองฮิดเดน มาร์คอฟ การหาลำดับที่ดีที่สุดสำหรับแต่ละสถานะของแบบจำลอง และการปรับค่าพารามิเตอร์เพื่อให้เหมาะสมที่สุดกับสัญญาณที่สังเกต

2.3.2.1. องค์ประกอบของแบบจำลองฮิดเดน มาร์คอฟ

องค์ประกอบของแบบจำลองฮิดเดน มาร์คอฟ ประกอบด้วยพารามิเตอร์ต่างๆ คือ

1) $\underline{N}$ คือจำนวนสถานะที่อยู่ภายในแบบจำลอง ซึ่งโดยทั่วไปแล้วแต่ละสถานะจะเชื่อมโยงถึงกันด้วยวิธีการที่จะทำให้สถานะใดๆ สามารถเข้าถึงสถานะอื่นๆ ได้ แต่ละสถานะแสดงได้ด้วย $S=\{S_1,S_2,\Lambda\ ,S_N\}$ โดยมีสถานะที่เวลา t แสดงได้ด้วย q_t

2) M คือจำนวนสัญลักษณ์ของค่าสังเกตต่อสถานะ ซึ่งสัญลักษณ์ของค่าสังเกตจะสัมพันธ์กับผลลัพธ์ขาออกทางกายภาพของระบบที่ถูกจำลอง แต่ละสัญลักษณ์สามารถแสดงได้ด้วย $V = \{v_1, v_2, \Lambda, v_M\}$

3) <u>การกระจายของความน่าจะเป็นในการเปลี่ยนแปลงสถานะ (State Transition Probability Distribution)</u> $A = \{a_{ij}\}$ เมื่อ

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \le i, j \le N \qquad (2.88)$$

ในกรณีเฉพาะที่สถานะใดๆ สามารถเข้าถึงสถานะอื่นได้ภายในขั้นตอนเดียวจะกำหนดให้ $a_{ij} > 0, \ \forall i, j$ ส่วนในกรณีอื่นนอกเหนือจากนี้จะกำหนดให้ $a_{ij} = 0$ สำหรับ (i, j) เพียงคู่เดียวหรือมากกว่า

4) <u>การกระจายของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต (Observation Symbol Probability Distribution)</u> $B = \{b_j(k)\}$ ในสถานะที่ j เมื่อ

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j], \quad \begin{matrix} 1 \le j \le N \\ 1 \le k \le M \end{matrix} \qquad (2.89)$$

ตารางที่ 2.5 รายละเอียดของขั้นตอนในการกำเนิดลำดับค่าสังเกต

ขั้นตอนที่ 1	เลือกสถานะเริ่มต้น $q_1 = S_i$ ที่สัมพันธ์กับการกระจายของสภาวะเริ่มต้น π
ขั้นตอนที่ 2	กำหนดให้ $t = 1$
ขั้นตอนที่ 3	เลือก $O_t = v_k$ ที่สัมพันธ์กับการกระจายของความน่าจะเป็นของสัญลักษณ์เมื่ออยู่ในสถานะ S_i เช่น $b_i(k)$
ขั้นตอนที่ 4	เคลื่อนย้ายไปยังสถานะใหม่ $q_{t+1} = S_j$ ที่สัมพันธ์กับการกระจายของความน่าจะเป็นในการเปลี่ยนแปลงสถานะสำหรับสถานะ S_i เช่น a_{ij}
ขั้นตอนที่ 5	กำหนดให้ $t = t + 1$ แล้วกลับไปทำซ้ำขั้นตอนที่ 3 ใหม่ถ้า $t < T$ นอกเหนือจากนี้ให้ยุติกระบวนการ

5) <u>การกระจายของสภาวะเริ่มต้น (Initial State Distribution)</u> $\pi = \{\pi_i\}$ เมื่อ

$$\pi_i = P[q_1 = S_i], \quad 1 \le i \le N \qquad (2.90)$$

โดยการกำหนดค่าที่เหมาะสมให้กับองค์ประกอบ N, M, A, B, π ของแบบจำลองฮิดเดน มาร์คอฟ ซึ่งใช้ในการกำเนิดลำดับค่าสังเกต เมื่อแต่ละค่าสังเกต O_t เป็นสัญลักษณ์ที่ได้จาก V และ T เป็นจำนวนค่าสังเกตทั้งหมดที่มีในลำดับ ซึ่งมีขั้นตอนวิธีการดังนี้

$$O = O_1\, O_2\, \Lambda\ \ O_T \qquad (2.91)$$

ขั้นตอนดังกล่าวนี้สามารถได้ทั้งการกำเนิดค่าสังเกต และเป็นแบบจำลองเพื่อบอกถึงความเหมาะสมในการกำเนิดลำดับค่าสังเกตด้วยแบบจำลองฮิดเดน มาร์คอฟ ดังนั้นการกำหนดคุณสมบัติเฉพาะของแบบจำลองฮิดเดน มาร์คอฟ ต้องการคุณสมบัติเฉพาะของพารามิเตอร์ของแบบจำลองสองค่า (N และ M) คุณสมบัติเฉพาะของสัญลักษณ์ของค่าสังเกต และคุณสมบัติเฉพาะของการวัดค่าความน่าจะเป็นได้แก่ A, B, π โดยเขียนอยู่ในรูปแบบย่อเพื่อบ่งบอกชุดของพารามิเตอร์ที่สมบูรณ์ของแบบจำลองดังนี้

$$\lambda = \left(A,\ B,\ \pi\right) \qquad (2.92)$$

2.3.2.2. ปัญหาพื้นฐานสามประการของแบบจำลองฮิดเดน มาร์คอฟ

ในการประยุกต์ใช้งานแบบจำลองฮิดเดน มาร์คอฟ ในทางปฏิบัตินั้น ก็คือการแก้ไขปัญหาพื้นฐานทั้งสามประการ โดยมีรายละเอียดแสดงในตารางที่ 2.6

- ปัญหาพื้นฐานข้อแรก คือปัญหาในการประเมินค่า ซึ่งก็คือการหาค่าความน่าจะเป็นของลำดับค่าสังเกตที่สร้างจากแบบจำลอง เมื่อกำหนดแบบจำลองและลำดับค่าสังเกตมาให้ หรืออีกนัยหนึ่งก็คือการแสดงว่าแบบจำลองที่กำหนดให้เข้าคู่กันได้กับลำดับค่าสังเกตที่กำหนดให้ได้ดีเพียงใด ตัวอย่างเช่น ถ้าในกรณีที่พิจารณาเลือกระหว่างแบบจำลองหลายแบบ ผลลัพธ์ของปัญหาพื้นฐานข้อแรกจะช่วยในการเลือกแบบจำลองที่เข้าคู่กันได้ดีที่สุดกับค่าสังเกต
- ปัญหาพื้นฐานข้อที่สอง คือความพยายามในการเปิดเผยส่วนที่แบบจำลองปิดบังไว้ ในเฉพาะกรณีของแบบจำลองที่ด้อยประสิทธิภาพจะไม่มีลำดับสถานะที่ถูกต้อง ดังนั้นในทางปฏิบัติจึงใช้กฎเกณฑ์ของความเหมาะสมที่สุดในการแก้ไขปัญหานี้ซึ่งมีด้วยกันหลายประเภท ดังนั้นการเลือกกฎเกณฑ์จึงเท่ากับเป็นการเปิดเผยลำดับสถานะที่ถูกปกปิดโดยแบบจำลอง
- ปัญหาพื้นฐานข้อที่สาม คือการทำให้พารามิเตอร์ของแบบจำลองมีประสิทธิภาพมากที่สุด เพื่อที่จะอธิบายลำดับค่าสังเกตได้ดีที่สุด ลำดับค่าสังเกตที่ใช้ในการปรับพารามิเตอร์ของแบบจำลองเรียกว่า "ลำดับฝึกฝน" (Training Sequences) เนื่องจากถูกใช้ในการฝึกฝนแบบจำลองฮิดเดน มาร์คอฟ ปัญหาในการฝึกฝนนี้จะช่วยให้ปรับแต่งพารามิเตอร์ของแบบจำลองให้เหมาะสมมากที่สุดกับข้อมูลฝึกฝนที่สังเกต

ในการประยุกต์ใช้งานแบบจำลองฮิดเดน มาร์คอฟ กับการรู้จำคำพูดนั้น เริ่มต้นจากการออกแบบสร้างแบบจำลองฮิดเดน มาร์คอฟ N สถานะสำหรับแต่ละคำของชุดคำศัพท์ W คำ สัญญาณเสียงพูดของแต่ละคำจะถูกแทนที่ด้วยลำดับเวลาของเวกเตอร์รหัสเชิงสเปกตรัม การเข้ารหัสจะอาศัยชุดรหัสเชิงสเปกตรัมที่ประกอบด้วยเวกเตอร์เชิงสเปกตรัม M เวกเตอร์ที่เป็นเอกลักษณ์ จึงทำให้แต่ละค่าสังเกตจะเป็นดรรชนีของเวกเตอร์เชิงสเปกตรัมที่ใกล้เคียงกับสัญญาณเสียงพูดต้นฉบับมากที่สุด ดังนั้นแต่ละคำศัพท์จะมีลำดับการฝึกฝนที่ประกอบด้วยจำนวนลำดับของดรรชนีชุดรหัสของคำ ขั้นตอนแรกเริ่มจากการสร้างแบบจำลองของแต่ละคำโดยการแก้ปัญหาพื้นฐานข้อที่ 3 เพื่อประมาณค่าพารามิเตอร์ของแบบจำลองให้เหมาะสมที่สุดสำหรับแต่ละแบบจำลอง ขั้นตอนที่สองเป็นการสร้างความเข้าใจในความหมายทางกายภาพของสถานะของแบบจำลองโดยการแก้ปัญหาพื้นฐานข้อที่ 2 เพื่อแบ่งแยกแต่ละลำดับฝึกฝนของคำไปยังแต่ละสถานะ และศึกษาถึงคุณสมบัติของเวกเตอร์เชิงสเปกตรัมที่ทำให้เกิดค่าสังเกตในแต่ละสถานะ โดยในขั้นตอนนี้จะทำการปรับแต่งแบบจำลองเพื่อเพิ่มพูนความสามารถในการจำลองแบบลำดับคำพูด ขั้นตอนสุดท้ายภายหลังจากการออกแบบชุดของแบบจำลองฮิดเดน มาร์คอฟ ทั้ง W ชุดพร้อมทั้งปรับให้มีประสิทธิภาพที่เหมาะสมแล้ว การรู้จำคำพูดที่ไม่รู้จักมาก่อนจะอาศัยการแก้ปัญหาพื้นฐานข้อที่ 1 เพื่อให้คะแนนแต่ละแบบจำลองของคำพูดด้วยลำดับค่าสังเกตที่ใช้ทดสอบและเลือกคำซึ่งมีแบบจำลองที่ให้คะแนนสูงที่สุด

ตารางที่ 2.6 รายละเอียดของปัญหาพื้นฐานสามประการของแบบจำลองฮิดเดน มาร์คอฟ

ปัญหาพื้นฐานข้อที่ 1	เมื่อกำหนดลำดับค่าสังเกต $O = O_1\ O_2\ \Lambda\ \ O_T$ และแบบจำลอง $\lambda = (A,\ B,\ \pi)$ จะทำการหาค่าความน่าจะเป็นของลำดับค่าสังเกต $P(O\|\lambda)$ ตามแบบจำลองที่กำหนดให้ได้อย่างไร
ปัญหาพื้นฐานข้อที่ 2	เมื่อกำหนดลำดับค่าสังเกต $O = O_1\ O_2\ \Lambda\ \ O_T$ และแบบจำลอง λ จะทำการเลือกลำดับสถานะที่สัมพันธ์กับ $Q = q_1\ q_2\ \Lambda\ \ q_T$ ซึ่งมีความเหมาะสมที่สุดกับแบบจำลองที่กำหนดให้ได้อย่างไร
ปัญหาพื้นฐานข้อที่ 3	จะทำการปรับค่าพารามิเตอร์ของแบบจำลอง $\lambda = (A,\ B,\ \pi)$ อย่างไรเพื่อให้ค่าความน่าจะเป็นของลำดับค่าสังเกต $P(O\|\lambda)$ มีค่ามากที่สุด

2.3.2.3. การแก้ไขปัญหาพื้นฐานสามประการของแบบจำลองฮิดเดน มาร์คอฟ

ในการแก้ไขปัญหาพื้นฐานทั้งสามประการของแบบจำลองฮิดเดน มาร์คอฟ เพื่อใช้ในการรู้จำคำพูดตามแนวทางดังกล่าวข้างต้นสามารถแสดงได้ดังนี้

2.3.2.3.1 การแก้ไขปัญหาพื้นฐานข้อที่ 1

เมื่อกำหนดแบบจำลอง l การหาค่าความน่าจะเป็น $P(O|\lambda)$ ของลำดับค่าสังเกต $O = O_1\ O_2\ \Lambda\ \ O_T$ จะอาศัยวิธีการหาค่าทุกลำดับสถานะความยาว T ที่เป็นไปได้ทั้งหมดตามลำดับ เมื่อ T เป็นจำนวนค่าสังเกต พิจารณาลำดับของสถานะที่มีค่าจำกัดดังนี้

$$Q = q_1 \quad q_2 \quad \Lambda \quad q_T \tag{2.93}$$

เมื่อ q_1 เป็นสถานะเริ่มต้น ค่าความน่าจะเป็นของลำดับค่าสังเกต O สำหรับลำดับสถานะในสมการที่ (2.88) ดังนี้

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda) \tag{2.94}$$

เมื่อกำหนดให้ค่าสังเกตเป็นชนิดไม่ขึ้นแก่กันในทางสถิติจะได้ว่า

$$P(O|Q,\lambda) = b_{q_1}(O_1)\cdot b_{q_2}(O_2)\ \Lambda\ \ b_{q_T}(O_T) \tag{2.95}$$

ดังนั้นค่าความน่าจะเป็นของลำดับสถานะ Q สามารถเขียนได้เป็น

$$P(Q|\lambda) = \pi_{q_1} a_{q_1q_2} a_{q_2q_3} \Lambda\ a_{q_{T-1}q_T} \tag{2.96}$$

ค่าความน่าจะเป็นร่วมระหว่าง O และ Q เป็นเพียงผลคูณของสมการข้างต้นดังนี้

$$P(O,Q|\lambda) = P(O|Q,\lambda)P(Q,\lambda) \tag{2.97}$$

ค่าความน่าจำเป็นของ O เมื่อกำหนดแบบจำลองให้ สามารถหาได้โดยผลรวมของค่าความน่าจะเป็นร่วมของลำดับสถานะทั้งหมดที่เป็นไปได้ q ดังนี้

$$P(Q|\lambda) = \sum_{\text{all } Q} P(O|Q,\lambda)\,P(Q,\lambda)$$

$$= \sum_{q_1, q_2, \Lambda, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \Lambda\ a_{q_{T-1} q_T} b_{q_T}(O_T) \quad \text{.......... (2.98)}$$

จากสมการที่ (2.98) เริ่มต้นที่เวลา t = 1 จะเริ่มที่สถานะ q_1 ด้วยค่าความน่าจะเป็น π_{q_1} และให้กำเนิดสัญลักษณ์ O_1 ภายในสถานะเดียวกันด้วยค่าความน่าจะเป็น $b_{q_1}(O_1)$ เมื่อเวลาเปลี่ยนจาก t เป็น t + 1 (t = 2) และเปลี่ยนไปยังสถานะ q_2 จากสถานะ q_1 ด้วยค่าความน่าจะเป็น $a_{q_1 q_2}$ และให้กำเนิดสัญลักษณ์ O_2 ภายในสถานะเดียวกันด้วยค่าความน่าจะเป็น $b_{q_2}(O_2)$ กระบวนการนี้จะดำเนินไปอย่างต่อเนื่องจนกระทั่งถึงการเปลี่ยนแปลงที่เวลา T จากสถานะ q_{T-1} ไปยังสถานะ q_T ด้วยค่าความน่าจะเป็น $a_{q_{T-1} q_T}$ และให้กำเนิดสัญลักษณ์ O_T ภายในสถานะเดียวกันด้วยค่าความน่าจะเป็น $b_{q_T}(O_T)$

ในการคำนวณค่าความน่าจะเป็น $P(O|\lambda)$ ตามสมการที่ (2.98) นั้นจะเกิดการคำนวณขึ้นด้วยอันดับประมาณ $2T \cdot N^T$ ครั้ง เนื่องจากทุกเวลา $t = 1, 2, \Lambda\ , T$ จะเกิดสถานะที่เป็นไปได้ N สถานะซึ่งเข้าถึงได้ และในแต่ละลำดับสถานะจะเกิดการคำนวณขึ้นประมาณ $2T$ ครั้งสำหรับแต่ละพจน์ในผลรวมของสมการที่ (2.98) ซึ่งจำแนกได้เป็นการคูณ $(2T-1)N^T$ ครั้งและการบวก $N^T - 1$ ครั้ง จึงทำให้การคำนวณค่าโดยใช้สมการนี้เป็นไปไม่ได้ถึงแม้ด้วย N และ T ค่าน้อยๆ ก็ตาม ดังนั้นจึงมีกระบวนการที่มีประสิทธิภาพในการคำนวณค่าที่เรียกว่า "กระบวนการไปหน้า-ย้อนกลับ" (Forward-Backward Procedure) ซึ่งประกอบด้วยกระบวนการไปหน้าและกระบวนการย้อนกลับร่วมกันดังนี้

1) **กระบวนการไปหน้า (Forward Procedure)**

พิจารณาตัวแปรไปหน้า $\alpha_t(i)$ ที่กำหนดไว้แล้วดังนี้

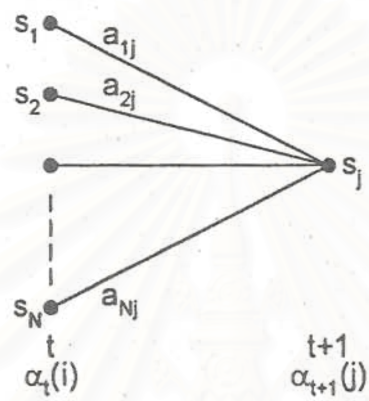
$$\alpha_t(i) = P(O_1\, O_2\, \Lambda\ O_t, q_t = S_i|\lambda) \quad \text{.......... (2.99)}$$

ค่าความน่าจะเป็นของลำดับค่าสังเกตบางส่วน $O_1\, O_2\, \Lambda\ O_t$ และสถานะ S_i ที่เวลา t เมื่อกำหนดแบบจำลอง ł จะสามารถหาค่าของ $\alpha_t(i)$ โดยอุปนัยได้ในตารางที่ 2.7 ซึ่งแสดงถึงรายละเอียดของกระบวนการไปหน้าดังนี้

ตารางที่ 2.7 รายละเอียดกระบวนการไปหน้า

ขั้นตอนที่ 1 **กระบวนการเริ่มต้น**	$\alpha_1(i) = \pi_i b_i(O_1),\quad 1 \le i \le N$ .... (2.100)
ขั้นตอนที่ 2 **กระบวนการอุปนัย**	$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}),\quad 1 \le t \le T-1,\ \ 1 \le i \le N$ .... (2.101)
ขั้นตอนที่ 3 **กระบวนการสิ้นสุด**	$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$ .... (2.102)

ขั้นตอนกระบวนการเริ่มต้น เป็นการกำหนดค่าเริ่มต้นให้กับค่าความน่าจำเป็นแบบไปหน้า โดยกำหนดให้เป็นความน่าจะเป็นร่วมของสถานะ S_i และค่าสังเกตเริ่มต้น O_1 ขั้นตอนกระบวนการอุปนัยถือเป็นหัวใจสำคัญของกระบวนการไปหน้าดังแสดงในรูปที่ 2.20 ซึ่งแสดงถึงการที่สถานะ S_j ที่สามารถเข้าถึงได้ที่เวลา t + 1 จาก N สถานะที่เป็นไปได้ $S_i, 1 \le i \le N$ ที่เวลา t เนื่องจาก $\alpha_t(i)$ เป็นค่าความน่าจะเป็นของเหตุการณ์ร่วมที่สังเกต $O_1\ O_2\ \Lambda\ O_t$ และสถานะที่เวลา t เป็น S_i ดังนั้นผลคูณ $\alpha_t(i)a_{ij}$ จึงเป็นค่าความน่าจะเป็นของเหตุการณ์ร่วมที่สังเกต $O_1\ O_2\ \Lambda\ O_t$ และเข้าถึงสถานะ S_j ที่เวลา t + 1 ผ่านทางสถานะ S_i ที่เวลา t ผลรวมผลคูณของ N สถานะทั้งหมดที่เป็นไปได้ $S_i, 1 \le i \le N$ ที่เวลา t ได้ผลเป็นค่าความน่าจะเป็น S_j ที่เวลา t + 1 ร่วมกับค่าสังเกตบางส่วนทั้งหมดที่สัมพันธ์กัน



รูปที่ 2.20 รายละเอียดลำดับกระบวนการในการคำนวณค่าตัวแปรไปหน้า $\alpha_t(i)$

เมื่อเสร็จสิ้นกระบวนการทั้งหมดและได้ค่า S_j โดยค่า $\alpha_{t+1}(j)$ จะได้มาจากค่าสังเกต O_{t+1} ในสถานะ j การคำนวณในสมการที่ (2.101) จะกระทำกับทุกสถานะ $j, 1 \le j \le N$ เมื่อกำหนดค่า t ให้ จากนั้นจะวนซ้ำกับทุก $t = 1, 2, \Lambda\ , T-1$ ขั้นตอนกระบวนการสิ้นสุด เป็นการคำนวณค่า $P(O|\lambda)$ ด้วยผลรวมของตัวแปรไปหน้าตัวสุดท้าย $\alpha_T(i)$ ซึ่งทำให้ $P(O|\lambda)$ เป็นเพียงผลรวมของ $\alpha_T(i)$ แต่ละตัวมีนิยามดังนี้

$$\alpha_T(i) = P(O_1\ O_2\ \Lambda\ O_T, q_T = S_i|\lambda) \quad \text{(2.103)}$$

เมื่อพิจารณาการคำนวณค่า $\alpha_t(j), 1 \le t \le T, 1 \le j \le N$ นั้น จะเกิดการคำนวณขั้นด้วยอันดับประมาณ N^2T ครั้ง เมื่อเปรียบเทียบกับ $2TN^T$ ครั้งเมื่อคำนวณโดยตรง ซึ่งจำแนกได้เป็นการคูณ $N(N+1)(T-1)+N$ ครั้งและการบวก $N(N-1)(T-1)$ ครั้ง

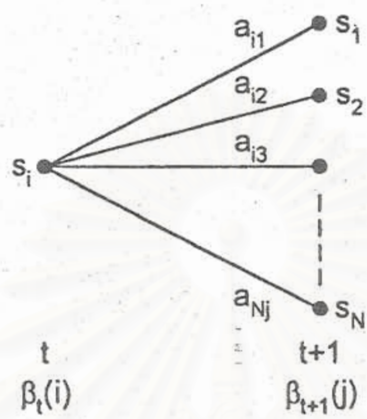
การคำนวณค่าความน่าจะเป็นแบบไปหน้าจะอยู่บนพื้นฐานของโครงสร้าง Lattice หรือ Trellis กุญแจสำคัญก็คือเนื่องจากมีเพียง N สถานะดังนั้นลำดับสถานะทั้งหมดที่เป็นไปได้จะรวมเข้ากับปม N ปมในโครงสร้างไม่ว่าลำดับของค่าสังเกตจะยาวเพียงใดก็ตาม ที่เวลา $t = 1$ ซึ่งเป็นช่องเวลาแรกของโครงสร้าง Lattice จะทำการคำนวณค่าของ $\alpha_1(i), 1 \le i \le N$ ที่เวลา $t = 2, 3, \Lambda\ , T$ จะเป็นเพียงการคำนวณค่าของ $\alpha_t(j), 1 \le j \le N$ โดยแต่ละครั้งจะคำนวณเพียงค่า $\alpha_{t-1}(i)$ จำนวน N ค่าก่อนหน้านี้เท่านั้น เนื่องจากแต่ละจุดตาราง N จุดสามารถเข้าถึงได้จาก N จุดตารางเดียวกันของช่องเวลาก่อนหน้านี้

2) กระบวนการย้อนกลับ (Backward Procedure)

พิจารณาตัวแปรย้อนกลับ $\beta_t(i)$ ที่กำหนดไว้แล้วดังนี้

$$\beta_t(i) = P\left(O_{t+1}\, O_{t+2} \Lambda\, O_T \middle| q_t = S_i, \lambda\right) \quad (2.104)$$

ค่าความน่าจะเป็นของลำดับค่าสังเกตบางส่วนจาก t + 1 จนสิ้นสุดเมื่อกำหนดสถานะ S_i ที่เวลา t ด้วยแบบจำลอง l จะสามารถหาค่าของ $\beta_t(i)$ โดยอุปนัยได้ดังนี้



รูปที่ 2.21 รายละเอียดลำดับกระบวนการในการคำนวณค่าตัวแปรย้อนกลับ $\beta_t(i)$

ตารางที่ 2.8 รายละเอียดกระบวนการย้อนกลับ

ขั้นตอนที่ 1 **กระบวนการเริ่มต้น**	$\beta_T(i) = 1, \quad 1 \le i \le N$	(2.105)
ขั้นตอนที่ 2 **กระบวนการอุปนัย**	$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j),$ $t = T-1, T-2, \Lambda, 1, \quad 1 \le i \le N$	(2.106)
ขั้นตอนที่ 3 **กระบวนการสิ้นสุด**	$P(O\|\lambda) = \sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i)$	(2.107)

ขั้นตอนกระบวนการเริ่มต้น เป็นการกำหนดค่าเริ่มต้นให้กับ $\beta_T(i)$ โดยกำหนดให้เป็น 1 ทั้งหมดทุกค่า i ขั้นตอนกระบวนการอุปนัยดังแสดงในรูปที่ 2.21 แสดงถึงการเข้าถึงสถานะ S_j ที่เวลา t และการให้รายละเอียดของลำดับค่าสังเกตตั้งแต่เวลา t + 1 เป็นต้นไป จะต้องพิจารณาสถานะ S_j ที่เป็นไปได้ทั้งหมด เพื่อนับรวมการเปลี่ยนแปลงจากสถานะ S_i ไปยัง S_j ของพจน์ a_{ij} รวมไปถึงค่าสังเกต O_{t+1} ในสถานะ j ของพจน์ $b_j(O_{t+1})$ และนับรวมถึงลำดับค่าสังเกตบางส่วนที่ยังเหลืออยู่จากสถานะ j ของพจน์ $\beta_{t+1}(j)$ ด้วย ซึ่งในการ

คำนวณค่า $\beta_t(i), 1 \le t \le T, 1 \le i \le N$ นั้น จะเกิดการคำนวณขึ้นด้วยอันดับประมาณ N^2T ครั้ง และสามารถคำนวณโดยอาศัยโครงสร้าง Lattice ได้เช่นเดียวกัน

2.3.2.3.2 การแก้ไขปัญหาพื้นฐานข้อที่ 2

ปัญหาพื้นฐานข้อที่ 2 เกี่ยวข้องกับการหาลำดับสถานะที่เหมาะสมที่สุดที่สัมพันธ์กับลำดับค่าสังเกตที่กำหนดให้ ความยากในการแก้ปัญหาขึ้นอยู่กับนิยามของลำดับสถานะที่เหมาะสมที่สุด ซึ่งมีกฎเกณฑ์ของความเหมาะสมที่สุดที่เป็นไปได้มากมาย กฎเกณฑ์หนึ่งที่เป็นไปได้ก็คือ การเลือกสถานะ q_t ซึ่งตัวต่อตัวแล้วคล้ายคลึงกันมากที่สุด กฎเกณฑ์ของความเหมาะสมที่สุดนี้จะทำให้ค่าประมาณของจำนวนสถานะที่ถูกต้องมีค่ามากที่สุด กำหนดให้ตัวแปรมีค่าดังนี้

$$\gamma_t(i) = P\left(q_t = S_i \middle| O, \lambda\right) \qquad (2.108)$$

ซึ่งเป็นค่าความน่าจะเป็นของการอยู่ในสถานะ S_i ที่เวลา t เมื่อกำหนดลำดับค่าสังเกต O และแบบจำลอง λ มาให้ สมการที่ (2.84) สามารถเขียนอยู่ในรูปของตัวแปรไปหน้าและตัวแปรย้อนกลับได้ดังนี้

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \qquad (2.109)$$

เมื่อ $\alpha_t(i)$ เป็นลำดับค่าสังเกตบางส่วน $O_1\, O_2 \Lambda\, O_t$ และสถานะ S_i ที่เวลา t ขณะที่ $\beta_t(i)$ เป็นลำดับค่าสังเกตที่คงเหลืออยู่ $O_{t+1}\, O_{t+2} \Lambda\, O_T$ เมื่อกำหนดสถานะ S_i ที่เวลา t ตัวประกอบที่ทำให้เป็นบรรทัดฐานเดียวกัน $P(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ ทำให้ $\gamma_t(i)$ กลายเป็นการวัดค่าความน่าจะเป็นดังนี้

$$\sum_{i=1}^{N}\gamma_t(i) = 1, \quad 1 \le t \le T \qquad (2.110)$$

โดยการใช้ $\gamma_t(i)$ จะสามารถหาค่าสถานะที่ตัวต่อตัวคล้ายคลึงกันมากที่สุด q_t ที่เวลา t ได้ดังนี้

$$q_t = \underset{1 \le i \le N}{\arg\max} \left[\gamma_t(i)\right], \quad 1 \le t \le T \qquad (2.111)$$

แม้ว่าสมการที่ (2.111) จะทำให้ค่าประมาณของจำนวนสถานะที่ถูกต้องมีค่ามากที่สุดโดยการเลือกสถานะที่คล้ายคลึงกันมากที่สุดสำหรับแต่ละ t ก็ตาม ก็ยังเกิดปัญหาขึ้นกับผลของลำดับสถานะที่ได้เนื่องจากสมการที่ (2.111) เป็นเพียงการหาสถานะที่คล้ายคลึงกันมากที่สุดในขณะหนึ่งเท่านั้น โดยไม่เกี่ยวข้องกับค่าความน่าจะเป็นในการปรากฏของลำดับสถานะ

วิธีการหนึ่งในการแก้ปัญหาก็คือการแก้ไขกฎเกณฑ์ของความเหมาะสมที่สุด โดยการคำนวณหาค่าลำดับสถานะที่ทำให้ค่าประมาณของคู่สถานะ (q_t, q_t) ที่ถูกต้องมีค่ามากที่สุด กฎเกณฑ์ของความเหมาะสมที่สุดที่นิยมใช้มากที่สุดก็คือ การหาลำดับสถานะที่ดีที่สุดเพียงลำดับเดียว ขั้นตอนวิธีการในการหาลำดับสถานะที่ดีที่สุดเพียงลำดับเดียวนี้ จะอยู่บนพื้นฐานของการโปรแกรมแบบพลวัตเรียกว่า "ขั้นตอนวิธีการ Viterbi"

ขั้นตอนวิธีการ Viterbi (Viberbi Algorithm)

ขั้นตอนวิธีการ Viterbi เป็นขั้นตอนวิธีในการหาลำดับสถานะที่ดีที่สุดเพียงลำดับเดียว $Q = \{q_1\, q_2 \Lambda\, q_T\}$ สำหรับลำดับค่าสังเกตที่กำหนดให้ $O = \{O_1\, O_2 \Lambda\, O_T\}$ จะกำหนดตัวแปรได้ดังนี้

$$\delta_t(i) = \max_{q_1, q_2 \Lambda, q_{t-1}} P\left[q_1\, q_2 \Lambda\, q_t = i, O_1\, O_2 \Lambda\, O_t | \lambda\right] \qquad (2.112)$$

โดยที่ $\delta_t(i)$ เป็นค่าความน่าจะเป็นที่มีค่าสูงที่สุดของเส้นทางเดี่ยวที่เวลา t ซึ่งเป็นค่าสังเกต t ค่าแรกและสิ้นสุดในสถานะ S_i ด้วยวิธีการอุปนัยจะได้ว่า

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] \cdot b_j\left(O_{t+1}\right) \qquad (2.113)$$

ในการเรียกใช้ค่าลำดับสถานะจำเป็นต้องติดตามค่าอาร์กิวเมนต์ที่ทำให้สมการที่ (2.94) มีค่ามากที่สุดสำหรับแต่ละค่า t และ j โดยอาศัยแถวลำดับ $\psi_t(j)$ ขั้นตอนวิธีการในการหาลำดับสถานะที่ดีที่สุดเพียงลำดับบเดียว ดังแสดงในตารางที่ 2.9 รายละเอียดของกระบวนการ Viterbi ดังนี้

ตารางที่ 2.9 รายละเอียดขั้นตอนวิธีการ Viterbi

ขั้นตอนที่ 1 **กระบวนการเริ่มต้น**	$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N$	(2.114ก)
	$\psi_1(i) = 0$	(2.114ข)
ขั้นตอนที่ 2 **กระบวนการวนซ้ำ**	$\delta_t(j) = \max\limits_{1\le i\le N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t), \quad 2 \le t \le T,\ 1 \le j \le N$	(2.115ก)
	$\psi_t(j) = \arg\max\limits_{1\le i\le N} \left[\delta_{t-1}(i) a_{ij}\right], \quad 2 \le t \le T,\ 1 \le j \le N$	(2.115ข)
ขั้นตอนที่ 3 **กระบวนการสิ้นสุด**	$P^* = \max\limits_{1\le i\le N} \left[\delta_T(i)\right]$	(2.116ก)
	$q_T^* = \arg\max\limits_{1\le i\le N} \left[\delta_T(i)\right]$	(2.116ข)
ขั้นตอนที่ 4 **กระบวนการย้อนกลับรอยเส้นทาง**	$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \Lambda\ , 1$	(2.117)

รายละเอียดขั้นตอนวิธีการ Viterbi นั้นคล้ายคลึงกับการคำนวณไปหน้าของสมการที่ (2.100) ถึง (2.102) ยกเว้นกระบวนการย้อนกลับรอยเส้นทางเท่านั้น ความแตกต่างที่สำคัญก็คือการหาค่าที่สูงที่สุดในสมการที่ (2.115ก) เหนือสถานะก่อนหน้าซึ่งใช้แทนผลรวมในสมการที่ (2.101) นั่นเอง

2.3.2.3.3 การแก้ไขปัญหาพื้นฐานข้อที่ 3

ปัญหาพื้นฐานข้อที่สามเกี่ยวข้องกับการค้นหาวิธีการปรับเปลี่ยนพารามิเตอร์ของแบบจำลอง $\lambda = (A, B, \pi)$ เพื่อทำให้ค่าความน่าจะเป็นของลำดับค่าสังเกตมีค่ามากที่สุดเมื่อกำหนดค่าพารามิเตอร์ของแบบจำลองมาให้ เนื่องจากไม่มีวิธีการที่แน่นอนในการวิเคราะห์เพื่อแก้ไขปัญหาสำหรับแบบจำลองที่จะให้ค่าความน่าจะเป็นของลำดับค่าสังเกตมีค่ามากที่สุด ถึงแม้ว่าจะกำหนดลำดับค่าสังเกตที่จำกัดให้เป็นข้อมูลฝึกฝนก็ตามก็ยังไม่มีวิธีการใดที่เหมาะสมที่สุดในการประมาณค่าพารามิเตอร์ของแบบจำลอง แต่จะสามารถเลือก $\lambda = (A, B, \pi)$ ที่ทำให้ $P(O|\lambda)$ มีค่ามากที่สุดโดยใช้กระบวนการวนซ้ำของ Baum-Welch

กระบวนการประมาณค่าซ้ำของ Baum-Welch

(Baum-Welch Reestimation Procedure)

กระบวนการประมาณค่าซ้ำของ Baum-Welch นี้อยู่บนพื้นฐานของหลักการของความน่าจะเป็นจริงสูงสุด ซึ่งจะช่วยปรับปรุงให้ค่าความน่าจะเป็นของลำดับค่าสังเกตให้มีค่าสูงขึ้น โดยมีขั้นตอนวิธีการดังนี้ (Picone, 1990)

ตารางที่ 2.10 รายละเอียดกระบวนการประมาณค่าซ้ำของ Baum-Walch

การกระจายของสภาวะเริ่มต้น	$\bar{\pi}_i = \alpha_1(i)\beta_1(i)$ .... (2.118ก)
การกระจายของความน่าจะเป็นในการเปลี่ยนแปลงสถานะ	$\bar{a}_{ij} = \dfrac{\sum_{t=1}^{T-1}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)}$ ....(2.118ข)
การกระจายของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต	$\bar{b}_j(k) = \dfrac{\sum_{t=1,\, O_t=v_k}^{T-1}\alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T-1}\alpha_t(j)\beta_t(j)}$ .... (2.118ค)

กระบวนการประมาณค่าซ้ำของ Baum-Walch นั้นอยู่บนพื้นฐานของความรู้ความเข้าใจในเรื่องของการประมาณค่าใหม่ของค่าความน่าจะเป็นในการเปลี่ยนแปลง ซึ่งอยู่บนพื้นฐานของจำนวนการเปลี่ยนแปลงจากสถานะ i ไปยังสถานะ j หารด้วยจำนวนการเปลี่ยนแปลงออกจากสถานะ i ในทำนองเดียวกันค่าความน่าจะเป็นใหม่ของสัญลักษณ์ขาออกสำหรับสัญลักษณ์ที่ k ที่สถานะ i ได้จากจำนวนครั้งของสัญลักษณ์ที่ออกจากสถานะหารด้วยจำนวนครั้งที่อยู่ในสถานะ

กระบวนการประมาณค่าซ้ำของ Baum-Welch แบบดัดแปลง

(Modified Baum-Welch Reestimation Procedure)

กระบวนการประมาณค่าซ้ำหรือการปรับปรุงและปรับให้ทันกาล ด้วยการวนซ้ำของค่าพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟนั้น เริ่มต้นจากการกำหนดให้ $\xi_t(i,j)$ เป็นค่าความน่าจะเป็นของการอยู่ในสถานะ S_i ที่เวลา t และสถานะ S_j ที่เวลา t + 1 เมื่อกำหนดแบบจำลองและลำดับค่าสังเกตให้จะได้ว่า (Rabiner, 1989)

$$\xi_t(i,j) = P\left(q_t = s_i, q_{t+1} = s_j \middle| O, \lambda\right) \text{ .... (2.119)}$$

ลำดับของเหตุการณ์ที่นำไปสู่เงื่อนไขที่จำเป็นตามสมการที่ (2.98) ดังแสดงในรูปที่ 2.20 จากนิยามของตัวแปรไปหน้าและตัวแปรย้อนกลับจะสามารถเขียน $\xi_t(i,j)$ ด้วยพจน์ของตัว

แปรทั้งสองได้ดังสมการที่ (2.113) โดยที่พจน์เศษเป็นเพียง $P(q_t = s_i, q_{t+1} = s_j, O|\lambda)$ หารด้วย $P(O|\lambda)$ ซึ่งให้วิธีการวัดค่าความน่าจะเป็นตามที่ต้องการ

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

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t-1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \quad \text{.......... (2.1220)}$$

จากการกำหนดให้ค่าของ $\gamma_t(i)$ เป็นค่าความน่าจะเป็นของการอยู่ในสถานะ S_i ที่เวลา t เมื่อกำหนดแบบจำลองและลำดับค่าสังเกตให้ ดังนั้นความสัมพันธ์ระหว่าง $\gamma_t(i)$ และ $\xi_t(i,j)$ เกิดจากผลรวมบน j ทั้งหมดดังนี้

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \quad \text{.......... (2.121)}$$

ผลรวมของ $\gamma_t(i)$ เหนือดรรชนีเวลา t จะเป็นจำนวนครั้งที่เข้าไปยังสถานะ S_i หรือเป็นจำนวนการเปลี่ยนแปลงไปจากสถานะ S_i เมื่อไม่รวมช่องเวลา t = T จากผลรวม ผลรวมของ $\xi_t(i,j)$ เหนือ t จาก t = 1 ถึง t = T - 1 จะเป็นจำนวนการเปลี่ยนแปลงจากสถานะ S_i ไปยังสถานะ S_j ดังนี้

$$\sum_{t=1}^{T-1}\gamma_t(i) = \text{จำนวนการเปลี่ยนแปลงไปจากสถานะ } S_i \quad \text{.......... (2.122ก)}$$

$$\sum_{t=1}^{T-1}\xi_t(i,j) = \text{จำนวนการเปลี่ยนแปลงไปจากสถานะ } S_i \text{ ไปยังสถานะ } S_j \quad \text{.......... (2.122ข)}$$

โดยอาศัยสมการทั้งสองข้างต้นจะได้กระบวนการประมาณค่าซ้ำสำหรับพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ (A, B, π) ดังแสดงในตารางที่ 2.11 เมื่อกำหนดแบบจำลองด้วย $\lambda = (A, B, \pi)$ เพื่อใช้ในการคำนวณค่าทางด้านขวาของสมการที่ (2.123ก) - (2.123ค) และกำหนดแบบจำลองที่ประมาณค่าใหม่ด้วย $\overline{\lambda} = (\overline{A}, \overline{B}, \overline{\pi})$ ตามค่าทางด้านซ้ายของสมการที่ (2.123ก) - (2.123ค) นอกจากนั้นสมการทั้งสามยังเป็นไปตามเงื่อนไขดังนี้ เงื่อนไขแรก แบบจำลองเริ่มต้น λ เป็นตัวกำหนดจุดวิกฤตของฟังก์ชันความน่าจะเป็นจริงสูงสุดในกรณีที่ $\overline{\lambda} = \lambda$ เงื่อนไขที่สอง แบบจำลอง $\overline{\lambda}$ มีความน่าจะเป็นจริงสูงกว่าแบบจำลอง λ ด้วยเงื่อนไข $P(O|\overline{\lambda}) > P(O|\lambda)$

ด้วยขั้นตอนพื้นฐานข้างต้นจะกระทำซ้ำโดยนำ $\overline{\lambda}$ มาแทนที่ λ ในการประมาณค่าใหม่ซึ่งเป็นการปรับปรุงค่าความน่าจะเป็นของ O ที่ถูกสังเกตจากแบบจำลองจนกระทั่งถึงจุดสิ้นสุดค่าหนึ่ง ผลลัพธ์ค่าสุดท้ายที่ได้จากกระบวนการประมาณค่าใหม่เรียกว่า "การประมาณค่าด้วยความน่าจะเป็นจริงสูงสุดของแบบจำลองฮิดเดน มาร์คอฟ" นอกจากนี้แล้วขั้นตอนวิธีการไปหน้า-ย้อนกลับจะเข้าสู่ค่าสูงสุดเฉพาะแห่งเท่านั้น ซึ่งเป็นปัญหาที่สำคัญเนื่องจากพื้นผิวของค่าที่เหมาะสมที่สุดมีความสลับซับซ้อนมากและเต็มไปด้วยค่าสูงสุดเฉพาะแห่งมากมาย

สมการ (2.123ก) - (2.123ค) ที่ใช้ในการคำนวณสามารถสร้างได้โดยตรงขึ้นมาด้วยการทำให้ฟังก์ชันเสริมของ Baum มีค่าสูงที่สุดเหนือ $\overline{\lambda}$ ดังนี้

$$Q(\lambda, \overline{\lambda}) = \sum_{Q} P(Q|O,\lambda)\log\left[P(O,Q|\overline{\lambda})\right] \quad \text{.......... (2.124)}$$

ซึ่งการทำให้ค่าของ $Q(\lambda, \overline{\lambda})$ มีค่าสูงที่สุดจะทำให้ความน่าจะเป็นจริงมีค่าสูงขึ้น โดยจะทำให้ฟังก์ชันความน่าจะเป็นจริงลู่เข้าสู่จุดวิกฤตจุดหนึ่งดังนี้

$$\max_{\overline{\lambda}}\left[Q(\lambda, \overline{\lambda})\right] \Rightarrow P(O|\overline{\lambda}) \geq P(O|\lambda) \quad \text{.......... (2.125)}$$

กระบวนการคำนวณค่าใหม่ก็คือขั้นตอนวิธีการ EM ในทางสถิติซึ่งประกอบด้วยสองขั้นตอนย่อย โดยที่ขั้นตอน E (Expectation) หรือการคำนวณค่าคาดหวัง เป็นการคำนวณฟังก์ชัน

ช่วย $Q(\lambda, \overline{\lambda})$ และขั้นตอน M (Modification) หรือการดัดแปลง เป็นการทำให้มีค่าสูงสุดเหนือค่าของ $\overline{\lambda}$ นอกจากนี้กระบวนการคำนวณค่าใหม่จะต้องเป็นไปตามเงื่อนไขเฟ้นสุ่ม (Stochastic Constraints) ของพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ ซึ่งกระบวนการคำนวณค่าใหม่นี้จะเป็นไปตามเงื่อนไขโดยอัตโนมัติดังแสดงในตารางที่ 2.12 รายละเอียดเงื่อนไขเฟ้นสุ่มของพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ ดังนี้

ตารางที่ 2.11 รายละเอียดกระบวนการประมาณค่าซ้ำของ Baum-Walch แบบดัดแปลง

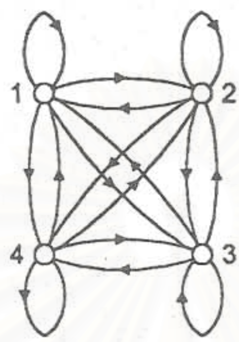
การกระจายของสภาวะเริ่มต้น	$\overline{\pi}_i = \gamma_1(i)$ .... (2.123ก)
การกระจายของความน่าจะเป็นในการเปลี่ยนแปลงสถานะ	$\overline{a}_{ij} = \dfrac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$ .... (2.123ข)
การกระจายของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต	$\overline{b}_j(k) = \dfrac{\sum_{t=1,\ s.t.\ O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$ .... (2.123ค)

ตารางที่ 2.12 รายละเอียดเงื่อนไขเฟ้นสุ่มของพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ

การกระจายของสภาวะเริ่มต้น	$\sum_{i=1}^{N} \overline{\pi}_i = 1$ .... (2.126ก)
การกระจายของความน่าจะเป็นในการเปลี่ยนแปลงสถานะ	$\sum_{j=1}^{N} \overline{a}_{ij} = 1, \quad 1 \le i \le N$ .... (2.126ข)
การกระจายของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต	$\sum_{k=1}^{M} \overline{b}_j(k) = 1, \quad 1 \le j \le N$ .... (2.126ค)

2.3.2.4. ประเภทของแบบจำลองฮิดเดน มาร์คอฟ

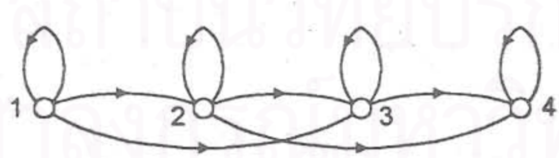
ขั้นตอนวิธีการของแบบจำลองฮิดเดน มาร์คอฟ ส่วนใหญ่เป็นการพิจารณาเพียงกรณีพิเศษของแบบจำลองประเภทเออร์กอดิก ซึ่งเป็นแบบจำลองฮิดเดน มาร์คอฟ ที่ทุกสถานะต่อเชื่อมถึงกันหมด โดยทุกสถานะของแบบจำลองสามารถเข้าถึงสถานะอื่นได้ในขั้นตอนเดียว ดังนั้นแบบจำลองแบบเออร์กอดิกจึงมีคุณสมบัติที่ทุกสถานะสามารถเข้าถึงได้จากสถานะอื่นด้วยขั้นตอนที่จำกัดแน่นอนดังแสดงในรูปที่ 2.21 ซึ่งเป็นแบบจำลองที่มีจำนวนสถานะ N = 4 สถานะ และมีคุณสมบัติเฉพาะที่สัมประสิทธิ์ a_{ij} ทั้งหมดมีค่าเป็นบวกซึ่งแสดงในสมการที่ (2.127) ดังนี้



รูปที่ 2.22 แบบจำลองแบบเออร์กอดิกที่มี 4 สถานะ

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad \text{(2.127)}$$

ในการประยุกต์ใช้งานแบบจำลองฮิดเดน มาร์คอฟกับงานเฉพาะอย่างนั้น ยังมีแบบจำลองประเภทอื่นที่เหมาะสมกับคุณสมบัติที่สังเกตของสัญญาณ ซึ่งถูกจำลองมากกว่าแบบจำลองประเภทเออร์กอดิกมาตรฐาน ดังแสดงในรูปที่ 2.22 ซึ่งเป็นแบบจำลองซ้าย-ขวา (Left-Right Model) หรือแบบจำลอง Bakis เนื่องจากลำดับสถานะที่อยู่ภายในที่สัมพันธ์กับแบบจำลองมีคุณสมบัติที่ ดัชนีของสถานะจะเพิ่มขึ้น หรือมีค่าเท่าเดิม เมื่อเวลาเพิ่มขึ้น เปรียบเสมือนกับสถานะดำเนินจากซ้ายไปขวา ดังนั้นแบบจำลองซ้าย-ขวานี้ จึงมีคุณสมบัติเหมาะสมในการจำลองแบบสัญญาณ ที่เปลี่ยนแปลงไปตามเวลา ดังเช่น สัญญาณเสียง เป็นต้น



รูปที่ 2.23 แบบจำลองแบบซ้าย-ขวาที่มี 4 สถานะ

คุณสมบัติพื้นฐานของทุกแบบจำลองฮิดเดน มาร์คอฟ ประเภทซ้าย-ขวานั้น สัมประสิทธิ์ของการเปลี่ยนสถานะจะต้องเป็นไปตามคุณสมบัติ ที่ไม่อนุญาตให้มีการเปลี่ยนแปลงสถานะไปยังสถานะ ที่มีดัชนีต่ำกว่าสถานะปัจจุบันตามสมการที่ (2.128) รวมทั้งค่าความน่าจะเป็นเริ่มต้นจะต้องมีคุณสมบัติเป็นไปตามสมการที่ (2.129) ดังนี้

$$a_{ij}=0, \quad j<i \quad \text{(2.128)}$$

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \quad \text{.......... (2.129)}$$

เนื่องจากลำดับสถานะจะต้องเริ่มต้นจากสถานะที่ 1 และสิ้นสุดในสถานที่ N ดังนั้นในแบบจำลองประเภทซ้าย-ขวาจึงต้องเพิ่มเติมเงื่อนไขบังคับให้กับสัมประสิทธิ์ของการเปลี่ยนสถานะ เพื่อไม่ให้เกิดการเปลี่ยนแปลงมากจนเกินไปดังนี้

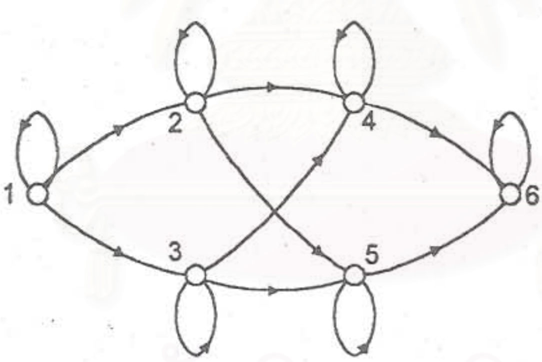
$$a_{ij} = 0, \quad j > i + \Delta \quad \text{.......... (2.130)}$$

จากรูปที่ 2.22 จะกำหนดให้ค่า $\Delta = 2$ ซึ่งไม่อนุญาตให้มีการข้ามสถานะเกินกว่า 2 สถานะ จึงได้เมตริกซ์ของการเปลี่ยนสถานะเป็นดังสมการที่ (2.132) และสถานะสุดท้ายของแบบจำลองประเภทซ้าย-ขวาจะมีสัมประสิทธิ์ของการเปลี่ยนสถานะมีค่าเฉพาะดังสมการที่ (2.131)

$$\begin{aligned} a_{NN} &= 1 \\ a_{Ni} &= 0, \quad i < N \end{aligned} \quad \text{.......... (2.131)}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix} \quad \text{.......... (2.132)}$$

นอกจากนั้นยังมีรูปแบบของแบบจำลองที่เป็นไปได้อีกมากมาย ตัวอย่างดังรูปที่ 2.23 ซึ่งแสดงถึงการเชื่อมต่อข้ามแบบจำลองประเภทขนานซ้าย-ขวาสองชุด แต่แบบจำลองนี้ก็ยังคงจัดอยู่ในแบบจำลองประเภทซ้าย-ขวาเพียงแต่มีความยืดหยุ่นมากยิ่งขึ้น อย่างไรก็ตามการข้ามสถานะตามเงื่อนไขของแบบจำลองนี้ไม่มีผลต่อกระบวนการประมาณค่าใหม่



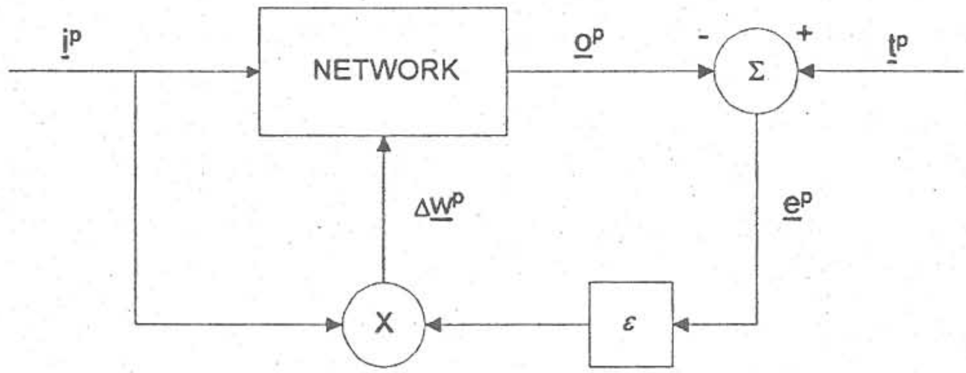
รูปที่ 2.24 แบบจำลองแบบเส้นทางขนานซ้าย-ขวาที่มี 6 สถานะ

2.3.3. นิวรอลเน็ตเวิร์ก

นิวรอลเน็ตเวิร์กแบ่งแยกตามลักษณะของการเรียนรู้ได้เป็น 2 ชนิดคือ unsupervised learning และ supervised learning ในวิทยานิพนธ์นี้เลือกใช้นิวรอลเน็ตเวิร์กแบบ multi-layer perceptron ซึ่งอยู่ในประเภท supervised learning

2.3.3.1. ขั้นตอนการฝึก (training) นิวรอลเน็ตเวิร์ก

Multi-layer perceptron neural network ใช้การฝึก (training) แบบ error backpropagation หรือ generalized delta rule ดังแสดงในรูปที่ 2.24



รูปที่ 2.25 โครงสร้างของการฝึก

โดยที่ $\underline{i}^p$ แทนค่าเวกเตอร์ input pattern ลำดับที่ p

$\underline{o}^p$ แทนค่าเวกเตอร์ output pattern ลำดับที่ p ที่ได้จากเน็ตเวิร์ก

$\underline{\omega}^p$ แทนค่าเวกเตอร์ network weights เมื่อใส่ค่าอินพุตลำดับที่ p เข้าสู่เน็ตเวิร์ก

$\underline{t}^p$ แทนค่าเวกเตอร์ output pattern ลำดับที่ p ที่ต้องการ

ε แทนค่า learning rate

$\underline{e}^p$ แทนค่าเวกเตอร์ความผิดพลาดของเอาต์พุตลำดับที่ p

นิวรอลเน็ตเวิร์กจะเรียนรู้จาก แต่ละตัวอย่างคู่ข้อมูลอินพุตเอาต์พุต $(\underline{i}^p, \underline{t}^p)$ ที่อยู่ในชุดฝึก (training set) ซึ่งมีขั้นตอนพื้นฐานดังนี้

- ป้อนค่าเวกเตอร์อินพุต (input vector) ให้กับระดับข้อมูลเข้า (input layer) ของนิวรอลเน็ตเวิร์ก
- 'Feed forward' หรือแพร่กระจายค่าอินพุต (input) เพื่อหาค่าเอาต์พุต (output) ของทุกโหนด
- เปรียบเทียบค่าเอาต์พุต $\underline{o}^p$ ในระดับข้อมูลออก (output layer) กับค่าเอาต์พุตที่ ต้องการ $\underline{t}^p$
- คำนวณและแพร่กระจายค่าความผิดพลาด ในทิศทางย้อนกลับ (เริ่มจากระดับข้อมูลออก (output layer)) ตลอดทั้งเน็ตเวิร์ก
- ลดค่าความผิดพลาดที่แต่ละระดับ (layer) โดยการปรับค่าน้ำหนัก (weight) ที่เชื่อมต่อระหว่างโหนดของแต่ละระดับ

การฝึกนี้จะถูกทำซ้ำไปเรื่อย ๆ จนกว่าค่าความผิดพลาดจะอยู่ในระดับที่ยอมรับได้จึงจะยุติการฝึก ค่าน้ำหนักการเชื่อมต่อ (connection weight) ของนิวรอลเน็ตเวิร์กที่ผ่านการฝึกแล้ว เปรียบเทียบได้กับความรู้ที่ได้รับจากการฝึกจากตัวอย่างคู่ข้อมูลอินพุตเอาต์พุต ดังนั้นถ้ามีตัวอย่างคู่ข้อมูลอินพุตเอาต์พุตที่หลากหลาย จะทำให้นิวรอลเน็ตเวิร์กมีความรู้เพียงพอที่จะใช้ในการเปรียบเทียบเสียง

รูปที่ 2.26 แสดงโครงสร้างของ multi-layer perceptron neural network ซึ่งประกอบด้วยระดับข้อมูลเข้า (input layer), ระดับซ่อนตัว (hidden layer) ซึ่งอาจมีมากกว่า 1 ระดับ และระดับข้อมูลออก (output layer) ระดับข้อมูลเข้า (input layer) มีจำนวนโหนดเท่ากับ d โหนด ระดับข้อมูลออก (output layer) มีจำนวนโหนดเท่ากับ c โหนด ที่แต่ละโหนดในชั้นใด ๆ จะมีค่าน้ำหนักการเชื่อมต่อที่เชื่อมต่อไปยังโหนดที่อยู่ในชั้นถัดไป ที่อยู่ติดกันเท่านั้น รูปที่ 2.27 แสดงรายละเอียดของโหนดในนิวรอลเน็ตเวิร์กโดยที่ ค่า net input ที่โหนด j แสดงได้ดังนี้

$$net_j^p = \sum_i \omega_{ji}\tilde{o}_i^p + bias_j \qquad (2.133)$$

เมื่อ $\widetilde{o}_i^p$ $= o_i^p$ ถ้าอินพุตเป็นค่าเอาต์พุตของโหนดในระดับ (layer) ที่อยู่ข้างหน้า (เมื่อ j เป็นโหนดในระดับซ่อนตัว (hidden layer) และระดับข้อมูลออก (output layer))

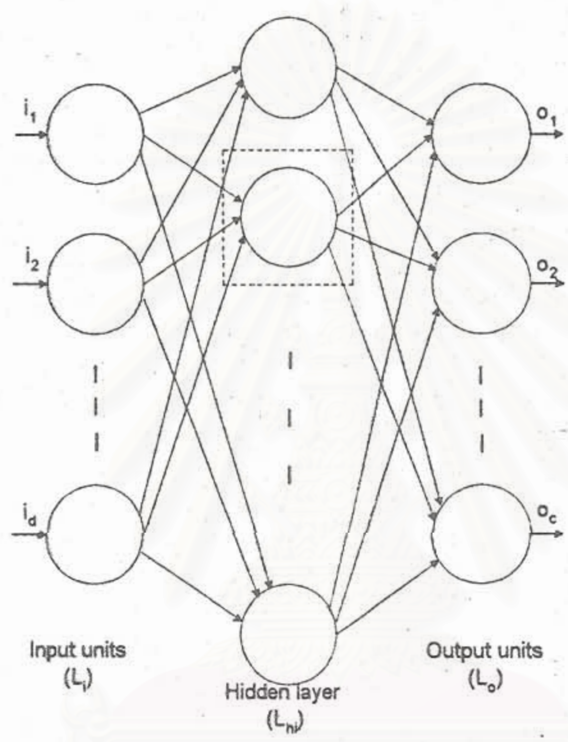
$= i_i^p$ ถ้าอินพุตเป็นค่าข้อมูลอินพุตที่ป้อนเข้าสู่เน็ตเวิร์ก

ω_{ji} เป็นค่าน้ำหนักการเชื่อมต่อที่เชื่อมต่อจากโหนด i ไปยังโหนด j ที่อยู่ในระดับถัดไป

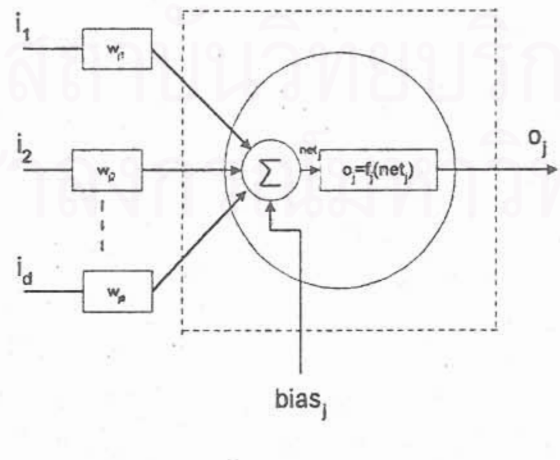
$bias_j$ เป็นค่าที่ใช้ปรับให้ net_j มีค่าไม่เท่ากับศูนย์ ในกรณีที่ $\widetilde{o}_i^p$ ทุกโหนดมีค่าเป็นศูนย์หมด

ค่าเอาต์พุต o_j ของโหนด j สามารถคำนวณได้จาก net_j ดังนี้

$$o_j^p = f_j(net_j) \qquad (2.134)$$



รูปที่ 2.26 โครงสร้างของ multi-layer perceptron neural network



รูปที่ 2.27 รายละเอียดของโหนดในนิวรอลเน็ตเวิร์ก

โดยที่ฟังก์ชันกระตุ้น (activation function) f_j เป็นฟังก์ชันเพิ่มและเป็นฟังก์ชันที่สามารถหาอนุพันธ์ได้(d ffe ent iabl e) ในงานวิจัยนี้เลือกใช้ฟังก์ชัน sigmoid

$$f_j(net_j) = \frac{1}{1+e^{-net_j}} \qquad (2.135)$$

2.3.3.2. การปรับค่าน้ำหนักการเชื่อมต่อ

เวกเตอร์ค่าความผิดพลาดของเอาต์พุต สำหรับตัวอย่างคู่ข้อมูลอินพุตเอาต์พุตที่ p กำหนดโดย

$$\underline{e}^p = \underline{t}^p - \underline{o}^p \qquad (2.136)$$

E_p แทนค่าความผิดพลาดของเอาต์พุต สำหรับตัวอย่างคู่ข้อมูลอินพุตเอาต์พุตที่ p หาได้จาก

$$E_p = \frac{1}{2}\sum_j (t_j^p - o_j^p)^2 \qquad (2.137)$$

หลักการของการปรับค่าน้ำหนักการเชื่อมต่อใน backpropagation training เริ่มต้นจากการคำนวณพื้นผิวของค่าความผิดพลาด E และคำนวณค่าเกรเดียนต์ (gradient) ของ E เทียบกับค่าน้ำหนักการเชื่อมต่อ $\partial E / \partial \omega_{ji}$ การปรับค่าน้ำหนักการเชื่อมต่อ $\Delta \omega_{ji}$ จะปรับค่าเป็นสัดส่วนกับ $-\partial E / \partial \omega_{ji}$ เพื่อให้การปรับน้ำหนักการเชื่อมต่อเป็นไปในทิศทางที่ลดค่าผิดพลาดลง สมการสำหรับการปรับค่าน้ำหนักการเชื่อมต่อแสดงได้ดังนี้

$$\Delta^p \omega_{ji} = \varepsilon \delta_j^p \tilde{o}_i^p \qquad (2.138)$$

เมื่อ $\tilde{o}_i^p$ มีค่าตามที่แสดงในสมการที่ (2.133)

ε คือค่า learning rate ซึ่งเป็นค่าคงที่และมีค่าเป็นบวก

สำหรับค่า δ_j^p คือค่าความไว (sensitivity) ของค่าความผิดพลาดเทียบกับค่า net input ที่โหนด j

$$\delta_j^p = -\frac{\partial E_p}{\partial net_j^p} \qquad (2.139)$$

สำหรับโหนดในระดับข้อมูลออก (output layer) $\delta_j^p = (t_j^p - o_j^p) f_j'(net_j^p)$

สำหรับโหนดใน internal layer $\delta_j^p = f_j'(net_j^p)\sum_n \delta_n^p \omega_{nj}$ เมื่อ δ_n^p เป็นค่า ความไว (sensitivity) ในชั้นถัดออกไป

อนุพันธ์ของฟังก์ชันกระตุ้นชนิด sigmoid อยู่ในรูปที่คำนวณได้ง่าย ซึ่งนับเป็นข้อดีของฟังก์ชันชนิดนี้ กำหนดโดย

$$f_j'(net_j^p) = o_j^p(1 - o_j^p) \qquad (2.140)$$

2.4. ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)

ขั้นตอนวิธีการตัดสินใจ เป็นขั้นตอนที่เกี่ยวข้องกับกฎเกณฑ์ที่ใช้ในการตัดสินเลือกรูปแบบที่มีความคล้ายคลึงกันมากที่สุดระหว่างคำพูดที่ไม่ทราบรูปแบบกับรูปแบบที่ได้จัดเก็บไว้ล่วงหน้า โดยอาศัยค่าความไม่คล้ายคลึงกันหรือค่าระยะทางที่ได้จากการทดสอบความคล้ายคลึงกันของรูปแบบที่มีค่ามากที่สุด เงื่อนไขในการตัดสินใจมีหลายแบบ ดังนี้

2.4.1. Nearest Neighbor rule (NN rule)

กำหนดให้มีรูปแบบอ้างอิงอยู่ V รูปแบบ R^i, $i = 1,2,\ldots,V$ โดยที่แต่ละแบบจะให้ average distance เป็น D^i ซึ่งได้จาก DTW algorithm ผลการรับรู้จากกฎการตัดสินใจนี้จะนำไปทำการเลือกแบบอ้างอิงที่มีระยะทางจากแบบน้อยที่สุด, R^{i^*}

$$i^{*} = \underset{i}{argmin}\left[D^{i}\right] \quad (2.141)$$

ซึ่งสามารถจัดเรียงลำดับของระยะการวัดใหม่ได้

$$D^{[1]} \leq D^{[2]} \leq \cdots \leq D^{[V]} \quad (2.142)$$

2.4.2. K-Nearest Neighbor rule (KNN rule)

ในกรณีที่รูปแบบอ้างอิงแต่ละรูปแบบ (pattern) มีด้วยกันหลายชุด กำหนดให้แต่ละรูปแบบอ้างอิง (reference pattern) V แบบมีอยู่ P ชุด ซึ่งจากการวัดระยะทางจาก DTW ของรูปแบบ (pattern) ที่ i จำนวน P ชุด แสดงได้ด้วย R^{ij} , $1 \leq i \leq V, 1 \leq j \leq P$ ซึ่งเราสามารถจัดเรียงได้ใหม่เป็น

$$D^{i,[1]} \leq D^{i,[2]} \leq \cdots \leq D^{i,[P]} \quad (2.143)$$

โดยที่ KNN rule จะหา average distance ได้จาก

$$r^{i} = \frac{1}{K}\sum_{k=1}^{K} D^{i,[k]} \quad (2.144)$$

โดยที่เราสามารถเลือกผลของการรู้จำได้จาก

$$i^{*} = \underset{i}{argmin}\left[r^{i}\right] \quad (2.145)$$

2.4.3. วิธีการของ Viterbi

ในขั้นตอนวิธีการแก้ปัญหาพื้นฐานทั้งสามประการของแบบจำลองฮิดเดน มาร์คอฟนั้น การแก้ปัญหาพื้นฐานข้อที่สองจัดเป็นขั้นตอนวิธีการตัดสินใจเลือกรูปแบบที่เหมาะสมที่สุดในการรู้จำ วิธีการหนึ่งที่ถูกนำมาใช้กับแบบจำลองฮิดเดน มาร์คอฟก็คือขั้นตอนวิธีการของ Viterbi (Rabiner and Levinson, 1981; Rabiner, 1989) โดยมีขั้นตอนวิธีการดังแสดงในการแก้ปัญหาพื้นฐานข้อที่สอง ส่วนรายละเอียดเชิงทฤษฎีมีดังนี้

รายละเอียดเชิงทฤษฎีของขั้นตอนวิธีการของ Viterbi

ขั้นตอนวิธีการของ Viterbi เป็นการแก้ปัญหาโดยการวนซ้ำเพื่อหาคำตอบที่เหมาะสมที่สุดของปัญหาในการประมาณค่าลำดับสถานะ หรือปัญหาในการประมาณค่าความน่าจะเป็นสูงสุดโดยการอุปนัย (Maximum a posteriori Probability, MAP) สำหรับกระบวนการมาร์คอฟที่มีสถานะจำกัดและเวลาไม่ต่อเนื่อง (Forney, Jr., 1973)

ภายใต้กระบวนการมาร์คอฟโดยเวลาไม่ต่อเนื่อง สถานะ x_k ที่เวลา k ซึ่งมีจำนวนจำกัด M ของสถานะ $m, 1 \leq m \leq M$ ขั้นแรกเริ่มต้นจากการสมมติให้กระบวนการอยู่ในช่วงเวลา 0 ถึง K เท่านั้นโดยมีสถานะเริ่มต้นเป็น x_0 และสถานะสิ้นสุดเป็น x_K ลำดับ สถานะจะแทนได้ด้วยเวกเตอร์จำกัด $\mathbf{x} = \left(x_0, \Lambda\ , x_K\right)$ ดังนั้นกระบวนการมาร์คอฟซึ่งมีค่าความน่าจะเป็น $P\left(x_{k+1} \middle| x_0, x_1, \Lambda\ , x_K\right)$ ของการอยู่ในสถานะ x_{k+1} ที่เวลา k+1 จะให้สถานะทั้งหมดจนถึงเวลา k โดยขึ้นอยู่กับสถานะ x_k ที่เวลา k เท่านั้น โดยที่ค่าความน่าจะเป็นของการเปลี่ยนแปลง $P\left(x_{k+1} \middle| x_k\right)$ อาจขึ้นกับเวลาดังนี้

$$P\left(x_{k+1} \middle| x_0, x_1, \Lambda\ , x_K\right) = P\left(x_{k+1} \middle| x_k\right) \quad (2.146)$$

กำหนดให้การเปลี่ยนแปลง ξ_k ที่เวลา k เป็นคู่สถานะ $\left(x_{k+1}, x_k\right)$ ดังนี้

$$\xi_k \cong \left(x_{k+1}, x_k\right) \quad (2.147)$$

กำหนดให้ Ξ เป็นชุดของการเปลี่ยนแปลง $\xi_k = \left(x_{k+1}, x_k\right)$ ซึ่ง $P\left(x_{k+1} \middle| x_k\right) \neq 0$ และมีค่าเป็น $|\Xi|$ โดยที่ $|\Xi| \leq M^2$ ดังนั้นจึงเป็นความสัมพันธ์แบบหนึ่งต่อหนึ่งระหว่างลำดับสถานะ $\mathbf{x}$ และลำดับการเปลี่ยนแปลง $\xi = \left(\xi_0, \Lambda\ , \xi_{K-1}\right)$ ซึ่งเขียนได้เป็น $\mathbf{x} \overset{1-1}{\leftrightarrow} \xi$ กระบวนการได้รับการสมมติให้ถูกสังเกต โดยประกอบไปด้วยลำดับ z ของค่าสังเกต z_k ซึ่งขึ้นอยู่กับค่าการเปลี่ยนแปลง ξ_k ที่เวลา k ในเชิงความน่าจะเป็นดังนี้

$$P(z|x) = P(z|\xi) = \prod_{k=0}^{K-1} P\left(z_k \middle| \xi_k\right) \quad (2.148)$$

ในกรณีที่มีการเปลี่ยนแปลงตามเวลาซึ่ง $P\left(z_k \middle| \xi_k\right)$ เป็นฟังก์ชันของ k เป็นกรณีพิเศษคือ

1) กรณีที่ z_k ขึ้นอยู่กับสถานะ x_k เท่านั้น

$$P(z|x) = \prod_k P\left(z_k \middle| \xi_k\right) \qquad (2.149)$$

2) กรณีที่ z_k ขึ้นอยู่กับขาออก y_k ของกระบวนการในเชิงความน่าจะเป็น ซึ่ง y_k เป็นฟังก์ชันที่สามารถหาค่าได้ของค่าการเปลี่ยนแปลง ξ_k หรือสถานะ x_k

หลักการของขั้นตอนวิธีการของ Viterbi

ขั้นตอนวิธีการของ Viterbi เป็นการหาคำตอบในการประมาณลำดับค่าความน่าจะเป็นสูงสุดโดยการอุปนัย ซึ่งเป็นวิธีเดียวกับการหาเส้นทางที่สั้นที่สุดภายในแผนภูมิ จากรูปที่ 2.14 เป็นแผนภาพสถานะของกระบวนการมาร์คอฟที่มีสถานะจำกัดและเวลาไม่ต่อเนื่อง โดยอาศัยปมแทนสถานะและกิ่งก้านสาขาแทนการเปลี่ยนแปลง ซึ่งภายในช่วงเวลาหนึ่งกระบวนการจะดำเนินไปตามเส้นทางจากสถานะหนึ่งไปยังอีกสถานะหนึ่งผ่านทางแผนภาพสถานะ

จากรูปที่ 2.15 แสดงแผนภาพ Trellis ซึ่งเป็นรายละเอียดของกระบวนการเดียวกันที่ซ้ำซ้อนมากยิ่งขึ้น โดยอาศัยปมแทนสถานะที่เวลาที่กำหนด และแต่ละกิ่งก้านสาขาแทนการเปลี่ยนแปลงไปยังสถานะใหม่ที่เวลาถัดไป แผนภาพ Trellis จะเริ่มต้นและสิ้นสุดลงที่สถานะ x_0 และ x_K ตามลำดับ โดยมีคุณสมบัติที่สำคัญซึ่งทุกลำดับสถานะ $\mathbf{x}$ ที่เป็นไปได้จะสัมพันธ์กับเส้นทางเอกลักษณ์ที่มีเพียงเส้นทางเดียวในแผนภาพ Trellis และเป็นจริงในทางกลับกัน

เมื่อกำหนดลำดับของค่าสังเกต $\mathbf{x}$ มาให้ ทุกเส้นทางอาจถูกกำหนดความยาวให้เป็นสัดส่วนกับ $-\ln P(\mathbf{x},\mathbf{z})$ เมื่อ $\mathbf{x}$ เป็นลำดับสถานะที่สัมพันธ์กับเส้นทางนั้น ซึ่งช่วยในการค้นหาลำดับสถานะซึ่ง $P(\mathbf{x}|\mathbf{z})$ มีค่าสูงสุด หรืออีกนัยหนึ่ง $P(\mathbf{x},\mathbf{z}) = P(\mathbf{x}|\mathbf{z})P(\mathbf{z})$ มีค่าสูงสุด โดยการค้นหาเส้นทางที่ความยาว $-\ln P(\mathbf{x},\mathbf{z})$ มีค่าน้อยที่สุด เนื่องจาก $\ln P(\mathbf{x},\mathbf{z})$ เป็นฟังก์ชันแบบไม่เปลี่ยนแปลงของ $P(\mathbf{x},\mathbf{z})$ ซึ่งมีความสัมพันธ์แบบหนึ่งต่อหนึ่งระหว่างเส้นทางและลำดับ ดังนั้นจึงสามารถสังเกต $P(\mathbf{x},\mathbf{z})$ ได้เป็นดังนี้

$$\begin{aligned} P(\mathbf{x},\mathbf{z}) &= P(\mathbf{x})P(\mathbf{z}|\mathbf{x}) \\ &= \prod_{k=0}^{K-1} P\left(x_{k+1} \middle| x_k\right) \prod_{k=0}^{K-1} P\left(z_k \middle| x_{k+1}, x_k\right) \end{aligned} \qquad (2.150)$$

ดังนั้นเมื่อกำหนดแต่ละกิ่งสาขาของการเปลี่ยนแปลงด้วยความยาวจะได้ว่า

$$\lambda(\xi_k) \cong -\ln P\left(x_{k+1} \middle| x_k\right) - \ln P\left(z_k \middle| \xi_k\right) \qquad (2.151)$$

ดังนั้นความยาวทั้งหมดของเส้นทางที่สัมพันธ์กับ $\mathbf{x}$ บางตัวจะมีค่าดังนี้

$$-\ln P(\mathbf{x},\mathbf{z}) = \sum_{k=0}^{K-1} \lambda(\xi_k) \qquad (2.152)$$

ในการค้นหาเส้นทางที่สั้นที่สุดบนแผนภูมินั้น มีพื้นฐานจากวิธีการของ Minty ซึ่งต้องการค่าสังเกตเพิ่มเติมอีกหนึ่งค่า เริ่มจากการกำหนดให้ $\mathbf{x}_0^k$ เป็นส่วนย่อยของ $\left(x_0, x_1, \Lambda\ , x_k\right)$ ที่ประกอบด้วยลำดับที่เวลา k ของลำดับสถานะ $\mathbf{x} = \left(x_0, x_1, \Lambda\ , x_k\right)$ ในแผนภาพ Trellis ที่ $\mathbf{x}_0^k$ สัมพันธ์กับส่วนย่อยของเส้นทางที่เริ่มต้นจากปม $\mathbf{x}_0$ และสิ้นสุดที่ปม $\mathbf{x}_k$ ดังนั้นสำหรับปม $\mathbf{x}_k$ ที่เฉพาะเวลา k ใดๆ จะประกอบด้วยส่วนย่อยของเส้นทางดังกล่าวหลายส่วนซึ่งมีความยาวเป็นดังนี้

$$\lambda\left(\mathbf{x}_0^k\right) = \sum_{i=0}^{k-1} \lambda(\xi_i) \qquad (2.153)$$

ส่วนของเส้นทางที่สั้นที่สุดจะสัมพันธ์กับปม x_k และแทนด้วย $\hat{\mathbf{x}}(x_k)$ ที่เวลา $k > 0$ ใดๆ จะประกอบด้วย M เส้นทางสำหรับแต่ละ x_k โดยที่ค่าสังเกตจะเป็นเส้นทางสมบูรณ์ที่สั้นที่สุด $\hat{\mathbf{x}}$ ซึ่งเริ่มต้นจากส่วนของเส้นทางที่สั้นที่

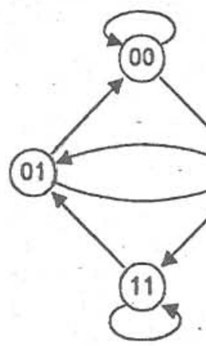
สุด ในทางกลับกันถ้าผ่านทางสถานะ x_k ที่เวลา k จะแทนที่ส่วนของเส้นทางเริ่มต้นด้วย $\hat{\mathbf{x}}(x_k)$ ซึ่งกลายเป็นเส้นทางที่สั้นที่สุด

ดังนั้นที่เวลา k ใดๆ จำเป็นต้องบันทึกไว้แต่เพียงเส้นทาง M เส้นทาง $\hat{\mathbf{x}}(x_k)$ พร้อมทั้งความยาว $\Gamma(x_k) \cong \lambda[\hat{\mathbf{x}}(x_k)]$ ของเส้นทางนั้น ที่เวลา k + 1 จะอาศัยการยืดขยายเส้นทางทั้งหมดที่เวลา k ออกไปหนึ่งหน่วยเวลา คำนวณหาความยาวในส่วนของเส้นทางที่ยืดขยายออกมา สำหรับแต่ละปม x_{k+1} จะเลือกส่วนของเส้นทางที่ยืดขยายออกมาที่สั้นที่สุดซึ่งสิ้นสุดที่ x_{k+1} โดยสัมพันธ์กับเส้นทางที่เวลา k + 1 กระบวนการวนซ้ำจะกระทำต่อไปจนกว่าจำนวนเส้นทางจะเกินกว่า M

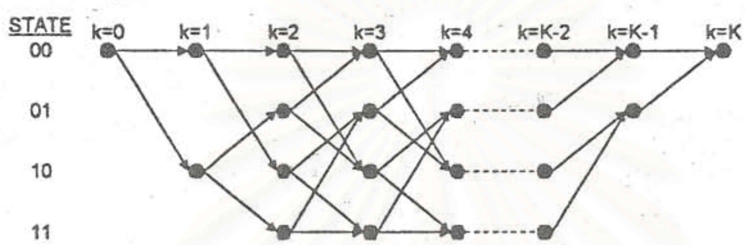
ตัวอย่างของขั้นตอนวิธีการ Viterbi ดังแสดงในรูปที่ 2.28 ด้วยแผนภาพ Trellis 4 สถานะที่ประกอบด้วย 5 หน่วยเวลา โดยบอกความยาวของแต่ละกิ่งสาขาไว้ด้วย ในรูปที่ 2.29 เป็นกระบวนการวนซ้ำ 5 รอบเพื่อค้นหาเส้นทางที่สั้นที่สุดจากปมเริ่มต้นไปยังปมสุดท้าย ในแต่ละชั้นจะแสดงเพียง 4 เส้นทางเท่านั้นพร้อมทั้งความยาวของแต่ละเส้นทาง ดังนั้นจากหลักการของขั้นตอนวิธีการดังแสดงในตารางที่ 2.13 จะได้ลำดับสถานะ **x** ที่มีจำนวนจำกัดซึ่งสิ้นสุดที่เวลา K ร่วมกับเส้นทางสมบูรณ์ที่สั้นที่สุด $\hat{\mathbf{x}}(x_K)$

ตารางที่ 2.13 หลักการเชิงทฤษฎีของขั้นตอนวิธีการ Viterbi

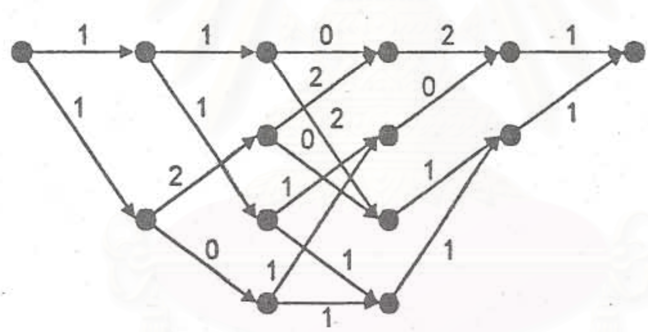
ขั้นตอนที่ 1 **กระบวนการกำหนดค่าเริ่มต้น**	k .... ดัชนีเวลา $\hat{\mathbf{x}}_k(x_k),\quad 1 \le x_k \le M$ .... เส้นทางที่สิ้นสุดที่ x_k $\Gamma(x_k),\quad 1 \le x_k \le M$ .... ความยาวของเส้นทาง
ขั้นตอนที่ 2 **กระบวนการเริ่มต้น**	k = 0 .... (2.122ก) $\hat{\mathbf{x}}(x_0) = x_0,\quad \hat{\mathbf{x}}(m)$ arbitrary, $\quad m \ne x_0$ .... (2.122ข) $\Gamma(x_0) = 0,\quad \Gamma(m) = \infty,\quad m \ne x_0$ .... (2.122ค)
ขั้นตอนที่ 3 **กระบวนการวนซ้ำ**	$\Gamma(x_{k+1}, x_k) \;\cong\; \Gamma(x_k) + \lambda[\xi_k = (x_{k-1}, x_k)],\quad \forall \xi_k = (x_{k+1}, x_k)$ .... (2.123ก) $\Gamma(x_{k+1}) = \min_{x_k} \Gamma(x_{k+1}, x_k)$, สำหรับแต่ละ x_{k-1} .... (2.123ข)
ขั้นตอนที่ 4 **กระบวนการสิ้นสุด**	จัดเก็บ $\Gamma(x_{k+1})$ และเส้นทาง $\hat{\mathbf{x}}(x_{k+1})$ ที่สัมพันธ์กัน กำหนดให้ k = k + 1 และวนซ้ำจะกระทั่ง k = K



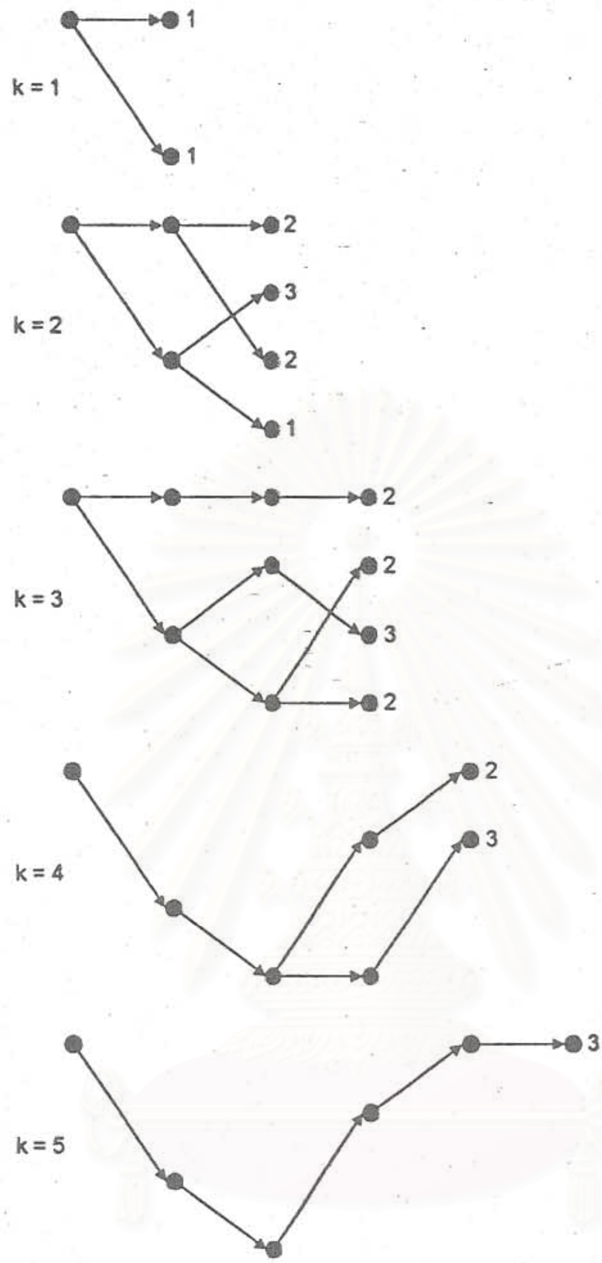
รูปที่ 2.28 แผนภาพสถานะ



รูปที่ 2.29 แผนภาพ Trellis



รูปที่ 2.30 แผนภาพ Trellis พร้อมแสดงขนาดความยาวของกิ่งสาขาเมื่อ M = 4 และ K = 5



รูปที่ 2.31 กระบวนการค้นหาเส้นทางที่สั้นที่สุดตามหลักการของขั้นตอนวิธีการ Viterbi

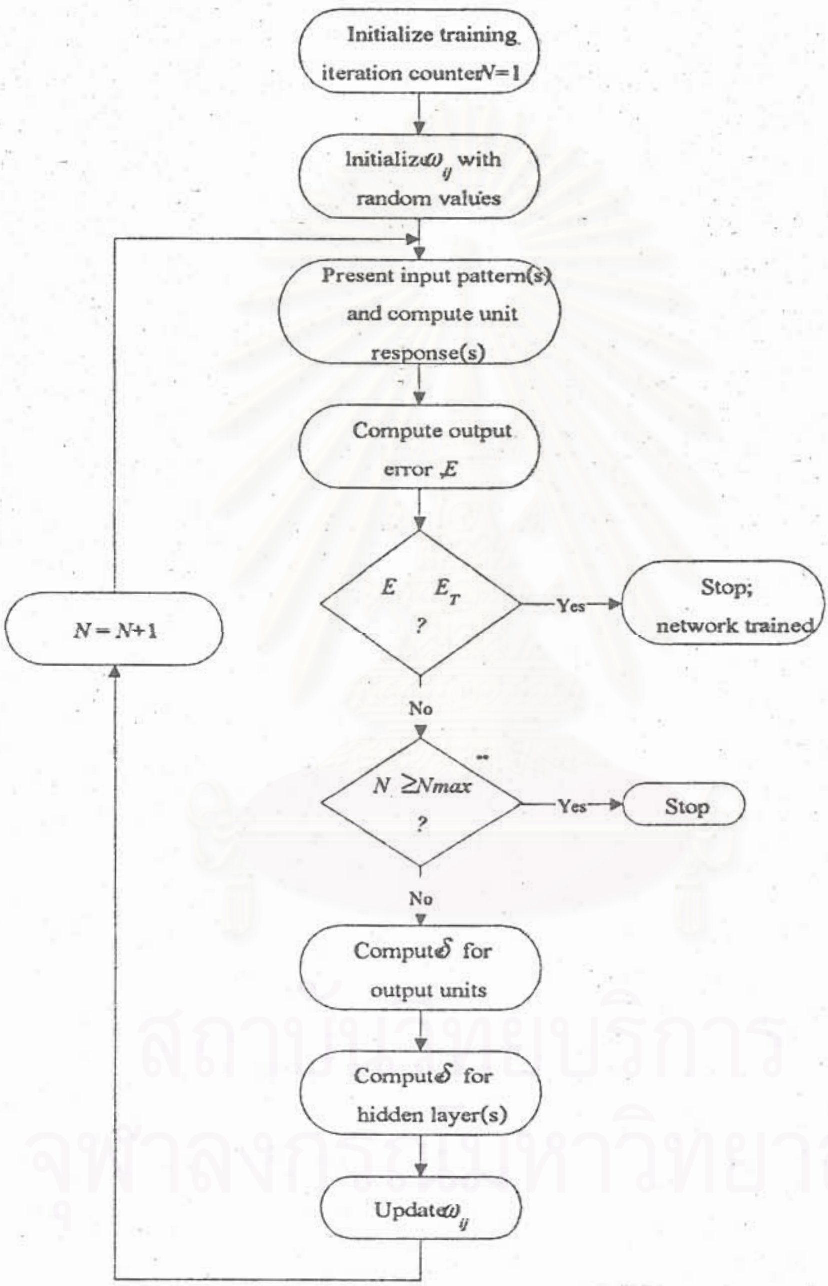
2.4.4. กฎเกณฑ์การตัดสินใจ (Decision Rule) ในกรรมวิธีนิวรอลเน็ตเวิร์ก

การใช้กฎเกณฑ์ใดตัดสินว่าคำที่เราไม่ทราบ (Unknown word) คือเสียงใด เราจะต้องคำนึงถึงวิธีการที่ใช้ในขั้นตอน Pattern similarity determination เพราะต้องมีความสอดคล้องกัน เนื่องจาก multi-layer perceptron neural network ที่ผ่านการฝึกแล้ว จะให้ค่าเอาต์พุตที่คล้ายคลึงกับค่าเอาต์พุตของตัวอย่างคู่ข้อมูลอินพุตเอาต์พุตที่มีค่าอินพุตของข้อมูลฝึกคล้ายกับค่าอินพุตที่ป้อนเข้ามา ดังนั้นเกณฑ์การตัดสินใจที่เลือกใช้คือการเลือกเสียงที่ตรงกับโหนดเอาต์พุตที่มีค่าเอาต์พุตสูงสุด อีกเหตุผลหนึ่งที่ใช้สนับสนุนเกณฑ์การตัดสินใจนี้คือ การทำงานของนิวรอลเน็ตเวิร์กขณะฝึกมีการปรับน้ำหนักการเชื่อมต่อในทิศทางที่ลดความผิดพลาดให้เหลือน้อยที่สุด ดังนั้นนิวรอลเน็ตเวิร์กที่ผ่านการฝึกแล้วจะให้ค่าเอาต์พุตที่มีความผิดพลาดน้อยที่สุดบนพื้นฐานความรู้ที่นิวรอลเน็ตเวิร์กได้รับจากการฝึก โหนดเอาต์พุตที่มีค่าเอาต์พุตสูงสุดคำนวณได้จาก

$$nodeoutput = i \qquad \text{เมื่อ } o_i = Max(o_1, o_2, \ldots, o_c) \quad (2.154)$$

โดยที่ c คือจำนวนกลุ่มของเสียงที่ต้องการรู้จำ

เราสามารถสรุปกระบวนการการเรียนรู้แบบ backpropagation ได้ดังแสดงในรูปที่ 2.32 และใช้สมการในตารางที่ 2.14



รูปที่ 2.32 กระบวนการการเรียนรูปแบบ backpropagation

โดยที่ N แทนจำนวนรอบในการเรียนรู้

N_{max} แทนจำนวนรอบสูงสุดที่ใช้ในการเรียนรู้

E แทนค่า output error

E_T แทนค่าระดับ output error ที่ต้องการเมื่อ E น้อยกว่าค่านี้ให้ยุติการ train

δ แทนค่า sensitivity ของ pattern error เทียบกับ net activation

ตารางที่ 2.14 สมการสำหรับการ train โดยใช้ backpropagation

(pattern) error measure	$E_p = \frac{1}{2}\sum_j (t_j^p - o_j^p)^2$
(pattern) weight correction	$\Delta^p \omega_{ji} = \in \delta_j^p \overline{o}_i^p$
(output units)	$\delta_j^p = (t_j^p - o_j^p) f_j'(net_j^p)$
(internal units)*	$\delta_j^p = f_j'(net_j^p)\sum_n \delta_n^p \omega_{nj}$
output derivative	$f_j'(net_j^p) = o_j^p(1 - o_j^p)$
(assumes sigmoidal characteristic)	

* Where δ_n^p are from the next (lower-numbered) layer.

บทที่ 3
ขั้นตอนวิธีในการดำเนินการวิจัยและผลการทดสอบ

3.0 บทนำ

รายละเอียดในบทที่ 3 นี้จะกล่าวถึงรายละเอียดของขั้นตอนวิธีในการดำเนินการวิจัย แต่เนื่องจากว่ากรรมวิธีที่งานวิจัยนี้ใช้มี 3 แนวทางคือ DTW HMM และ NN รายละเอียดในแต่ละขั้นตอนของกระบวนการจะแตกต่างกันเพื่อให้ผลลัพธ์ที่ได้ออกมาดีที่สุดเท่าที่จะทำได้สำหรับกรรมวิธีหนึ่ง ๆ อย่างไรก็ตาม จะมีการเปรียบเทียบผลการทดสอบในตอนท้ายของบทนี้ ทั้งนี้จะเริ่มอธิบายขั้นตอนวิธีในการดำเนินการวิจัยของ DTW ก่อน จากนั้นจะเป็นของ HMM และท้ายสุดจึงเป็นของ NN ผลการทดสอบของแต่ละกรรมวิธีจะแสดงในหัวข้อต่อไปเรียงตามลำดับเช่นเดียวกับขั้นตอนวิธีในการดำเนินการวิจัย และท้ายสุดเป็นการเปรียบเทียบผลการทดสอบ

3.1. วิธีการดำเนินการวิจัย

เนื่องจากงานวิจัยนี้ พยายามนำกรรมวิธีที่นิยมใช้กัน 3 แนวทางมาศึกษาร่วมกัน ทั้ง ๆ ที่มีรายละเอียดแตกต่างกันค่อนข้างมาก ดังนั้นวิธีการดำเนินการวิจัยที่ใช้ในแต่ละกรรมวิธีจะต้องปรับแต่งเพื่อให้ผลลัพธ์ที่ได้สำหรับกรรมวิธีนั้น ๆ ดีที่สุด อย่างไรก็ตาม มีขั้นตอนการเตรียมข้อมูลที่จะเหมือนกัน กล่าวคือ ในขั้นตอนเก็บข้อมูล การเก็บตัวอย่างข้อมูลเสียงพูดจะอาศัยการเก็บบันทึกข้อมูลไว้ในเครื่องคอมพิวเตอร์ โดยทำการบันทึกเสียง ณ ห้องปฏิบัติการวิจัยประมวลผลสัญญาณดิจิตอล ห้อง 303 ชั้น 3 ตึกวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ซึ่งได้รับการควบคุมสภาพแวดล้อมขณะทำการบันทึกเสียงให้คล้ายคลึงกับสภาพแวดล้อมของสถานที่ทำงานทั่วไปแต่พยายามให้มีเสียงรบกวนน้อยที่สุด โดยการบันทึกเสียงในเวลาปลอดคนทำงานอื่น เสียงพูดที่บันทึกไว้จะจัดเก็บด้วยตัวอย่างขนาด 8 บิตและมีอัตราการชักตัวอย่าง 8 KHz เนื่องจากเสียงพูดของคนเราส่วนมากจะอยู่ในช่วงความถี่ไม่เกิน 4 KHz จึงใช้อัตราการชักตัวอย่างที่ค่าความถี่สูงกว่าสองเท่าของค่า 4 KHz นี้อย่างน้อย 2 เท่า ตามทฤษฎี Nyquist

3.1.1. กฎเกณฑ์ในการคัดเลือกผู้บอกภาษา

คุณสมบัติของผู้บอกภาษาที่จะได้รับการบันทึกเสียง ต้องเป็นไปตามกฎเกณฑ์ดังนี้

1) เป็นผู้ที่ใช้ภาษาไทยกรุงเทพเป็นภาษาพูดทั่วไปและมีอายุระหว่าง 18 - 25 ปี
2) เป็นผู้ที่มีการออกเสียงเป็นปกติและตรงตามหลักการออกเสียงพูดภาษาไทย

3.1.2. อุปกรณ์เครื่องมือที่ใช้ในการบันทึกเสียง

รายละเอียดอุปกรณ์ที่ใช้ในการเก็บบันทึกเสียงดังนี้

1) เครื่องคอมพิวเตอร์ 80486DX2-66 พร้อมหน่วยความจำขนาด 8 MB พร้อมด้วย Harddisk ขนาด 1.2 GB และ Floppy Disk 3.5" ขนาด 1.44 MB
2) การ์ดเสียง Sound Blaster Pro ของบริษัท Creative Technology
3) ไมโครโฟน Philips Uni-directional Dynamic Microphone รุ่น SBC 465
4) ระบบปฏิบัติการ MS-DOS Version 6.22 ภาษาไทย
5) โปรแกรมพัฒนาซอฟต์แวร์ Borland C Version 3.1

ในงานวิจัยนี้ ทำการเก็บตัวอย่างเสียงพูดจำนวน 55 คน แบ่งเป็นชาย 50 คนและหญิง 10 คน ทำการบันทึกเสียงพูดตัวเลขศูนย์ถึงเก้า คนละ 3 ครั้ง ข้อมูลที่บันทึกไว้จะถูกนำไปใช้ในแต่ละกรรมวิธีรู้จำแตกต่างกันไป แต่โดยรวม จะแบ่งข้อมูลออกเป็น 2 กลุ่ม กลุ่มแรกจะแบ่งย่อยออกเป็น 2 กลุ่มย่อย A1 และ A2 กลุ่มย่อย A1 จะเป็นกลุ่มที่ใช้ฝึกฝน (Training set) และทดสอบความถูกต้องของกระบวนการในขั้นตอนต่าง ๆ ตั้งแต่การตัวหัวท้ายคำ ไปจนถึงกรรมวิธีรู้จำที่พัฒนาขึ้น ในขณะที่กลุ่มย่อย A2 จะเป็นกลุ่มที่ใช้ทดสอบเบื้องต้น (Test set 1) เพื่อตรวจสอบความผิดปกติของข้อมูลตัวอย่าง และตรวจสอบการทำงานของระบบโดยรวม ในขณะที่กลุ่มที่สอง B จะเป็นกลุ่มที่ใช้ทดสอบระบบ (Test set 2) จำนวนข้อมูลในแต่ละกลุ่มจะแปรเปลี่ยนไปตามกรรมวิธี ดังตารางที่ 3.1 เพื่อให้ได้อัตราการรู้จำของกรรมวิธีนั้น ๆ ที่ดีที่สุดสำหรับทุก ๆ ลักษณะของการทดสอบ ดังจะกล่าวถึงในภายหลัง

ตารางที่ 3.1 จำนวนข้อมูลแต่ละกลุ่มของกรรมวิธีรู้จำแบบต่าง ๆ ชุดละ 10 คำ ศูนย์ถึงเก้า

กรรมวิธี	Training set, A1	Test set 1, A2	Test set 2, B
ไดนามิก ไทม์วาร์ปิง	20 ตัวอย่าง ๆ ละ 1 ชุด	20 ตัวอย่าง ๆ ละ 2 ชุด	20 ตัวอย่าง ๆ ละ 2 ชุด
ฮิดเดน มาร์คอฟ	45 ตัวอย่าง ๆ ละ 1 ชุด	45 ตัวอย่าง ๆ ละ 1 ชุด	10 ตัวอย่าง ๆ ละ 1 ชุด
นิวรอลเน็ตเวิร์ก	30 ตัวอย่าง ๆ ละ 2 ชุด	30 ตัวอย่าง ๆ ละ 1 ชุด	12 ตัวอย่าง ๆ ละ 3 ชุด

3.1.3. ขั้นตอนในการดำเนินการวิจัยในแต่ละกรรมวิธี

3.1.3.1. กรรมวิธีไดนามิก ไทม์วาร์ปิง

รายละเอียดของขั้นตอนในการรู้จำคำพูดของทั้งสามกรรมวิธี แบ่งออกได้เป็น 2 ขั้นตอน ได้แก่ขั้นตอนการฝึกฝนระบบการรู้จำคำพูด และขั้นตอนการทดสอบระบบการรู้จำคำพูด

ก) ขั้นตอนการฝึกฝนระบบการรู้จำคำพูด

ขั้นตอนการฝึกระบบในกรรมวิธีไดนามิก ไทม์วาร์ปิง จะประกอบด้วยขั้นตอนต่าง ๆ คือ

ก.1) ขั้นตอนการประมวลสัญญาณเบื้องต้น ที่มีกรรมวิธีย่อยต่าง ๆ ดังนี้

ก.1.1) กรรมวิธีหาจุดสิ้นสุดเสียงพูด

กรรมวิธีหาจุดสิ้นสุดเสียงพูดที่ใช้ในกรรมวิธีไดนามิก ไทม์วาร์ปิง เป็นกรรมวิธีหาจุดสิ้นสุดเสียงพูดโดยใช้ค่าพลังงาน ในหัวข้อที่ 2.1.3 ที่มีการใช้ค่าพลังงานสามระดับเทียบกับค่าพลังงานสูงสุด ตามสมการที่ (2.6) โดยใช้ค่า a, b, และ c เป็น 0.3, 0.1, และ 0.5 ตามลำดับ (ระพีพัฒน์ เพ็ญศิริ 2538)

ก.1.2) กรรมวิธีวางกรอบขนาดสัญญาณ ในขั้นตอนการประมวลผลสัญญาณเบื้องต้นนั้น เนื่องจากสัญญาณเสียงพูดมีความแปรเปลี่ยนตามเวลา (Time-varying) และไม่เสถียร (Nonstationary) อีกทั้งยังเป็นสัญญาณสุ่มที่ไม่มีความเป็นเออร์กอดิก (Non-Ergodic) และไม่เป็นสัญญาณเฟ้นสุ่ม (Non-stochastic Signal) อีกด้วย ดังนั้นในการประยุกต์ใช้งานขั้นตอนวิธีการต่างๆ กับสัญญาณเสียงพูดจึงต้องแบ่งสัญญาณเสียงพูดออกเป็นส่วนย่อย (Rabiner and Levinson, 1981; Furui, 1985) เรียกว่า "กรอบเสียงพูด" (Speech Frame) โดยแต่ละกรอบเสียงพูดจะมีความยาวประมาณ 10 - 40 มิลลิวินาที (ms) ขึ้นอยู่กับความถี่ในการสุ่มตัวอย่าง (Sampling Frequency) ซึ่งในงานวิจัยนี้กำหนดให้ขนาดของกรอบเสียงพูดมีความยาว 20 มิลลิวินาที เพื่อให้สอดคล้องกับความถี่ในการสุ่มตัวอย่างของสัญญาณเสียงพูดที่ 8 KHz ทำให้ได้จำนวนข้อมูล 160 ค่าต่อหนึ่งกรอบเสียงพูดเพื่อใช้ในการประมวลผลต่อไป

ก.2) ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ

กรรมวิธีไดนามิก ไทม์วาร์ปิง ใช้ผลการแปลงฮาร์ตเลย์ เพื่อวิเคราะห์สัญญาณเชิงความถี่และวัดค่าลักษณะสำคัญตามสมการที่ (2.15) ในหัวข้อที่ 2.2.1 จากนั้นนำค่าพารามิเตอร์ที่ได้ไปสร้างแบบอ้างอิงตามสมการที่ (3.1)

$$R^i = \frac{\sum_{j=0}^{J-1} R_j^i}{J} \qquad (3.1)$$

โดยที่ j แทนจำนวนบุคคลที่จะนำมาสร้างแบบอ้างอิง, $j = 0,1,2,\ldots,J-1$

i แทนหมายเลขแบบอ้างอิง

R_j^i แทนรูปแบบของเสียงที่ i ของคนที่ j

R^i เป็นแบบอ้างอิงที่ i

ข) ขั้นตอนการทดสอบระบบการรู้จำคำพูด

ขั้นตอนการทดสอบระบบในกรรมวิธีไดนามิก ไทม์วาร์ปิง จะประกอบด้วยขั้นตอนต่าง ๆ คล้ายคลึงกับขั้นตอนการฝึกฝนระบบ กล่าวคือ

ข.1) ขั้นตอนการประมวลสัญญาณเบื้องต้น ที่มีกรรมวิธีย่อยต่าง ๆ ดังนี้

ก.1.1) กรรมวิธีหาจุดสิ้นสุดเสียงพูด

ก.1.2) กรรมวิธีวางกรอบขนาดสัญญาณ

ข.2) ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ

จะเหมือนกับขั้น ก.2 ที่แตกต่างกันคือการนำค่าพารามิเตอร์ที่ได้ไปสร้างแบบทดสอบ ตามสมการที่ (3.1)

ข.3) ขั้นตอนการวัดค่าหรือการทดสอบความคล้ายคลึงกัน

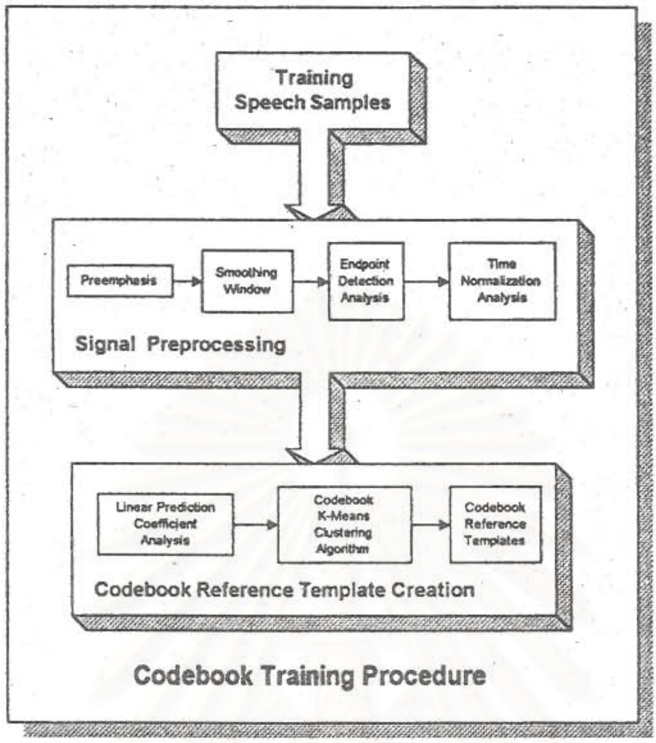
จะนำแบบที่จะทดสอบที่ได้ในขั้นตอน ข.2 ไปเปรียบเทียบกับแบบอ้างอิงในขั้นตอน ก.2 และวัดหา distance ตามสมการที่ (2.16) โดยใช้ Nearest Neighbor Rule ตามหัวข้อ 2.4.1 เป็นเงื่อนไขในการตัดสินใจ

กรรมวิธีไดนามิก ไทม์วาร์ปิง ในขั้นตอนการประมวลสัญญาณเบื้องต้น ไม่มีกรรมวิธีเน้นล่วงหน้า เพราะอาจลดทอนรายละเอียดหรือลักษณะสำคัญที่จำเป็นต่อการรู้จำ และไม่มีกรรมวิธีปรับบรรทัดฐานเชิงเวลาด้วยเช่นกัน เพราะหลักการของไดนามิก ไทม์วาร์ปิง เสมือนกับมีกรรมวิธีปรับบรรทัดฐานเชิงเวลาอยู่ในตัว

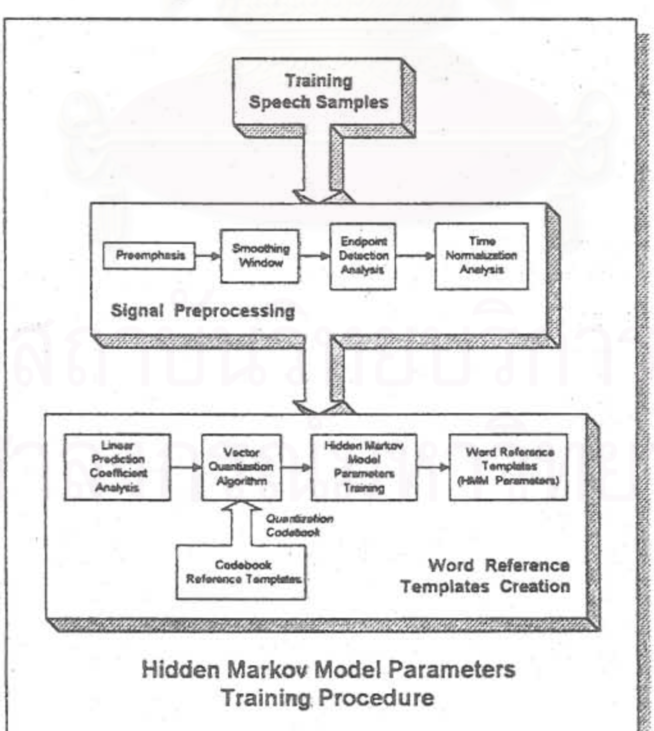
3.1.3.2. กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

ก) ขั้นตอนการฝึกฝนระบบการรู้จำคำพูด

ขั้นตอนการฝึกฝนระบบการรู้จำคำพูดนี้ จัดเป็นขั้นตอนในการสร้างชุดรหัสและชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ เพื่อใช้ในการควอนไทซ์แบบเวกเตอร์และการรู้จำคำพูดตามลำดับ โดยมีรายละเอียดของขั้นตอนการฝึกฝนระบบการรู้จำคำพูดแบ่งออกได้เป็น 2 ขั้นตอน ได้แก่ กระบวนการสร้างและฝึกฝนชุดรหัส และกระบวนการสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ ดังแสดงในรูปที่ 3.1 และ 3.2 ตามลำดับ ซึ่งมีรายละเอียดในแต่ละขั้นตอนดังนี้



รูปที่ 3.1 รายละเอียดกระบวนการสร้างและฝึกฝนชุดรหัส



รูปที่ 3.2 รายละเอียดกระบวนการสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ

ก.1) การสร้างและฝึกฝนชุดรหัส (Codebook Training Procedure)

กระบวนการสร้างและฝึกฝนชุดรหัส จัดเป็นกระบวนการสร้างชุดรหัสด้วยขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนของข้อมูลเสียงพูดเพื่อใช้ในการควอนไทซ์แบบเวกเตอร์ ซึ่งประกอบไปด้วย 2 ขั้นตอนได้แก่ ขั้นตอนการประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing) และขั้นตอนการสร้างชุดรหัสอ้างอิง (Codebook Reference Template Creation) โดยมีรายละเอียดแต่ละขั้นตอน ดังนี้

ก.1.1) ขั้นตอนการประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

ขั้นตอนการประมวลผลสัญญาณเบื้องต้นเป็นกรรมวิธีในการจัดเตรียมข้อมูลจากข้อมูลเสียงที่ได้จากการบันทึกเสียงซึ่งเป็นข้อมูลดิบ นำมาผ่านกรรมวิธีประมวลผลสัญญาณเชิงเลขโดยแบ่งออกได้เป็น 4 กรรมวิธีย่อยได้แก่ กรรมวิธีเน้นล่วงหน้า (Preemphasis) กรรมวิธีวางกรอบขนาดสัญญาณ (Smoothing Window) กรรมวิธีหาจุดสิ้นสุดเสียงพูดพร้อมทั้งจุดสิ้นสุดพยางค์ (Endpoint Detection) และกรรมวิธีปรับบรรทัดฐานเชิงเวลา (Time Normalization) ตามลำดับ

ก.1.1.1) กรรมวิธีเน้นล่วงหน้า (Preemphasis)

ขั้นตอนกรรมวิธีเน้นล่วงหน้านี้ จะนำสัญญาณผ่านตัวกรองเชิงเลขลำดับหนึ่ง (First-Order Digital Filter) ที่มีฟังก์ชันถ่ายโอน H(z) ดังสมการที่ (2.1) และ (2.2) โดยกำหนดให้ค่าสัมประสิทธิ์ของตัวกรอง a มีค่าเข้าใกล้ 1 ในงานวิจัยนี้กำหนดให้มีค่า a = 0.95 เนื่องจากเป็นค่าที่ให้ผลดีที่สุดในการคำนวณหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น (Rabiner, Levinson, Rosenberg, and Wilpon, 1979)

ก.1.1.2) กรรมวิธีวางกรอบขนาดสัญญาณ (Smoothing Window)

ขั้นตอนกรรมวิธีวางกรอบขนาดสัญญาณ สำหรับในงานวิจัยนี้เลือกใช้ฟังก์ชันกรอบชนิด Hamming สำหรับวิเคราะห์เสียงพูดโดยเฉพาะดังสมการที่ (2.3) และ (2.4)

ก.1.1.3) กรรมวิธีหาจุดสิ้นสุดเสียงพูด (Endpoint Detection)

ขั้นตอนกรรมวิธีหาจุดสิ้นสุดเสียงพูด ใช้วิธีการเดียวกับของกรรมวิธีไดนามิก ไทม์วาร์ปิง

ก.1.1.4) กรรมวิธีปรับบรรทัดฐานเชิงเวลา (Time Normalization)

ขั้นตอนกรรมวิธีปรับบรรทัดฐานเชิงเวลา เป็นขั้นตอนในการเพิ่มหรือลดขนาดความยาวของสัญญาณในเชิงเวลา เนื่องจากสัญญาณเสียงพูดของแต่ละบุคคลมีความยาวไม่เท่ากัน จึงจำเป็นต้องปรับแต่งขนาดความยาวของสัญญาณเสียงพูดให้เหมาะสม เพื่อใช้ในการหาค่าลักษณะสำคัญและการเปรียบเทียบลักษณะสำคัญของสัญญาณเสียงต่อไป อย่างไรก็ตาม งานวิจัยนี้เน้นการรู้จำเสียงพูดตัวเลขภาษาไทย 0-9 ซึ่งเป็นพยางค์โดด ๆ ความแตกต่างของช่วงสัญญาณเสียงที่มีข้อสนเทศของแต่ละบุคคลไม่มีนัยสำคัญพอที่จะต้องดำเนินกรรมวิธีปรับบรรทัดฐาน (เสาวลักษณ์ อารีย์พงศา, 2538)

ก.1.2) ขั้นตอนการสร้างชุดรหัสอ้างอิง (Codebook Reference Template Creation)

ขั้นตอนการการสร้างชุดรหัสอ้างอิง เป็นกรรมวิธีในการฝึกฝนชุดรหัสที่สร้างขึ้นมาโดยการสุ่มจากชุดตัวอย่างเสียงพูดทั้งหมด แล้วนำชุดรหัสเริ่มต้นที่ได้จากการสุ่ม (Randomly Initialized Codebook) มาทำการฝึกฝนกับชุดตัวอย่างเสียงพูดทั้งหมดด้วยขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วน (K-Means Clustering Algorithm) ซึ่งจัดอยู่ในประเภทขั้นตอนวิธีการแบ่งกลุ่มแบบวนซ้ำ (Iterative Clustering Algorithm) เนื่องจากขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนมีประสิทธิภาพใกล้เคียงกับการแบ่งกลุ่มแบบทวิภาคเมื่อจำนวนข้อมูลเวกเตอร์ฝึกฝนมีมากเพียงพอ (Makhoul, Roucos, and Gish, 1985) ขั้นตอนการสร้างชุดรหัสอ้างอิงแบ่งออกเป็น 2 กรรมวิธีย่อยได้แก่ ขั้นตอนการหาค่าสัมประสิทธิ์ของการประมาณ

พันธะเชิงเส้น (Linear Prediction Coefficient Analysis) และขั้นตอนการฝึกฝนชุดรหัสอ้างอิง (Codebook Reference Template Training) ตามลำดับ

ก.1.2.1) ขั้นตอนการหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น

ขั้นตอนการหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น จัดเป็นขั้นตอนในการลดจำนวนข้อมูลโดยการแสดงลักษณะของรูปคลื่นด้วยค่าพารามิเตอร์เพียงไม่กี่ค่าได้อย่างมีประสิทธิภาพ โดยนำข้อมูลที่ผ่านกรรมวิธีหาจุดสิ้นสุดเสียงพูดมาทำการวิเคราะห์ สำหรับงานวิจัยนี้จะอาศัยวิธีการทางอัตสหสัมพันธ์ด้วยขั้นตอนวิธีการ Levinson-Durbin (O'Shaughnessy, 1988) ในการหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นสำหรับแต่ละเสียงพูดนั้น จะหาค่าสัมประสิทธิ์เฉพาะแต่ละกรอบข้อมูล โดยการแทนที่แต่ละกรอบข้อมูลด้วยเวกเตอร์ของสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น สำหรับงานวิจัยนี้จะใช้ค่าลำดับของสัมประสิทธิ์ที่ 10 ลำดับ เนื่องจากเป็นค่าลำดับที่ให้อัตราการรู้จำสูงที่สุดในการทดลองเมื่อเปรียบเทียบกับค่าลำดับอื่น (เสาวลักษณ์ อารีย์พงศา, 2538)

ก.1.2.2) ขั้นตอนการฝึกฝนชุดรหัสอ้างอิง

ขั้นตอนการฝึกฝนชุดรหัสอ้างอิง จัดเป็นขั้นตอนในการสร้างและฝึกฝนชุดรหัสโดยอาศัยชุดตัวอย่างเสียงพูดในการฝึกฝน การสร้างชุดรหัสอ้างอิงเริ่มต้นจะทำการสุ่มตัวอย่าง จากชุดเวกเตอร์ของสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นในแต่ละตัวอย่างเสียงพูด สำหรับการฝึกฝนกับชุดตัวอย่างเสียงพูดเพื่อสร้างชุดรหัสอ้างอิงที่สมบูรณ์ การฝึกฝนชุดรหัสอ้างอิงจะอาศัยขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วนโดยใช้วิธีการวัดค่าความเพี้ยนแบบค่าความเพี้ยนกำลังสองเฉลี่ย (Mean Square Error, MSE) สำหรับงานวิจัยนี้จะทำการสร้างและฝึกฝนชุดรหัสจำนวน 10 ชุด และนำมาเฉลี่ยพร้อมทั้งฝึกฝนเป็นชุดรหัสอ้างอิงเฉลี่ยเพื่อนำไปทำการควอนไทซ์แบบเวกเตอร์ต่อไป

การสุ่มตัวอย่างชุดรหัสเริ่มต้นที่แตกต่างกันไป จะช่วยให้ชุดรหัสที่ได้เข้าสู่ค่าที่เหมาะสมที่สุดที่ครอบคลุมทั้งหมด (Global Optimum) และมีค่าความเพี้ยนต่ำที่สุด สำหรับในงานวิจัยนี้จะใช้ชุดรหัสขนาด 64 เวกเตอร์ของสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ 10 ลำดับ ซึ่งเป็นขนาดชุดรหัสที่เหมาะสมสำหรับเก็บข้อมูลเสียงพูดที่มีปริมาณไม่มาก (Rabiner, Levinson, and Sondhi, 1982; เสาวลักษณ์ อารีย์พงศา, 2538)

ก.2) ขั้นตอนการสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Model Parameters Training)

รายละเอียดกระบวนการสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ แสดงในรูปที่ 3.2 ประกอบไปด้วย 3 ขั้นตอนได้แก่ ขั้นตอนการวิเคราะห์หาสัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear Prediction Coefficient Analysis) ขั้นตอนการควอนไทซ์แบบเวกเตอร์ (Vector Quantization) และขั้นตอนการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Model Parameters Training) ภายหลังจากการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟจะได้ชุดรูปร่องต้นแบบอ้างอิงของเสียงพูดแต่ละคำ (Word Reference Template)

ก.2.1) ขั้นตอนการวิเคราะห์หาสัมประสิทธิ์การประมาณพันธะเชิงเส้น

รายละเอียดอยู่ในขั้นตอนการสร้างและฝึกฝนชุดรหัสหัวข้อที่ ก.1.2

ก.2.2) ขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์ (Vector Quantization Algorithm)

ขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์ จัดเป็นขั้นตอนในการลดขนาดจำนวนข้อมูลลงให้มีความเพี้ยนน้อยที่สุด โดยการแทนที่เวกเตอร์สัมประสิทธิ์ของการประมาณพันธะเชิงเส้นด้วยเวกเตอร์ของชุดรหัสที่ให้ค่าความเพี้ยนกำลังสองเฉลี่ยมีค่าน้อยที่สุด ชุดรหัสที่ใช้ในการควอนไทซ์แบบเวกเตอร์จะเป็นชุดรหัสที่ได้จากกระบวนการฝึกฝน

ชุดรหัส (Makhoul, Roucos, and Gish, 1985) ผลลัพธ์ที่ได้จะเป็นลำดับหมายเลขเวกเตอร์ของชุดรหัสที่ใช้ในการควอนไทซ์ สำหรับงานวิจัยนี้จะใช้ชุดรหัสขนาด 64 เวกเตอร์ของสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ 10 ลำดับ

ก.2.3) ขั้นตอนการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ

(Hidden Markov Model Parameters Training)

ขั้นตอนวิธีการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ จัดเป็นกระบวนการสร้างชุดรูปร่องต้นแบบอ้างอิงของเสียงพูดแต่ละคำ โดยทำการปรับเปลี่ยนค่าพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ $\lambda = (A, B, \pi)$ ให้เป็นไปตามเสียงพูดแต่ละคำ ซึ่งอาศัยการแก้ไขปัญหาพื้นฐานข้อที่ 1 และข้อที่ 3 ของแบบจำลองฮิดเดน มาร์คอฟด้วยกระบวนการไปหน้า-ย้อนกลับ (Forward-Backward Procedure) และการประมาณค่าซ้ำของ Baum-Welch (Baum-Welch Reestimation Procedure)

สำหรับการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟในงานวิจัยนี้ จะใช้ข้อมูลเสียงพูดที่ผ่านกระบวนการควอนไทซ์แบบเวกเตอร์เป็นข้อมูลขาเข้า โดยใช้เสียงพูดแต่ละคำของผู้พูดทุกคนมาฝึกฝนร่วมกันในแต่ละครั้ง ทำให้ชุดพารามิเตอร์ที่ได้เป็นตัวแทนหนึ่งเดียวของทุกเสียงพูดทั้งหมด เพื่อใช้เป็นชุดรูปร่องต้นแบบอ้างอิงสำหรับเสียงพูดแต่ละคำในการรู้จำต่อไป

ข) ขั้นตอนการทดสอบระบบการรู้จำคำพูด

ขั้นตอนในการทดสอบระบบการรู้จำคำพูดโดยใช้แบบจำลองฮิดเดน มาร์คอฟประกอบด้วย 3 ขั้นตอน ได้แก่ขั้นตอนการประมวลผลสัญญาณเบื้องต้น ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ และขั้นตอนการจำแนกรูปแบบร่วมกับขั้นตอนวิธีการตัดสินใจ ตามลำดับ เนื่องจากในขั้นตอนการประมวลผลสัญญาณเบื้องต้นมีรายละเอียดอยู่ในขั้นตอนการสร้างและฝึกฝนชุดรหัสหัวข้อที่ ก.1 ในหัวข้อนี้จึงแสดงแต่เพียงรายละเอียดของสองขั้นตอนที่เหลือเท่านั้นดังนี้

ข.1) ขั้นตอนการประมวลผลสัญญาณเบื้องต้น

รายละเอียดอยู่ในขั้นตอนการสร้างและฝึกฝนชุดรหัสหัวข้อที่ ก.1.2

ข.2) ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ

ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญประกอบด้วย 2 ขั้นตอน ได้แก่ขั้นตอนการหาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น และขั้นตอนวิธีการควอนไทซ์แบบเวกเตอร์ตามลำดับ โดยอาศัยชุดรหัสต้นแบบอ้างอิง (Codebook Reference Templates) ที่ได้จากการฝึกฝนชุดรหัสอ้างอิงมาใช้ในการควอนไทซ์แบบเวกเตอร์ ซึ่งขั้นตอนทั้งสองนี้เป็นขั้นตอนเดียวกับขั้นตอนย่อยของการสร้างและฝึกฝนชุดรหัส กับการสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟตามลำดับ ดังมีรายละเอียดแสดงในหัวข้อที่ ก.1.2 และ ก.2 ตามลำดับ

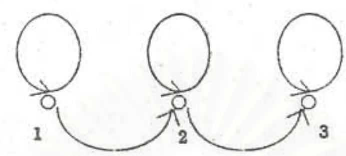
ข.3) ขั้นตอนการจำแนกรูปแบบร่วมกับขั้นตอนวิธีการตัดสินใจ

(Pattern Classification and Decision Making Algorithm)

ขั้นตอนการจำแนกรูปแบบของเสียงพูดอาศัยแบบจำลองฮิดเดน มาร์คอฟ ร่วมกับชุดรูปร่องต้นแบบอ้างอิงของเสียงพูดแต่ละคำ ซึ่งได้จากขั้นตอนการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ เพื่อการจำแนกความแตกต่างของเสียงพูดแต่ละคำโดยการเปรียบเทียบเสียงพูด ที่นำมาทดสอบกับชุดรูปร่องต้นแบบของเสียงพูดเพื่อใช้ในขั้นตอนวิธีการตัดสินใจ สำหรับในงานวิจัยนี้ ขั้นตอนการวิธีการตัดสินใจจะอาศัยแบบจำลองฮิดเดน มาร์คอฟ โดยการแก้ไขปัญหาพื้นฐานข้อที่ 2 ด้วยขั้นตอนวิธีการ Viterbi (Viterbi Algorithm) ผลลัพธ์ที่ได้จากขั้นตอนนี้จะเป็นชุดของเสียงพูดที่รู้จำได้ซึ่งมีค่าความน่าจะเป็นสูงที่สุด (Rabiner and Juang, 1986; Rabiner, 1989)

แบบจำลองฮิดเดน มาร์คอฟที่ใช้สำหรับงานวิจัยนี้ เป็นประเภทแบบจำลองซ้าย-ขวา (Left-Right Model) หรือแบบจำลองอนุกรม (Serial Model) ที่มี 3 สถานะซึ่งมีการเชื่อมโยงระหว่างสถานะดังมีตัวอย่างแสดงในรูปที่ 3.3

เนื่องจากจำนวนสถานะที่เลือกใช้มีความเพียงพอสำหรับใช้กับคำโดด และการเชื่อมโยงระหว่างสถานะไม่จำเป็นต้องเชื่อมโยงกันทั้งหมด ดังนั้นแบบจำลองที่ใช้จึงมีเพียงการเชื่อมโยงในสถานะ การเชื่อมโยงระหว่างสถานะ และการเชื่อมโยงข้ามสถานะเท่านั้นดังรูป (Rabiner and Levinson, 1985; Bahl, Brown, Souza, Mercer, and Picheny, 1988; Lee, Hon, and Reddy, 1990; เสาวลักษณ์ อารีย์พงศา, 2538)



รูปที่ 3.3 แบบจำลองฮิดเดน มาร์คอฟ แบบซ้าย-ขวา ที่มี 3 สถานะ

3.1.3.3. กรรมวิธีนิวรอลเน็ตเวิร์ก

ขั้นตอนการฝึกฝนระบบการรู้จำคำพูด และขั้นตอนการทดสอบระบบการรู้จำคำพูด ที่ใช้ในกรรมวิธีนิวรอลเน็ตเวิร์ก จะมีกระบวนการคล้ายคลึงกัน กล่าวคือ

ก) ขั้นตอนการประมวลผลสัญญาณเบื้องต้น

ก.1) กรรมวิธีหาจุดสิ้นสุดเสียงพูด

กรรมวิธีหาจุดสิ้นสุดเสียงพูดตามกรรมวิธีนิวรอลเน็ตเวิร์ก เป็นไปตามหัวข้อที่ 2.1.3 โดยที่ ค่าพารามิเตอร์ a, b , m, และ n มีค่าเป็น 0.5 เท่า 0.3 เท่า 3 ส่วน และ 2 ส่วน ตามลำดับ ซึ่งเป็นค่าที่ได้จากการใช้กรรมวิธีลองผิดลองถูก (trial and error) กับเสียงพูดที่มีทั้ง หนึ่ง สอง และสาม พยางค์ (เพื่อรองรับการนำนิวรอลเน็ตเวิร์กไปใช้ในการรู้จำเสียงพูดอื่น ๆ ต่อไป) โดยทดสอบกับเสียงพูดรวม 880 เสียง (วุฒิพงษ์ พรสุขจันทรา, 2539) จากจุดสิ้นสุดเสียงพูดทั้งหัวและท้ายคำที่ได้ จาก a_1 ถึง b_1 จะเก็บข้อสนเทศในส่วนหน้าของหัวและส่วนตามของท้ายไว้ด้วยโดยการขยายช่วงพิจารณาออกไปเป็นจาก a_2 ถึง b_2 เป็นระยะ 6/26 และ 1/26 ตามลำดับ เมื่อเทียบกับช่วงจาก a_1 ถึง b_1 (วุฒิพงษ์ พรสุขจันทรา, 2539)

ก.2) กรรมวิธีปรับบรรทัดฐานเชิงเวลา

กรรมวิธีปรับบรรทัดฐานเชิงเวลาที่ใช้ในกรรมวิธีนิวรอลเน็ตเวิร์ก คือวิธีการประมาณค่าในช่วงเชิงเส้นในการปรับบรรทัดฐานของสัญญาณเสียงพูด เนื่องจากเป็นวิธีการที่ต้องการเวลาในการประมวลต่ำ และต้องการหน่วยความจำไม่มากในการเก็บข้อมูล ในขณะที่ให้ผลลัพธ์ใกล้เคียงกับวิธีการเปลี่ยนอัตราชักตัวอย่าง ที่ใช้เวลาและหน่วยความจำมากกว่าซึ่งแปรตามอัตราการชักตัวอย่าง จำนวนค่าตัวอย่างของเสียงพูดหนึ่งพยางค์โดดถูกกำหนดไว้ที่ 4000 ค่า

ก.3) กรรมวิธีวางกรอบขนาดสัญญาณ

กรอบเสียงพูดคือ 20 มิลลิวินาที ที่ 8 KHz รวมค่าตัวอย่าง 160 ค่าต่อกรอบ เช่นเดียวกับสองกรรมวิธีข้างต้น การเหลื่อมของส่วนย่อยมีค่าเท่ากับ 1/4 ส่วนย่อย (เสาวลักษณ์ อารีย์พงศา,2538) ดังนั้นจำนวนค่าตัวอย่างใน 1 ส่วนย่อยที่ไม่เหลื่อมกับข้อมูลในส่วนย่อยอื่นจะเท่ากับ 120 ค่า ทำให้ได้ค่า L ที่เป็นจำนวนกรอบในสัญญาณเสียงพูดแต่ละพยางค์ (คำโดด) มีจำนวนกรอบ 33 กรอบ

ข) ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ

งานวิจัยนี้อาศัยวิธีการทางอัตสหสัมพันธ์ เพื่อหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น และใช้ลำดับของค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นเป็น 10 เท่ากับ ที่ใช้ในกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

ค) ขั้นตอนการทดสอบหรือการวัดความคล้ายคลึงกัน

นิวรอลเน็ตเวิร์กที่ใช้ในงานวิจัยนี้ประกอบด้วย ระดับข้อมูลเข้า (input layer) ซึ่งใช้เป็นที่เก็บค่าอินพุตที่ป้อนให้กับเน็ตเวิร์ก, ระดับซ่อนตัว (hidden layer) และระดับข้อมูลออก (output layer) ซึ่งใช้สำหรับแสดงค่าเอาต์พุตของเน็ตเวิร์ก โดยที่ระดับซ่อนตัว (hidden layer) มีเพียง 1 ระดับ เพราะนิวรอลเน็ตเวิร์ก 3 ระดับสามารถประมาณฟังก์ชันต่อเนื่องใด ๆ ได้ (Schalkoff,1992; KUNG,1993)

เงื่อนไขในการเลือกจำนวนโหนดในระดับต่าง ๆ มีดังต่อไปนี้ จำนวนโหนดในระดับข้อมูลเข้า (input layer) กำหนดโดยขนาดของข้อมูลอินพุตที่ใช้แทนเสียงพูด 1 ค่า คือ L ส่วนย่อยและ 1 ส่วนย่อยประกอบด้วยสัมประสิทธิ์ LPC p ค่า ดังนั้นจำนวนโหนดในระดับข้อมูลเข้ามีค่าเท่ากับ pL โหนด จำนวนโหนดในระดับข้อมูลออก (output layer) กำหนดโดยจำนวนกลุ่มข้อมูลที่ต้องการแบ่งแยก ในที่นี้คือจำนวนเสียงพูดที่ต้องการรู้จำ สำหรับจำนวนโหนดในระดับซ่อนตัว (hidden layer) ซึ่งเป็นตัวกำหนดความยืดหยุ่นของขอบเขตการตัดสินใจ จะต้องทดลองเพื่อหาค่าที่เหมาะสม เพราะถ้ากำหนดให้มีจำนวนโหนดมากก็จะเสียเวลาในการฝึกนาน และอาจทำให้นิวรอลเน็ตเวิร์กจำลักษณะเฉพาะของเสียงในข้อมูลฝึกมากเกินไป ถ้ากำหนดให้มีจำนวนโหนดน้อยเกินไป นิวรอลเน็ตเวิร์กอาจไม่มีความสามารถพอสำหรับการจำลักษณะทั่ว ๆ ไปของเสียงในข้อมูลฝึก ดังนั้นจะทดลองฝึกนิวรอลเน็ตเวิร์กโดยใช้จำนวนโหนดในระดับซ่อนตัว (hidden layer) มีค่าหลาย ๆ ค่า แล้วค่อย ๆ ลดจำนวนโหนดลงเพื่อหาจำนวนโหนดที่น้อยที่สุดที่นิวรอลเน็ตเวิร์กยังมีความสามารถในการจำลักษณะทั่ว ๆ ไปของเสียงได้

การทำงานของนิวรอลเน็ตเวิร์กในขณะฝึก (training) สามารถแบ่งเป็น 2 ลักษณะคือ การแพร่กระจายแบบไปข้างหน้า (feed forward) และการแพร่กระจายในทิศย้อนกลับ (back propagation) โดยในส่วนแรกจะเป็นการป้อนชุดของสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น เป็นข้อมูลอินพุตให้กับนิวรอลเน็ตเวิร์ก จากนั้นนิวรอลเน็ตเวิร์กจะทำการคำนวณจากระดับข้อมูลเข้า (input layer) ไปยังระดับข้อมูลออก (output layer) เพื่อหาค่าเอาต์พุตของทุกโหนด และในส่วนที่สองเป็นการคำนวณในทิศย้อนกลับจากเอาต์พุตมายังอินพุต โดยเป็นการหาค่าความผิดพลาดระหว่างค่าเอาต์พุตที่ได้กับค่าเอาต์พุตที่ต้องการ และลดค่าความผิดพลาดโดยการปรับค่าน้ำหนักการเชื่อมต่อ (connection weight) ที่เชื่อมต่อระหว่างโหนดแต่ละระดับ (layer) การทำงานทั้งสองลักษณะจะถูกทำซ้ำไปเรื่อย ๆ โดยการใช้ตัวอย่างคู่ข้อมูลอินพุตเอาต์พุตในชุดฝึก จนกว่าจะได้ผลตามที่ต้องการ ส่วนการทำงานของนิวรอลเน็ตเวิร์กในขณะทดสอบ จะมีแต่การแพร่กระจายแบบไปข้างหน้า เพื่อหาค่าเอาต์พุตของทุกโหนดเท่านั้นและใช้กฎเกณฑ์ในการตัดสินใจ เพื่อเลือกว่าเสียงพูดเป็นเสียงใด กระบวนการเรียนรู้แบบ backpropagation ที่ได้กล่าวไปแล้วนั้น แสดงในรูปที่ 3.4

โดยที่ N แทนจำนวนรอบในการเรียนรู้

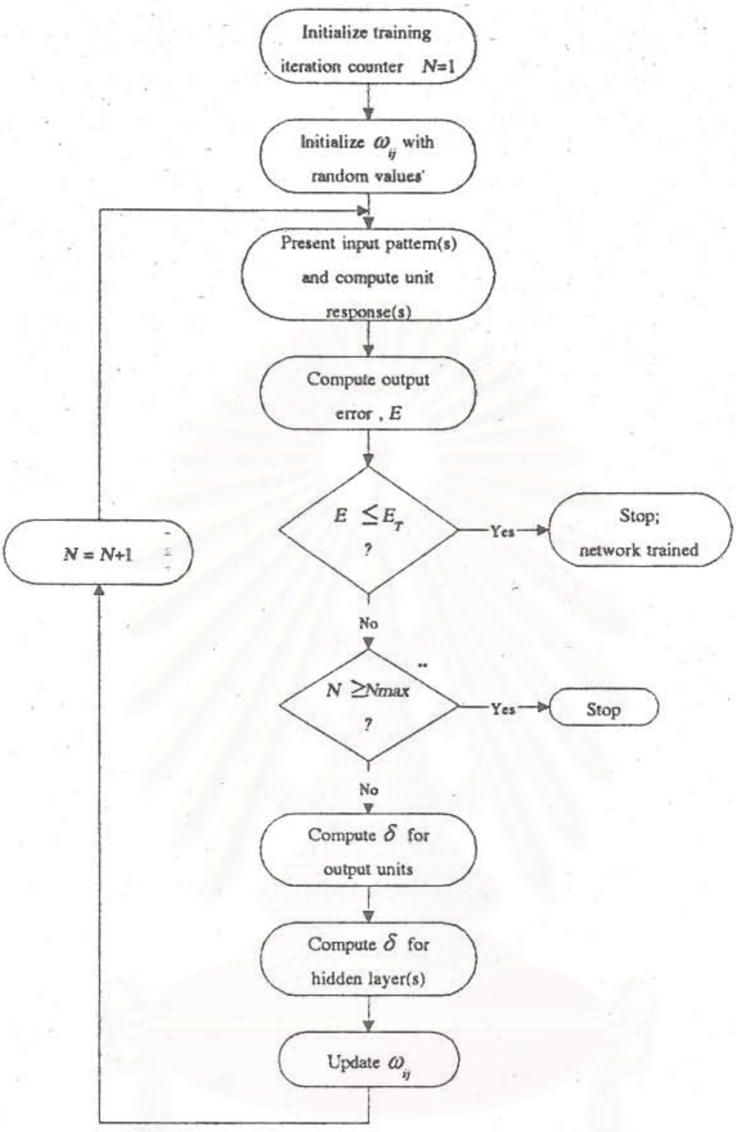
N_{max} แทนจำนวนรอบสูงสุดที่ใช้ในการเรียนรู้

E แทนค่า output error

E_T แทนค่าระดับ output error ที่ต้องการเมื่อ E น้อยกว่าค่านี้ให้ยุติการฝึก

δ แทนค่าความไว (sensitivity) ของ pattern error เทียบกับ net activation

ค่า E_T ในรูปที่ 3.6 เป็นค่าระดับความผิดพลาดที่ต้องการ ซึ่งจะใช้เป็นเงื่อนไขในการหยุดการฝึก (training) ถ้าค่า E_T มีค่ามากเกินไปนิวรอลเน็ตเวิร์กจะไม่สามารถแบ่งแยกข้อมูลแต่ละกลุ่มได้ ถ้าค่า E_T มีค่าน้อยมากนิวรอลเน็ตเวิร์กอาจมีการจำลักษณะเฉพาะของเสียงในข้อมูลฝึกมากเกินไปและจะใช้เวลาในการฝึกนาน ค่าตัวแปรสำคัญในการฝึก (training) นิวรอลเน็ตเวิร์กคือ learning rate และ momentum



รูปที่ 3.4 ขั้นตอนกระบวนการการเรียนรูปแบบ backpropagation

ค่า learning rate กำหนดสัดส่วนในการปรับค่าน้ำหนักการเชื่อมต่อ โดยทั่วไปเป็นค่าสุ่มมีค่าตั้งแต่ 0 ถึง 1 การเลือกค่า learning rate ที่เหมาะสมขึ้นกับคุณลักษณะของพื้นผิวความผิดพลาด (error surface) ถ้าพื้นผิวมีการเปลี่ยนแปลงอย่างรวดเร็ว ควรเลือกค่า learning rate ให้มีค่าน้อย ๆ ถ้าพื้นผิวค่อนข้างราบเรียบควรเลือกค่า learning rate มากขึ้น เพื่อลดเวลาที่ใช้ในการฝึก (training) แต่ถ้ามีค่ามากเกินไปอาจทำให้เกิดการ oscillation และทำให้การลู่เข้าของนิวรอลเน็ตเวิร์กช้าหรือไม่สำเร็จ เนื่องจากคุณลักษณะของพื้นผิวความผิดพลาดเป็นสิ่งที่ไม่ทราบ ดังนั้นกฎเกณฑ์ในการเลือกค่า learning rate คือเลือกค่ามากที่สุดที่ใช้ได้ และไม่ทำให้เกิด oscillation

ค่า momentum เป็นการนำค่าการปรับค่าน้ำหนักการเชื่อมต่อในรอบก่อน ไปใช้ในการคำนวณการปรับค่าน้ำหนักการเชื่อมต่อในรอบปัจจุบัน ค่า momentum อาจช่วยไม่ให้เกิด oscillation และป้องกันการติด local minimum ในขณะฝึก ค่า momentum เหมือนกับค่า learning rate ตรงที่ค่าที่เหมาะสมจะขึ้นกับคุณลักษณะของพื้นผิวความผิดพลาด การใช้ค่า momentum ทำให้สมการสำหรับการปรับค่าน้ำหนักการเชื่อมต่อในรอบที่ n+1 ถูกดัดแปลงเป็น

$$\Delta^p \omega_{ji}(n+1) = \varepsilon(1-\alpha)\delta_j^p \tilde{o}_i^p + \alpha\Delta^p \omega_{ji}(n) \qquad (3.6)$$

ขั้นตอนการให้ค่าเริ่มต้นแก่ค่าน้ำหนักการเชื่อมต่อด้วยการสุ่ม ทำเพื่อป้องกันการสมมาตร (symmetry) ของค่าน้ำหนักการเชื่อมต่อในนิวรอลเน็ตเวิร์ก (Schalkoff,1992) ถ้าหากนิวรอลเน็ตเวิร์กมีค่าเริ่มต้นของน้ำหนักการเชื่อมต่อทุกค่าเท่ากัน การปรับค่าน้ำหนักการเชื่อมต่อของแต่ละโหนดในระดับซ่อนตัว (hidden layer) จะมีค่าเท่ากัน ทำให้นิวรอลเน็ตเวิร์กทำตัวเสมือนมีโหนดในระดับซ่อนตัว (hidden layer) เพียงโหนดเดียว

การเก็บตัวอย่างคู่ข้อมูลอินพุตเอาต์พุต เนื่องจากนิวรอลเน็ตเวิร์กต้องใช้ตัวอย่างข้อมูลในการฝึกจำนวนมาก ทำให้ต้องใช้หน่วยความจำจำนวนมากในการเก็บข้อมูล จึงแก้ปัญหานี้โดยการเก็บข้อมูลในแฟ้มข้อมูล เมื่อต้องการใช้ตัวอย่างคู่ข้อมูลอินพุต เอาต์พุตคู่ใด ก็จะอ่านค่าจากแฟ้มข้อมูลเข้ามาสู่หน่วยความจำของระบบ เพื่อป้อนค่าเข้าสู่นิวรอลเน็ตเวิร์ก

ง) ขั้นตอนการตัดสินใจ

เป็นขั้นตอนที่มีเฉพาะในขั้นตอนการทดสอบระบบการรู้จำคำพูดเท่านั้น กฎเกณฑ์การตัดสินใจที่เลือกใช้คือ เลือกเสียงที่ตรงกับโหนดเอาต์พุตที่มีค่าเอาต์พุตสูงสุด ซึ่งโหนดเอาต์พุตที่มีค่าเอาต์พุตสูงสุดคำนวณได้จากสมการที่ (2.154) จากนั้นนำค่าโหนดเอาต์พุตที่เลือกไปเปิดตารางเพื่อดูว่าโหนดเอาต์พุตที่เลือกตรงกับเสียงคำว่าอะไรในจำนวนเสียงพูดที่ต้องการรู้จำ C คำ

กรรมวิธีนิวรอลเน็ตเวิร์ก ไม่มีการทำกรรมวิธีเน้นล่วงหน้า ในขั้นตอนการประมวลผลสัญญาณเบื้องต้นด้วยเหตุผลเดียวกับของกรรมวิธีไดนามิก ไทม์วาร์ปิง และไม่มีความจำเป็นในการทำกรรมวิธีวางกรอบขนาดสัญญาณ เพราะโครงสร้างเครือข่ายที่เราออกแบบ จะรองรับรายละเอียดหรือลักษณะสำคัญของเสียงพูดทั้งพยางค์

สรุปขั้นตอนของการรู้จำเสียงพูดของ กรรมวิธีไดนามิก ไทม์วาร์ปิง กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ และกรรมวิธีนิวรอลเน็ตเวิร์ก ถูกแสดงเปรียบเทียบกัน ในตารางที่ 3.2 จะเห็นได้ว่า กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ มีขั้นตอนที่สลับซับซ้อนมากที่สุด ทั้งในขั้นตอนของการฝึกฝนระบบการรู้จำ และ ขั้นตอนการทดสอบระบบการรู้จำ ทำให้เป็นกรรมวิธีที่ต้องการเวลาในการประมวลผลในแต่ละขั้นตอนมาก ในขณะที่กรรมวิธีไดนามิก ไทม์วาร์ปิง มีความสลับซับซ้อนของขั้นตอนทั้งสองขั้น ใกล้เคียงกับกรรมวิธีนิวรอลเน็ตเวิร์ก แต่กรรมวิธีนิวรอลเน็ตเวิร์กต้องการเวลาในขั้นตอนการฝึกฝนระบบการรู้จำสูงสุด โดยเฉพาะเมื่อขนาดของนิวรอลเน็ตเวิร์กเพิ่มขึ้น เวลาที่ต้องใช้ในการฝึกฝนจะยิ่งมากขึ้นในอัตราก้าวหน้า แต่ในขั้นตอนการทดสอบระบบการรู้จำ กรรมวิธีนิวรอลเน็ตเวิร์กจะใช้เวลาน้อยที่สุดในทั้งสามกรรมวิธี สำหรับกรรมวิธีไดนามิก ไทม์วาร์ปิง จะใช้เวลาในขั้นตอนของการฝึกฝนระบบการรู้จำน้อยที่สุด และเป็นอันดับสองในขั้นตอนการทดสอบระบบการรู้จำ เนื่องจากต้องใช้เวลาในการหาผลการแปลงฮาร์ตเลย์ ที่น่าสนใจคือทั้งสามกรรมวิธีต้องผ่านการประมวลสัญญาณเบื้องต้นในสองขั้นตอน และขั้นตอนย่อยในการประมวลสัญญาณเบื้องต้นที่มีเหมือน ๆ กัน คือ กรรมวิธีหาจุดสิ้นสุดเสียงพูด และกรรมวิธีวางกรอบสัญญาณ ซึ่งเป็นผลมาจากคุณลักษณะของสัญญาณเสียงพูด ในขณะที่ขั้นตอนย่อยอื่นจะแตกต่างกันไป ขึ้นกับธรรมชาติของแต่ละกรรมวิธี ส่วนขั้นตอนการวิเคราะห์และวัดค่าสำคัญ และขั้นตอนการวัดค่าความคล้ายคลึงกัน ของทั้งสามกรรมวิธีจะมีความแตกต่างกันทั้งหมด เนื่องจากเป็นส่วนของหลักการสำคัญของแต่ละกรรมวิธี

ตารางที่ 3.2 สรุปเชิงเปรียบเทียบขั้นตอนของการรู้จำเสียงพูดของกรรมวิธี ไดนามิก ไทม์วาร์ปิง, แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

ขั้นตอนหลักของการรู้จำ	ไดนามิก ไทม์วาร์ปิง	แบบจำลองฮิดเดน มาร์คอฟ	นิวรอลเน็ตเวิร์ก
การฝึกฝนระบบการรู้จำ	การประมวลสัญญาณเบื้องต้น -หาจุดสิ้นสุดเสียงพูด -วางกรอบขนาดสัญญาณ การวิเคราะห์และวัดค่าสำคัญ -ผลการแปลงฮาร์ตเลย์ -สร้างแบบอ้างอิงตามสมการที่ (3.1)	การสร้างและฝึกฝนชุดรหัส -การประมวลสัญญาณเบื้องต้น -เน้นส่วงหน้า -หาจุดสิ้นสุดเสียงพูด -วางกรอบขนาดสัญญาณ -การสร้างชุดรหัสอ้างอิง -หาสัมประสิทธิ์ LPC -ฝึกฝนชุดรหัสอ้างอิง การสร้างและฝึกฝนชุดพารามิเตอร์แบบจำลองฮิดเดน มาร์คอฟ -วิเคราะห์หาสัมประสิทธิ์ LPC -ควอนไตซ์แบบเวกเตอร์ -ฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ	การประมวลสัญญาณเบื้องต้น -หาจุดสิ้นสุดเสียงพูด -ปรับบรรทัดฐานเชิงเวลา -วางกรอบขนาดสัญญาณ การวิเคราะห์และวัดค่าสำคัญ -หาสัมประสิทธิ์ LPC -สร้างแบบอ้างอิงโดยการฝึกฝนด้วยนิวรอลเน็ตเวิร์ก
การทดสอบระบบการรู้จำ	การประมวลสัญญาณเบื้องต้น -หาจุกสิ้นสุดเสียงพูด -วางกรอบขนาดสัญญาณ การวิเคราะห์และวัดค่าสำคัญ -ผลการแปลงฮาร์ตเลย์ -สร้างแบบทดสอบตามสมการที่ (3.1) การวัดค่าความคล้ายคลึงกัน -วัดหาความคลาดเคลื่อนตามสมการที่ (2.16) ระหว่างแบบทดสอบและแบบอ้างอิง	การประมวลสัญญาณเบื้องต้น -เน้นส่วงหน้า -หาจุดสิ้นสุดเสียงพูด -วางกรอบขนาดสัญญาณ การวิเคราะห์และวัดค่าลักษณะสำคัญ -หาสัมประสิทธิ์ LPC -การควอนไตซ์แบบเวกเตอร์ การจำแนกรูปแบบและการตัดสินใจ -Viterbi algorithm	การประมวลสัญญาณเบื้องต้น -หาจุดสิ้นสุดเสียงพูด -ปรับบรรทัดฐานเชิงเวลา -วางกรอบขนาดสัญญาณ การวิเคราะห์และวัดค่าสำคัญ -หาสัมประสิทธิ์ LPC -ฝึกฝนด้วยนิวรอลเน็ตเวิร์ก การวัดค่าความคล้ายคลึงกัน -วัดหาความใกล้เคียงแบบอ้างอิงสูงสุดจากนิวรอลเน็ตเวิร์ก

3.2 ผลการทดสอบ

การทดสอบระบบการรู้จำค่าผิดของแต่ละกรรมวิธี จะแสดงผลอัตราการรู้จำของแต่ละกลุ่มตัวอย่าง A1, A2, และ B ตามลำดับ และเรียงตามกรรมวิธีตั้งแต่ ไดนามิก ไทม์วาร์ปิง ฮิดเดน มาร์คอฟ และ นิวรอลเน็ตเวิร์ก ทั้งนี้จำนวนตัวอย่างในแต่ละกลุ่มจะแตกต่างกันไปดังตารางที่ 3.1 เนื่องจากต้องทำการปรับแต่งแต่ละกรรมวิธีให้ทำงานได้อย่างมีประสิทธิภาพสูงสุด ตัวอย่างที่ใช้ในกลุ่ม A1 อาจต้องเพิ่มขึ้นหรือลดลงได้ ส่งผลกระทบกับกลุ่ม A2 และ B ด้วย อย่างไรก็ตาม ได้เปรียบเทียบผลลัพธ์ของการรู้จำเมื่อมีการแปรค่าจำนวนกลุ่มตัวอย่าง A1 ที่ใช้เป็นต้นแบบ สำหรับแต่ละกรรมวิธีไว้ในท้ายบทนี้

3.3.1. ผลการทดสอบของกรรมวิธีไดนามิก ไทม์วาร์ปิง

ในตารางที่ 3.2, 3.3, และ 3.4 แสดงตัวอย่างผลการรู้จำของกลุ่มตัวอย่าง A1, A2, และ B ตามลำดับ เมื่อกลุ่มตัวอย่าง A1 มีจำนวนคน 20 คน บันทึกเสียงศูนย์ถึงเก้าคนละ 1 ครั้ง รวม 200 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 20 คำ

นำไปใช้เป็นแบบอ้างอิง และทดสอบดังแสดงในตารางที่ 3.2 กลุ่มตัวอย่าง A2 เป็นกลุ่มคนกลุ่มเดียวกับ A1 แต่ทำการบันทึกเสียงแยกออกจากกลุ่ม A1 โดยบันทึกเสียงศูนย์ถึงเก้าคนละ 2 ครั้ง รวม 400 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 40 คำ ไว้ใช้ทดสอบดังแสดงในตารางที่ 3.3 และกลุ่มตัวอย่าง B เป็นกลุ่มคนอีกกลุ่มหนึ่งมีจำนวนรวม 20 คน ทำการบันทึกเสียงพูดเช่นเดียวกับกลุ่มตัวอย่าง A2 กล่าวคือ บันทึกเสียงศูนย์ถึงเก้าคนละ 2 ครั้ง รวม 400 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 40 คำ

ตารางที่ 3.3 แสดงผลการรู้จำของกลุ่ม A1 จำนวน 20 คน ๆ ละ 1 ชุดเสียง รวม 200 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 20 คน ๆ ละ 1 ชุดเสียง รวม 200 คำ ของกรรมวิธีไดนามิก ไทม์วาร์ปิง

คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (A1)	0	1	2	3	4	5	6	7	8	9
0	19				1					
1	1	18						1		
2			20							
3				18		1				1
4	2				18					
5				2		18				
6	1	1					18			
7		2					1	17		
8				1					18	1
9				2		1				17

ผลการทดสอบจากตารางที่ 3.3 ถึง 3.5 จะได้ว่าอัตราการรู้จำของกลุ่ม A1, A2 และ B เป็น 90.50, 86.50 และ 79.25 ตามลำดับ ความผิดพลาดที่เกิดขึ้นส่วนมากกระจายแบบสุ่มในทุกกลุ่ม ยกเว้นความผิดพลาดในการรู้จำระหว่างเสียง "สาม" และ "เก้า" ที่ควรจะมาจากรูปคลื่นสัญญาณที่คล้ายคลึงกัน แสดงให้เห็นว่า กรรมวิธีไดนามิก ไทม์วาร์ปิง ที่เป็นกรรมวิธีในการเข้าคู่ต้นแบบโดยอาศัยการปรับรูปคลื่นสัญญาณเชิงเวลา ไม่เหมาะสมที่จะนำมาใช้กับกระบวนการรู้จำเสียงพูดภาษาไทย เพราะแม้แต่นำกลุ่มตัวอย่างที่ใช้สร้างแบบอ้างอิงเอง ยังได้ผลการรู้จำที่ผิดพลาดในเกือบทุกเสียง และอัตราการรู้จำโดยรวมของกลุ่มนี้เองมีค่าต่ำ เพื่อยืนยันให้เห็นข้อด้อยของกรรมวิธีนี้ ผู้วิจัยได้ทดลองเปลี่ยนจำนวนแบบอ้างอิงเป็น 5, 10, 15, และ 20 คน จากนั้นทำการทดสอบการรู้จำของแต่ละกลุ่ม เมื่อแปรค่าจำนวนตัวอย่างของแบบอ้างอิง และได้ผลลัพธ์ของค่าอัตราการรู้จำรวมของแต่ละกลุ่มแสดงในตารางที่ 3.6 และแสดงเป็นกราฟในรูปที่ 3.5 ซึ่งชี้ให้เห็นว่ายิ่งเพิ่มจำนวนแบบอ้างอิงอัตราการรู้จำโดยตัวของแบบอ้างอิงเองกลับลดลง คำอธิบายของสาเหตุนี้ อยู่ที่ถ้าการจัดทำแบบอ้างอิงไม่ดีพอ การเพิ่มจำนวนตัวอย่างในการทำแบบอ้างอิง จะเพิ่มความแปรปรวนของแต่ละแบบอ้างอิงมากขึ้นตาม อย่างไรก็ดี การเพิ่มจำนวนแบบอ้างอิงจะเพิ่มอัตราการรู้จำของกลุ่มที่ 2 และ 3 แต่ในอัตราที่น้อยกว่าการลดลงของอัตราการรู้จำของกลุ่มที่ 1

สาเหตุหลักที่คาดว่าทำให้ได้แบบอ้างอิงที่ไม่เหมาะสม คือเสียงพูดภาษาไทยเป็นเสียงดนตรีประการหนึ่ง อีกประการหนึ่งสังคมไทยเราไม่เข้มงวดในเรื่องการออกเสียงที่ถูกต้อง

ตารางที่ 3.4 แสดงผลการรู้จำของกลุ่ม A2 จำนวน 20 คน ๆ ละ 2 ชุดเสียง รวม 400 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 20 คน ๆ ละ 1 ชุดเสียง รวม 200 คำ ของกรรมวิธีไดนามิก ไทม์วาร์ปิง

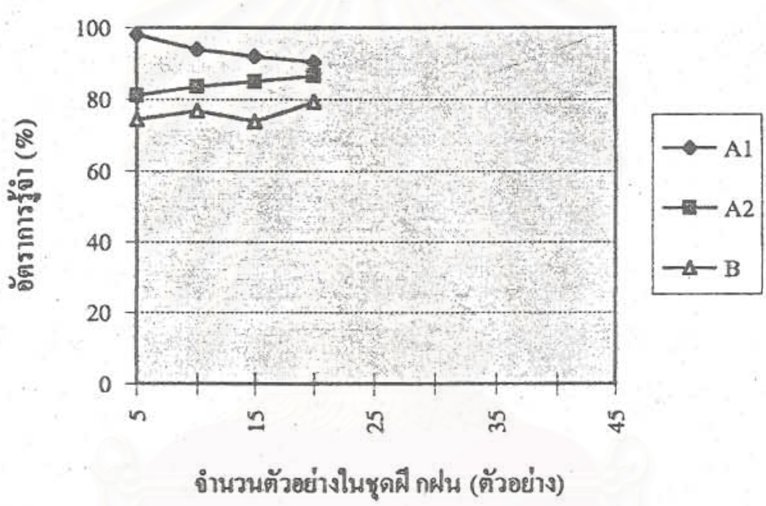
คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (A2)	0	1	2	3	4	5	6	7	8	9
0	33	2			4			1		
1	2	34			1			3		
2			39	1						
3			2	33		3				2
4	8				31		1			
5				3		34			1	2
6		3					37			
7		2					2	36		
8				2		1			36	1
9			1	5					1	33

ตารางที่ 3.5 แสดงผลการรู้จำของกลุ่ม B จำนวน 20 คน ๆ ละ 2 ชุดเสียง รวม 400 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 20 คน ๆ ละ 1 ชุดเสียง รวม 200 คำ ของกรรมวิธีไดนามิก ไทม์วาร์ปิง

คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (B)	0	1	2	3	4	5	6	7	8	9
0	26	1			13					
1		32	1		4		3			
2			36	1			1		2	
3				24		2		1		13
4	1				38	1				
5				3		37				
6	2	3	1				32	2		
7		4					2	34		
8				4		4			31	1
9				11		2				27

ตารางที่ 3.6 แสดงผลการรู้จำของเสียงพูดกลุ่มต่าง ๆ (0 - 9) เมื่อใช้จำนวนแบบอ้างอิงต่างกัน ของกรรมวิธีไดนามิก ไทม์วาร์ปิง

กลุ่ม ทดสอบ	อัตราการรู้จำ (%)			
	จำนวนตัวอย่างที่ใช้ในการฝึกฝนเป็นแบบอ้างอิง			
	5	10	25	30
A1	98.00	94.00	92.00	90.50
A2	81.00	83.50	85.00	86.50
B	74.25	76.25	73.75	79.25



รูปที่ 3.5 อัตราการรู้จำเทียบกับจำนวนตัวอย่างในชุดฝึกฝนของกรรมวิธีไดนามิก ไทม์วาร์ปิง

3.2.2 ผลการทดสอบของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

ในตารางที่ 3.7, 3.8, และ 3.9 แสดงตัวอย่างผลการรู้จำของกลุ่มตัวอย่าง A1, A2, และ B ตามลำดับ เมื่อกลุ่มตัวอย่าง A1 มีจำนวนคน 45 คน ๆ ละ 10 เสียง รวม 450 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 10 คำ นำไปใช้เป็นแบบอ้างอิง และทดสอบดังแสดงในตารางที่ 3.2 กลุ่มตัวอย่าง A2 เป็นกลุ่มคนกลุ่มเดียวกับ A1 แต่ทำการบันทึกเสียงแยกออกจากกลุ่ม A1 โดยบันทึกเสียงศูนย์ถึงเก้าคนละ 1 ครั้ง รวม 450 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 10 คำ ไว้ใช้ทดสอบดังแสดงในตารางที่ 3.3 และกลุ่มตัวอย่าง B เป็นกลุ่มคนอีกกลุ่มหนึ่งมีจำนวนรวม 10 คน ทำการบันทึกเสียงพูดเช่นเดียวกับกลุ่มตัวอย่าง A2 กล่าวคือ เป็นเสียงศูนย์ถึงเก้าเสียงคนละ 1 ครั้ง รวม 100 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 10 คำ

ตารางที่ 3.7 แสดงผลการรู้จำของกลุ่ม A1 จำนวน 45 คน ๆ ละ 1 ชุดเสียง รวม 450 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 45 คน ๆ ละ 1 ชุดเสียง รวม 450 คำ ของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (A1)	0	1	2	3	4	5	6	7	8	9
0	44						1			
1	1	42						1	1	
2	1		43				1			
3				45						
4	1				42			2		
5				2		40				3
6						1	44			
7					1			44		
8						2			41	2
9							1			44

ตารางที่ 3.8 แสดงผลการรู้จำของกลุ่ม A2 จำนวน 45 คน ๆ ละ 1 ชุดเสียง รวม 450 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 45 คน ๆ ละ 1 ชุดเสียง รวม 450 คำ ของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (A2)	0	1	2	3	4	5	6	7	8	9
0	41		2		1					
1	2	39		1	2			1		
2	2		40				2			1
3			1	41		2				1
4	2	1			38			4		
5				2		37			3	3
6			1				44			
7		1			1	1	2	39	1	
8						3			41	1
9						1				44

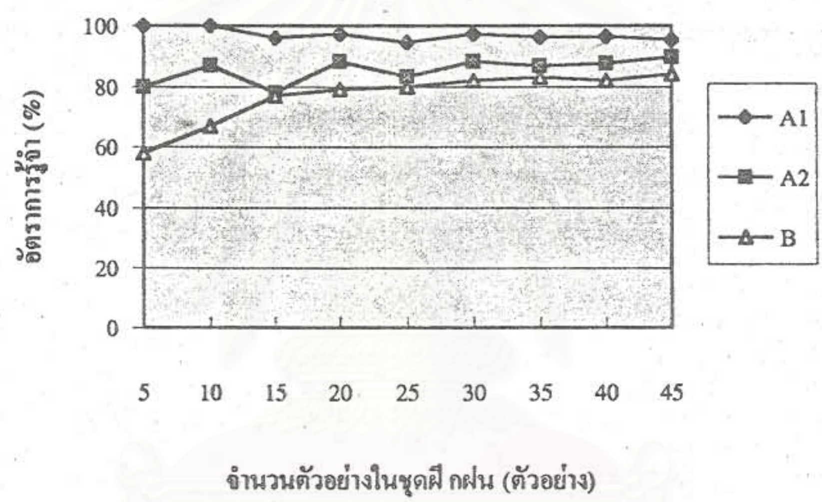
ตารางที่ 3.9 แสดงผลการรู้จำของกลุ่ม B จำนวน 10 คน ๆ ละ 1 ชุดเสียง รวม 100 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 45 คน ๆ ละ 1 ชุดเสียง รวม 450 คำ ของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

คำในกลุ่มทดสอบ (B)	คำที่วิเคราะห์ได้									
	0	1	2	3	4	5	6	7	8	9
0	9		1							
1	1	8			1					
2	3		7							
3				8		1			1	
4	1				9					
5		1				8				1
6							9			1
7				1				9		
8				1					8	1
9	1									9

ผลการทดสอบจากตารางที่ 3.7 ถึง 3.9 จะได้ว่าอัตราการรู้จำของกลุ่ม A1, A2 และ B เป็น 95.30, 89.70 และ 84.00 ตามลำดับ ซึ่งให้ค่าที่ดีกว่าของกรรมวิธีไดนามิก ไทม์วาร์ปิง แต่ใช้ตัวอย่างทำแบบอ้างอิงมากกว่า ความผิดพลาดที่เกิดขึ้นกระจายแบบสุ่มในทุกกลุ่ม ผู้วิจัยได้ทดลองเปลี่ยนจำนวนแบบอ้างอิงเป็น 5, 10, 15, 20, 25, 30, 35, 40 และ 45 คน จากนั้นทำการทดสอบการรู้จำของแต่ละกลุ่ม และได้ผลลัพธ์ของค่าอัตราการรู้จำรวมของแต่ละกลุ่มแสดงในตารางที่ 3.10 และแสดงเป็นกราฟในรูปที่ 3.6 ที่มีแนวโน้มคล้ายคลึงกับกรรมวิธีไดนามิก ไทม์วาร์ปิง แต่ในสภาพที่ดีกว่า กล่าวคือ อัตราการรู้จำของกลุ่มที่ 1 ลดลงช้า ๆ ในขณะที่ของกลุ่มที่ 2 และ 3 มีอัตราการรู้จำเพิ่มขึ้นอย่างมีนัยสำคัญ (ด้วยอัตราการเพิ่มที่สูงกว่าอัตราการลดลงของกลุ่มที่ 1) โดยเฉพาะกลุ่มที่ 3 ที่เพิ่มสูงมากในช่วงต้นของกราฟ อย่างไรก็ตาม อัตราการเพิ่มของทั้ง 2 กลุ่มจะลดลง แต่คาดว่าถ้ามีการกำหนดกลุ่มตัวอย่างเพื่อทำแบบอ้างอิงที่ดีพอ การเพิ่มจำนวนตัวอย่าง ไม่ควรจะทำให้อัตราการรู้จำของกลุ่มที่ 1 ลดลงมากนัก และจะทำให้อัตราการรู้จำของกลุ่มที่ 2 และ 3 เพิ่มสูงขึ้นต่อไปได้ ซึ่งชี้ให้เห็นว่ากรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ควรจะเป็นกรรมวิธีที่มีความเหมาะสมที่จะนำไปพัฒนาเป็นกรรมวิธีหลักกรรมวิธีหนึ่งของกระบวนการรู้จำเสียงพูดภาษาไทยได้

ตารางที่ 3.10 แสดงผลการรู้จำของเสียงพูดกลุ่มต่าง ๆ (0 - 9) เมื่อใช้จำนวนแบบอ้างอิงต่างกันของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

กลุ่มทดสอบ	อัตราการรู้จำ (%)								
	จำนวนตัวอย่างที่ใช้ในการฝึกฝนเป็นแบบอ้างอิง								
	5	10	15	20	25	30	35	40	45
A1	100	100	96.00	97.30	94.50	97.30	96.28	96.50	95.30
A2	80.00	87.00	78.00	88.30	83.20	88.30	86.85	87.75	89.70
B	58.00	67.00	77.00	79.00	80.00	82.00	83.00	82.00	84.00



รูปที่ 3.6 อัตราการรู้จำเทียบกับจำนวนตัวอย่างในชุดฝึกฝนของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ

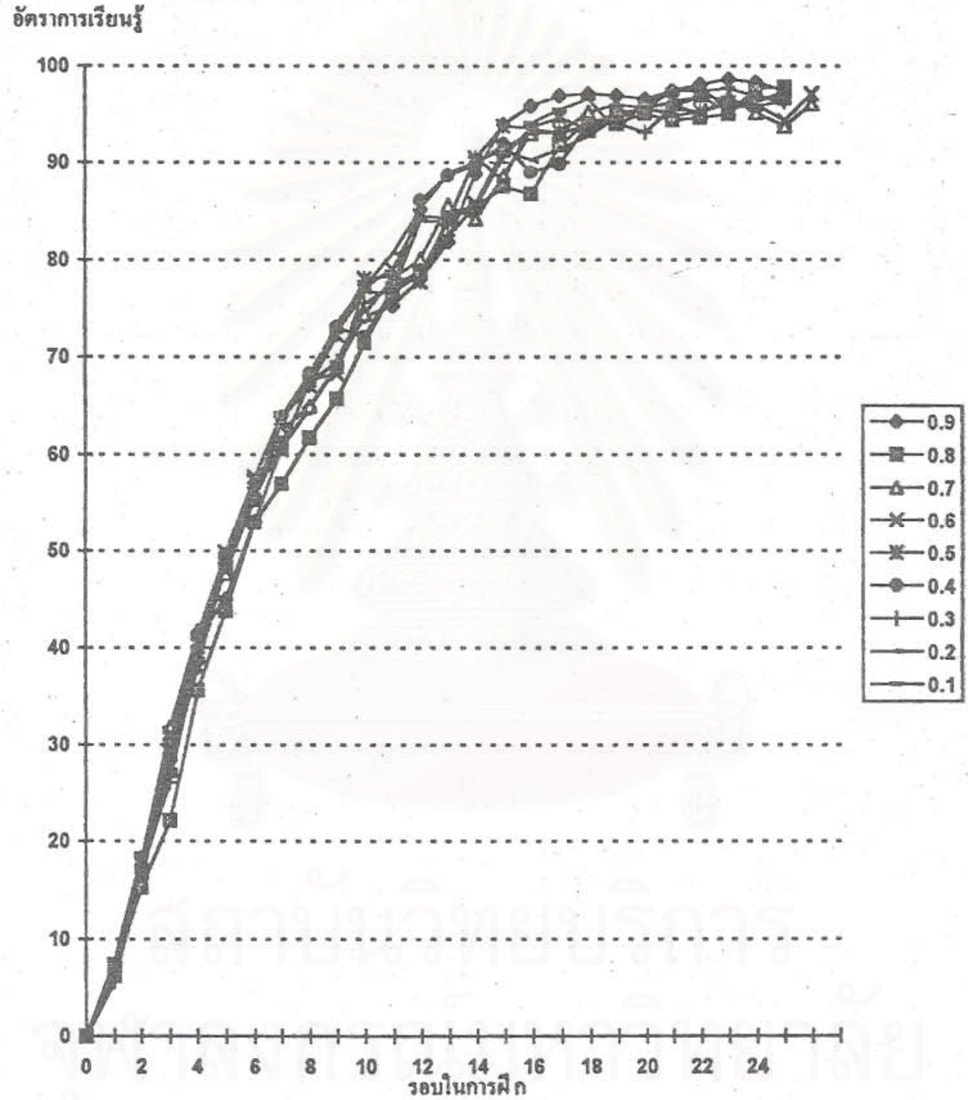
3.2.3. ผลการทดสอบของกรรมวิธีนิวรอลเน็ตเวิร์ก

เนื่องจากค่า learning rate และ momentum เป็นค่าที่ขึ้นกับคุณลักษณะของพื้นผิวความผิดพลาด จึงต้องทำการทดลองเพื่อหาค่าที่เหมาะสม และเนื่องจากค่าทั้งสองมีผลกับเวลาที่ใช้ในการฝึกนิวรอลเน็ตเวิร์กจึงทำการทดลองเพื่อหาค่าที่ทำให้ใช้เวลาในการฝึกน้อยที่สุด จากสมการที่ 3.6 พบว่าค่า momentum กำหนดสัดส่วนของค่า learning rate ในการปรับค่าน้ำหนักการเชื่อมต่อด้วย จึงทำการทดลองโดยกำหนดให้ค่า learning rate มีค่าคงที่เท่ากับ 1 เพื่อจำกัดขอบเขตการทดลอง อีกเหตุผลหนึ่งที่เลือกใช้ค่านี้คือถ้า momentum มีค่าเท่ากับ 0 ค่า learning rate เท่ากับ 1 จะทำให้สัดส่วนการปรับค่าน้ำหนักการเชื่อมต่อมีค่ามากและใช้เวลาในการฝึกน้อย การปรับค่า momentum มีผลกับเวลาที่ใช้ในการฝึกน้อยมาก (วุฒิพงษ์ พรสุขจันทรา, 2539) จึงนำอัตราการเรียนรู้ในแต่ละรอบของการฝึก มาช่วยพิจารณา ดังแสดงในรูปที่ 3.7

จากรูปที่ 3.7 พบว่าค่า momentum ที่เหมาะสมคือ 0.9 เพราะให้อัตราการเรียนรู้ที่สูงกว่าค่าอื่น และในช่วงท้ายของการฝึกให้อัตราการเรียนรู้ที่ค่อนข้างสม่ำเสมอ

จำนวนโหนดในระดับข้อมูลเข้า ถูกกำหนดโดยขนาดของข้อมูลอินพุตที่ใช้แทนเสียงพูด 1 คำ มีจำนวนโหนด 330 โหนดซึ่งตรงกับจำนวนกรอบในเสียงพูด 1 คำ คือ 33 กรอบ และ 1 กรอบประกอบด้วยสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น 10 ค่า

จำนวนโหนดในระดับข้อมูลออก ถูกกำหนดโดยจำนวนกลุ่มข้อมูลที่ต้องการแบ่งแยก มีจำนวนโหนด 10 โหนดสำหรับนิวรอลเน็ตเวิร์กชุดแรกและมีจำนวนโหนด 12 โหนดสำหรับนิวรอลเน็ตเวิร์กชุดที่สอง



รูปที่ 3.7 อัตราการเรียนรู้ในแต่ละรอบของการฝึกฝนของกรรมวิธีนิวรอลเน็ตเวิร์ก

การทดลองเพื่อหาจำนวนโหนดในระดับซ่อนตัว เนื่องจากจำนวนโหนดในระดับซ่อนตัวกำหนดความสามารถในการเรียนรู้ของนิวรอลเน็ตเวิร์ก จึงออกแบบการทดลองโดยฝึกนิวรอลเน็ตเวิร์กที่มีจำนวนโหนดในระดับซ่อนตัวหลาย ๆ ค่า และดูผลอัตราความถูกต้องในการรู้จำเสียงของเสียงพูดในกลุ่ม B โดยที่กำหนดให้ค่า learning rate และ momentum มีค่าคงที่เท่ากับ 1 และ 0.9 ตามลำดับ เพื่อหาจำนวนโหนดที่ไม่มากเกินไปและมีความสามารถในการรู้จำเสียงได้ จากผลการทดลองที่ได้ในตารางที่ 3.13 พบว่าอัตราความถูกต้องในการรู้จำเสียงมีแนวโน้มที่สูงขึ้นเมื่อเพิ่มจำนวนโหนดในระดับซ่อนตัวขึ้น

และจำนวนโหนดที่เหมาะสมคือ 50 โหนดเพราะใช้หน่วยความจำไม่มากเกินไป และให้ผลอัตราความถูกต้องในการรู้จำเสียงค่อนข้างสูง

ตารางที่ 3.11 ความสัมพันธ์ของจำนวนโหนดในระดับซ่อนตัวและอัตราการรู้จำ

จำนวนโหนดในระดับซ่อนตัว	อัตราการรู้จำเสียง (%)
30	87.5
40	85.6
50	88.9
60	90.6
70	90.6
80	91.7
90	91.7

ตารางที่ 3.12 แสดงผลการรู้จำของกลุ่ม A1 จำนวน 30 คน ๆ ละ 2 ชุดเสียง รวม 600 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 30 คน ๆ ละ 2 ชุดเสียง รวม 600 คำ ของกรรมวิธีนิวรอลเน็ตเวิร์ก

คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (A1)	0	1	2	3	4	5	6	7	8	9
0	59	1	0	0	0	0	0	0	0	0
1	0	60	0	0	0	0	0	0	0	0
2	1	0	56	2	0	0	0	1	0	0
3	0	0	0	59	0	0	0	0	1	0
4	0	1	0	0	58	0	1	0	0	0
5	1	0	0	0	0	58	0	0	1	0
6	0	0	0	0	0	0	60	0	0	0
7	0	0	0	0	0	0	0	60	0	0
8	0	0	0	0	0	0	0	0	60	0
9	0	0	0	1	0	0	0	0	0	59

ในตารางที่ 3.12, 3.13, และ 3.14 แสดงตัวอย่างผลการรู้จำของกลุ่มตัวอย่าง A1, A2, และ B ตามลำดับ เมื่อกลุ่มตัวอย่าง A1 มีจำนวนคน 30 คน บันทึกเสียงศูนย์ถึงเก้าคนละ 2 ครั้ง รวม 600 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 20 คำ นำไปใช้เป็นแบบอ้างอิง และทดสอบดังแสดงในตารางที่ 3.2 กลุ่มตัวอย่าง A2 เป็นกลุ่มคนกลุ่มเดียวกับ A1 แต่ทำการบันทึกเสียงแยกออกจากกลุ่ม A1 โดยบันทึกเสียงศูนย์ถึงเก้าคนละ 2 ครั้ง รวม 600 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 20 คำ

ไว้ใช้ทดสอบดังแสดงในตารางที่ 3.3 และกลุ่มตัวอย่าง B เป็นกลุ่มคนอีกกลุ่มหนึ่งมีจำนวนรวม 12 คน ทำการบันทึกเสียงพูดเช่นเดียวกับกลุ่มตัวอย่าง A2 กล่าวคือ เป็นเสียงศูนย์ถึงเก้าเสียงคนละ 3 ครั้ง รวม 360 คำ เป็นเสียงศูนย์ถึงเก้าเสียงละ 36 คำ

ตารางที่ 3.13 แสดงผลการรู้จำของกลุ่ม A2 จำนวน 30 คน ๆ ละ 1 ชุดเสียง รวม 300 คำ และแบบอ้างอิงสร้างจากกลุ่ม A1 จำนวน 30 คน ๆ ละ 2 ชุดเสียง รวม 600 คำ ของกรรมวิธีนิวรอลเน็ตเวิร์ก

คำในกลุ่ม	คำที่วิเคราะห์ได้									
ทดสอบ (A2)	0	1	2	3	4	5	6	7	8	9
0	25	1	3	0	0	1	0	0	0	0
1	0	27	0	0	2	0	1	0	0	0
2	5	0	17	4	0	1	1	0	0	2
3	1	1	2	24	0	1	1	0	0	0
4	0	1	0	0	26	0	2	1	0	0
5	0	0	1	1	1	25	0	0	1	1
6	0	0	0	0	0	0	30	0	0	0
7	0	0	0	0	1	1	2	26	0	0
8	0	0	0	1	0	1	2	0	24	2
9	0	0	0	0	0	1	0	0	0	29

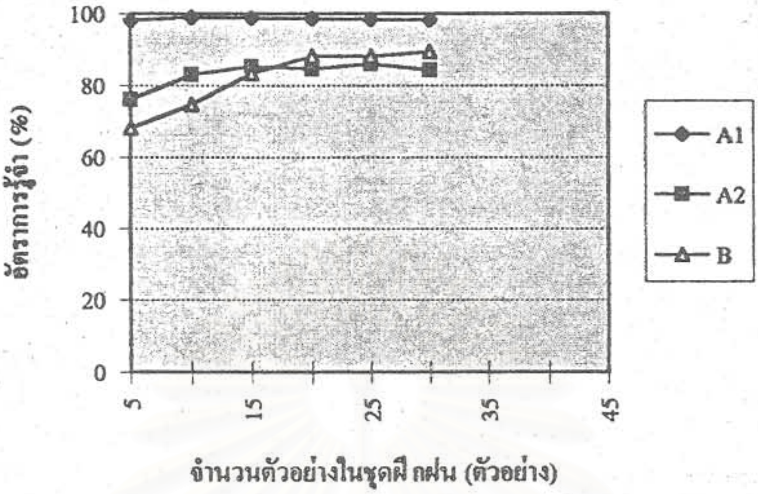
ผลการทดสอบจากตารางที่ 3.12 ถึง 3.14 จะได้ว่าอัตราการรู้จำของกลุ่ม A1, A2 และ B เป็น 98.20, 84.30 และ 89.40 ตามลำดับ ซึ่งให้ค่าที่ดีกว่าทั้งของกรรมวิธีไดนามิก ไทม์วาร์ปิง และกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ในกลุ่มที่ 1 และ 3 แต่กลุ่มที่ 3 กลับให้ผลลัพธ์ที่ต่ำกว่า ทั้งนี้อาจเป็นผลมาจากผู้วิจัยพยายามเลือกเสียงพูดที่มีสัญญาณรบกวนน้อยและมีรูปคลื่นสม่ำเสมอเพื่อใช้เป็นเสียงที่ใช้ในการฝึกฝนของกลุ่มที่ 1 เสียงที่เหลือของผู้พูดในชุดฝึกฝน จะถูกนำกลับไปใช้ในกลุ่มที่ 2 ทำให้เสียงในกลุ่มที่ 2 เป็นเสียงที่มีคุณภาพไม่ดี ผู้วิจัยได้ทดลองเปลี่ยนจำนวนแบบอ้างอิงเป็น 5, 10, 15, 20, 25 และ 30 คน จากนั้นทำการทดสอบการรู้จำของแต่ละกลุ่ม และได้ผลลัพธ์ของค่าอัตราการรู้จำรวมของแต่ละกลุ่มแสดงในตารางที่ 3.15 และแสดงเป็นกราฟในรูปที่ 3.8 ที่มีแนวโน้มคล้ายคลึงกับกรรมวิธีไดนามิก ไทม์วาร์ปิง และกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ แต่มีสภาพที่ดีกว่าทั้ง 2 กลุ่มแรก กล่าวคือ อัตราการรู้จำของกลุ่มที่ 1 ลดลงน้อยมากจนถือได้ว่าคงที่ ในขณะที่ของกลุ่มที่ 2 และ 3 มีอัตราการรู้จำเพิ่มขึ้นอย่างมีนัยสำคัญ (ด้วยอัตราการเพิ่มที่สูงกว่าอัตราการลดลงของกลุ่มที่ 1) โดยเฉพาะกลุ่มที่ 3 ที่เพิ่มสูงมากในช่วงต้นของกราฟ เช่นเดียวกับของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ อย่างไรก็ตาม อัตราการเพิ่มของทั้ง 2 กลุ่มจะลดลงเช่นเดียวกับ 2 กรรมวิธีแรก แต่ถ้ามีการกำหนดกลุ่มตัวอย่างเพื่อทำแบบอ้างอิงที่ดีพอ การเพิ่มจำนวนตัวอย่าง ควรจะทำให้การลดลงของอัตราการรู้จำของกลุ่มที่ 1 น้อยมาก และจะทำให้อัตราการรู้จำของกลุ่มที่ 2 และ 3 เพิ่มสูงขึ้น ได้มากกว่า 2 กรรมวิธีแรก ซึ่งชี้ให้เห็นว่ากรรมวิธีแบบนิวรอลเน็ตเวิร์ก ควรจะเป็นอีกกรรมวิธีหนึ่งที่มีความเหมาะสมที่จะนำไปพัฒนาเป็น กรรมวิธีหลักของกระบวนการรู้จำเสียงพูดภาษาไทยได้ เช่นเดียวกับกรรมวิธีฮิดเดน มาร์คอฟ

ตารางที่ 3.14 แสดงผลการรู้จำของกลุ่ม B จำนวน 12 คน ๆ ละ 3 ชุดเสียง รวม 360 คำ และแบบอ้างอิงสร้างจากผู้พูด A1 จำนวน 30 คน ๆ ละ 2 ชุดเสียง รวม 600 คำ ของกรรมวิธีนิวรอลเน็ตเวิร์ก

คำในกลุ่ม ทดสอบ (B)	คำที่วิเคราะห์ได้									
	0	1	2	3	4	5	6	7	8	9
0	32	1	1	1	0	0	1	0	0	0
1	0	36	0	0	0	0	0	0	0	0
2	5	0	27	3	0	0	0	0	0	1
3	2	0	1	33	0	0	0	0	0	0
4	0	0	0	0	36	0	0	0	0	0
5	1	2	0	5	0	28	0	0	0	0
6	0	0	0	0	1	0	34	0	0	1
7	0	0	0	0	0	0	1	34	1	0
8	0	0	0	2	0	2	0	0	31	1
9	0	0	2	1	0	2	0	0	0	31

ตารางที่ 3.15 แสดงผลการรู้จำของเสียงพูดกลุ่มต่าง ๆ (0 - 9) เมื่อใช้จำนวนแบบอ้างอิงต่างกัน ของกรรมวิธีนิวรอลเน็ตเวิร์ก

กลุ่ม ทดสอบ	อัตราการรู้จำ (%)					
	จำนวนตัวอย่างที่ใช้ในการฝึกฝนเป็นแบบอ้างอิง					
	5	10	15	20	25	30
A1	98.00	99.00	98.70	98.50	98.40	98.20
A2	76.00	83.00	85.30	84.50	86.00	84.30
B	68.10	74.70	83.30	88.10	88.10	89.40



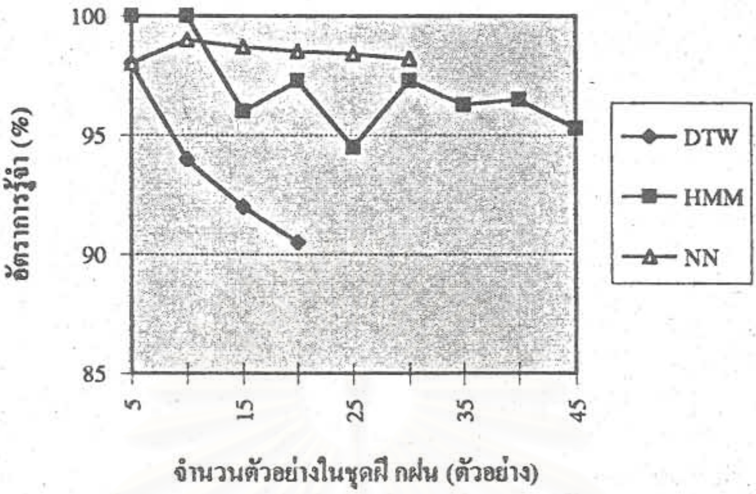
รูปที่ 3.8 อัตราการรู้จำเทียบกับจำนวนตัวอย่างในชุดฝึกฝนของกรรมวิธีนิวรอลเน็ตเวิร์ก

เพื่อให้เห็นถึงสมรรถนะทางด้านการรู้จำของแต่ละกรรมวิธี ได้ชัดเจนยิ่งขึ้น จึงนำเอาผลลัพธ์จากตารางที่ 3.6, 3.10 และ 3.15 มาเปรียบเทียบอัตราการรู้จำในแต่ละกลุ่ม A1, A2 และ B ของแตะกรรมวิธี ในตารางที่ 3.16, 3.17 และ 3.18 และแสดงเป็นกราฟในรูปที่ 3.9, 3.10 และ 3.11 ตามลำดับ

จากตารางที่ 3.16 และกราฟในรูปที่ 3.9 จะเห็นจุดอ่อนของกรรมวิธีไดนามิก ไทม์วาร์ปิง อย่างชัดเจนว่า ยิ่งเพิ่มจำนวนตัวอย่างของแบบอ้างอิง อัตราการรู้จำด้วยกลุ่มตัวเองลดลงอย่างมาก ในขณะที่กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ มีอัตราการรู้จำของกลุ่มตัวเองสูงถึงร้อยละ 100 ที่จำนวนตัวอย่างของแบบอ้างอิงน้อย ๆ (ความแปรปรวนน้อย) และจะลดลงเมื่อจำนวนตัวอย่างเพิ่มขึ้น ส่วนกรรมวิธีนิวรอลเน็ตเวิร์กมีการเปลี่ยนแปลงของอัตราการู้จำของกลุ่มตัวเองน้อยมากจนถือได้ว่าคงที่

ตารางที่ 3.16 การเปรียบเทียบอัตราการรู้จำของกลุ่ม A1 เมื่อแปรจำนวนแบบอ้างอิงของกลุ่ม A1 ระหว่างกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

กลุ่มทดสอบ	อัตราการรู้จำ (%)								
	จำนวนตัวอย่างที่ใช้ในการฝึกฝนเป็นแบบอ้างอิง								
	5	10	15	20	25	30	35	40	45
DTW	98.00	94.00	92.00	90.50					
HMM	100	100	96.00	97.30	94.50	97.30	96.28	96.50	95.30
NN	98.00	99.00	98.70	98.50	98.40	98.20			

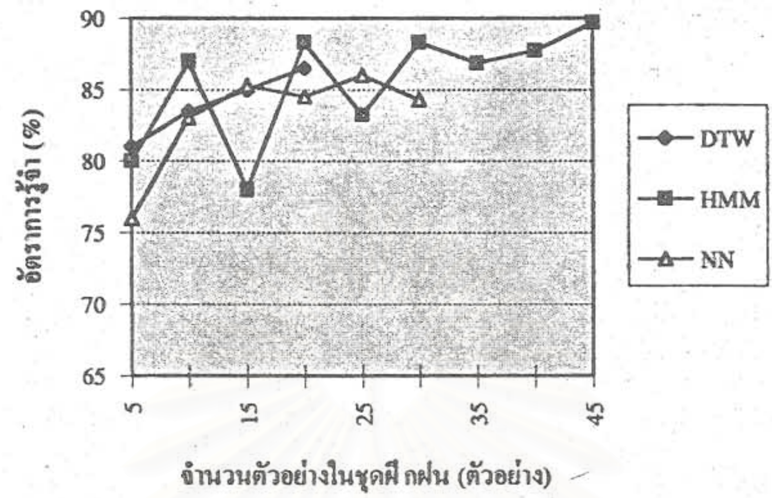


รูปที่ 3.9 กราฟแสดงการเปรียบเทียบอัตราการรู้จำของกลุ่ม A1 เมื่อแปรจำนวนแบบอ้างอิงของกลุ่ม A1 ระหว่างกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

จากตารางที่ 3.17 และกราฟในรูปที่ 3.10 อัตราการรู้จำของกลุ่มที่ 2 ที่ใช้กรรมวิธีนิวรอลเน็ตเวิร์กให้ผลลัพธ์โดยรวมต่ำสุด ด้วยเหตุผลที่ได้ชี้แจงตั้งแต่ต้นแล้ว ในขณะที่ของ 2 กรรมวิธีที่เหลือมีผลลัพธ์ที่ใกล้เคียงกัน แต่ของกรรมวิธีไดนามิก ไทม์วาร์ปิง จะพบกับขีดจำกัดจากอัตราการลดลงในการรู้จำของกลุ่มที่ 1 ที่จะเป็นเพดานกำหนดอัตราการรู้จำของกลุ่มที่ 2 ไว้ หมายความว่ากรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ควรจะให้ผลลัพธ์ที่ดีที่สุด แต่ต้องไม่ลืมว่าอัตราการลดลงในการรู้จำของกลุ่มที่ 1 ของกรรมวิธีนี้ก็จะเป็นตัวกำหนดเพดานไว้เช่นกัน เพียงแต่เพดานที่ได้จากกรรมวิธีนี้สูงกว่าของกรรมวิธีไดนามิก ไทม์วาร์ปิง แต่ต่ำกว่าของแบบจำลองนิวรอลเน็ตเวิร์ก ดังนั้นถ้ากลุ่มที่ 2 ที่ใช้ในกรรมวิธีนิวรอลเน็ตเวิร์กมีคุณภาพของข้อมูลดีพอ ผลลัพธ์ที่ได้จากกลุ่มนี้น่าจะดีเท่าเทียมหรือดีกว่าของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ได้

ตารางที่ 3.17 การเปรียบเทียบอัตราการรู้จำของกลุ่ม A2 เมื่อแปรจำนวนแบบอ้างอิงของกลุ่ม A1 ระหว่างกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

กลุ่มทดสอบ	อัตราการรู้จำ (%)								
	จำนวนตัวอย่างที่ใช้ในการฝึกฝนเป็นแบบอ้างอิง								
	5	10	15	20	25	30	35	40	45
DTW	81.00	83.50	85.00	86.50					
HMM	80.00	87.00	78.00	88.30	83.20	88.30	86.85	87.75	89.70
NN	76.00	83.00	85.30	84.50	86.00	84.30			

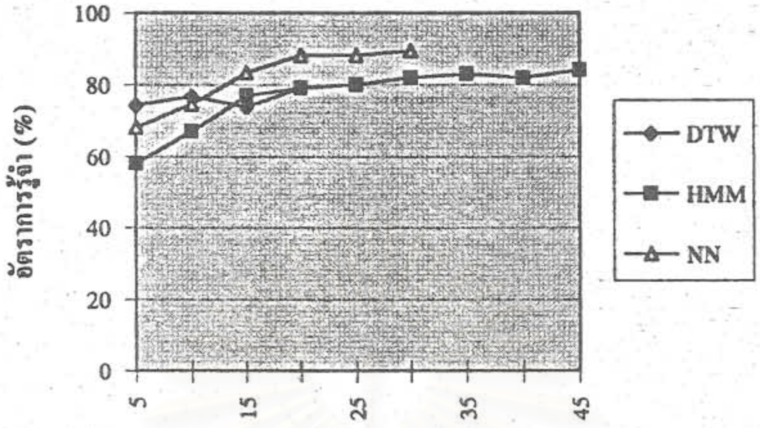


รูปที่ 3.10 กราฟแสดงการเปรียบเทียบอัตราการรู้จำของกลุ่ม A2 เมื่อแปรจำนวนแบบอ้างอิงของกลุ่ม A1 ระหว่างกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

จากตารางที่ 3.18 และกราฟในรูปที่ 3.11 อัตราการรู้จำของกลุ่มที่ 3 ที่ใช้กรรมวิธีนิวรอลเน็ตเวิร์กให้ผลลัพธ์โดยรวมสูงสุด ด้วยเหตุผลที่ได้ชี้แจงตั้งแต่ต้นแล้ว ในขณะที่ของ 2 กรรมวิธีที่เหลือมีผลลัพธ์ที่ใกล้เคียงกัน แต่ของกรรมวิธีไดนามิก ไทม์วาร์ปิง จะพบกับขีดจำกัดจากอัตราการลดลงในการรู้จำของกลุ่มที่ 1 ที่จะเป็นเพดานกำหนดอัตราการรู้จำของกลุ่มที่ 2 ไว้ หมายความว่ากรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ควรจะให้ผลลัพธ์ที่ดีที่สุด แต่ต้องไม่ลืมว่าอัตราการลดลงในการรู้จำของกลุ่มที่ 1 ของกรรมวิธีนี้ก็จะเป็นตัวกำหนดเพดานไว้เช่นกัน เพียงแต่เพดานที่ได้จากกรรมวิธีนี้สูงกว่าของกรรมวิธีไดนามิก ไทม์วาร์ปิง แต่ต่ำกว่าของแบบจำลองนิวรอลเน็ตเวิร์ก ดังนั้นถ้ากลุ่มที่ 2 ที่ใช้ในกรรมวิธีนิวรอลเน็ตเวิร์กมีคุณภาพของข้อมูลดีพอ ผลลัพธ์ที่ได้จากกลุ่มนี้น่าจะดีเท่าเทียมหรือดีกว่าของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ได้

ตารางที่ 3.18 การเปรียบเทียบอัตราการรู้จำของกลุ่ม B เมื่อแปรจำนวนแบบอ้างอิงของกลุ่ม A1 ระหว่างกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

กลุ่ม ทดสอบ	อัตราการรู้จำ (%) จำนวนตัวอย่างที่ใช้ในการฝึกฝนเป็นแบบอ้างอิง								
	5	10	15	20	25	30	35	40	45
DTW	74.25	76,75	73.75	79.25					
HMM	58.00	67.00	77.00	79.00	80.00	82.00	83.00	82.00	84.00
NN	68.10	74.70	83.30	88.10	88.10	89.40			



รูปที่ 3.11 กราฟแสดงการเปรียบเทียบอัตราการรู้จำของกลุ่ม B เมื่อแปรจำนวนแบบอ้างอิงของกลุ่ม A1 ระหว่างกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก

จากผลการทดสอบ จะเห็นได้ว่าแบบอ้างอิงของกลุ่ม A1 ที่ใช้ในกรรมวิธีรู้จำแบบต่างๆ มีจำนวนไม่เท่ากัน เนื่องจากแต่ละกรรมวิธีมีข้อจำกัดที่แตกต่างกัน เช่น กรรมวิธีไดนามิก ไทม์วาร์ปิง เมื่อเพิ่มจำนวนแบบอ้างอิงตั้งแต่ 15 คนขึ้นไป อัตราการรู้จำของกลุ่ม A2 จะเพิ่มในอัตราที่ลดลงจนเกือบมีค่าคงที่ ในขณะที่กรรมวิธี แบบจำลองฮิดเดน มาร์คอฟ นั้น อัตราการรู้จำของกลุ่ม A2 ยังเพิ่มขึ้นต่อไปตามการเพิ่มขึ้นของจำนวนแบบอ้างอิงของกลุ่ม A1 จนกระทั่งถึง 30 คน ที่อัตราการรู้จำมีแนวโน้มคงที่ ส่วนกรรมวิธีนิวรอลเน็ตเวิร์ก อัตราการรู้จำของกลุ่ม A2 จะมีการแกว่งค่าในลักษณะเพิ่มขึ้นช้ามาก ๆ ตั้งแต่จำนวนแบบอ้างอิงของกลุ่ม A1 เป็น 15 คน อย่างไรก็ตาม เมื่อพิจารณาอัตราการรู้จำของกลุ่ม B จะเห็นได้ชัดเจนยิ่งขึ้นว่า จำนวนแบบอ้างอิงของกลุ่ม A1 ที่มีความเหมาะสมสำหรับกรรมวิธี ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก คือ 20 45 และ 30 คน ตามลำดับ

เมื่อพิจารณาโดยรวมแล้ว ในการรู้จำเสียงตัวเลขภาษาไทย กรรมวิธีนิวรอลเน็ตเวิร์ก ควรเป็นกรรมวิธีที่มีสมรรถนะในการรู้จำสูงสุด ทั้งในด้านอัตราการรู้จำและความเร็วในการรู้จำ แต่ต้องใช้เวลาในการฝึกฝนสูงสุด โดยมีกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ มีสมรรถนะใกล้เคียงกัน ในเรื่องอัตราการรู้จำ โดยที่ความเร็วในการรู้จำต่ำสุด และเวลาที่ต้องใช้ในการฝึกฝนสูงเป็นอันดับสอง ในขณะที่กรรมวิธีไดนามิก ไทม์วาร์ปิง ให้สมรรถนะต่ำสุด ในเรื่องอัตราการรู้จำ แต่ความเร็วในการรู้จำเป็นอันดับสอง และเวลาที่ใช้ในการฝึกฝนเป็นอันดับหนึ่ง อย่างไรก็ตาม ได้มีการปรับแต่งแบบจำลองฮิดเดน มาร์คอฟ รวมถึงค่าพารามิเตอร์ต่าง ๆ ในขั้นตอนการหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น และการควอนไตซ์แบบเวกเตอร์ [Ahkuputra, V. et al., 1997] ทำให้อัตราการรู้จำเสียงตัวเลขภาษาไทยตามกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ สูงขึ้นเทียบเท่ากับที่ได้จากกรรมวิธีนิวรอล เน็ตเวิร์ก คือประมาณร้อยละ 89 และคาดว่าถ้าทำการวิเคราะห์และวัดค่าลักษณะสำคัญได้ชัดเจนขึ้น อัตราการรู้จำทั้งของแบบจำลองฮิดเดน มาร์คอฟ และแบบนิวรอล เน็ตเวิร์ก ควรจะมีค่าสูงขึ้นได้อีก

บทที่ 4

สรุปผลการวิจัยและข้อเสนอแนะ

4.1. สรุปผลการวิจัย

งานวิจัยนี้เป็นการศึกษาระบบการรู้จำเสียงพูดภาษาไทยที่เป็นคำโดดแบบไม่ขึ้นกับผู้พูด ที่เน้นเสียงตัวเลข และอยู่บนพื้นฐานของกรรมวิธีที่แตกต่างกัน 3 แบบคือ ไดนามิก ไทม์วาร์ปิง แบบจำลองฮิดเดน มาร์คอฟ และนิวรอลเน็ตเวิร์ก การทำงานของระบบมี 4 ขั้นตอนคือ ขั้นตอนการประมวลสัญญาณเบื้องต้น ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ ขั้นตอนการทดสอบหรือวัดความคล้ายคลึงกันของรูปแบบ และขั้นตอนการตัดสินใจ

ขั้นตอนการประมวลสัญญาณเบื้องต้น แยกย่อยได้เป็น กรรมวิธีเน้นล่วงหน้า กรรมวิธีหาจุดสิ้นสุดเสียงพูด กรรมวิธีวางกรอบขนาดสัญญาณ และกรรมวิธีปรับบรรทัดฐานเชิงเวลา ซึ่งกรรมวิธีไดนามิก ไทม์วาร์ปิงต้องดำเนินการกรรมวิธีย่อยที่ 2 และ 3 ในขณะที่กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ต้องดำเนินการกรรมวิธีย่อยที่ 1, 2 และ 3 ส่วนกรรมวิธีนิวรอลเน็ตเวิร์ก ต้องดำเนินการกรรมวิธีย่อยที่ 2, 3 และ 4

ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ สำหรับกรรมวิธีไดนามิก ไทม์วาร์ปิง ใช้ผลการแปลงฮาร์ตเลย์ ในการคำนวณหาค่าพารามิเตอร์ที่เหมาะสม ในขณะที่กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ทำการหาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น จากนั้นทำการควอนไตซ์แบบเวกเตอร์ และฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ ส่วนกรรมวิธีนิวรอลเน็ตเวิร์ก ทำการหาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น จากนั้นฝึกฝนแบบจำลองด้วย นิวรอลเน็ตเวิร์ก

ขั้นตอนการตัดสินใจ สำหรับกรรมวิธีไดนามิก ไทม์วาร์ปิง ใช้การเปรียบเทียบเข้าคู่ต้นแบบและใช้เงื่อนไขในการวัดหาค่าผิดพลาดแบบ Nearest Neighbor ในขณะที่กรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ ใช้อัลกอริทึม Viterbi และใช้เงื่อนไขในการวัดหาค่าผิดพลาดแบบค่าผิดพลาดกำลังสองเฉลี่ย ส่วนกรรมวิธีนิวรอลเน็ตเวิร์ก ใช้เงื่อนไขในการวัดหาความใกล้เคียงแบบอ้างอิงสูงสุด

ในการทดสอบของทั้งสามกรรมวิธี จะกำหนดกลุ่มตัวอย่าง 3 กลุ่ม กลุ่มฝึกฝน กลุ่มทดสอบ 1 และ 2 กลุ่มฝึกฝนและกลุ่มทดสอบ 1 เป็นผู้พูดชุดเดียวกัน เพื่อใช้กลุ่มทดสอบ 1 ในการปรับแต่งระบบการรู้จำให้ทำงานได้ดีที่สุดในแต่ละกรรมวิธี ดังนั้นจำนวนตัวอย่างในแต่ละกลุ่มของแต่ละกรรมวิธีจะแตกต่างกัน ดังนี้ กลุ่มฝึกฝน กรรมวิธี DTW, HMM, และ NN จะมีจำนวนตัวอย่างเป็น 20, 45, และ 30 ตัวอย่างตามลำดับ โดยที่ของ NN นั้นแต่ละตัวอย่างจะทำการบันทึกเสียงพูดไว้ 2 ชุด กลุ่มทดสอบ 1 กรรมวิธี DTW, HMM, และ NN จะมีจำนวนตัวอย่างเป็น 20, 45, และ 30 ตัวอย่างตามลำดับ โดยที่ของ DTW นั้นแต่ละตัวอย่างจะทำการบันทึกเสียงพูดไว้ 2 ชุด กลุ่มทดสอบ 2 กรรมวิธี DTW, HMM, และ NN จะมีจำนวนตัวอย่างเป็น 20, 10, และ 12 ตัวอย่างตามลำดับ โดยที่ของ DTW นั้นแต่ละตัวอย่างจะทำการบันทึกเสียงพูดไว้ 2 ชุด ส่วนของ NN นั้นแต่ละตัวอย่างจะทำการบันทึกเสียงพูดไว้ 3 ชุด

ผลการทดสอบระบบการรู้จำของแต่ละกรรมวิธี จะทำการเปลี่ยนแปลงจำนวนตัวอย่างของกลุ่มฝึกฝน และวัดหาอัตราการรู้จำของแต่ละกลุ่ม ปรากฏว่าทุก ๆ กรรมวิธี เมื่อเพิ่มจำนวนตัวอย่างที่ใช้ฝึกฝน อัตราการรู้จำของกลุ่มฝึกฝนเองจะลดลง โดยที่ของกรรมวิธี DTW ลดลงมากที่สุด และของ NN ลดลงน้อยที่สุดจนอาจกล่าวได้ว่ามีค่าคงที่ ในขณะที่อัตราการรู้จำของกลุ่มทดสอบทั้ง 2 กลุ่มเพิ่มขึ้น โดยที่ของกลุ่มทดสอบ 2 อัตราการรู้จำของ NN ดีที่สุด ตามมาด้วยของ HMM ส่วนของกลุ่มทดสอบ 1 ของทั้ง 3 กรรมวิธี มีความแตกต่างกันไม่มากนัก และเมื่อพิจารณาบนเงื่อนไขที่ดีที่สุดของแต่ละกรรมวิธี จะได้ว่า กรรมวิธี DTW ให้อัตราการรู้จำของกลุ่มฝึกฝน กลุ่มทดสอบ 1 และกลุ่มทดสอบ 2 เป็นร้อยละ 90.50 86.50 และ 79.25

ตามลำดับ เมื่อใช้ตัวอย่างจำนวน 20 ตัวอย่างในการฝึกฝน กรรมวิธี HMM ให้อัตราการรู้จำ ของแต่ละกลุ่มเป็นร้อยละ 95.30 89.70 และ 84.00 ตามลำดับ เมื่อใช้ตัวอย่างจำนวน 45 ตัวอย่างในการฝึกฝน และกรรมวิธี NN ให้อัตราการรู้จำของแต่ละกลุ่มเป็นร้อยละ 98.20 84.30 และ 89.40 ตามลำดับ เมื่อใช้ตัวอย่างจำนวน 30 ตัวอย่าง ๆ ละ 2 ชุด ในการฝึกฝน

ความผิดพลาดในการรู้จำที่เกิดขึ้นส่วนใหญ่เป็นแบบสุ่ม ซึ่งคาดว่ามาจากสาเหตุหลัก ๆ ดังนี้ จากกรรมวิธีหาจุดสิ้นสุดเสียงพูดที่ยังไม่ดีพอ ทำให้ข้อสนเทศที่ปรากฏในส่วนหน้าของเสียงคำพูดถูกตัดทิ้งไปได้ จากกรรมวิธีวัดและหาค่าลักษณะสำคัญที่ยังไม่สมบูรณ์เพียงพอ เช่นไม่ได้นำเรื่องค่าความถี่ของสัญญาณเสียงพูดมาประเมิน ทั้ง ๆ ที่เสียงพูดภาษาไทยเป็นเสียงดนตรี เป็นเหตุให้ข้อสนเทศนี้ถูกละเลย การรู้จำจึงอาจผิดพลาดได้ จากการออกเสียงที่ผิดพลาดของผู้พูด ที่อาจทำให้รูปคลื่นสัญญาณผิดเพี้ยนได้ หรือจากสัญญาณรบกวนที่เกิดขึ้นในระหว่างการบันทึกเสียง เพราะกระทำกันในสภาวะแวดล้อมปกติ สัญญารรบกวนที่เกิดขึ้นอาจไปทำให้รูปคลื่นสัญญาณบิดเบี้ยวไปได้

4.2. ข้อเสนอแนะ

ข้อเสนอแนะสำหรับปรับปรุงกระบวนการรู้จำเสียงพูดภาษาไทยมีดังนี้

4.2.1. ควรทำการวิเคราะห์สัญญาณเสียงพูดเชิงความถี่ เพื่อหาลักษณะสำคัญเพิ่มเติม เนื่องจากลักษณะเฉพาะของเสียงพูดภาษาไทย ที่เป็นเสียงดนตรี (Tonal Language) ทั้งนี้มีงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์เสียงวรรณยุกต์ (ธีระ ภัทราพรนันท์, 2538) แนวทางแก้ปัญหาที่เป็นไปได้ เช่น การวิเคราะห์ Cepstrum

4.2.2. ควรเปลี่ยนขนาดข้อมูลในการบันทึกจาก 8 บิต เป็น 16 บิต เพื่อเพิ่มอัตราส่วนสัญญาณต่อสัญญาณรบกวน

4.2.3. ควรปรับปรุงกรรมวิธีหาจุดสิ้นสุดเสียงพูดให้ดีขึ้น เช่น การนำค่า Pitch Period มาใช้ หรือ การพิจารณาการเกิด Zero Crossing เป็นต้น

4.2.4. ควรเพิ่มขนาดของชุดรหัสในกรรมวิธี HMM เพื่อรองรับคำศัพท์ที่จะเพิ่มขึ้น และเพิ่มสถานะของแบบจำลองให้มากขึ้นด้วยเช่นกัน เพื่อเพิ่มประสิทธิภาพของระบบ เช่น งานวิจัยของ วิศรุต อาชุบุตร (2539) ได้ทดลองเพิ่มขนาดของชุดรหัสเป็น 128 และสถานะของแบบจำลองเป็น 5 ปรากฏว่าอัตราการรู้จำเสียงพูดตัวเลขภาษาไทยสูงขึ้นเป็นร้อยละ 84.25 ทั้ง ๆ ที่เป็นระบบการรู้จำเสียงพูดที่มีคำศัพท์รวม 70 คำ ซึ่งถ้าปรับให้เหลือเพียง 10 คำตามเสียงศูนย์ถึงเก้า อัตราการรู้จำควรจะสูงขึ้นกว่าค่าดังกล่าวอย่างมีนัยสำคัญ

4.2.5. ควรนำข้อดีของกรรมวิธีแบบจำลองฮิดเดน มาร์คอฟ และของกรรมวิธีนิวรอลเน็ตเวิร์กมาผสมกัน ที่เรียกว่า Hybrid HMM-NN และอาจเพิ่มส่วนของฟัซซี เข้าไปในขั้นตอนการตัดสินใจต่าง ๆ

4.2.6. ควรมีการศึกษาโครงสร้างและรูปแบบการออกเสียงของภาษาไทยอย่างจริงจัง เพื่อนำมาช่วยในการรู้จำ

4.2.7. ควรสร้างฐานข้อมูลเสียงพูดที่เป็นมาตรฐาน เพื่อลดความสับสนในการศึกษาเชิงเปรียบเทียบของเทคนิคหรือของกรรมวิธีหรือของผลการทดลองต่าง ๆ

4.2.8. ควรควบคุมสภาพแวดล้อมในระหว่างทำการบันทึกเสียงของผู้พูดที่จะใช้ฝึกฝนเป็นรูปแบบอ้างอิง เพื่อลดผลของสัญญาณรบกวนให้มากที่สุด

รายการอ้างอิง

ภาษาไทย

- สุเธียร เกียรติสุนทร. การประมาณพันธะเชิงเส้นเสียงพูด (LINEAR PREDICTION OF SPEECH). วิทยานิพนธ์ปริญญามหาบัณฑิต สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง, 2525.
- ทวิ ประทุมทาน. การตรวจรู้เสียงพูดภาษาไทยโดยใช้หน่วยพยางค์. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2530.
- ไพศาล ธรรมโพธิ์ทอง, ระบบการรับรู้เสียงพูดแบบต่างบุคคล. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2533.
- สุนิสา จันทวีกูล. การวิเคราะห์สเปกตรัมกำลังของสัญญาณโดยใช้ดีสครีตฮาร์ตเลย์ทรานส์ฟอร์ม.. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2536.
- ณัฐกร ทับทอง. การรู้จำคำพูดภาษาไทยโดยใช้ลักษณะบ่งความต่างของหน่วยเสียง. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2538.
- เสาวลักษณ์ อารีย์พงศา. การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดน มาร์คอฟ โมเดล และเวกเตอร์ควอนไตซ์เซชัน. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2538.
- ระพีพัฒน์ เพ็ญศิริ. การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นต่อผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปิง. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2538.
- ธีระ พัทราพรนันท์. การรู้จำเสียงพูดสระภาษาไทยโดด ๆ ไม่ขึ้นกับผู้พูดโดยการวัดสเปกตรัมดิสแตนซ์และใช้ไดนามิกไทม์วาร์ปปิง. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2538.
- วิศรุต อาชุบุตร. ระบบการรู้จำคำไทยหลายพยางค์แบบไม่ขึ้นกับผู้พูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟ. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2539.
- วุฒิพงษ์ พรสุขจันทรา. การรู้จำเสียงตัวเลขภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยใช้แอลพีซี และนิวรอลเน็ตเวิร์กแบบแบ็กพรอพาเกชัน. วิทยานิพนธ์ปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2539.
- ชัยศรี เอี่ยมอำไพ การตรวจหาจุดเริ่มต้นและจุดสิ้นสุดของคำโดดๆ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, ม.ป.ท.
- กิตติพงษ์ เจนวิถีสุข. การรู้จำตัวอักษรพิมพ์ภาษาไทยโดยใช้นิวรอลเน็ตเวิร์กและวิธีซินแทกติก. วิทยานิพนธ์ปริญญารมหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย, 2538.

ภาษาอังกฤษ

- Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W., and Laksaneeyanawin, S., A Speaker Independent Thai Polysyllabic Word Recognition System Using Hidden Markov Model, IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing (PACRIM'97), Victoria, B.C., Canada, August 20-22, 1997, pp. 593-599.
- Allen, J. B., How Do Humans Process and Recognize Speech ?, IEEE Transaction on Speech and Audio Processing, 2 (October 1994): 567-577.
- Areeponsa, S. and Jitapunkul, S., Speaker Independent Thai Numeral Speech Recognition Using Hidden Markov Model and Vector Quantization, International Symposium on Natural Language Processing 1995: SNLP'95, Bangkok, Thailand, 2-4 August, 1995, pp.370-378.
- Bahl, L. R., Brown, P. F., Souza, P. V., Mercer, R. L. and Picheny, M. A., Acoustic Markov Models Used in the Tangara Speech Recognition System, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 497-500.
- Bahl, L. R., Bakis, R., Souza, P. V., and Mercer, R. L., Obtaining Candidate Words by Polling in a Large Vocabulary Speech Recognition System, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 489-492.
- Bahl, L. R., Brown, P. F., Souza, P. V. and Mercer, R. L., Estimating Hidden Markov Model Parameters so as to Maximize Speech Recognition Accuracy, IEEE Transaction on Speech and Audio Processing, 1 (January 1993): 77-83.
- Bahl, L. R., Gennaro, S. V., Gopalakrishnan, P. S., Mercer, R. L., A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition, IEEE Transaction on Speech and Audio Processing, 1 (January 1993): 59-67.
- Bedworth, Comparison of Neural and Conventional Classifiers on a Speech Recognition Problem, First IEE International Conference or Artificial Neural Networks, (October 1989): 86-89.
- Beth A. Carlson, and Mark A. Clements, A Projection-Based Likelihood Measure for Speech Recognition in Noise, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 97-102.
- Bocchieri, E. L., and Doddington, G. R., Frame-Specific Statistical Features for Speaker Independents Speech Recognition, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-34, No. 4, August 1986, pp. 755-764.
- Bocchieri, E. L., Doddington, G. R., Speaker Independent Connected Digit Recognition with Frame Specific Distance Measures, Digital Signal Processing - 87, (1987): 534-538.
- Bourlard, H. and Morgan, N., Continuous Speech Recognition by Connectionist Statistical Methods, IEEE Transaction on Neural Networks, 4 (November 1993): 893-909.

- Carre, R., A Summary of Speech Research Activities in France, IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-22, No. 4, August 1974, pp. 268-272.
- Chantawekul, S., and Jitapunkul, S., Power Spectrum Analysis by Discrete Hartley Transform, 16th Electrical Engineering Conference, Bangkok, November, 25-26, 1993, pp. 646-649.
- Chen, S.H., and Wang, Y.R., Tone Recognition of Continuous Mandarin Speech Based on Neural Networks, IEEE Transactions on Speech and Audio Processing, Vol 3 (March 1995): 146-150
- Christian Dugast, Laurence Devillers, and Xavier Aubert, Combining TDNN and HMM in a Hybrid System for Improved Continuous-Speech Recognition, IEEE trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 217-223.
- Clifford J. Weistein, Opportunities for Advanced Speech Processing in Military Computer-Based Systems, IEEE Proc., Vol. 79, No. 11, Nov. 1991, pp. 1626-1641.
- Dante, H. M., and Sarma, V. V. S., Automatic Speaker Identification for a Large Population, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, No. 3, June 1979, pp. 255-263.
- Das, S. K., Some Experiments with a Supervised Classification Procedure, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 461-464.
- Das, S. K., Some Experiments in Discrete Utterance Recognition, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-30, No. 5, October 1982, pp. 766-770.
- David Mansour, and Biing Hwang Juang, The Short-Time Modified Coherence Representation and Noisy Speech Recognition, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 37, NO. 6, June 1989, pp. 795-804.
- Deller, J. R., Jr., Proakis, J. G., and Hansen, J. H. L., Discrete-Time Processing of Speech Signals,. MacMillan, 1993.
- Deng, L., Lennig, M., Gupta, V. N., Mermelstein, P., A Modeling Acoustic-Phonetic Detail in an HMM-Based Large Vocabulary Speech Recognizer, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 509-512.
- Dermatas, E. S., A New Algorithm for Optimum Reference Templates Creation in Speech Recognition Systems, Digital Signal Processing - 87, (1987).
- Dermatas, E. S., Fakotakis, N. D., and Kokkinakis, G. K., Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment, 1991 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1991): 733-736.
- Elvira, and Carrasco, R.A., Neural Network Architectures for Speech Recognition, IEE Colloquium on Telecommunications, Consumer and Industrial Applications of Speech Technology, (May 1992): 4/1-5.
- Embree, P. M., C Algorithms for Real-time DSP, Prentice-Hall International Inc. 1995.

- Eng-Fong Huang, and Hsiao-Chuan Wang, An Efficient Algorithm for Syllable Hypothesization in Continuous Mandarin Speech Recognition, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 446-449.
- Eng-Fong Huang, Hsiao-Chuan Wang, and Frank K. Soong, A Fast Algorithm for Large Vocabulary Keyword Spotting Application, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 449-452.
- Euler, S. and Wolf, D., Speaker Independent Isolated Word Recognition Based on Continuous Hidden Markov Models Using Multidimensional Spherically Invariant Functions, Digital Signal Processing - 87, (1987): 539-542.
- Evangelos S. Dermatas, Nikos D. Fakotakis, and George K. Kokkinakis, Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment, CH2977-7/91/0000-0733, 1991 IEEE, pp. 733-736.
- Fissore, L., Laface, P., Micca, P., and Pieraccini, R., Lexical Access to Large Vocabularies for Speech Recognition, IEEE Transaction on Acoustic, Speech, and Signal Processing, 36 (May 1988): 1197-1213.
- Forney Jr., G. D., The Viterbi Algorithm, Proceedings of the IEEE, 61 (March 1973): 268-278.
- Furui, S., A Training Procedure for Isolated Word Recognition Systems, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, No. 2, April 1980, pp. 129-136.
- Furui, S., Speaker-Independent Isolated Word Recognition Using Dynamic Feature of Speech Spectrum, IEEE Trans. on Acoustics Speech and Signal Processing, Vol. Assp-34, No.1, February 1986, pp. 52-59.
- Furui, S., Digital Speech Processing, Synthesis and Recognition, New York and Basel:Marcel Dekker, 1989.
- Furui, S., Advances in Speech Signal Processing, Marcel Dekker, New York, 1989.
- Furui, S., and Sondhi, M. M., Digital Speech Processing Synthesis and Recognition, Marcel Dekker, New York, 1991.
- Furui, S., and Sondhi, M. M., Advances in Speech Signal Processing, Marcel Dekker, 1992.
- Gerhard Rigoll, Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition Systems, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 175-184.
- Gopalakrishnan, P. S., and Nahamoo, D., Models and Algorithms for Continuous Speech Recognition: A Brief Tutorial, 1993 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1993): 1535-1538.
- Gray, R. M., Vector Quantization, IEEE ASSP Magazine, (April 1984): 4-29.

- Gupta, N., Bryan, J. K., and Gowdy, J. N., A Speaker-Independent Speech-Recognition System Based on Linear Prediction, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-26 (February 1978): 27-33.
- Haiyan, H., and Chengyi, W., Art2-Based Multiple MLPs Neural Network for Speaker Independent Recognition of Isolated Words, IEEE Proceedings 11 th IAPR International Conference on Pattern Recognition, Vol 2 (September 1992): 590-593.
- Hamid Sheikhzadeh, and Li Deng, Waveform-Based Speech Recognition Using Hidden Filters : Parameter Selection and Sensitivity to Power Normalization, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 80-89.
- Huang, Phoneme Classification Using Semicontinuous Hidden Markov Models, IEEE Trans. Signal Processing, Vol. 40, No. 5, May 1992, pp.1062-1067.
- Hunt, M. J., Evaluating the Performance of Connected-Word Speech Recognition System, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 457-460.
- Jianing Dai, Iain G. MacKenzie, and Jon E. M. Tyler, Stochastic Modeling of Temporal Information in Speech for Hidden Markov Models, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 102-104.
- Jin, G., and Chung, L.H., A Multilayer Perceptron Postprocessor to Hidden Markov Modeling for Speech Recognition, IEEE International Conference on Acoustics,Speech and Signal Processing, Vol 2 (April 1993): 263-266.
- John Huang, and Ben-Dau Tseng, A Walsh Transform Based Endpoint Detection of Isolated Utterances, 1058-6393/91, 1991 IEEE, pp. 335-338.
- Jung-Kuei Chen, and Frank K. Soong, An N-Best Candidates-Based Discriminative Training for Speech Recognition Applications, IEEE trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 206-216.
- Kammerer, B.R., and Kupper, W.A., Design of Hierarchical Perceptron Structures and Their Application to the Task of Isolated-Word Recognition, International Joint Conference on Neural Networks, Vol 1 (June 1989): 243-249.
- Kannan, M. Ostendorf, and J.R. Rohlicek, Maximum Likelihood Clustering of Gaussians for Speech Recognition, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 453-455.
- Keh-Yih Su, and Chin-Hui Lee, Speech Recognition Using Weighted HMM and Subspace Projection Approaches, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 69-79.
- Kevin R. Farrell, Richard J. Mammone, and Khaled T. Assaleh, Speaker Recognition Using Neural Networks and Conventional Classifiers, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 194-205.

- Kim, H. R., and Lee, H. S., Postprocessor Using Fuzzy Vector Quantizer in HMM-Based Speech Recognition, Electronics Letter 24th, 27 (October 1991).
- Koo, J. M., and Un, C. K, Fuzzy Smoothing of HMM Parameters in Speech Recognition, Electronics Letter 24th, 26 (May 1990): 734-744.
- Kuc, R., Introduction to Digital Signal Processing, McGraw-Hill, 1988.
- Kuhn, M. H., and Tomaschewski, H. H., Improvements in Isolated Word Recognition, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-31 (February 1983): 157-167.
- Kung, S.Y., Digital neural networks, USA:PRENTICE HALL, 1993.
- Lamel, L. F., Rabiner, L. R., Rosenberg, A. E. and Wilpon, J. G., An Improved Endpoint Detector for Isolated Word Recognition, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-29 (August 1981): 777-785.
- Lee, Chin-hui, On Robust Linear Prediction of Speech, IEEE Transaction on Acoustic, Speech, and Signal Processing, 36 (May 1988): 642-650.
- Lee, K., and Hon, H., Large-Vocabularty Speaker-Independent Continuous Speech Recognition Using HMM, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 123-126.
- Lee, Kai-Fu., and Hon, Hsiao-Wuen, Speaker-Independent Phone Recognition Using Hidden Markov Models, IEEE Transaction on Acoustics, Speech and Signal Processing, 37 (Nov. 1989): 1641-1648.
- Lee, K., Hon, H., and Reddy, R., An Overview of the SPHINX Speech Recognition System, IEEE Transaction on Acoustic, Speech, and Signal Processing, 38 (January 1990): 35-45.
- Levinson, S. E., Rabiner, L. R., and Sondhi, M. M., An Introduction to the Application of the Theory of Probability Functions of a Markov Process to Automatic Speech Recognition, The Bell System Technical Journal, 62 (April 1983).
- Levinson, S. E., Ljolje, A., Miller L. G., Large Vocabulary Speech Recognition Using a Hidden Markov Model for Acoustic/Phonetic Classification, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 505-508.
- Levinson, S. E., Roe D. B., A Perspective on Speech Recognition, IEEE Communications Magazine, 28 (January 1990): 28-34.
- Levinson, S. E. and Ljolje, A., Development of an Acoustic-Phonetic Hidden Markov for Continuous Speech Recognition, IEEE Transaction on Signal Processing, 39 (January 1994): 29-39.
- Lleida, E., Nadeu, C., Monte, E., Marino, J. B., Statistical Feature Selection for Isolated Word Recognition, 1990 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1990): 757-760.

- Lubensky, D., Word Recognition Using Neural Nets, Multi-State Guassian and k-nearest Neighbor Classifiers, IEEE Intenational Conference on Acoustics.speech and Signal Processing, Vol 1 (May 1991): 141-144.
- Luksaneeyanawin, S., Linguistics Research and Thai Speech Technology, The 5th International Conference on Thai Studies, School of Oriental and African Studies, University of London (July 1993).
- Makhoul, J., Roucos, S., and Gish, H., Vector Quantization in Speech Coding, Proceedings of the IEEE, 73 (November 1985): 1551-1588.
- Mariani, J., Recent Advances in Speech Recognition, 1989 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1989): 429-440.
- Markel, J. D., Oshika, B. T., and Gray, A. H., Long-Term Feature Averaging for Speaker Recognition, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, No. 4, August 1977, pp. 330-337.
- Mattia, M., and Giachin, E. P., Experimental Result on Large-Vocabulary Continuous Speech Recognition and Understanding, 1988 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), (1988): 691-694.
- McInnes, F. R., Jack, M. A., and Laver, J., Template Adaptation in a Isolated Word-Recognition System, IEE Proceedings, 136 (April 1989): 119-126.
- Morgan, N. and Bourlard, H., Continuous Speech Recognition, IEEE Signal Processing Magazine, (May 1995): 25-42.
- Myers, C., Rabiner, L. R., and Rosenberg, A. E., Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition, IEEE Trans.Acoust.,Speech, Signal Processing, Vol. ASSP-28, No. 6, December 1980, pp. 623-635.
- Nejat Ince, A., Digital Speech Processing: Speech Coding, Synthesis and Recognition, Kluwer Academic Publishers 1992.
- Neri Merhav, and Yariv Ephraim, A Bayesian Classification Approoach with Application to Speech Recognition, IEEE Trans. Siganl Processing, Vol. 39, No. 10, October 1991, pp. 2157-2166.
- Oded Ghitza, Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition, IEEE trans. Speech and Audio Processing, Vol. 2, No. 1, Part II, January 1994, pp. 115- 132.
- Oppenheim, A. V., Schafer, R. W., Discrete-Time Signal Processing, Prentice Hall 1989.
- O'Shaughnessy, Douglas , Linear Predictive Coding, IEEE Potentials, (February 1988): 29-32.
- Owens, F. J., Signnal Processing of Speech, MaccMillan, 1993.
- Pan K., Soong, F. K., and Rabiner, L. R., A Vector-Quantization-Based Preprocessor for Speaker-Independent Isolated Word Recognition, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-33 (June 1985): 546-560.

- Peacocke, R. D. and Graf, D. H., An Introduction to Speech and Speaker Recognition, IEEE Computer Magazine, (August 1990): 26-33.
- Peinado, A. M., Lopez, J. M., Sanchez, V. E., Segura, J. C., Ruio Ayuso, A. J., Improvements in HMM-Based Isolated Word Recognition System, IEE Proceedings, 138 (June 1991): 201-206.
- Pelton, G.E., Voice Processing, Singapore : McGraw-Hill, 1993.
- Pensiri, R., and Jitapunkul, S., Speaker-Independent Thai Numeral Voice Recognition by Using Dynamic Time Warping, 18th Electrical Engineering Conference, Pattaya, November, 22-24, 1995, pp. 977-981.
- Phatrapornnant, T., and Jitapunkul, S. Speaker-Independent Isolated Thai Spoken Vowel Recognition by Using Spectrum Distance Measurement and Dynamic Time Warping, 18th Electrical Engineering Conference, Pattaya, November, 22-24, 1995, pp. 988-993.
- Philippe Le Cerf, Weiye Ma, and Dirk Van Compernolle, Multilayer Perceptrons as Labelers for Hidden Markov Models, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 185-193.
- Picone, J.W., Continuous Speech Recognition Using Hidden Markov Model, IEEE ASSP Magazine, (July 1990): 26-41.
- Picone, J. W., Signal Modeling Techniques in Speech Recognition, Proceedings of the IEEE, 81 (September 1993): 1215-1247.
- Pornsukjantra, W., Jitapunkul, S., Ahkuputra, V., and Wutiwiwatchai, C., Speaker-Independent Thai Numeral Speech Recognition Using LPC and the Back Propagation Neural Network, International Symposium on Natural Language Processing 1997: SNLP'97, Phuket, Thailand, 1997, pp.
- Pornsukjantra, and S. Jitapunkul, Speaker-Independent Thai Numeral Speech Recognition Using LPC and the Back Propagation Neural Network, 19th. Electrical Engineering Conference, Khonkaen, November, 7-8, 1996, pp. DS-67-DS-72.
- Rabiner, L. R., and Sambur, M. R., Some Preliminary Experiments in the Recognition of Connected Digits, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 2, April 1976, pp. 170-182.
- Rabiner, L. R., On Creating Reference Templates for Speaker-Independent Recognition of Isolated Words, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-26 (February 1978): 34-42.
- Rabiner, L. R., RosenBerg, A. E., and Levinson, S. E., Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition, IEEE Trans. Acoust.,Speech, Signal Processing, Vol. ASSP-26, No. 6, December 1978, pp. 575-582.
- Rabiner, L. R., and Schafer, R. W., Digital Processing of Speech Signals, Prenttice-Hall, 1978.

- Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G., Speaker-Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-27 (August 1979): 336-349.
- Rabiner, L. R., and Wilpon, J. G., A Speaker-Independent Isolated Word Recognition for a Moderate Size (54 Words) Vocabulary, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-27 (December 1979): 583-587.
- Rabiner, L. R., and Schmidt, C. E., Application of Dynamic Time Warping to Connected Digit Recognition, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 4, August 1980, pp. 377-388.
- Rabiner, L. R., and Levinson, S. E., Isolated and Connected Word Recognition - Theory and Selected Applications, IEEE Trans. on Comm., Vol. COM-29, No. 5, May 1981, pp. 621-659.
- Rabiner, L. R., Levinson, S. E., and Sondhi, M. M., On the Application of Vector Quantization and Hidden Markov Model to Speaker Independent Isolated Word Recognition, The Bell System Technical Journal, 62 (April 1983) : 1075-1106.
- Rabiner, L. R. and Levinson, S. E., A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based on Hidden Markov Models and Level Building, IEEE Transaction on Acoustic, Speech, and Signal Processing, ASSP-33 (June 1985): 561-573.
- Rabiner, L. R., and Juang, B. H., An Introduction to Hidden Markov Models, IEEE ASSP Magazine, (January 1986): 4-16.
- Rabiner, L. R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 77 (February 1989): 257-286.
- Rabiner, L. R., Wilpon, J. G., and Soong, F. K., High Performance Connected Digit Recognition Using Hidden Markov Models, IEEE Transaction on Acoustic, Speech, and Signal Processing, 37 (August 1989): 1214-1225.
- Rabiner, L. R., Application of Voice Processing to Telecommunications, Proceedings of the IEEE, 82 (February 1994): 199-228.
- Rabiner, L. R., The Role of Voice Processing in Telecommunication, Proceedings on 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA 94), (1994): 1-8.
- Rashwan, M. A., and Fahmy, M. M., New Technique for Speaker-Independent Isolated-Word Recognition, IEE Proceedings, 135 (June 1988): 251-257.
- Reynolds, and Tarassenko, L., Isolated Word Recognition with the Radial Basis Function Classifier, IEE Second International Conference on Artificial Neural Networks, (November 1991): 345-349.
- Roe, D. B. and Wilpon, J. G., Whither Speech Recognition: The Next 25 Years, IEEE Communications Magazine, 31 (November 1993): 54-62.

- Schaikoff, R. J., Pattern Recognition: Statistical, Structural, and Neural Approaches, John Wiley & Sons Inc. 1992.
- Schurer, T., An Experimental Comparison of Different Feature Extraction and Classification Methods for Telephone Speech, Proceedings Second IEEE Workshop in Interactive Voice Technology for Telecommunications Applications, (September 1994): 93-96.
- Sedgewick, R., Algorithms, 2nd ed. Massachusetts : Addison Wesley, 1989.
- Silverman, H. F., and Morgan, D. P., The Application of Dynamic Programming to Connected Speech Recognition, IEEE ASSP Magazine, (July 1990): 6-25.
- Silverman, H. F., and Dixon N. R., A Comparison of Several Speech-Spectra Classification Methods, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 4, August 1976, pp. 289-295.
- Steve Renals, Nelson Morgan, Herve Bourlard, Michael Cohen, and Horacio Franco, Connectionist Probability Estimators in HMM Speech Recognition, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 161-174.
- Tomoko Matsui, and Sadaoki Furui, Comparision of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 456-459.
- Ukita, T., Saito E., Nitta, T., and Watanabe, S., A Speaker-Independent Connected Digit Recognition System Concatenating Statistically Discriminated Words, IEEE Trans. Signal Proessing, Vol.40, No. 40, October 1992, pp. 2414-2424.
- Waibel, A., Hanazawa, T., Hinton, G.,Shikano, K., and Lang, K. Phoneme Recognition: Neural Networks vs. Hidden Markov Models, IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1 (April 1988): 107-110.
- Warren, J. H., A Pattern Classification Technique for Speech Recognition, IEEE Trans., Audio, Electroacoustics, Vol. AU-19, No. 4, December 1971, pp. 281-285.
- White, G. M., and Neely, R. B., Speech Recognition Experiments with Linear Predication, Bandpass Filtering, and Dynamic Programming, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 2, April 1976, pp. 183-188.
- Ying, C.D. Mitchell, and L.H. Jamieson, Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement, 0-7803-0946-4/93, 1993 IEEE, pp. II-732-II-735.
- Ying Hao, and Ditang Fang, Speech Recognition Using Speaker Adaptation by System Parameter Transformation, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 63-68.
- Yunxin Zhao, An Acoustic-Phonetic-Based Speaker Adaptation Technique for Improving Speaker-Independent Continuous Speech Recognition, IEEE trans. Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 380-394.

- Zavaliagkos, Y. Zhao, R. Schwartz, and J. Makhoul, A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 1, Part I, January 1994, pp. 151-160.
- Zelinski, R., and Class, F., A Segmentation Algorithm for Connected Word Recognition Based on Estimation Principles, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-31, No. 4, August 1983, pp. 818-827.
- Zeng, H., and Yu, T., Parallel Sequential Running Neural Network and Its Application to Automatic Speech Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1 (March 1992): 429-432.
- Sound Blaster User Reference Manual, Creative Labs, Inc. 1989.
- The Developer Kit for Sound Blaster Series, Creative Labs, Inc. 1991.