การรู้จำการอ่านริมฝีปากโดยการใช้เทคนิคการวิเคราะห์สัญญาณแปรตามเวลาและนิวรอลเน็ตเวิร์ก

นายปกิต ศีลประชาวงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2543
ISBN 974-346-442-5

LIPREADING RECOGNITION USING TIME-VARYING SIGNAL ANALYSIS AND NEURAL NETWORKS

Mr. Pakit Silprachawong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2000
ISBN 974-346-442-5

หัวข้อวิทยานิพนธ์	การรู้จำการอ่านริมฝีปากโดยการใช้เทคนิคการวิเคราะห์สัญญาณแปรตามเวลาและนิวรอลเน็ตเวิร์ก
โดย	นายปกิต ศีลประชาวงศ์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อ.ดร.บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร.บุญเสริม กิจศิริกุล)

..... กรรมการ
(รองศาสตราจารย์ ดร.วันชัย ริ้วไพบูลย์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สาธิต วงศ์ประทีป)

ปกิต ศีลประชาวงศ์ : การรู้จำการอ่านริมฝีปากโดยการใช้เทคนิคการวิเคราะห์สัญญาณแปรตามเวลาและนิวรอลเน็ตเวิร์ก (LIPREADING RECOGNITION USING TIME-VARYING SIGNAL ANALYSIS AND NEURAL NETWORKS) อ. ที่ปรึกษา:อ.ดร.บุญเสริม กิจศิริกุล, 49 หน้า. ISBN 974-346-442-5.

งานวิจัยนี้เสนอวิธีการสำหรับการรู้จำการอ่านริมฝีปาก (Lipreading Recognition) โดยใช้ข้อมูลที่เป็นลำดับภาพที่ได้จากภาพเทาของริมฝีปากของผู้พูด โดยในขั้นตอนการรู้จำมีการดึงข้อมูลของแต่ละภาพและนำมาเข้าโมเดลโดยการใช้ การเปลี่ยนแปลงของความเข้มของแต่ละจุดเทียบกับเวลาเป็นสัญญาณหลัก และมีการใช้การแปลงแบบฟูเรียร์ (Fourier transform) เพื่อแทนสัญญาณ จากนั้นจะดึงค่าค่าสัมประสิทธิ์ฟูเรียร์ (Fourier coefficients) เพื่อใช้เป็นคุณลักษณะ (feature) ให้กับนิวรอลเน็ตเวิร์ก (Neural Networks) สำหรับขั้นตอนการรู้จำ เราทำการทดลองกับฐานข้อมูล 2 ชุด คือชุดตัวเลขและชุดตัวอักษรภาษาอังกฤษ ผลการทดลองแสดงให้เห็นถึงประสิทธิผลของวิธีที่นำเสนอ

ภาควิชา วิศวกรรมคอมพิวเตอร์	**ลายมือชื่อนิสิต**.....
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์	**ลายมือชื่ออาจารย์ที่ปรึกษา**.....
ปีการศึกษา 2543	**ลายมือชื่ออาจารย์ที่ปรึกษาร่วม**.....

4170392721 : MAJOR COMPUTER SCIENCE
KEYWORD :RECOGNITION, LIPREADING, FEATURE EXTRACTION, TIME-VARYING SIGNAL ANALYSIS, NEURAL NETWORKS, PATTERN RECOGNITION
PAKIT SILPRACHAWONG : LIPREADING RECOGNITION USING TIME-VARYING SIGNAL ANALYSIS AND NEURALNETWORKS. THESIS ADVISOR : BOONSERM KIJSIRIKUL, Ph.D. 49 pp. ISBN 974-346-442-5.

This thesis presents an approach for lipreading recognition based on visual features extracted from gray level image sequences of the speaker's lips. The recognition is done by extracting visual information from each image, and the extracted information is modeled by using the intensity curve of pixels along the time axis as the primary signal. Fourier transform is then applied to this signal. Therefore, the Fourier coefficients of a signal curve encode the motion information in a compact manner and are used as features to Neural Networks for the recognition. We run experiments on two databases of English digits and letters. The results show the effectiveness of our method.

Department Computer Engineering **Student's signature**.....
Field of Study Computer Science **Advisor's signature**.....
Academic year 2000 **Co-advisor's signature**.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างยิ่งของ อ.ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้ความรู้ คำแนะนำ ข้อคิดเห็น และข้อเสนอแนะต่าง ๆ ตลอดการวิจัยในครั้งนี้ ขอขอบคุณ อาจารย์นวลวรรณ สุนทรภิษัชและสมาชิก MIND LAB ที่ให้ความรู้เพิ่มเติมและข้อคิดเห็นต่าง ๆในระหว่างการทำวิจัย จึงขอขอบพระคุณทุกท่านเป็นอย่างสูงมา ณ ที่นี้ด้วย

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดา มารดา ครูอาจารย์ ทุกท่าน ซึ่งได้อบรมสั่งสอนและสนับสนุนผู้วิจัยเรื่อยมาจนสำเร็จการศึกษา

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ง
บทคัดย่อภาษาอังกฤษ..... จ
กิตติกรรมประกาศ..... ฉ
สารบัญ..... ช
สารบัญตาราง..... ฌ
สารบัญภาพ..... ญ
บทที่
1 บทนำ..... 1
1.1 ความสำคัญและความเป็นมาของปัญหา..... 1
1.2 วัตถุประสงค์ของการวิจัย..... 2
1.3 ขอบเขตของการวิจัย..... 2
1.4 ขั้นตอนและวิธีดำเนินงานวิจัย..... 2
1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย..... 3
2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง..... 4
2.1 ทฤษฎีที่เกี่ยวข้อง..... 4
2.1.1 การเรียนรู้ของเครื่อง..... 5
2.1.2 การวิเคราะห์ตัวประกอบสำคัญ..... 6
2.1.3 การเฉลี่ยจุดภาพ..... 8
2.1.4 การวิเคราะห์สัญญาณแปรตามเวลา..... 9
2.1.5 การแปลงฟูเรียร์..... 10
2.1.6 แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก..... 11
2.2 งานวิจัยที่เกี่ยวข้อง..... 13
3 วิธีการรู้จำการอ่านริมฝีปาก..... 14
3.1 โครงสร้างของระบบ..... 14
3.1.1 ขั้นตอนการเรียนรู้..... 14
3.1.2 ขั้นตอนการรู้จำ..... 17
3.2 การประมวลผลขั้นต้น..... 18
3.2.1 การแยกคำออกจากคำพูดที่ต่อเนื่อง..... 18
3.2.2 การดึงลักษณะสำคัญ..... 20

สารบัญ (ต่อ)

บทที่

3.3 การเรียนรู้โดยนิวรอลเน็ตเวิร์ก..... 24
3.3.1 โครงสร้างนิวรอลเน็ตเวิร์ก..... 25
3.3.2 การเรียนรู้เพื่อสร้างนิวรอลเน็ตเวิร์ก..... 27
3.4 ขั้นตอนการรู้จำ..... 28
4 การทดลองและผลการทดลอง..... 29
4.1 การทดลองกับชุดข้อมูลที่ 1..... 29
4.2 การทดลองกับชุดข้อมูลที่ 2..... 31
4.3 การทดลองกับชุดข้อมูลที่ 3..... 32
4.4 สรุปผลการทดลองและเปรียบเทียบผลที่ได้กับงานวิจัยก่อนหน้า..... 33
5 สรุปการวิจัยและข้อเสนอแนะ..... 36
5.1 สรุปการวิจัย..... 36
5.2 ข้อเสนอแนะ..... 37
รายการอ้างอิง..... 38
ภาคผนวก 40
ภาคผนวก ก..... 41
ภาคผนวก ข..... 45
ภาคผนวก ค..... 47
ภาคผนวก ง..... 48
ประวัติผู้วิจัย..... 49

สารบัญตาราง

หน้า

ตารางที่

4.1 แสดงความถูกต้องจากการใช้วิธีวิเคราะห์ตัวประกอบสำคัญของข้อมูลชุดที่ 1..... 30
4.2 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพแบบช่วงของข้อมูลชุดที่ 1..... 30
4.3 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพโดยใช้กริดของข้อมูลชุดที่ 1..... 30
4.4 แสดงความถูกต้องจากการใช้วิธีวิเคราะห์ตัวประกอบสำคัญของข้อมูลชุดที่ 2..... 31
4.5 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพแบบช่วงของข้อมูลชุดที่ 2..... 31
4.6 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพโดยใช้กริดของข้อมูลชุดที่ 2..... 32
4.7 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพแบบช่วงของข้อมูลชุดที่ 3..... 32
4.8 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพโดยใช้กริดของข้อมูลชุดที่ 3..... 33
4.9 แสดงความถูกต้องจากงานวิจัยของ K. Yu, X. Jiang, and H. Bunke [10]..... 34

สารบัญภาพ

หน้า

รูปที่

2.1 รูปแบบการทำงานของการอ่านริมฝีปาก โดยใช้ HMMs..... 4
2.2 การมองเวกเตอร์แทนรูปภาพ..... 6
2.3 แสดงลำดับภาพของการพูดตัวอักษร H..... 8
2.4 ค่าความเข้มของจุดภาพเทียบกับเวลา..... 9
2.5 องค์ประกอบซิกมอยด์..... 12
3.1 ขั้นตอนในการเรียนรู้เพื่อใช้ในการรู้จำการอ่านริมฝีปาก..... 17
3.2 ขั้นตอนในการรู้จำการอ่านริมฝีปาก..... 17
3.3 แสดงค่าที่ได้จากการใช้สมการการหาค่าเฉลี่ยของผลต่างของภาพในแต่ละเฟรม. 19
3.4 แสดงผลที่ได้จากการใช้สมการเกาเซียนในการปรับฟังก์ชัน..... 19
3.5 แสดงตัวอย่างนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้ข้อมูลชุดที่ 1..... 25
3.6 แสดงตัวอย่างนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้ข้อมูลชุดที่ 2..... 26
3.7 แสดงตัวอย่างนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้ข้อมูลชุดที่ 3..... 26

บทที่ 1
บทนำ

1.1 ความสำคัญและความเป็นมาของปัญหา

ในการติดต่อสื่อสารกันนอกจากสัญญาณเสียงแล้วคนเรายังมองใบหน้าของผู้พูดซึ่งอาจจะเป็นส่วนริมฝีปากเพื่อใช้เป็นข้อมูลช่วยในการสื่อสารกรณีที่อยู่ในสภาพแวดล้อมที่มีเสียงรบกวน การใช้ข้อมูลภาพในการรับรู้คำพูดของคนเรา หรือที่เราเรียกว่าการอ่านริมฝีปาก (Lipreading) สามารถเพิ่มประสิทธิภาพในการรับรู้ของคนเราได้มากขึ้น นอกจากนี้คนที่ประสาทหูใช้การไม่ได้ยังสามารถเข้าใจได้โดยใช้การอ่านการเคลื่อนไหวของริมฝีปาก

โดยทั่วไปในการรู้จำเสียงพูดจะมีแหล่งที่มาของข้อมูล 2 ทางด้วยกันคือ สัญญาณเสียงและสัญญาณภาพโดยที่สัญญาณเสียงมาจากเสียงที่เปล่งโดยผู้พูดและสัญญาณภาพประกอบด้วยการเคลื่อนไหวของริมฝีปาก สีหน้า ท่าทาง โดยพื้นฐานแล้วแหล่งข้อมูลที่ใช้ในการรู้จำเสียงพูดมาจากสัญญาณเสียง การวิจัยเพื่อให้คอมพิวเตอร์รู้จำเสียงพูด ได้มีการค้นคว้าเป็นเวลานานแล้ว แต่ส่วนใหญ่จะมุ่งไปที่การใช้สัญญาณเสียงเท่านั้น ปัญหาที่พบในการวิจัยเหล่านั้นคือ ความถูกต้องในการรู้จำจะลดน้อยลงในสภาพแวดล้อมที่มีเสียงรบกวน

แหล่งที่มาของเสียงรบกวนอาจแบ่งได้ 3 อย่างคือจากไมโครโฟนที่ใช้ทดสอบ สิ่งแวดล้อมและเสียงรบกวนที่เกี่ยวข้องกับเสียงที่ใช้ทดสอบ เสียงรบกวนจากไมโครโฟนอาจเกิดจากข้อจำกัดในเรื่องแถบความถี่(frequency bandwidth) และคุณภาพของไมโครโฟน เสียงรบกวนจากสิ่งแวดล้อม อย่างเช่น สำนักงาน สนานบิน โรงงาน เสียงโทรศัพท์ดัง และเสียงรถวิ่งผ่านไปมา ส่วนเสียงรบกวนที่เกี่ยวกับเสียงที่ทดสอบอาจเกิดจากเสียงสะท้อนภายในห้องที่ใช้ทดสอบ

ดังนั้นเราจะใช้สัญญาณภาพซึ่งก็คือภาพการเคลื่อนไหวของริมฝีปากมาเป็นข้อมูลในการให้คอมพิวเตอร์รู้จำ ซึ่งจะไม่ถูกกระทบโดยสัญญาณเสียงรบกวน และ การเคลื่อนไหวของริมฝีปากยังสามารถที่จะให้ข้อมูลเพิ่มเติมสำหรับการแยกแยะคำพูดต่าง ๆ และที่สำคัญการใช้ข้อมูลทั้งสองส่วนคือข้อมูลเสียงและภาพร่วมกันก็จะสามารถเพิ่มความถูกต้องในการรู้จำเสียงพูดมากขึ้น

ในการอ่านริมฝีปากนั้นมีส่วนที่สำคัญสองส่วนหลักคือ การดึงลักษณะสำคัญ (feature extraction) และการสร้างตัวแยกแยะ (classifier) ในการดึงข้อมูลออกจากภาพมีที่สังเกตได้ 2 วิธีคือ การใช้ภาพโดยตรง (image based approach) และ การสร้างโมเดลเพื่อแทนภาพ (model based approach) ในการใช้ภาพโดยตรง ค่าความเข้มของภาพจะถูกผ่านขั้นตอนประมวลผลเบื้องตน โดยทั่วไปจะเป็นการทำฟิลเตอร์ และลดขนาดของภาพ ข้อได้เปรียบของวิธีนี้คือ ไม่มี

การทิ้งข้อมูลไป ส่วนข้อเสียเปรียบคือ เราจะต้อง แก้ปัญหาเรื่อง ตำแหน่ง ขนาด มุม และ แสงสว่าง ต่าง ๆที่เกิดขึ้นกับภาพ ส่วนข้อเสียเปรียบอื่นก็คือคุณลักษณะที่ได้มีขนาดใหญ่

ในการสร้างโมเดลเพื่อแทนภาพ โดยทั่วไปจะใช้ เส้นรูปร่างของริมฝีปากและมีการแทนรูปร่างด้วยเซตของ พารามิเตอร์ ข้อได้เปรียบคือ คุณลักษณะที่ได้อยู่ในมิติที่มีขนาดเล็ก และไม่ถูกกระทบจากการเปลี่ยน ตำแหน่ง ขนาด มุม และ แสงสว่าง ข้อเสียเปรียบคือ โมเดลที่ได้อาจไม่สามารถแทนข้อมูลคำพูดทั้งหมดได้

งานวิจัยเกี่ยวกับการอ่านริมฝีปากเท่าที่พบใช้ตัวแยกแยะหลัก ๆด้วยกันคือ ฮิดเดนมาร์คอฟโมเดล(Hidden Markov Models-HMMs) และ นิวรอลเน็ตเวิร์ก (Neural Networks-NNs) และวิธีการอื่น ๆ อย่างเช่น การแยกแยะโดยใช้ เทมเพลตแมตชิ่ง (template matching) การแยกแยะโดยใช้สถิติ โดยที่วิธีการเหล่านี้ส่วนใหญ่แล้วคุณลักษณะที่ได้มาจากแต่ละภาพและการเคลื่อนไหวของริมฝีปากจะนำไปเข้าโมเดล HMMs NNs และวิธีการอื่นที่คล้าย ๆกัน

ในงานวิจัยนี้เราดึงข้อมูลโดยการใช้ภาพโดยตรง โดยมองการเปลี่ยนแปลงของความเข้มของแต่ละจุดเทียบกับเวลาเป็นสัญญาณหลักโดยที่เราจะมีการลดข้อมูลภาพ 2 วิธีคือ (1)วิธีการวิเคราะห์ตัวประกอบสำคัญ(Principal Component Analysis-PCA) และ (2)การหาค่าเฉลี่ยความเข้มของจุดภาพ จากนั้นจะมีการใช้การแปลงแบบฟูเรียร์ (Fourier transform) เพื่อแทนสัญญาณ และจะดึงค่าสัมประสิทธ์ฟูเรียร์ (Fourier coefficients) [10][15][18] เพื่อใช้เป็นคุณลักษณะ (feature) ให้กับแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก (Back Propagation Neural Networks-BPNNs) [16] [17] สำหรับขั้นตอนการรู้จำ

1.2วัตถุประสงค์ของการวิจัย

เพื่อศึกษาและพัฒนาวิธีสำหรับการรู้จำการอ่านริมฝีปาก (Lipreading recognition) โดยใช้เทคนิคการวิเคราะห์สัญญาณแปรตามเวลาและนิวรอลเน็ตเวิร์ก

1.3ขอบเขตของการวิจัย

1.3.1 ข้อมูลที่ใช้เป็นข้อมูลที่เป็นรูปภาพโดยไม่รวมข้อมูลที่เป็นเสียง

1.3.2 รูปที่ใช้จะตัดเอาเฉพาะส่วนที่เป็นริมฝีปากของผู้พูด

1.3.3 คำที่ใช้ทดสอบมีลักษณะเป็นคำสั้น ๆ คำเดียวภาษาอังกฤษ โดยมีการทดสอบกับตัวเลข(0-9) และตัวอักษร (A-J)

1.3.4 ข้อมูลที่ใช้ในการทดสอบ เป็นข้อมูลที่เป็นการเคลื่อนไหวของริมฝีปากโดยเก็บเป็นเฟรมโดยในแต่ละเฟรมเป็นภาพเทา 256 ระดับ (gray scale)

1.3.4.1 ชุดข้อมูล Tulips1 [11][12] ซึ่งเป็นการพูด ตัวเลขภาษาอังกฤษ 1-4 โดยคน 12 คนโดยที่แต่ละคนพูดคนละ 2 ครั้ง และรูปแบบไฟล์รูปภาพที่ใช้เป็น PGM(Portable gray map)

1.3.4.2 ชุดข้อมูลจากการวิจัย ของ K. Yu, X. Jiang, and H. Bunke [10] ซึ่งเป็นการพูดตัวเลขภาษาอังกฤษ 0-9 โดยคน 2 คน โดยแต่ละคนพูดคนละ 19 ครั้งต่อหนึ่งตัว

1.3.4.3 ชุดข้อมูลจาก Computer Vision Lab , Computer Science Department,University of Central Florida , Orlando [19] ซึ่งเป็นการพูดตัวอักษรภาษาอังกฤษ10 ตัวจาก A – J โดยคนเดียว และแต่ละตัวอักษรมีการพูด 18 ครั้ง

1.4 ขั้นตอนและวิธีดำเนินงานวิจัย

1.4.1 ศึกษาแนวคิดและทฤษฎีการวิเคราะห์ตัวประกอบสำคัญ

1.4.2 ศึกษาแนวคิดการวิเคราะห์สัญญาณตามเวลา

1.4.3 ศึกษาแนวคิดและทฤษฎีการแปลงฟูเรียร์ (Fourier transform)

1.4.4 ศึกษาแนวคิดและทฤษฎีการเรียนรู้ของวิธีการแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก

1.4.5 ศึกษางานวิจัย การวิเคราะห์สัญญาณตามเวลาในการรู้จำการอ่านริมฝีปาก

1.4.6 พัฒนาวิธีการดึงลักษณะสำคัญจากภาพโดยวิธีการวิเคราะห์ตัวประกอบสำคัญหรือการหาค่าเฉลี่ยความเข้มของจุดภาพ

1.4.7 พัฒนาการแปลงฟูเรียร์เพื่อใช้เป็นโมเดลเพื่อใช้ในขั้นตอนการรู้จำ

1.4.8 ออกแบบโครงสร้างของแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กเพื่อใช้ในขั้นตอนการรู้จำ

1.4.8 สร้างแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กเพื่อใช้ในขั้นตอนการรู้จำ

1.4.9 นำแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กมาทำการทดสอบหาอัตราการรู้จำการอ่านริมฝีปากกับกลุ่มตัวอย่างที่ใช้ในการทดสอบ

1.4.11 สรุปผลการวิจัย ข้อเสนอแนะ พร้อมทั้งแนวทางการวิจัยต่อไป

1.5 ประโยชน์ที่คาดว่าได้รับจากงานวิจัย

1.5.1 ได้เข้าใจถึงเทคนิควิธีในการทำการรู้จำการอ่านริมฝีปาก

1.5.2 สามารถนำไปรวมกับการรู้จำด้วยเสียงเพื่อความถูกต้องมากขึ้น

1.5.3 เป็นแนวทางในการพัฒนาการรู้จำในเรื่องอื่น

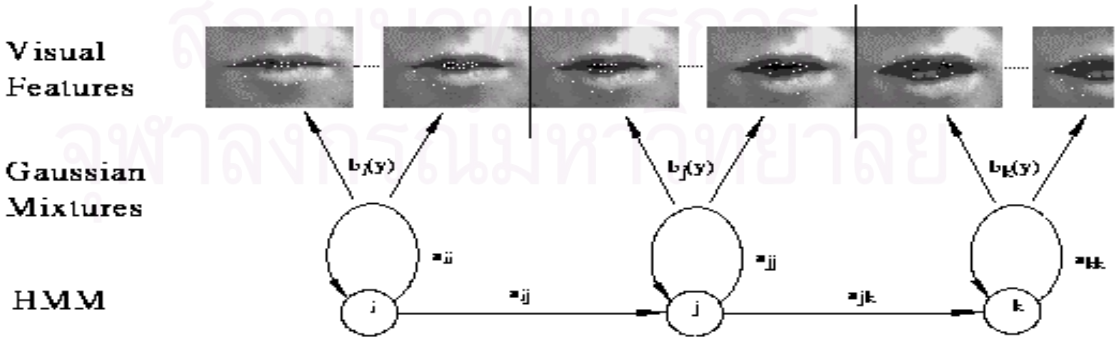
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะได้กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องในงานวิจัยนี้ โดยทฤษฎีที่เกี่ยวข้อง มีเนื้อหาเกี่ยวกับทฤษฎีที่ใช้ในงานวิจัย ซึ่งประกอบด้วย การเรียนรู้ของเครื่อง การวิเคราะห์ตัวประกอบสำคัญ การหาค่าเฉลี่ยความเข้มของจุดภาพ การวิเคราะห์สัญญาณตามเวลา การแปลงฟูเรียร์ และ แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กส่วนงานวิจัยที่เกี่ยวข้องประกอบด้วยงานวิจัยในด้านการรู้จำการอ่านริมฝีปากซึ่งเท่าที่พบ มีวิธีการที่ใช้หลักๆ ด้วยกันคือ ฮิดเดนมาร์คอฟโมเดล (Hidden Markov Models-HMMs) และ นิวรอลเน็ตเวิร์ก (Neural Networks-NNs) และวิธีการอื่นๆ อย่างเช่น การแยกแยะโดยใช้ เทมเพลตแมตชิง (template matching) การแยกแยะโดยใช้สถิติ

2.1 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องในงานวิจัยนี้ประกอบด้วย การเรียนรู้ของเครื่อง (Machine Learning) การวิเคราะห์ตัวประกอบสำคัญ (Principal Component Analysis) การหาค่าเฉลี่ยความเข้มของจุดภาพ การวิเคราะห์สัญญาณตามเวลา (Time-Varying Signal Analysis) การแปลงฟูเรียร์ (Fourier transform) และวิธีการแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก (Back Propogation Neural Network)

จากงานวิจัยที่ผ่านมาเกี่ยวกับการอ่านริมฝีปากส่วนใหญ่จะใช้วิธี HMMs ,NNs และอื่นๆ โดยที่วิธีการเหล่านี้ส่วนใหญ่แล้วคุณลักษณะที่ได้มาจากแต่ละภาพและการเคลื่อนไหวของริมฝีปากจะนำไปเข้าโมเดล HMMs , NNs และวิธีการอื่นที่คล้ายๆกัน ซึ่งสามารถแสดงได้ดังรูป



รูปที่ 2.1 รูปแบบการทำงานของการอ่านริมฝีปาก โดยใช้ HMMs

สำหรับการอ่านริมฝีปากโดยการใช้เทคนิคการวิเคราะห์สัญญาณแปรตามเวลาและนิวรอลเน็ตเวิร์ก จะทำการปรับปรุงวิธีการของ K. Yu, X. Jiang, and H. Bunke [10] โดยการนำเทคนิคการเรียนรู้แบบนิวรอลเน็ตเวิร์กมาใช้ในการรู้จำ รวมทั้งมีการลดข้อมูลภาพที่มีขนาดใหญ่โดยการใช้เทคนิคการวิเคราะห์ตัวประกอบสำคัญ(Principal Component Analysis) หรือเรียกอีกชื่อหนึ่งว่า การแปลงแบบเค-เอล (K-L Transform) [13][14][15] และอีกวิธีหนึ่งก็คือการเฉลี่ยจุดภาพ

2.1.1 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องคือการพยายามทำให้คอมพิวเตอร์สามารถพัฒนาความขีดความสามารถของตัวเองได้ โดยใช้ประสบการณ์ที่ได้จากการเรียนรู้จากภายนอกหรือจากผู้สอน ได้มีการศึกษาถึงเรื่องการเรียนรู้ของเครื่องมีอย่างแพร่หลาย แต่ในปัจจุบันการเรียนรู้ของเครื่องยังไม่สามารถทำได้ดีเท่ากับการเรียนรู้ของมนุษย์ แต่ก็ได้มีการคิดค้นและศึกษาเพื่อหา วิธี และขั้นตอนใหม่ๆ เพื่อให้สามารถใช้งานกับการเรียนรู้หลายๆ แบบได้อย่างมีประสิทธิภาพ ในด้านการประยุกต์ใช้ โปรแกรมหลายโปรแกรมใช้ประโยชน์จากการพัฒนาให้ใช้การเรียนรู้ได้เป็นอย่างดี เช่น การรู้จำเสียงพูด การทำนายอัตราการฟื้นตัวของผู้ป่วยโรคปอดอักเสบ การสร้างกลยุทธ์หลากหลายในการเล่นเกม สำหรับในด้านการศึกษาแขนงใหม่ ได้มีการศึกษาถึงวิธีการเรียนรู้ของเครื่องในด้านต่างๆ มากขึ้น เช่นวิธีการทำเหมืองข้อมูล (Data Mining) เป็นวิธีการหนึ่งซึ่งพัฒนาขึ้นจากการเรียนรู้ของเครื่องเพื่อทำ การค้นพบความรู้ (Knowledge Discovery) ที่ซ่อนอยู่ภายในระบบฐานข้อมูลที่มีขนาดใหญ่ เราจะเห็นได้ว่าวิธีการเรียนรู้ของเครื่องได้กลายเป็นปัจจัยสำคัญอย่างยิ่งในการพัฒนาโปรแกรมคอมพิวเตอร์ในปัจจุบัน

ตัวอย่างการนำการเรียนรู้ของเครื่องไปใช้งานในด้านต่างๆได้อย่างมีประสิทธิภาพ

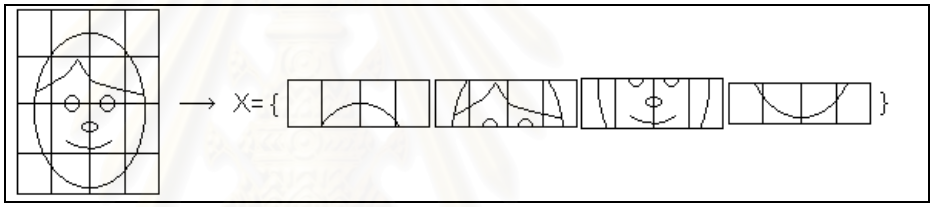
- การเรียนรู้เพื่อทำการรู้จำเสียงพูด
  ระบบที่ทำการรู้จำเสียงพูดที่ประสบความสำเร็จส่วนใหญ่ได้นำวิธีการเรียนรู้ของเครื่องไปใช้ในรูปแบบต่างๆ เช่น ระบบ SPHINX [20] ได้ทำการเรียนรู้เพื่อการทำการรู้จำองค์ประกอบพื้นฐานของเสียงพูด (phonemes) และคำศัพท์ (words) จากสัญญาณเสียงที่ได้ทำการเรียนรู้ไว้ก่อนจากนั้นใช้ นิวรอลเน็ตเวิร์ก และแบบจำลองฮิดเด็นมาคอฟ เพื่อทำการรู้จำเสียงของผู้พูดแต่ละคน
- การเรียนรู้เพื่อบังคับพาหนะโดยอัตโนมัติ
  ใช้การเรียนรู้ของเครื่องเพื่อทำการสอนรถยนต์ที่ควบคุมด้วยเครื่องคอมพิวเตอร์ ให้ควบคุมได้อย่างถูกต้องเมื่อขับเคลื่อนอยู่บนถนนหลายรูปแบบ ดังตัวอย่างเช่น ระบบ ALVINN [21] สามารถทำการขับเคลื่อนรถ

ยนต์ด้วยความเร็ว 70 ไมล์ต่อชั่วโมง เป็นระยะทาง 90 ไมล์ บนทางหลวงในขณะที่มีรถยนต์คันอื่นกำลังเคลื่อนที่อยู่ด้วย

- การเรียนรู้เพื่อทำการจำแนกโครงสร้างใหม่ทางดาราศาสตร์
  มีการนำวิธีการเรียนรู้ของเครื่องไปประยุกต์ใช้เพื่อทำการค้นหาลักษณะพื้นฐานในระบบฐานข้อมูลขนาดใหญ่ เช่น องค์การอวกาศแห่งสหรัฐอเมริกา หรือ NASA ได้ใช้ขั้นตอนวิธีการเรียนรู้ต้นไม้การตัดสินใจ (decision tree learning algorithm) ทำการเรียนรู้เพื่อจำแนกวัตถุที่อยู่บนท้องฟ้าจากภาพที่มีขนาดใหญ่มาก [22]

2.1.2 การวิเคราะห์ตัวประกอบสำคัญ(Principal Component Analysis)

ในขั้นแรกจะนำภาพตัวอย่างทั้งหมดมาผ่านขั้นตอนประมวลผลขั้นต้นโดยการเก็บแต่ละภาพในรูปของเวกเตอร์ **x** ขนาด N โดยที่นำแต่ละจุดมาต่อกันดังรูป



รูปที่ 2.2 การมองเวกเตอร์แทนรูปภาพ

จากนั้นภาพทั้งหมดจะถูกนำไปหา ค่าเฉลี่ยเวกเตอร์ (mean vector) –$\mathbf{m}_x$ และโคแวเรียนซ์เมตริกซ์ (covariance matrix) $\mathbf{C}_x$ เนื่องจาก **x** มีขนาดเท่ากับ N ดังนั้น $\mathbf{C}_x$ จึงมีขนาด NxN ตาม สมการ ต่อไปนี้

$$\mathbf{m}_x = \frac{1}{t}\sum_{i=1}^{t}\mathbf{x}_i$$

$$\mathbf{C}_x = \frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k\mathbf{x}_k^T - \mathbf{m}_x\mathbf{m}_x^T$$

โดยที่ t เป็นจำนวนตัวอย่างทั้งหมด

จากนั้นจะใช้ โคแวเรียนซ์เมตริกซ์ $\mathbf{C}_x$ ที่ได้ไปหา ไอเกนเวกเตอร์เมตริกซ์ (eigenvectors matrix) –**A**(ขนาด NxN)

โดยนิยามถ้า C เป็น เมตริกซ์จัตุรัส(Square Matrix) ขนาด NxN เราสามารถหาค่าไอเกน(eigenValue–λ) และไอเกนเวกเตอร์(eigen vector)ได้ตามสมการ

$$Ce_i = \lambda e_i \quad \text{โดยที่ i=1,2,...,N}$$

โดยเราสามารถหาค่าไอเกนได้จากสมการ

$$|A - \lambda I| = 0$$

และ ไอเกนเวกเตอร์สามารถหาได้จาก $(A - \lambda I)x = 0$

ตัวอย่าง

$$A = \begin{bmatrix} 2 & 5 \\ 6 & 1 \end{bmatrix}$$

$$\det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 5 \\ 6 & 1-\lambda \end{vmatrix}$$

$$= \lambda^2 - 3\lambda - 28$$

$$= (\lambda - 7)(\lambda + 4)$$

ค่าไอเกนที่ได้คือ $\lambda = 7$ และ $\lambda = -4$

จากนั้นเราสามารถหาไอเกนเวกเตอร์ที่สอดคล้องกับ ค่าไอเกนแต่ละตัวได้โดยใช้สมการ $(A - \lambda I)x = 0$

ตัวย่างการหาไอเกนเวกเตอร์โดยใช้ค่าไอเกน = 7

$$\begin{bmatrix} 2 & 5 \\ 6 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 7\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

จากเราจะได้สมการ 2 ตัวแปรคือ

$$2x + 5y = 7x$$
$$6x + y = 7y$$

เมื่อแก้สมการแล้วเราจะได้ $x = y$ ซึ่งหมายความว่าค่า x และ y จะเป็นค่าใดก็ได้แต่ต้องมีค่าเท่ากันซึ่งเราจะมีไอเกนเวกเตอร์ได้หลาย ๆค่าและเราจะเลือกตัวที่นอร์มอลไลซ์

โดยการนำไอเกนเวกเตอร์ที่ได้จากค่าไอเกนที่เป็นค่ามากสุดเก็บไว้ที่แถวแรกและนำ ไอเกนเวกเตอร์ ซึ่งได้จาก ค่าไอเกน ค่าถัดมาเก็บไว้ที่แถวถัดไปตามลำดับ ดังนั้นแถวสุดท้ายจะเก็บ ไอเกนเวกเตอร์ ซึ่งได้จาก ค่าไอเกน ที่เป็นค่าที่น้อยที่สุด

จากนั้นจะนำ เมตริกซ์ – A ที่ได้ไปใช้สำหรับการแปลงด้วยวิธีการแปลงแบบ PCA ได้เวกเตอร์ **y** (ขนาด N) ตามสมการ $\mathbf{y = A(x - m_x)}$

แทนที่เราจะใช้ ไอเกนเวกเตอร์ ทุกตัวของ $\mathbf{C_x}$ เราจะใช้ ไอเกนเวกเตอร์ เพียง K ตัวแรกซึ่งได้จากค่าไอเกนที่มากที่สุด K ตัวแรก ดังนั้นเราจะได้ เมตริกซ์ – $\mathbf{A_K}$ ซึ่งมีขนาด KxN และเวกเตอร์ **y** ที่ได้จะมีขนาด K เนื่องจากค่าไอเกนจะลดลงเมื่อลำดับมีค่าสูงขึ้น ดังนั้นเราจะใช้เพียงแค่ ค่าไอเกน แรก ๆที่มีค่าสูงและจะทิ้งตัวหลัง ๆที่มีค่าน้อย ๆซึ่งถือว่าไม่มีความสำคัญมากนัก ดังนั้นเราสามารถลดขนาดข้อมูล จาก N มาเป็น K ได้

2.1.3 <u>การเฉลี่ยจุดภาพ</u>

เราเสนอวิธีในการเฉลี่ยจุดภาพ 2 วิธีคือ (1) การเฉลี่ยจุดภาพเป็นช่วงตามแนวความกว้างของภาพ และ (2) การเฉลี่ยจุดภาพโดยใช้กริด

2.1.3.1 <u>การเฉลี่ยจุดภาพเป็นช่วง</u>

เราจะนำค่า ๆหนึ่งมาเฉลี่ยภาพเป็นช่วง ๆ เพื่อลดขนาดของภาพให้เล็กลง ถ้าเรามีภาพ ขนาด N จุด และต้องการหาค่าเฉลี่ยทุก ๆ M จุด นั่นคือในทุก ๆ M จุดเราจะหาผลรวมของความเข้มของจุดภาพ จากนั้นนำไปหารด้วย M ดังนั้นเราจะได้ค่าเฉลี่ยของจุดภาพทุก ๆ M ดังสมการ

$$\overline{X} = \frac{1}{M} \sum_{i=1}^{M} x_i$$

เพราะฉะนั้นขนาดของข้อมูลที่ได้จะลดลงเป็น N/M ยกตัวอย่างเช่น เรามีภาพในแต่ละเฟรมที่มีขนาด 48x38 ซึ่งมีจำนวนจุดทั้งหมด 1824 จุด ดังนั้น ค่าที่เป็นไปได้เพื่อนำไปเฉลี่ยจุดภาพ คือ2,4,6,8,12,16,24,48 โดยที่เราจะใช้ค่าที่หารความกว้างของภาพได้ลงตัวเท่านั้นเนื่องเราต้องการเฉลี่ยจุดภาพในแนวความกว้างของภาพ

2.1.3.2 <u>การเฉลี่ยจุดภาพโดยใช้กริด</u>

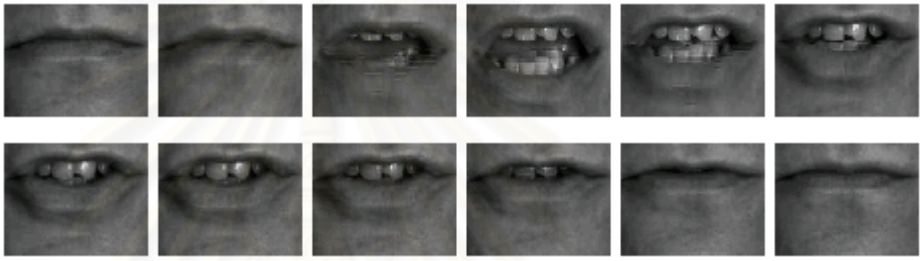
เราสร้างกริดสี่เหลี่ยม ขนาด wxh เพื่อใช้ในการเฉลี่ยจุดภาพภายในบริเวณกริด ถ้าเรามีภาพ ขนาด N จุด และต้องการหาค่าเฉลี่ยทุก ๆ wxh จุด นั่นคือในทุก ๆ wxh จุดเราจะหาผลรวมของความเข้มของจุดภาพ จากนั้นนำไปหารด้วย wxh ดังนั้นเราจะได้ค่าเฉลี่ยของจุดภาพ ดังสมการ

$$\overline{X} = \frac{1}{w * h} \sum_{i=1}^{h} \sum_{j=1}^{w} x_{ij}$$

เพราะฉะนั้นขนาดของข้อมูลที่ได้จะลดลงเป็น N/(w*h) ยกตัวอย่างเช่น เรามีภาพในแต่ละเฟรมที่มีขนาด 48x38 ซึ่งมีจำนวนจุดทั้งหมด 1824 จุด ดังนั้น ขนาดของกริดที่นำไปเฉลี่ยจุดภาพ คือ2x2,4x2,6x2,8x2,12x2,16x2,24x2,48x2โดยที่เราจะใช้ขนาดของกริดที่หารกว้างและความสูงของภาพได้ลงตัวเท่านั้น

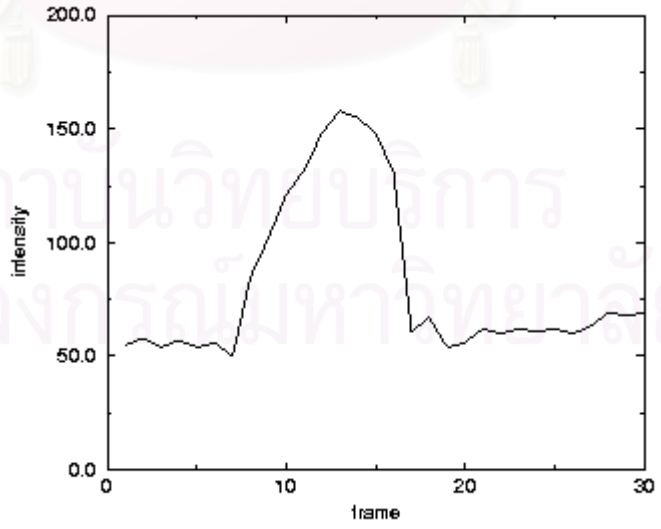
2.1.4 การวิเคราะห์สัญญาณแปรตามเวลา(Time-Varying Signal Analysis)

ข้อมูลที่เป็นการเคลื่อนไหวของริมฝีปากจะมีการเก็บเป็นเฟรมโดยในแต่ละเฟรมเป็นภาพเทา 256 ระดับ (gray scale) ตัวอย่างของลำดับภาพการเคลื่อนไหวของริมฝีปากแสดงได้ รูปที่ 2.3 ซึ่งเป็นลำดับภาพของการพูดตัวอักษร H



รูปที่ 2.3 แสดงลำดับภาพของการพูดตัวอักษร H

จากงานวิจัยของ K.Yu ,X.Jiang ,and H. Bunke [10]ได้มีการพิจารณาสัญญาณเทียบกับเวลาโดยที่พิจารณาความเข้มของจุดภาพ I(n) ของทุก ๆจุดในภาพ โดย n ที่เป็น ลำดับของภาพ ดังแสดงในรูปที่ 2.3



รูปที่ 2.4 ค่าความเข้มของจุดภาพเทียบกับเวลา

แต่เนื่องจากในงานวิจัยนี้จะทำการปรับปรุง โดยใช้เทคนิคการวิเคราะห์ตัวระกอบสำคัญเพื่อลดขนาดของข้อมูล หรือ ใช้วิธีหาค่าเฉลี่ยความเข้มของจุดภาพ ดังนั้นข้อมูลที่เราจะนำมาใช้ในการวิเคราะห์เทียบกับเวลาจึงไม่ใช่ข้อมูลที่เป็นความเข้มของจุดภาพโดยตรง แต่จะเป็นฟังก์ชันเทียบกับเวลาของข้อมูลที่ผ่านการทำ PCA หรือ ฟังก์ชันเทียบกับเวลาของข้อมูลที่ผ่านการเฉลี่ยความเข้มของจุดภาพ จากนั้นเราจะหาคุณลักษณะที่สามารถแทนฟังก์ชันโดยรวมได้ ซึ่งวิธีการที่ใช้คือ การแปลง ฟูเรียร์ [10][15][18] เพื่อลดขนาดของข้อมูลของสัญญาณ

2.1.5 การแปลงฟูเรียร์(Fourier Transform)

ถ้า f(x) เป็นฟังก์ชันต่อเนื่องของตัวแปรที่เป็นจำนวนจริง x ผลการแปลงฟูเรียร์ของฟังก์ชัน f(x) คือ

$$F(u) = \int_{-\infty}^{\infty} f(x)\exp[-j2\pi ux]dx$$

โดยที่ $j = \sqrt{-1}$

การแปลง ฟูเรียร์ ที่ได้จะอยู่ในรูปของตัวเลขเชิงซ้อน นั่นคือ

$$F(u) = R(u) + jI(u)$$

โดยที่ R(u) คือส่วนจำนวนจริง(real) และ I(u) คือส่วนจินตภาพ(imaginary)

โดยทั่วไปเราสามารถแทนสมการข้างต้นให้อยู่ในรูปของเอกซ์โปเนนเชียล (exponential)ได้ นั่นก็คือ

$$F(u) = |F(u)|e^{j\phi(u)}$$

โดยที่

$$|F(u)| = \sqrt{R^2(u) + I^2(u)}$$

และ

$$\phi(u) = \tan^{-1}\left[\frac{I(u)}{R(u)}\right]$$

เราได้แมกนิจูด ฟังก์ชัน |F(u)| ซึ่งเรียกว่า ฟูเรียร์สเปคตรัมของ f(x) และ $\phi(u)$ ซึ่งเรียกว่ามุมเฟส

เนื่องจากว่าฟังก์ชันที่เราใช้ไม่ใช่ ฟังก์ชันแบบต่อเนื่อง เราจึงต้องใช้การแปลงฟูเรียร์ แบบไม่ต่อเนื่อง(Discrete Fourier Transform–DFT)

$$c(k) = \frac{1}{N}\sum_{n=0}^{N-1} f(n)\exp[-j2\pi kn/N] \quad , \quad k=0,1,2,\ldots,N-1$$

โดยที่เรานำการแปลงฟูเรียร์แบบไม่ต่อเนื่องมาใช้กับฟังก์ชันที่เราได้เทียบกับเวลาโดย N คือจำนวนของเฟรมของลำดับภาพ และ c(k) คือค่าสัมประสิทธิ์เชิงซ้อน โดยที่เราจะใช้เพียงค่าสัมประสิทธิ์ในช่วงต้น ๆ จำนวนหนึ่ง มาประมาณค่าของฟังก์ชัน และจะใช้ค่า แมกนิจูด แทนที่จะใช้ ค่าสัมประสิทธิ์โดยตรง

2.1.6 <u>แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก(Backpropagation Neural Networks)</u>

แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กแบบหลายชั้น (multilayer backpropagation neural network) เป็นรูปแบบหนึ่งของนิวรอลเน็ตเวิร์ก ซึ่งมีความสามารถในการสร้างดิซิชันเซอร์เฟสแบบไม่เป็นเชิงเส้น(nonlinear decision surface) ซึ่งต่างจากเพอร์เซบตรอน(perceptron)ซึ่งเป็นดิซิชันเซอร์เฟสเชิงเส้น(linear decision surface)

นิวรอลเน็ตเวิร์กแบบหลายชั้นใช้ฟังก์ชันกระตุ้น (activation function)ที่สามารถหาค่าอนุพันธ์ได้ โดยใช้ องค์ประกอบซิกมอยด์(sigmoid unit) ดังรูปที่ 2.5 ซึ่งเอาต์พุต จะเป็นฟังชันต่อเนื่องของอินพุตดังนี้

$$o = \sigma(\vec{w}.\vec{x})$$

โดย

$$\sigma(y) = \frac{1}{1+e^{-y}}$$

เมื่อ

o – เอาต์พุต

$\vec{x}$ – อินพุต

$\vec{w}$ – ค่าน้ำหนักของอินพุตนั้น ๆ

σ – ฟังก์ชันซิกมอยด์ (sigmoid function)ซึ่งให้ค่าเอาต์พุตระหว่าง 0 และ 1

ขั้นตอนการปรับน้ำหนักจะทำโดยการใช้ เกรเดียนเดซเซนต์ (gradient descent) เพื่อหาค่าต่ำสุดของกำลังสองของค่าผิดพลาดระหว่างเอาต์พุตที่ได้จากเน็ตเวิร์กและค่าเป้าหมาย (target value) โดยมีขั้นตอนสำหรับปรับน้ำหนักดังต่อไปนี้

กำหนดให้ตัวอย่างที่ใช้ในการเรียนรู้แต่ละตัวอย่างอยู่ในรูป $(\vec{x}, \vec{t})$ เมื่อ $\vec{x}$ เป็นเวกเตอร์ของอินพุตของเน็ตเวิร์ก และ $\vec{t}$ เป็นเวกเตอร์ของเป้าหมายของเอาต์พุตของเน็ตเวิร์ก

η เป็นค่าอัตราการเรียนรู้ (learning rate)

อินพุตขององค์ประกอบ j ซึ่งมาจากองค์ประกอบ i แทนด้วย x_{ji} และค่าน้ำหนักขององค์ประกอบ j ซึ่งมาจากองค์ประกอบ i แทนด้วย w_{ji}

1. สร้างนิวรอลเน็ตเวิร์กตามโครงสร้างที่ต้องการ กำหนดจำนวนนิวรอลของแต่ละชั้น
2. กำหนดค่าน้ำหนักเริ่มต้นแบบสุ่มให้มีค่าน้อย ๆ (เช่น ระหว่าง -0.05 ถึง 0.05)
3. ทำการปรับค่าน้ำหนักด้วยขั้นตอนวิธีดังนี้

– สำหรับ$(\vec{x}, \vec{t})$ แต่ละตัวในตัวอย่างของการเรียนรู้ให้ทำดังนี้

1. อินพุต x *ให้กับเน็ตเวิร์กและคำนวณเอาต์พุต* o_u *ของทุก ๆ โหนด* u ในเน็ตเวิร์ก
   สำหรับเอาต์พุต k *แต่ละตัว* คำนวณค่าความผิดพลาด δ_k

$$\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k)$$

2. คำนวณค่าความผิดพลาด δ_h ของแต่ละโหนดในชั้นซ่อน h

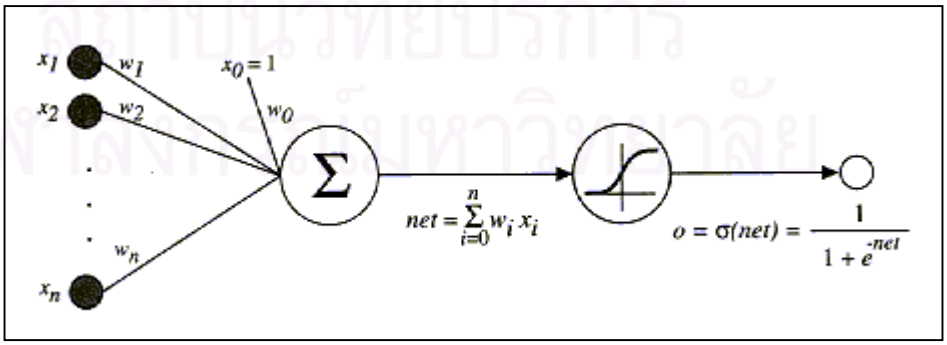
$$\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in outputs} w_{kh}\delta_k$$

3. ทำการปรับค่าน้ำหนัก w_{ji}

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$$

โดย

$$\Delta w_{ji} = \eta \delta_j x_{ji}$$



รูปที่ 2.5 องค์ประกอบซิกมอยด์ (sigmoid unit)

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยเกี่ยวกับการอ่านริมฝีปากเท่าที่พบมีวิธีการที่ใช้หลัก ๆด้วยกันคือ ฮิดเดนมาร์คอฟโมเดล(Hidden Markov Models-HMMs) และ นิวรอลเน็ตเวิร์ก Neural Networks-NNs) และวิธีการอื่น ๆ อย่างเช่น การแยกแยะโดยใช้ เทมเพลตแมตชิง (template matching) การแยกแยะโดยใช้สถิติ

งานวิจัยที่ใช้ HMMs

A. Adjoudani and C. Benoit [1] ใช้คุณลักษณะที่ได้จากริมฝีปากของผู้พูดซึ่งมีการทำให้เป็นสีฟ้า(เพื่อความง่ายในการเซ็กเมนต์) เพื่อใช้เรียนรู้และทดสอบ HMMs คุณลักษณะที่ใช้คือความกว้างและความสูงของริมฝีปาก โดยที่ข้อมูลที่ใช้ในการทดลองประกอบไปด้วยคำ 450 คำ โดยแต่ละคำจะมีการพูดซ้ำกัน 9 ครั้ง โดยมีการใช้ 7 ตัวอย่างในการเรียนรู้และเหลือไว้ 2 ตัวอย่างสำหรับการทดสอบ และมีความถูกต้อง 78%

Brooke et al. [2] ใช้ การวิเคราะห์ตัวประกอบสำคัญ (Principal Component Analysis-PCA) ในการลดข้อมูล โดยใช้สัมประสิทธิ์ 10 ตัวแรก แทนภาพแต่ละภาพ โดยที่ ข้อมูลที่ใช้ทดสอบ ประกอบด้วยตัวเลข 10 ตัวของภาษาอังกฤษ และใช้ HMMs ในการการเรียนรู้

Goldschem et al. [3] ใช้ HMMs สำหรับการเรียนรู้ โดยมีการใช้คุณลักษณะ 35 อย่าง ยกตัวอย่างเช่น พื้นที่ ความกว้าง ความสูง บริเวณช่องปาก ของผู้พูด และมีการใช้ PCA เพื่อลดขนาดข้อมูล เป็น 13 ตัว โดยที่ข้อมูลที่ใช้ในการทดลองประกอบด้วย ประโยค 50 ประโยค และมีการสุ่ม 150 ตัวอย่างเพื่อใช้ทดสอบ โดยส่วนที่เหลือใช้สำหรับการเรียนรู้ และมีความถูกต้อง 25.3%

Luettin and Thacker. [4] ดึงคุณลักษณะจากภาพโดยใช้ แอคทีฟเชพโมเดล (Active Shape Model-ASR) โดยคุณลักษณะที่ได้เป็นพารามิเตอร์ที่แทนรูปปาก จากนั้นนำคุณลักษณะที่ได้ไปทำการเรียนรู้โดยใช้ HMMs โดยข้อมูลที่ใช้ มีชื่อว่า Tulips ประกอบไปด้วย ลำดับของภาพของผู้พูด 12 คน โดยแต่ละคนพูดตัวเลขของภาษาอังกฤษ 2 ครั้ง ในการทดลองมีการเหลือตัวอย่างเดียวสำหรับทดสอบและส่วนที่เหลือใช้สำหรับการเรียนรู้(leave-one-out) และมีความถูกต้อง 88.5 %

งานวิจัยที่ใช้ NNs

Bregler et al. [5] ใช้มัลติสเตทไทม์ดีเลย์นิวรอลเน็ตเวิร์ก (Multi-state Time-delay Neural Network - MS-TDNN) ในการรู้จำโดยในขั้นแรกจะมีการใช้ การแปลงฟาสต์ฟูเรียร์ (Fast Fourier Transform - FFT) กับภาพขนาด 64x64 และใช้ ค่า ลอคแมกนิจูด (log magnitude) ของ FFT ของ 13x13 ตัวแรก เพื่อ ทำให้ค่าอยู่ในช่วง [-1.0,1.0] โดยที่ข้อมูลที่ใช้ประกอบด้วยลำดับของการสะกดตัวอักษรในภาษาเยอรมัน ซึ่งมีข้อมูล 2 ชุด โดยในชุดแรกใช้ 144 ตัว และชุดที่สองใช้ 350 ตัว ซึ่งสะกดโดยคน 2 คน ตามลำดับ และจะมีการแบ่งข้อมูลชุดแรกเป็นส่วนของการเรียนรู้ 75 ตัว และที่เหลืออีก 39 ตัวใช้สำหรับทดสอบ ส่วนในชุดที่สองจะใช้ 200 ตัวในการเรียนรู้และใช้อีก 150 ตัวที่เหลือเพื่อการทดสอบ และมีความถูกต้อง 50.2 %

P. Duchnowski, U. Meie, and A. Waibel [6] ใช้ MS-TDNN ในการรู้จำการสะกดตัวอักษร โดยมีการใช้ คุณลักษณะต่างๆกัน 4 แบบ ดังต่อไปนี้ คือ ใช้ สัมประสิทธิ์ PCA 32 ค่า ใช้ ลีเนียร์ดิสคริมิแนนต์(Linear Discriminants -LDs) 32 ค่า ใช้สัมประสิทธิ์ 29 ตัวของการแปลงฟูเรียร์แบบไม่ต่อเนื่อง 2 มิติ (2D Discreate Fourier Transform - 2D-DFT) และใช้ ค่าความเข้มโดยตรงของภาพ (gray level) ขนาด 24x16 ในการทดลองใช้ ภาษาเยอรมัน 200 คำ โดยแต่ละคำมีประมาณ 6 ตัวอักษร ซึ่งพูดเพียงคนเดียว และใช้ 170 ตัวอย่างสำหรับการเรียนรู้ และใช้ 15 ตัวอย่างสำหรับ cross-validation และใช้อีก 15 ตัวอย่างที่เหลือสำหรับการทดสอบ และมีความถูกต้องโดยใช้วิธี PCA 48% วิธี LDs 57% วิธี 2D-DFT 44% และวิธี Grey Level 49% ตามลำดับ

Wu et al. [7] ใช้ แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก (Back Propagation Neural Networks) สำหรับการรู้จำ สระ 5 ตัวของภาษาญี่ปุ่นโดยใช้ผู้พูด 14 คน โดยข้อมูลที่นำเข้าได้แก่ ค่าความเข้มของภาพเทา ภาพขาวดำ และ รูปร่างทางเรขาคณิต ตามลำดับ และมีความถูกต้อง 50%

งานวิจัยที่ใช้วิธีอื่นๆ

Kirby et al. [8] ใช้ PCA เพื่อลดข้อมูลของภาพ โดยภาพที่ใช้ประกอบไปด้วย P ลำดับ และแทนด้วย Matrix ขนาด QxP โดยที่แต่ละคอลัมน์ของ เมตริกซ์ แทนเวกเตอร์ที่โปรเจคชันบน Q ไอเกนลิปส์ (Eigenlips) และในการแยกแยะใช้วิธีเทมเพลตแมตชิง (template matching) โดยคำที่ใช้ในการทดสอบคือคำภาษาเยอรมัน 11 คำ ซึ่งพูดเพียงคนเดียว

Petajan et al. [9] ใช้พื้นที่ในส่วนของปากที่กำลังเปิดเพื่อสร้าง โค้ดบุค (codebook) โดยในขั้นแรกจะมีการให้ภาพปากเป็นภาพไบนารี ซึ่งมีแต่สีขาวกับดำ ดังนั้นในส่วนปากที่กำลังเปิดจะเป็นสีดำ จากนั้นจะมีการลดข้อมูลของภาพซึ่งมีขนาดใหญ่โดยการทำให้เป็นส่วนย่อย ๆ 255 ส่วน จากนั้นจึงเก็บแต่ละส่วนย่อยไว้ใน โค้ดบุค ซึ่งมีการเรียงลำดับตามขนาดและแยกแยะโดยใช้ค่า ดัชนี(index) จากนั้นจึงทำเวกเตอร์ควอนไทเซชั่นเพื่อแทนลำดับของภาพจากค่าดัชนีของภาพที่ใกล้เคียงที่สุดใน โค้ดบุค ในขั้นตอนการรู้จำมีการใช้การคำนวณเวกเตอร์ระยะห่างจากเทมเพลต ที่มีอยู่ และคำศัพท์ที่ใช้ประกอบด้วยตัวเลข 10 ตัวและตัวอักษร 26 ตัวซึ่งพูดเพียงคนเดียว และมีความถูกต้องในการรู้จำ 94%กับการทดสอบด้วยตัวเลข และ 81%สำหรับตัวอักษรตามลำดับ

K. Yu, X. Jiang, and H. Bunke [10] เสนอวิธีแบบใหม่โดยที่มีการพิจารณาสัญญาณของความเข้มของจุดภาพเทียบกับเวลา และมีการใช้การแปลง 2 แบบคือ เวฟเลต (Wavelet) และ ฟูเรียร์ (Fourier) แบบ 1 มิติ เพื่อลดข้อมูลของสัญญาณ และใช้ค่าสัมประสิทธิ์ของแต่ละแบบแทนคุณลักษณะเพื่อใช้ในการรู้จำ โดยแทนคุณลักษณะที่ได้เป็น เมตริกซ์ ขนาด h x w ซึ่งสมาชิกของ เมตริกซ์ เป็นเวกเตอร์ที่มีขนาด k โดยที่ h และ w แทนความสูงและความกว้างของภาพตามลำดับ และสมาชิกของ เมตริกซ์ แทนค่าสัมประสิทธิ์ k ตัวแรกของการแปลงแบบ เวฟเลต หรือค่า แมกนิจูด(Magnitude)ของสัมประสิทธิ์ k ตัวแรกของการแปลงแบบ ฟูเรียร์ และวิธีที่ใช้ในการแยกแยะมีการใช้รูปแบบทางสถิติซึ่งใช้ 5 แบบคือ Mean Distribution, Gaussian Distribution , Mahalanobis Distance, Nearest Neighbor และ Individual Nearest Neighbor ตามลำดับ โดยข้อมูลที่ใช้ในการทดลองมี 2 ชุด โดยในชุดแรกคือตัวเลขภาษาอังกฤษ 10 ตัว (0-9) ซึ่งพูดโดยคน 2 คน ซึ่งประกอบด้วยลำดับภาพ 19 บล็อค ต่อคน โดยในแต่ละบล็อค มีตัวเลข 10 ตัวต่อเนื่องกัน ส่วนข้อมูลชุดที่สองเป็นตัวอักษรภาษาอังกฤษ 10 ตัวตั้งแต่ A ถึง J และประกอบไปด้วย 18 บล็อค โดยแต่ละบล็อคมีตัวอักษร 10 ตัวต่อเนื่องกัน โดยในการทดสอบใช้วิธี เหลือหนึ่งตัวสำหรับการทดสอบ ส่วนที่เหลือใช้สำหรับการเรียนรู้ (leave one out)

บทที่ 3
วิธีการรู้จำการอ่านริมฝีปาก

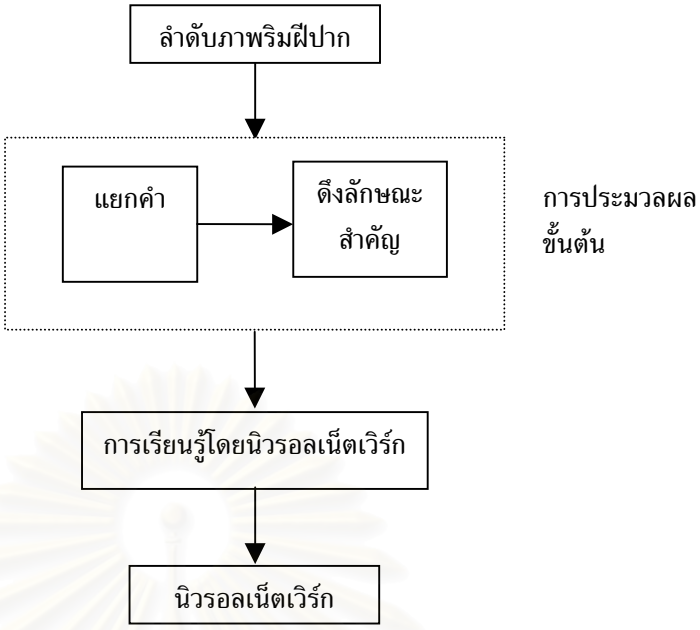
งานวิจัยนี้ใช้การวิเคราะห์สัญญาณตามเวลาและแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กในการรู้จำการอ่านริมฝีปาก ซึ่งมีขั้นตอนหลายขั้นตอน ในบทนี้จะได้กล่าวถึงโครงสร้างของระบบโดยรวมและรายละเอียดของแต่ละขั้นตอน การแยกคำ (word segmentation) การวิเคราะห์ตัวประกอบสำคัญ (Principal Component Analysis) การเฉลี่ยจุดภาพ การดึงลักษณะสำคัญ (feature extraction) การวิเคราะห์สัญญาณตามเวลา (Time-Varying Signal Analysis) การแปลงฟูเรียร์ (Fourier transform) และการเรียนรู้และรู้จำโดยวิธีการแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก

3.1 โครงสร้างของระบบ

หลักสำคัญของงานวิจัยนี้คือการเรียนรู้โดยใช้แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กจากกลุ่มตัวอย่างซึ่งเป็นลักษณะสำคัญที่ดึงได้จากภาพริมฝีปาก แล้วนำโครงสร้างนั้นไปใช้ในการรู้จำตัวอย่างอื่น ๆ แต่ก่อนที่จะนำภาพริมฝีปากมาผ่านขั้นตอนการเรียนรู้และรู้จำ จะต้องทำการประมวลผลขั้นต้นก่อน (preprocess) ก่อน ซึ่งขั้นตอนการประมวลผลขั้นต้นประกอบด้วยการแยกคำ การดึงลักษณะสำคัญ แล้วจึงนำลักษณะสำคัญที่ได้ไปใช้ในการเรียนรู้และรู้จำต่อไปโดยใช้แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก เนื่องจากข้อมูลที่เราใช้ในการทดลองบางส่วนเป็นลักษณะของคำพูดหลาย ๆคำต่อ ๆกันและมีการแยกโดยช่วงเวลาที่มีการหยุดพูดหรือปากปิดอยู่ซึ่งเราจะต้องแยกคำเหล่านี้เป็นคำ ๆก่อนเนื่องจากเราจะทำการเรียนรู้และรู้จำเป็นคำ ๆ จากนั้นจะเป็นการดึงลักษณะสำคัญและนำลักษณะสำคัญที่ได้ไปใช้ทำการเรียนรู้ และการรู้จำต่อไป ดังนั้นขั้นตอนหลักในงานวิจัยนี้มี 2 ขั้นตอน คือ ขั้นการตอนเรียนรู้ และขั้นตอนการรู้จำ

3.1.1 ขั้นตอนการเรียนรู้

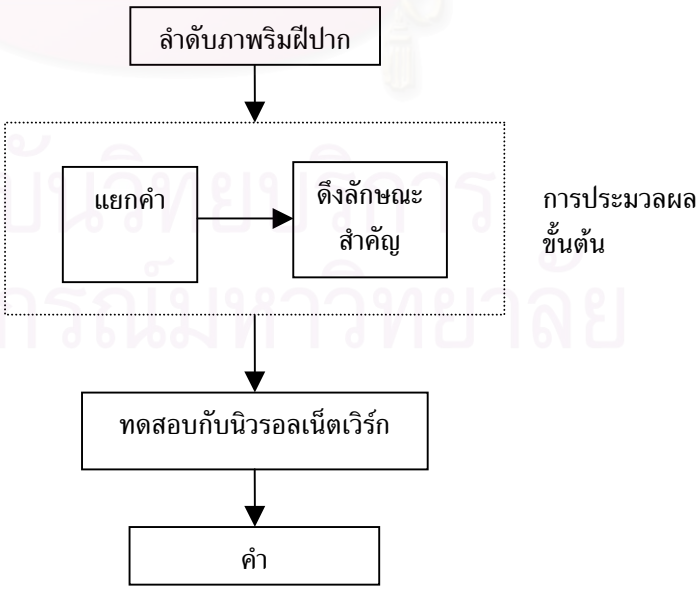
ในขั้นตอนการเรียนรู้ประกอบด้วยวิธีการแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก โดยลักษณะสำคัญที่ได้จากการประมวลผลขั้นต้นซึ่งจะมีการแบ่งเป็นขั้นตอนย่อย 2 ขั้นตอนคือการแยกคำและการดึงลักษณะสำคัญ จากนั้นจะถูกใช้เป็นตัวอย่างในการเรียนรู้ ดังรูปที่ 3.1



รูปที่ 3.1 ขั้นตอนในการเรียนรู้เพื่อใช้ในการรู้จำการอ่านริมฝีปาก

3.1.2 <u>ขั้นตอนการรู้จำ</u>

ขั้นตอนการรู้จำการอ่านริมฝีปากเมื่อสร้างนิวรอลเน็ตเวิร์กจากการเรียนรู้แล้ว เราสามารถนำนิวรอลเน็ตเวิร์กนี้ไปใช้ทำการรู้จำการอ่านริมฝีปากโดยตัวอย่างที่ใช้ในขั้นตอนนี้จะผ่านการประมวลผลขั้นต้นก่อน เช่นเดียวกับกระบวนการเรียนรู้ เพื่อให้ได้ลักษณะสำคัญ แล้วจึงนำลักษณะสำคัญที่ได้นี้ไปทำการทดสอบกับนิวรอลเน็ตเวิร์กที่ถูกสร้างไว้แล้ว ดังขั้นตอนในรูปที่ 3.2



รูปที่ 3.2 ขั้นตอนในการรู้จำการอ่านริมฝีปาก

3.2 การประมวลผลขั้นต้น

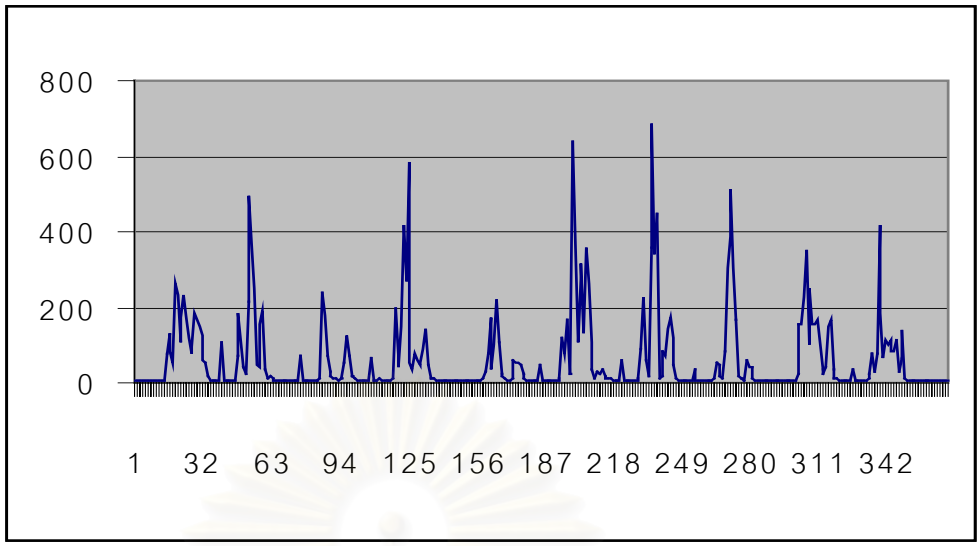
จุดประสงค์หลักของการประมวลผลขั้นต้นคือการหาลักษณะสำคัญของภาพริมฝีปากโดย ขั้นตอนในการประมวลผลขั้นต้นนี้ประกอบด้วย การแยกคำจากคำพูดที่มีความต่อเนื่อง และการดึงลักษณะสำคัญ โดยในการดึงลักษณะสำคัญเราจะมีการดึงข้อมูลออกจากภาพริมฝีปากในแต่ละเฟรม ซึ่งเราได้นำเสนอ 2 วิธีด้วยกันคือ วิธีการวิเคราะห์ตัวประกอบสำคัญ และวิธีการเฉลี่ยจุดภาพ จากนั้นเราจะนำข้อมูลที่ได้ไปทำการวิเคราะห์ให้อยู่ในรูปสัญญาณแปรตามเวลา และ ทำการแปลงสัญญาณที่ได้ให้อยู่ในรูปของสัมประสิทธิ์ของฟูเรียร์โดยเราจะดึงค่าแมกนิจูดของฟูเรียร์เพื่อแทนลักษณะสำคัญ จากนั้นข้อมูลเหล่านี้จะนำไปเป็นตัวอย่างให้กับนิวรอลเน็ตเวิร์กเพื่อการเรียนรู้และใช้เป็นข้อมูลเพื่อการรู้จำต่อไป

3.2.1 การแยกคำออกจากคำพูดที่ต่อเนื่อง[10][19]

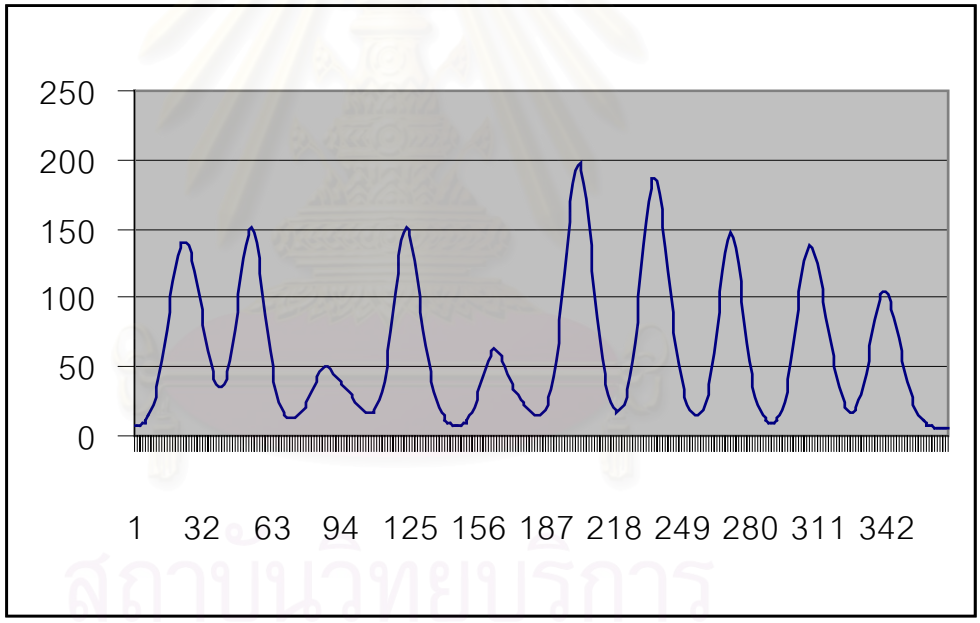
เนื่องจากในขั้นตอนการเรียนรู้และรู้จำเรามีการทดสอบกับคำพูดเป็นคำ ๆ และข้อมูลบางส่วนเป็นลักษณะของคำพูดหลาย ๆคำที่ต่อเนื่องกันและมีการแบ่งแยกโดยช่วงเวลาที่ปากปิดอยู่ ดังนั้นในเราจะมีการแยกคำเหล่านั้นออกมาเป็นคำ ๆ โดยวิธีที่เราใช้จะขึ้นอยู่กับการเปลี่ยนแปลงของลำดับภาพ เราใช้วิธีการหาค่าเฉลี่ยของผลต่างของภาพในแต่ละเฟรมดังสมการ

$$f(n) = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} (I_n(x,y) - I_{n-1}(x,y))^2$$

โดยที่ M และ N คือจำนวนแถวและจำนวนคอลัมน์ของภาพ และ $I_n(x,y)$ เป็นความเข้มของภาพ ลำดับที่ n ณ ตำแหน่ง (x,y) ถ้าเรานำค่าที่ได้จากการคำนวณโดยใช้สมการข้างต้นมาแสดงด้วยกราฟเราจะได้ดังภาพที่ 3.3 และซึ่งแสดงถึงค่าเฉลี่ยของผลต่างของภาพในแต่เฟรมของคำพูดต่อเนื่อง 10 คำ จะสังเกตได้ว่าช่วงกราฟที่มีค่าต่ำ ๆจะแสดงถึงช่วงที่ริมฝีปากปิดซึ่งก็คือเกิดการเปลี่ยนแปลงของความเข้มของภาพน้อยมาก



รูป 3.3 แสดงค่าที่ได้จากการใช้สมการการหาค่าเฉลี่ยของผลต่างของภาพในแต่ละเฟรม



รูป 3.4 แสดงผลที่ได้จากการใช้สมการเกาเซียนในการปรับฟังก์ชัน

จากรูปที่3.4 แสดงผลจากการใช้ สมการ เกาเซียน (Gaussian smoothed) เพื่อปรับฟังก์ชันให้มีความราบเรียบมากขึ้น ซึ่งเราจะสังเกตเห็นการแบ่งคำได้ชัดเจนมากขึ้นจากจุดต่ำสุดของฟังก์ชันในแต่ละช่วง จากการใช้จุดที่ได้เป็นตัวแบ่ง ลำดับภาพจะถูกแบ่งเป็นลำดับย่อยๆ ซึ่งจะสอดคล้องกับคำแต่ละคำ

3.2.2 <u>การดึงลักษณะสำคัญ</u>

ลำดับภาพที่ได้จากการแยกคำจะถูกนำมาวิเคราะห์เพื่อดึงลักษณะสำคัญของภาพนั้น ซึ่งลักษณะสำคัญที่ใช้ในงานวิจัยนี้ประกอบด้วย การคำนวณโดยวิธีการวิเคราะห์ตัวประกอบสำคัญและการเฉลี่ยจุดภาพซึ่งรายละเอียดของการดึงลักษณะสำคัญจากภาพ มีดังนี้

3.2.2.1 <u>การวิเคราะห์ตัวประกอบสำคัญ</u>

วิธีการที่เราใช้ในการคำนวณจะแตกต่างกับที่กล่าวถึงในบทที่ 2 เล็กน้อย เนื่องจากว่าวิธีนี้จะคำนวณได้เร็วกว่า และทำได้ง่ายกว่า ซึ่งจะได้ผลลัพธ์เหมือนกัน ซึ่งรายละเอียดการคำนวณจะเป็นดังนี้ คือ เราจะนำลำดับภาพทั้งหมดที่เตรียมไว้สำหรับขั้นตอนการเรียนรู้มาเปลี่ยนให้อยู่ในรูปของเวกเตอร์ ดังที่กล่าวไว้ในบทที่2 โดยที่เวกเตอร์ 1 ตัว จะแทนภาพ 1 ภาพ จากนั้นเรานำเวกเตอร์เหล่านี้มาสร้างเมตริกซ์ A ขึ้นมา โดยข้อมูลในแนวคอลัมน์จะแทนเวกเตอร์แต่ละตัวหรือเรียกอีกอย่างว่าคอลัมน์เวกเตอร์ และจำนวนแถวของเมตริกซ์จะมีขนาดเท่ากับจำนวนภาพทั้งหมดที่ใช้ในขั้นตอนการเรียนรู้

$$A=[u^1,u^2,...,u^t]$$

โดยที่ $u^i=(I_i(1,1),...I_i(M,N))$ เป็น คอลัมน์เวกเตอร์

I_i คือ ความเข้มของภาพ ที่จุดใด ๆในภาพและ M , N คือความกว้างและความสูงของภาพ

จากนั้นเราจะหา โคริเลย์ชันเมตริกซ์ (correlation matrix) $L=AA^T$ และหา ไอเกนเวกเตอร์ซึ่งมีนิยามว่า

$$L\phi_i=\lambda_i\phi_i,1\le i\le MN,$$

$$\lambda_1\le\lambda_2\le...\le\lambda_{MN}$$

และ

โดยที่ ϕ_i คือไอเกนเวกเตอร์ที่ i–th หรือเรียกอีกอย่างหนึ่งว่า ไอเกนลิป (Eigenlip) และ λ_i คือค่าไอเกนที่สอดคล้องกัน

ภาพใด ๆสามารถแทนให้อยู่ในรูปสมการผลรวมของไอเกนลิปได้ดังสมการ

$$u^x = \sum_{i=1}^{MN} a_i \phi_i$$

โดยที่

$$a_i = u_x^T \phi_i, 1 \le i \le MN$$

และเราจะใช้ สัมประสิทธิ์ เพียงแค่ n ตัวแรก โดยที่ n<<MN เพื่อแทนลักษณะสำคัญของ u^x

ในการคำนวณ PCA เมตริกซ์ A จะมีขนาด I x K โดยที่ I แทนจำนวนจุดทั้งหมดของภาพ และ K คือ จำนวนภาพทั้งหมด ดังนั้นในการหาขนาด ของ โครีเลย์ชันเมตริกซ์ $L = AA^T$ ที่ได้จะมีขนาดเท่ากับ I X I ถ้า I มีขนาดมากกว่า K มาก ๆ อย่างเช่น ในข้อมูลชุดที่ 1 Tulips1 [11][12] ขนาดของภาพคือ 100x75 ซึ่งจะได้ค่า I เท่ากับ 7500 ส่วนค่า K มีขนาดประมาณ 800(จำนวนภาพทั้งหมดสำหรับการเรียนรู้) ดังนั้นขนาดของ AA^T คือ I X I (7500x7500) ซึ่งจะเป็นการเสียเวลาและใช้หน่วยความจำมากในการคำนวณ ดังนั้นเราจะใช้วิธีการคำนวณที่มีประสิทธิภาพดีกว่า คือ เราจะคำนวณหา อินเนอร์โปรดัก(inner product) $A^T A$ แทนที่เราจะหาเอาต์เตอร์โปรดัก(outer product) AA^T และเราสามารถหาไอเกนเวกเตอร์ได้จาก อินเนอร์โปรดักนั้น

ถ้าให้

เมตริกซ์ V เป็น อินเนอร์โปรดัก $A^T A$

เมตริกซ์ Q เป็น ไอเกนเวกเตอร์เมตริกซ์ของ V

เมตริกซ์ P เป็นไอเกนเวกเตอร์เมตริกซ์ของเอาต์เตอร์โปรดัก AA^T และ เมตริกซ์ทะแยง (diagonal matrix) Λ เป็น เมตริกซ์ของค่าไอเกนของอินเนอร์โปรดักและมีค่าเท่ากับค่าไอเกนของเอาเตอร์โปรดัก

เราจะได้สมการในการหา ไอเกนเวกเตอร์เมตริกซ์ จาก อินเนอร์โปรดักคือ

$$P = A.Q.\Lambda^{-\frac{1}{2}}$$

เราสามารถตรวจสอบได้ว่า P ที่ได้จากสมการข้างต้น คือไอเกนเวกเตอร์ของ AA^T ดังต่อไปนี้

$$A * A^T * P = \Lambda * P$$

เนื่องจาก Q เป็นไอเกนเวกเตอร์เมตริกซ์ของ V จะได้ว่า

$$A^T * A * Q = \Lambda * Q$$

ฉะนั้น

$$A * A^T * A * Q * \Lambda^{-\frac{1}{2}} = \Lambda * A * Q * \Lambda^{-\frac{1}{2}}$$

และ

$$A * \Lambda * Q * \Lambda^{-\frac{1}{2}} = \Lambda * A * Q * \Lambda^{-\frac{1}{2}}$$

ดังนั้น P ที่ได้จากสมการข้างต้นคือไอเกนเวกเตอร์ของ AA^T ดังสมการข้างต้น

เราใช้เทคนิคนี้ในการคำนวณ PCA เฉพาะกับข้อมูลชุดที่ 1 Tulips1[11][12] ส่วนในข้อมูลชุดที่ 2 ขนาดของภาพคือ 48x38 ซึ่งจะได้ค่า I เท่ากับ 1824 ส่วนค่า K มีขนาดประมาณ 6000(จำนวนคำในการเรียนรู้คือ 180 คำ และแต่ละคำประกอบด้วยภาพประมาณ 35 เฟรม) ซึ่งจะเห็นว่าค่า I << K มากซึ่งเราจะไม่ใช้เทคนิคที่กล่าวถึงนี้กับข้อมูลชุดที่สอง เนื่องจากถ้าเราหาอินเนอร์โปรดักเราจะได้ขนาดของเมตริกซ์เป็น 6000x6000 แต่ขนาดของเอาต์เตอร์โปรดักเพียงแค่ 1824x1824 ส่วนในข้อมูลชุดที่ 3 ขนาดของภาพคือ190x150 ซึ่งจะได้ค่า I เท่ากับ 28500 ส่วนค่า K มีขนาดประมาณ 4000(จำนวนคำในการเรียนรู้คือ 170 คำ และแต่ละคำประกอบด้วยภาพประมาณ 25 เฟรม) ไม่ว่าเราจะหาไอเกนเวกเตอร์จาก อินเนอร์โปรดัก หรือ เอาต์เตอร์โปรดัก ก็ต้องใช้ เวลาในการคำนวณและหน่วยความจำสูง เราจึงไม่ได้ใช้ PCA ในการคำนวณ กับข้อมูลชุดที่ 3

หลังจากขั้นตอนการคำนวณ PCA แล้วภาพแต่ละภาพจะถูกนำมาแปลงโดยวิธีการแปลง PCA โดยนำข้อมูลความเข้มของภาพไปคูณด้วยไอเกนเวกเตอร์ ซึ่งจะทำให้ข้อมูลความเข้มถูกเปลี่ยนไปอยู่ในอีกมิติหนึ่ง ซึ่งเราจะเลือกเอาไอเกนเวกเตอร์ช่วงต้นๆ ที่สอดคล้องกับค่าไอเกนที่สูง ๆมาใช้ ทำให้สามารถลดขนาดข้อมูลลงได้

3.2.2.2 <u>การเฉลี่ยจุดภาพ</u>

ในการเฉลี่ยจุดภาพนั้น เป็นวิธีที่ง่ายและคงรักษารูปร่างของภาพไว้ได้พอสมควร โดยเราจะมีการเฉลี่ยจุดภาพ 2 แบบคือ แบบช่วงตามความกว้างของภาพ และแบบใช้กริด เพื่อลดขนาดภาพ ภาพทั้งหมดที่เตรียมไว้สำหรับการเรียนรู้ จะถูกนำเฉลี่ยจุดภาพทีละภาพเพื่อนำไปใช้ในขั้นตอนต่อไป วิธีการเฉลี่ยจุดภาพนั้นเป็นการคำนวณเฉพาะในแต่ละภาพโดยที่ไม่เกี่ยวข้องกับข้อมูลอื่นๆ ซึ่งต่างจากการคำนวณโดยใช้ PCA ที่ต้องนำข้อมูลทั้งหมดของตัวอย่างสำหรับการเรียนรู้มาใช้ด้วย ดังนั้นในการคำนวณโดยวิธีเฉลี่ยจุดภาพจึงง่ายและเร็วกว่ามาก

ตัวอย่าง โปรแกรมวิธีการคำนวณโดยการเฉลี่ยจุดภาพแบบช่วง

```
ImageFeatures.setSize(NUM_OF_PIXELS/DIVIDER,NUM_FRAMES_WORD);
for(j=0;j< NUM_FRAMES_WORD;j++) // Read until Number of Frame
{
    ReadImage(data);      //Read Image Data
    for(k=0;k<NUM_OF_PIXELS;k+=DIVIDER)
    {
        fSum = 0;
        for(kk=0;kk<DIVIDER;kk++)
        {
                fSum = fSum + data[k+kk];   // Sum intensity
        }
        ImageFeatures.cell(k/DIVIDER,j) = fSum/DIVIDER;
    }
}
```

ตัวอย่าง โปรแกรมวิธีการคำนวณโดยการเฉลี่จุดภาพโดยใช้กริด

```
ImageFeatures.setSize(NUM_OF_PIXELS/DIVIDER,NUM_FRAMES_WORD);
for(j=0;j< NUM_FRAMES_WORD;j++) // Read until Number of Frame
{
        ReadImage(data);            //Read Image Data
        for(ii=0;ii<YSIZE;ii+=Y_GRID)
        {
           for(jj=0;jj<XSIZE;jj+=X_GRID)
           {
              fSum=0;
              for(row=ii;row<ii+Y_GRID;row++)
             {
                  for(col=jj;col<jj+X_GRID;col++)
                  fSum = fSum + g_img.data[XSIZE*row+col];
              }
                   ImageFeatures.cell(kk++,j) = fSum/DIVIDER;

           }
        }
}
```

3.2.2.3 <u>การวิเคราะห์สัญญาณแปรตามเวลาและการแปลงฟูเรียร์</u>

ข้อมูลที่ผ่านการทำ PCA หรือ การเฉลี่ยจุดภาพจะถูกนำมาวิเคราะสัญญาณแปรตามเวลาและแปลงโดยใช้การแปลงฟูเรียร์

เราใช้สมการการแปลง ฟูเรียร์ แบบไม่ต่อเนื่อง(Discrete Fourier Transform-DFT)

$$c(k) = \frac{1}{N}\sum_{n=0}^{N-1} f(n)\exp[-j2\pi kn/N] \quad , \quad k=0,1,2,\ldots,N-1$$

โดยที่เรานำการแปลงฟูเรียร์แบบไม่ต่อเนื่องมาใช้กับฟังก์ชันที่เราได้เทียบกับเวลา โดย N คือจำนวนของเฟรมของลำดับภาพ และ $c(k)$ คือค่าสัมประสิทธิ์เชิงซ้อน

การแปลง ฟูเรียร์ ที่ได้จะอยู่ในรูปของตัวเลขเชิงซ้อน นั่นคือ

$$c(u) = R(u) + jI(u)$$

โดยที่ c(u) คือส่วนจำนวนจริง(real) และ I(u) คือส่วนจินตภาพ(imaginary)

จำนวนค่าสัมประสิทธิ์ที่ใช้จะเลือกมาจำนวนหนึ่งในช่วงต้น ๆ เท่านั้นมาประมาณค่าของฟังก์ชัน และใช้ค่าแมกนิจูด แทนที่จะใช้ค่าสัมประสิทธิ์โดยตรง

$$|c(u)| = \sqrt{R^2(u) + I^2(u)}$$

3.3 การเรียนรู้โดยนิวรอลเน็ตเวิร์ก

เราจะแทนลักษณะสำคัญที่ได้ด้วยเมตริกซ์ C ขนาด KxN

$$C = [C_i]_{KxN}$$

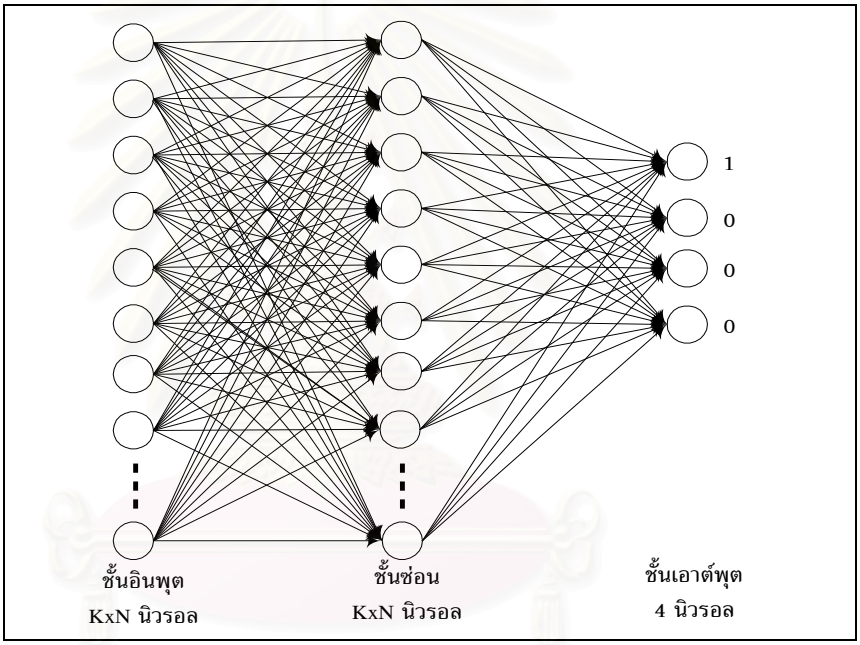
โดยที่ K คือขนาดของข้อมูลในแต่ละเฟรมของภาพ โดยในกรณีเราใช้วิธีการวิเคราะห์ตัวประกอบสำคัญค่านี้จะเป็นค่าที่สอดคล้องกับค่าไอเก้น K ตัวแรกที่มาค่าสูง ๆ ส่วนในกรณีที่เราใช้วิธีการเฉลี่ยจุดภาพค่านี้จะเป็นขนาดของข้อมูลที่ผ่านการเฉลี่ยจุดภาพแล้ว และ N คือค่า ของสัมประสิทธิ์ N ตัวแรกของ ฟูเรียร์ ส่วนค่าที่อยู่ในเมตริกซ์จะเป็นค่าแมกนิจูดที่ได้จากการแปลงฟูเรียร์ แทนที่เราจะใช้ค่าแมกนิจูดที่ได้โดยตรง เราจะมีการแปลงโดยใช้ ลอกการิทึม (logarithmic transformation) เพื่อลดความแตกต่างของสัมประสิทธิ์ที่ได้จากการคำนวณฟูเรียร์และทำให้การคำนวณในนิวรอลเน็ตเวิร์กลู่เข้า(converge)เร็วขึ้น ดังสมการข้างล่าง

$$c'_i = \begin{cases} \log(c_i + 1) \\ -\log(-c_i + 1) \end{cases} \begin{matrix} c_i \geq 0 \\ c_i < 0 \end{matrix}$$

3.3.1 <u>โครงสร้างนิวรอลเน็ตเวิร์ก</u>

จากนั้นเมตริกซ์ ที่ได้จะถูกนำเข้าสู่ขั้นตอนการตัดสินใจแยกแยะ (classification) โดยใช้แบ็คพรอพาเกชันนิวรอลเน็ตเวิร์ก (Backpropagation Neural Networks) โดยในชั้นอินพุตจะมีขนาดเท่ากับ KxN และในชั้นซ่อนจะใช้ขนาดน้อยกว่าหรือเท่ากับชั้น อินพุต และในชั้นเอาต์พุตจะมีขนาดขึ้นอยู่กับชุดตัวอย่างข้อมูลที่เราใช้ทดสอบโดยที่จะขึ้นอยู่กับจำนวนคำที่ใช้ทดสอบของข้อมูลชุดนั้น ๆ ซึ่งเรามีข้อมูลที่ใช้ทดสอบอยู่ 3 ชุดคือ

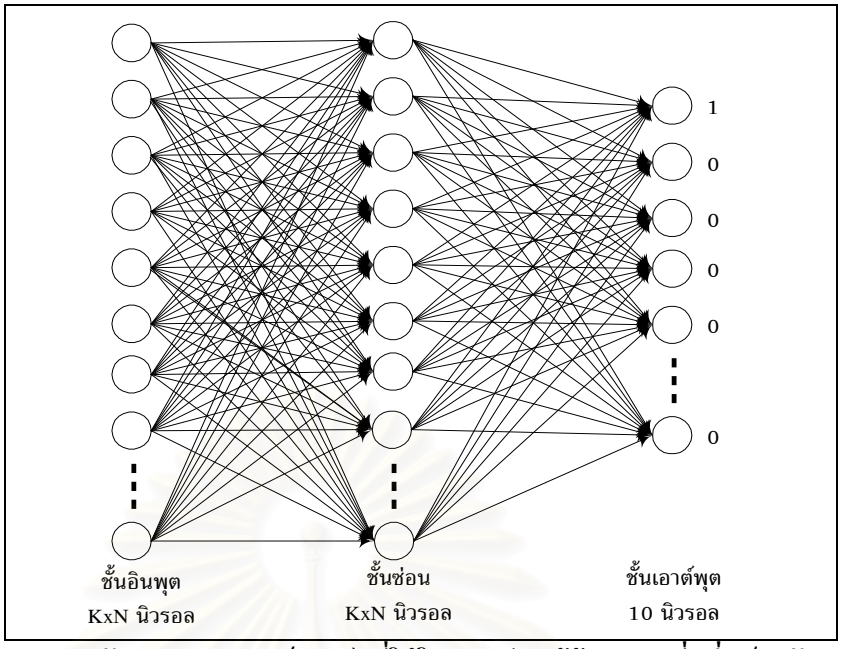
ชุดที่ 1 Tulips1[11][12] เป็นตัวเลขภาษาอังกฤษ 1-4 เราจะได้ ชั้นเอาต์พุต 4 นิวรอล ดังแสดงในรูปที่ 3.5



รูปที่ 3.5 แสดงตัวอย่างนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้ข้อมูลชุดที่1ซึ่งเป็นตัวเลข1-4

โดยในรูปจะเป็นโครงสร้างของคำพูดตัวเลข 0 ในขณะที่เราสอนนิวรอลเน็ตเวิร์กนั้นเราจะแทน โหนดบนสุดด้วย 1 ส่วนโหนดที่เหลือจะเป็น 0 หมด ส่วนในกรณีของการเรียนรู้คำพูดตัวเลข 1 เราจะแทนโหนด ที่ 2 ด้วย 1 ส่วนโหนดที่เหลือจะเป็น 0 หมด ตามลำดับ ในกรณีของการทดสอบนิวรอลเน็ตเวิร์กโหนดเอาต์พุตที่ให้ค่าสูงสุดจะถูกเลือกเป็นคำตอบ

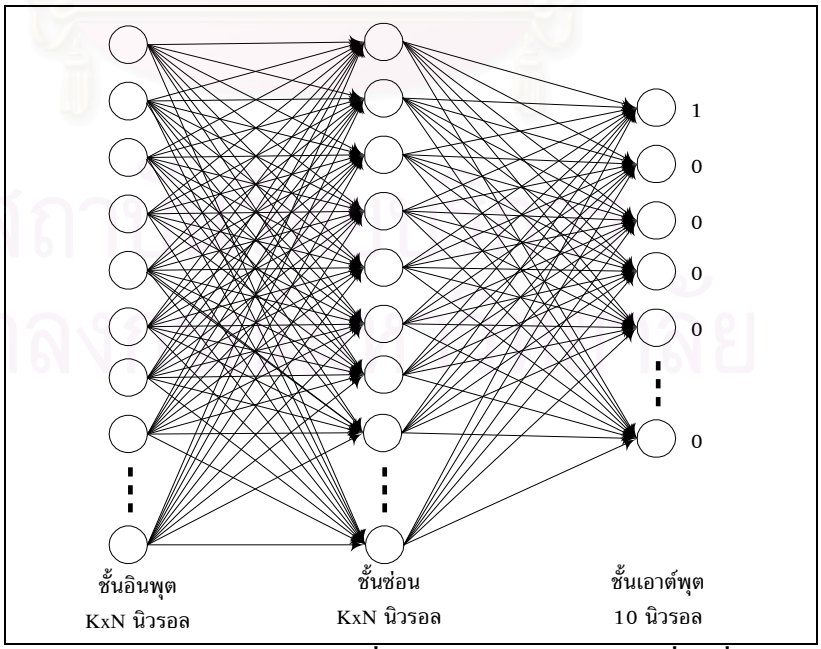
ชุดที่ 2 เป็นตัวเลขภาษาอังกฤษ 0-9 เราจะได้ ชั้นเอาต์พุต 10 นิวรอล ดังรูปที่ 3.6



รูปที่ 3.6 แสดงตัวอย่างนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้ข้อมูลชุดที่2ซึ่งเป็นตัวเลข 0-9

โดยในรูปจะเป็นโครงสร้างของคำพูดตัวเลข 0 ในขณะที่เราสอนนิวรอลเน็ตเวิร์กนั้นเราจะแทน โหนดบนสุดด้วย 1 ส่วนโหนดที่เหลือจะเป็น 0 หมด ส่วนในกรณีของการเรียนรู้คำพูดตัวเลข 1 เราจะแทนโหนด ที่ 2 ด้วย 1 ส่วนโหนดที่เหลือจะเป็น 0 หมด ตามลำดับ ในกรณีของการทดสอบนิวรอลเน็ตเวิร์กโหนดเอาต์พุตที่ให้ค่าสูงสุดจะถูกเลือกเป็นคำตอบ

ชุดที่ 3 เป็นตัวอักษรภาษาอังกฤษ A-J ซึ่งมี 10 ตัวเราจะได้ชั้นเอาต์พุต 10 นิวรอล



รูปที่ 3.7 แสดงตัวอย่างนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้ข้อมูลชุดที่ 3 ซึ่งเป็นตัวอักษร A - J

โดยในรูปจะเป็นโครงสร้างของคำพูดตัวอักษร A ในขณะที่เราสอนนิวรอลเน็ตเวิร์กนั้น เราจะแทน โหนดบนสุดด้วย 1 ส่วนโหนดที่เหลือจะเป็น 0 หมด ส่วนในกรณีของการเรียนรู้คำพูดตัวอักษร B เราจะแทนโหนด ที่ 2 ด้วย 1 ส่วนโหนดที่เหลือจะเป็น 0 หมด ตามลำดับ ในกรณีของการทดสอบนิวรอลเน็ตเวิร์กโหนดเอาต์พุตที่ให้ค่าสูงสุดจะถูกเลือกเป็นคำตอบ

3.3.2 <u>การเรียนรู้เพื่อสร้างนิวรอลเน็ตเวิร์ก</u>

ในการเรียนรู้เพื่อสร้างนิวรอลเน็ตเวิร์กด้วยขั้นตอนวิธีแบ็คพรอพาเกชัน มีขั้นตอนและการกำหนดข้อมูลที่ใช้ในการทดสอบดังนี้

ข้อมูลชุดที่ 1 ซึ่งเป็นการพูด ตัวเลขภาษาอังกฤษ 1-4 โดยคน 12 คนโดยที่แต่ละคน พูดคนละ 2 ครั้ง โดยเราจะมีการแบ่ง 11 คนไว้สำหรับการเรียนรู้ และ เหลือไว้ 1 คน สำหรับทดสอบ ดังนั้นในขั้นตอนการเรียนรู้จะมีตัวอย่างทั้งหมด 88 ตัวอย่าง ส่วนในการทดสอบจะมี 8 ตัวอย่าง

ข้อมูลชุดที่ 2 ซึ่งเป็นการ พูดตัวเลขภาษาอังกฤษ 0-9 โดยคน 2 คน ลักษณะของข้อมูลจะมีการเก็บเป็นบล๊อก 19 บล๊อก โดยในแต่ละบล๊อกจะเป็นคำพูดต่อเนื่อง 10 คำ โดยเราจะมีทดลองแยกแต่ละคน และในการทดลองเราเก็บไว้ 1 บล๊อกสำหรับทดสอบ ส่วนที่เหลือไว้สำหรับการเรียนรู้ ดังนั้นจึงมีตัวอย่างทั้งหมด 180 ตัวอย่างในขั้นตอนการเรียนรู้ และ 10 ตัวอย่างในการทดสอบ

ข้อมูลชุดที่ 3 ซึ่งเป็นการ พูดตัวอักษรภาษาอังกฤษ A-J โดยคน 1 คน ลักษณะของข้อมูลจะมีการเก็บเป็นบล๊อก 18 บล๊อก โดยในแต่ละบล๊อกจะเป็นคำพูดต่อเนื่อง 10 คำ และในการทดลองเราเก็บไว้ 1 บล๊อกสำหรับทดสอบ ส่วนที่เหลือไว้สำหรับการเรียนรู้ ดังนั้นจึงมีตัวอย่างทั้งหมด 170 ตัวอย่างในขั้นตอนการเรียนรู้ และ 10 ตัวอย่างในการทดสอบ

เมื่อจัดตัวอย่างทั้งสำหรับการทดลองเสร็จจึงป้อนให้กับเน็ตเวิร์กเพื่อทำการเรียนรู้เพื่อหาค่าน้ำหนักของลิงค์ (link) และไบแอส (bias) ของนิวรอน จะได้เป็นนิวรอลเน็ตเวิร์กที่เสร็จเรียบร้อย สามารถนำไปใช้รู้จำตัวอักษรได้ ต่อไป

3.4 ขั้นตอนการรู้จำ

การทดลองในงานวิจัยนี้ทำโดยการนำภาพตัวอย่างที่เตรียมไว้สำหรับทดสอบที่ได้มีการแบ่งไว้ก่อนหน้าแล้ว โดยมีการนำตัวอย่างทั้งหมดมาผ่านการประมวลผลขั้นต้นเพื่อหาลักษณะสำคัญของภาพ เช่นเดียวกับการประมวลผลขั้นต้นในการเรียนรู้และการสร้างเน็ตเวิร์ก ลักษณะสำคัญที่ได้จะต่างไปจากตัวอย่างที่ทำการเรียนรู้ จากนั้นจึงนำลักษณะสำคัญเหล่านี้ไปทำการทดสอบกับนิวรอลเน็ตเวิร์กที่สร้างขึ้นไว้แล้ว เพื่อทำการรู้จำว่าตัวอย่างเป็นคำพูดใด และเปรียบเทียบกับวิธีการอื่น ๆ ต่อไป

บทที่ 4
การทดลองและผลการทดลอง

ในบทนี้จะได้กล่าวถึงข้อมูลที่ใช้ในการเรียนรู้และการทดสอบ วิธีการเตรียมข้อมูลเหล่านี้และผลการทดลองเปรียบเทียบกับวิธีอื่น ข้อมูลที่ใช้ในการทดลอง เป็นข้อมูลแสดงการเคลื่อนไหวของริมฝีปาก ข้อมูลจะถูกเก็บเป็นเฟรม โดยในแต่ละเฟรมเป็นภาพเทา 256 ระดับ (gray scale) และแบ่งเป็น 3 ชุดดังต่อไปนี้

(1) ชุดข้อมูล Tulips1 [11][12] ซึ่งเป็นการพูด ตัวเลขภาษาอังกฤษ 1-4 โดยคน 12 คนโดยที่แต่ละคนพูดคนละ 2 ครั้ง และขนาดของภาพคือ 100x75

(2) ส่วนชุดข้อมูลชุดที่ 2 เป็นชุดข้อมูลที่ได้จากงานวิจัย ของ K. Yu, X. Jiang, and H. Bunke [10] ซึ่งเป็นการพูดตัวเลขภาษาอังกฤษ 0-9 โดยคน 2 คน โดยแต่ละคนพูดคนละ 19 ครั้งต่อหนึ่งตัว และขนาดของภาพคือ 48x38

(3) ชุดข้อมูลชุดที่ 3 เป็นชุดข้อมูลจาก Computer Vision Lab, Computer Science Department,University of Central Florida, Orlando [19] ซึ่งเป็นการพูดตัวอักษรภาษาอังกฤษ 10 ตัวจาก A – J โดยคนเดียว และแต่ละตัวอักษรมีการพูด 18 ครั้ง ขนาดของภาพคือ 190x150

ผลการทดลองกับข้อมูลทั้ง 3 ชุดเป็นดังด้านล่างนี้

4.1 การทดลองกับชุดข้อมูลที่ 1

เราแบ่งข้อมูลออกเป็นสองส่วนคือส่วนสำหรับการเรียนรู้และส่วนทดสอบ โดยที่จะเก็บข้อมูลของ 1 คนไว้สำหรับทดสอบ และที่เหลือ 11 คนไว้สำหรับการเรียนรู้ เพื่อให้ผลการทดลองไม่โน้มเอียงกับการแบ่งชุดข้อมูลเราจึงทำการทดสอบทั้งหมด 12 ครั้งโดยในแต่ละครั้งเลือกคนสำหรับทดสอบต่าง ๆกันไป ผลการทดลองโดยรวมจะเป็นค่าเฉลี่ยของการทดสอบทั้ง 12 ครั้ง ในการทดลองนี้ เราได้เปรียบเทียบวิธีดึงคุณลักษณะที่เสนอทั้ง 2 แบบ คือ การวิเคราะห์ตัวประกอบสำคัญ และการเฉลี่ยจุดภาพโดยใช้สัมประสิทธิ์ของฟูเรียร์เท่ากันเท่ากับ 5 ผลการทดลองแสดงในตารางที่ 4.1 และ 4.2

ตารางที่ 4.1 แสดงความถูกต้องจากการใช้วิธีวิเคราะห์ตัวประกอบสำคัญของข้อมูลชุดที่ 1

จำนวน สัมประสิทธิ์ไอเก้น	จำนวน ประสิทธ์ฟูเรียร์	จำนวน อินพุต โหนด	จำนวน ฮิดเดน โหนด	ความถูกต้อง (%)
40	**5**	**200**	**200**	**51.04**
60	5	300	300	50
80	5	400	400	48.95

ตารางที่ 4.2 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพแบบช่วงของข้อมูลชุดที่ 1

ความกว้าง ของช่วง	จำนวน คุณลักษณะจาก ภาพ	สัมประสิทธิ์ฟู เรียร์	จำนวน อินพุต โหนด	จำนวน ฮิดเดน โหนด	ความถูกต้อง (%)
25	**300**	**5**	**1500**	**1500**	**58.3**
50	150	5	750	750	50
100	75	5	375	375	55.2

ตารางที่ 4.3 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพโดยใช้กริดของข้อมูลชุดที่ 1

ขนาดของ กริด	จำนวน คุณลักษณะจาก ภาพ	สัมประสิทธิ์ฟู เรียร์	จำนวน อินพุต โหนด	จำนวน ฮิดเดน โหนด	ความถูก ต้อง (%)
5x5=25	300	5	1500	1500	60.41
10x5=50	150	5	750	750	61.45
20x5=100	**75**	**5**	**375**	**375**	**70.83**
25x5=125	60	5	300	300	66.66

ผลการทดลองแสดงให้เห็นว่า เมื่อค่าที่นำไปหารมีค่าเท่ากับ 25 หรือสัมประสิทธิ์ไอเก้นมีค่าเป็น 40 ให้ผลที่ดีที่สุดในแต่ละวิธี (แสดงด้วยตัวอักษรหนาในตาราง) ซึ่งในกรณีนี้แสดงให้เห็นว่าสำหรับวิธีการเฉลี่ยจุดภาพนั้น การเฉลี่ยจุดภาพโดยใช้กริดให้ผลดีที่สุด นอกจากนั้นความถูกต้องที่ได้จากวิธีการเฉลี่ยจุดภาพให้ผลที่ดีกว่าการใช้วิธีวิเคราะห์ตัวประกอบสำคัญ

4.2 การทดลองกับชุดข้อมูลที่ 2

ในการทดลองชุดนี้ ข้อมูลถูกแยกออกเป็นสองส่วนสำหรับผู้พูดคนที่หนึ่งกับคนที่สอง ข้อมูลของแต่ละคนจะถูกแบ่งไว้สำหรับการเรียนรู้และการทดสอบ โดยที่จะเก็บไว้ 1 ครั้งสำหรับทดสอบ และเหลือ 18 ครั้งไว้สำหรับการเรียนรู้ และทำการทดสอบทั้งหมด 19 ครั้งเพื่อไม่ให้ผลโน้มเอียงกับการแบ่งข้อมูลด้วยวิธีเดียวกันกับในการทดลองกับชุดข้อมูลที่ 1 เนื่องจากข้อมูลของแต่ละคนเป็นการพูดตัวเลข 10 ตัว ดังนั้นจึงมีการทดสอบทั้งหมด 190 ครั้ง แล้วนำค่าเฉลี่ยที่ได้เป็นผลการทดลองของข้อมูลของแต่ละคน ในการทดลองได้มี การทดลองเปรียบเทียบวิธีดึงคุณลักษณะ 2 แบบ คือ การวิเคราะห์ตัวประกอบสำคัญ และการเฉลี่ยจุดภาพ โดยที่สัมประสิทธิ์ของฟูเรียร์ใช้ค่าเท่ากับ 5 ผลที่ได้แสดงในตารางที่ 4.3 และ 4.4

ตารางที่4.4 แสดงความถูกต้องจากการใช้วิธีวิเคราะห์ตัวประกอบสำคัญของข้อมูลชุดที่ 2

จำนวนสัมประสิทธิ์ไอเก้น	จำนวนสัมประสิทธิ์ฟูเรียร์	จำนวนอินพุตโหนด	จำนวนฮิดเดนโหนด	ความถูกต้องสำหรับผู้พูดคนที่1 (%)	ความถูกต้องสำหรับผู้พูดคนที่2 (%)
30	5	150	150	97.89	75.78
50	**5**	**250**	**250**	97.89	**76.31**
60	5	300	300	96.31	73.68
90	**5**	**450**	**450**	**98.42**	75.26

ตารางที่ 4.5 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพแบบช่วงของข้อมูลชุดที่ 2

ความกว้างของช่วง	จำนวนคุณลักษณะจากภาพ	จำนวนสัมประสิทธิ์ฟูเรียร์	จำนวนอินพุตโหนด	จำนวนฮิดเดนโหนด	ความถูกต้องสำหรับผู้พูดคนที่1(%)	ความถูกต้องสำหรับผู้พูดคนที่2(%)
8	**228**	**5**	**1140**	**1140**	**100**	94.73
12	**152**	**5**	**760**	**760**	98.94	**95.78**
16	114	5	570	570	98.42	95.26
24	76	5	380	380	97.36	94.73
48	38	5	190	190	98.42	93.15

ตารางที่ 4.6 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพโดยใช้กริดของข้อมูลชุดที่ 2

ขนาดของกริด	จำนวนคุณลักษณะจากภาพ	จำนวนสัมประสิทธิ์ฟูเรียร์	จำนวนอินพุตโหนด	จำนวนฮิดเดนโหนด	ความถูกต้องของผู้พูดคนที่1 (%)	ความถูกต้องของผู้พูดคนที่2 (%)
4x2=8	**228**	**5**	**1140**	**1140**	**98.94**	**97.89**
6x2=12	**152**	**5**	**760**	**760**	**100.00**	96.31
8x2=16	**114**	**5**	**570**	**570**	**100.00**	96.31
12x2=24	76	5	380	380	99.47	95.78
24x2=48	38	5	190	190	97.89	94.73

ผลการทดลองได้ผลในทำนองเดียวกับการทดลองกับข้อมูลชุดที่ 1 กล่าวคือวิธีการเฉลี่ยจุดภาพให้ผลการทดลองที่ดีกว่าวิธีวิเคราะห์ตัวประกอบสำคัญและ การเฉลี่ยจุดภาพโดยใช้กริดให้ผลดีที่สุด สำหรับกรณีของวิธีวิเคราะห์ตัวประกอบสำคัญนั้นจำนวนสัมประสิทธิ์ไอเก้นที่มากคือ 90 ตัวให้ผลที่ดีที่สุดสำหรับผู้พูดคนที่ 1 และ จำนวนสัมประสิทธิ์ไอเก้น 50 ตัว สำหรับผู้พูดคนที่ 2

4.3 การทดลองกับชุดข้อมูลที่ 3

สำหรับการทดลองกับข้อมูลชุดที่ 3 ข้อมูลได้ถูกแบ่งไว้สำหรับการเรียนรู้และการทดสอบโดยที่จะเก็บไว้ 1 ครั้งสำหรับทดสอบและที่เหลือ 17 ครั้งไว้สำหรับการเรียนรู้ ทำการทดสอบทั้งหมด 18 ครั้ง โดยผลที่ได้จะเป็นค่าเฉลี่ยของ 18 ครั้ง ดังนั้นจึงมีการทดสอบทั้งหมด 180 ครั้ง เนื่องจากในแต่ละครั้งเป็นการพูดตัวอักษร10 ตัว ในการทดลองได้ใช้วิธีการเฉลี่ยจุดภาพอย่างเดียวเนื่องจากว่าขนาดของภาพที่ค่อนข้างใหญ่จึงทำให้ใช้วิธีการใช้การวิเคราะห์ตัวประกอบสำคัญต้องใช้หน่วยความจำสูงมากและเวลาในการคำนวณมาก จำนวนสัมประสิทธิ์ของฟูเรียร์ใช้ค่าเท่ากับ 5

ตารางที่ 4.7 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพแบบช่วงของข้อมูลชุดที่ 3

ความกว้างของช่วง	จำนวนคุณลักษณะจากภาพ	จำนวนสัมประสิทธิ์ฟูเรียร์	จำนวนอินพุตโหนด	จำนวนฮิดเดนโหนด	ความถูกต้อง (%)
95	300	5	1500	1500	91.11
190	**150**	**5**	**750**	**750**	**92.22**

ตาราง 4.8 แสดงความถูกต้องจากการใช้วิธีเฉลี่ยจุดภาพโดยใช้กริดของข้อมูลชุดที่ 3

ขนาดของกริด	จำนวนคุณลักษณะจากภาพ	จำนวนสัมประสิทธิ์ฟูเรียร์	จำนวนอินพุตโหนด	จำนวนฮิดเดนโหนด	ความถูกต้อง (%)
19x5=95	300	5	1500	1500	95.00
10x10=100	**285**	**5**	**1425**	**750**	**95.55**
10x15=150	190	5	950	950	94.44
19x10=190	150	5	750	750	93.88
19x15=285	100	5	500	500	93.88

ผลการทดลองแสดงให้เห็นว่าค่าความถูกต้องจากการเฉลี่ยจุดภาพโดยใช้กริดให้ผลดีที่สุด โดยที่ค่าที่ดีที่สุดคือกรณีที่จำนวนคุณลักษณะมีค่าเท่ากับ 285 และใช้ กริดขนาด 10x10

4.4 สรุปผลการทดลองและเปรียบเทียบผลที่ได้กับงานวิจัยของ ของ K. Yu, X. Jiang, and H. Bunke

การทดลองในหัวข้อที่ 4.1 และ 4.2 แสดงให้เห็นว่าผลที่ได้จากวิธีการเฉลี่ยจุดภาพดีกว่าการใช้วิธีการวิเคราะห์ตัวประกอบสำคัญ เราสามารถอธิบายได้ว่าการคำนวณของวิธีการวิเคราะห์ตัวประกอบสำคัญจะมีการแปลงข้อมูลความเข้มของจุดภาพไปอยู่อีกมิติหนึ่ง ซึ่งจะทำให้เราสูญเสียข้อมูลที่แสดงถึงการเปลี่ยนแปลงของความเข้มของจุดภาพ ณ เวลาใดๆ เนื่องจากข้อมูลของเรามีลักษณะเป็นเฟรมซึ่งแสดงการต่อเนื่องของภาพ แต่วิธีการเฉลี่ยจุดภาพยังคงข้อมูลที่เป็นความเข้มของจุดภาพอยู่แม้ว่าจะมีการลดขนาดด้วยค่าเฉลี่ยแล้วก็ตาม

เมื่อเปรียบเทียบผลของการทดลอง 4.1 4.2 และ 4.3 พบว่า การทดลองกับข้อมูลชุดที่ 1 ให้ความถูกต้องต่ำกว่าข้อมูลชุดอื่น ข้อมูลของการทดลองชุดที่ 1 นี้มีจำนวนของเฟรมในการพูดแต่ละคำมีค่อนข้างน้อยซึ่งโดยเฉลี่ยมีประมาณ 8 เฟรม ซึ่งต่างจากข้อมูลชุดที่ 2 และ 3 ซึ่งมีขนาดประมาณ 30 เฟรม ทำให้การนำไปหาสัมประสิทธิ์ฟูเรียร์มีความคลาดเคลื่อน และในข้อมูลชุดที่ 1 ไม่มีการควบคุมในเรื่องของแสงซึ่งต่างจากข้อมูลชุดที่ 2 และ 3 ซึ่งเป็นผลที่ได้ของการทดลอง 4.1 มีค่าน้อยกว่าการทดลองอื่น

ในวิธีการเฉลี่ยจุดภาพเราเสนอวิธี 2 วิธีด้วยกันคือ การเฉลี่ยแบบช่วงตามแนวความกว้างของภาพ และการเฉลี่ยโดยใช้กริด ในการเฉลี่ยแบบช่วงความกว้างของช่วงจะเป็นตัวเลขที่สามารถ

หารความกว้างของภาพได้ลงตัวเนื่องจากว่าเราใช้วิธีการเฉลี่ยจุดภาพในแต่ละแถว และในการเฉลี่ยโดยใช้กริดขนาดของกริดจะเป็นตัวเลขที่หารความกว้างและความสูงของภาพได้ลงตัว

เมื่อสังเกตจากผลการทดลอง 4.2 พบว่าเปอร์เซ็นต์ความถูกต้องที่ได้ผู้พูดคนที่ 1 ดีกว่าผู้พูดคนที่ 2 เนื่องจากว่า ในระหว่างการบันทึกภาพ ผู้พูดทั้งสองคนมีการเคลื่อนที่ศีรษะต่างกัน โดยที่คนที่ 1 ค่อย ๆเคลื่อนศีรษะจากซ้ายไปขวา ในขณะที่คนที่ 2 ค่อย ๆเคลื่อนศีรษะลง การเคลื่อนที่ศีรษะลงในแนวดิ่งทำให้เกิดการเปลี่ยนแปลงความเข้มเป็นพื้นที่กว้างมากกว่าการเคลื่อนที่ในแนวราบ ดังนั้นการเคลื่อนที่ของศีรษะในแนวดิ่งจะมีผลต่อเปอร์เซ็นต์ความถูกต้องอย่างมากกับวิธีการนี้ ผลที่ได้ สนับสนุนผลการการทดลองของ K. Yu, X. Jiang, and H. Bunke [10] เช่นกัน ซึ่งเราจะเห็นได้จากหัวข้อต่อไปจากตารางที่ 4.9 จะเห็นว่า เปอร์เซ็นต์ความถูกต้องของผู้พูดคนที่ 2 น้อยกว่าของคนที่ 1

เนื่องจากข้อมูล (4.2 และ 4.3) ที่ใช้ทดสอบได้มาจากงานวิจัยของ K. Yu, X. Jiang, and H. Bunke [10] เราจึงทำการเปรียบเทียบผลที่ได้กับผลการทดลองจากงานวิจัย โดยใช้ข้อมูลชุดเดียวกันและการแบ่งข้อมูลสำหรับการเรียนรู้และการทดสอบแบบเดียวกัน และใช้สัมประสิทธิ์ฟูเรียร์เท่ากันเท่ากับ 5 วิธีที่ใช้ในการแยกแยะมีการใช้รูปแบบทางสถิติซึ่งใช้ 4 แบบคือ Mean Distribution, Gaussian Distribution, Mahalanobis Distance และ Nearest Neighbor ตามลำดับ ผลที่ได้แสดงในตารางที่ 4.9

ตารางที่ 4.9 แสดงความถูกต้องจากงานวิจัยของ K. Yu, X. Jiang, and H. Bunke [10]

วิธีการของ ของ K. Yu, X. Jiang, and H. Bunke	ความถูกต้องจาก ผู้พูดคนที่1 ของข้อมูลชุดที่ 2(%)	ความถูกต้องจาก ผู้พูดคนที่2(%) ของข้อมูลชุดที่ 2(%)	ความถูกต้องจาก ข้อมูลชุดที่ 3(%)
Mean	98.42	87.89	87.77
Gaussian	49	40	44
Mahalanobis	95	76	83
Nearest Neighbor	**99.47**	**93**	**91**
วิธีการที่นำเสนอที่ได้ผลที่ดีที่สุดโดยใช้การเฉลี่ยจุดภาพโดยใช้กริด	**100**	**97.89**	**95.55**

ตารางที่ 4.9 แสดงให้เห็นว่าวิธีการ Nearest Neighbor ให้ความถูกต้องดีที่สุด (แสดงด้วยตัวอักษรหนาในตาราง) คือ 99.47% 93% และ 91% สำหรับผลที่ได้จากผู้พูดคนที่1 ผลที่ได้จากผู้พูดคนที่2 และผลที่ได้จากข้อมูลชุดที่ 3 ตามลำดับ ส่วนผลการทดลองโดยวิธีการที่นำเสนอด้วยวิธีการเฉลี่ยจุดภาพโดยใช้กริดให้ความถูกต้องดีที่สุดคือ 100% 97.89% และ 95.55% สำหรับข้อมูล ผู้พูดคนที่1 ผู้พูดคนที่2 และ ข้อมูลชุดที่ 3 ตามลำดับ ผลที่ได้นี้แสดงให้เห็นถึงเปอร์เซ็นต์ความถูกต้องที่สูงขึ้นเมื่อเทียบกับวิธี Nearest Neighbor ที่ใช้ข้อมูลชุดเดียวกัน

บทที่ 5
สรุปการวิจัย และข้อเสนอแนะ

5.1 สรุปการวิจัย

งานวิจัยนี้ได้นำเสนอการรู้จำการอ่านริมฝีปากโดยใช้เทคนิคการวิเคราะห์สัญญาณแปรตามเวลาเราได้เสนอวิธี 2 วิธีการในการลดขนาดข้อมูลคือ วิธีวิเคราะห์ตัวประกอบสำคัญและวิธีการเฉลี่ยจุดภาพ ผลการทดลองแสดงให้เห็นว่าวิธีการเฉลี่ยจุดภาพให้ผลที่ดีกว่า เราสามารถอธิบายได้ว่าการคำนวณของวิธีการวิเคราะห์ตัวประกอบสำคัญจะมีการแปลงข้อมูลความเข้มของจุดภาพไปอยู่อีกมิติหนึ่ง ซึ่งจะทำให้เราสูญเสียข้อมูลที่แสดงถึงการเปลี่ยนแปลงของความเข้มของจุดภาพ ณ เวลาใดๆ เนื่องจากข้อมูลของเรามีลักษณะเป็นเฟรมซึ่งแสดงถึงความต่อต่อเนื่องของภาพ และเราใช้การวิเคราะห์สัญญาณแปรตามเวลาเพื่อดูถึงการเปลี่ยนแปลงของข้อมูลในแต่ละเฟรม ซึ่งการเปลี่ยนแปลงที่ดีที่สุดที่เราควรพิจารณา คือการเปลี่ยนแปลงความเข้มของภาพเป็นเฟรมต่อเฟรม ส่วนวิธีการเฉลี่ยจุดภาพยังคงข้อมูลที่เป็นความเข้มของจุดภาพอยู่แม้ว่าจะมีการลดขนาดด้วยค่าเฉลี่ยแล้วก็ตามดังนั้นสามารถสรุปได้ว่าวิธีการลดขนาดภาพโดยการใช้วิธีวิเคราะห์ตัวประกอบสำคัญไม่เหมาะที่จะนำมาใช้ร่วมกับการวิเคราะสัญญาณแปรตามเวลา และการนำนิวรอลเน็ตเวิร์กมาใช้สำหรับการแยกแยะร่วมกับวิธีการเฉลี่ยจุดภาพให้ผลการทดลองที่สูงกว่าวิธีการอื่นที่นำมาเปรียบเทียบในการทดลอง งานในอนาคตที่เราจะทำวิจัยคือนำผลที่ได้ไปใช้ร่วมกับการรู้จำด้วยเสียงเพื่อเพิ่มประสิทธิภาพในการรู้จำให้ดีมากขึ้น

5.2 ข้อเสนอแนะ

5.2.1 การเรียนรู้โดยใช้วิธีการแบบอื่น

นอกจากวิธีการแบ็คพรอพาเกชันนิวรอลเน็ตเวิร์กซึ่งใช้ในการทดลองนี้แล้ว ยังมีวิธีการอื่นที่น่าสนใจในการนำมาเรียนรู้อีก เช่น แบบจำลองฮิดเด็นมาคอฟ วิธีการฟัซซีโลจิก การเรียนรู้แบบพรอพอซิชันนอล ซี 4.5 (C4.5) ต้นไม้การตัดสินใจ ไอดี 3 (ID3) เป็นต้น

5.2.2 การประมวลผลขั้นต้น

การประมวลผลขั้นต้น ซึ่งประกอบด้วยวิธีการแยกคำ และการดึงลักษณะสำคัญ ในขั้นตอนการลดข้อมูลภาพแต่ละเฟรมโดยใช้การเฉลี่ยจุดภาพอาจทำให้สูญเสียข้อมูลความเข้มของจุดภาพไปบางส่วน หากสามารถนำวิธีการอื่นมาใช้โดยที่ยังสามารถคงข้อมูลความเข้มจุดภาพไว้ให้ได้มากที่สุด อาจทำให้เปอร์เซ็นต์ความถูกต้องสูงขึ้นได้ ในขั้นตอนการแปลงข้อมูลที่อยู่ในรูป

สัญญาณแปรตามเวลาโดยใช้การแปลงฟูเรียร์เราใช้แค่แมกนิจูดของสัมประสิทธ์ของฟูเรียร์เท่านั้น เราอาจสามารถใช้ข้อมูลเฟส(Phase Information)ที่ได้จากการแปลงฟูเรียร์มาใช้ร่วมกันได้

5.2.3 การแยกแยะโดยใช้หลายวิธีร่วมกัน(classifier combination)

ในงานวิจัยนี้เราใช้วิธีแยกแยะโดยใช้การเรียนรู้ของนิวรอลเน็ตเวิร์กเพียงแค่วิธีเดียว ในงานที่ซับซ้อนอย่างเช่นการรู้จำการอ่านริมฝีปากนี้ การใช้วิธีแยกแยะโดยใช้เพียงแค่วิธีเดียวอาจได้ผลไม่ถูกต้องมากนัก เราสามารถใช้การแยกแยะหลาย ๆวิธีร่วมกันได้ อย่างเช่นการใช้วิธีการนิวรอลเน็ตเวิร์กร่วมกับวิธีการทางสถิติเพื่อเพิ่มประสิทธิภาพในการรู้จำ

รายการอ้างอิง

1. A. Adjoudani and C. Benoit. On the integration of Auditory and Visual Parameters in and HMM-based ASR. Speechreading by Human and Machines, D. G. Stork and M. E. Hennecke(Eds) , pp 461-471,1996.
2. N. M. Brooke , M. J. Tomlinson, and R. K. Moore. Automatic Speech Recognition that Includes Visual Speech Cues. Proc. Of Institute of Acoustics,Vol. 16,NO. 5,pp. 15-22,Windermere,1994.
3. A. J. Goldschen ,O.N. Garcia, and E. Petajan. Continuous Optical Automatic Speech Recognition by Lipreading. Proc. Of 28th Annual Asilomar Conf. On Signals , Systems and Computers,pp. 572-577 , 1995.
4. J. Luettin and N. A. Thacker. Speechreading Using Probabilistic Models. Computer Vision and Image Understanding. Vol. 65 ,No. 2,pp - 163-178,1997.
5. C. Bregler, S. Manke, H Hild,and A. Waibel. Bimodal Sensor Integration on the Example of 'Speech-Reading'. Proc. Of IEEE Int. Conf on Neural Networks,pp. 667-671,San Francisco,1993.
6. P. Ducnowski, U. Meier, and A. Waibel. See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-reading. Int. Conf. On Sopken Language Processing, Yokohama,Japan ,1994.
7. J. Wu, S. Tamura, H. Mitsumoto, H. Kawai, K. Kurosu, and K.Okazaki. Neural Network Vowel Recognition Jointly Using Voice Features and Mouth Shape Image. Pattern Recognition, Vol. 24,No.10 ,pp.921-927,1991.
8. M. Kirby, F. Weisser, and G. Dangelmayr. A Model Problem in the Representation of Digital Image Sequences. Pattern Recognition, Vol. 26, No .1 ,pp.63-73, 1993.
9. E. Pedajan, B. Bischoff, and D. Bodoff. An Improved Automatic Lipreading System to Enhance Speech Recognition. SIGCHI'88: Human Factors in Computering systems,pp. 19-25, Oct. 1998.
10. K. Yu, X. Jiang, and H. Bunke. A New Approach to Lipreading Using Time-Varying Signal Analysis. Pattern Recognition ,1999.
11. J. Luettin ,N. A. Thacker ,and S W.beet. Statistical Lip Modelling For Visual Speech Recognition. Appears in VIII European Signal Processing Conference, Trieste, Italy,1996.

12. J. Luettin ,N. A. Thacker ,and S W.beet. Visual Speech Recognition Using Active Shape Model And Hidden Markov Models. Appears in IEEE ICASSP,Atlanta GA,May 1996.

13. Qing Jlang. Principal Component Analysis and Neural Network Based Face Recognition.

14. Sami Romdhani. Face Recognition using PCA

15. Rafael C. Gonzalez and Richard E. Woods. Digital Image Processing. Addison Wesley,1993

16. Tom M. Mitchell. Machine Learning. The Mcgraw-Hill Companies,Inc.,1997.

17. Y. Anzai. Pattern Recognition and Machine Learning. Academic Press,Inc.,1989.

18. Roman Kuc. Introduction To Digital Signal Processing. Mcgraw-Hill Internation Editions ,1982.

19. Nan Li,shawn Dettmer , and Mubarak Shah. Lipreading Using Eigensequences. Computer Vision Lab,Computer Science Department , University of Central Florida Orlando,1995.

20. Lee, K. Automatic speech recognition: The development of the Sphinx system. Boston: Kluwer Academic Publishers, 1989.

21. Pomerleau, D. A. ALVINN: An autonomous land vehicle in a neural network (Technical Report CMU-CS-89-107) Pittsburgh, PA: Carnegie Mellon University, 1989.

22. Fayyad, U. M., Smyth, P., Weir, N. and Djorgovski, S. Automated analysis and exploration of image databases: Results, progress, and challenges. Journal of Intelligent Information Systems, 4, pp 1-19, 1995.

ภาคผนวก

ภาคผนวก ก
รูปแบบไฟล์รูปภาพที่ใช้ที่ใช้ในการวิจัย

รูปแบบไฟล์รูปภาพ PGM (ข้อมูลชุดที่ 1)

1. รูปแบบไฟล์รูปภาพ PGM แบบ แอสกี (Ascii graymap)

รูปแบบ PGM(portable gray map) เป็นการเก็บข้อมูลรูปภาพในรูปแบบภาพเทา (grayscale) โดยมีรายละเอียดดังต่อไปนี้

- “magic number” เป็นตัวบอกประเภทของไฟล์ มี 2 ตัวคือ “P2”
- ช่องว่างอย่างเช่น blanks ,**TABS, CRs , LFs**
- ความกว้าง รูปแบบเป็น ตัวเลข ฐาน 10
- ช่องว่าง
- ความสูง รูปแบบเป็น ตัวเลข ฐาน 10
- ค่าความเข้มเทาสูงสุด รูปแบบเป็น ตัวเลข ฐาน 10 ค่านี้ต้องเป็น 255
- ช่องว่าง
- ค่าความเข้มเทา ความกว้าง * ความสูง มีค่าอยู่ระหว่าง 0 ถึงความเข้มเทาสูงสุด โดยมีการแบ่งแยกด้วยช่องว่าง เริ่มต้นจากซ้ายบนของภาพ มีลำดับการเก็บเหมือนการอ่านภาษาอังกฤษตือจากซ้ายไปขวา และบนลงล่าง ค่า 0 คือ สีดำ และค่าสูงสุดคือสีขาว
- บรรทัดที่เริ่มต้นด้วย # คือคอมเมนต์
- ความยาวแต่ละบรรทัดจะไม่เกิน 70 ตัวอักษร

ตัวอย่าง

```
P2
#Created by Pakit Silprachawong
24 7
255
0  0  0  0  0  0  0   0  0  0  0  0  0   0  0  0  0  0  0  0   0  0  0  0
0  3  3  3  3  0  0  77  7  7  7  0  0 111 11 11 11  0  0 215 15 15 15  0
0  3  0  0  0  0  0  77  0  0  0  0  0 111  0  0  0  0  0 215  0  0 15  0
0  3  3  3  0  0  0  77  7  7  0  0  0 111 11 11  0  0  0 215 15 15 15  0
0  3  0  0  0  0  0  77  0  0  0  0  0 111  0  0  0  0  0 215  0  0  0  0
0  3  0  0  0  0  0  77  7  7  7  0  0 111 11 11 11  0  0 215  0  0  0  0
0  0  0  0  0  0  0   0  0  0  0  0  0   0  0  0  0  0  0  0   0  0  0  0
```

2. รูปแบบไฟล์รูปภาพ PGM แบบไบนารี (Raw graymap)

ไฟล์รูปอีกรูปแบบหนึ่งของ PGM

- “magic number” คือ “P5” แทนที่จะเป็น “P2”
- ค่าความเข้มเทาจะเก็บในรูปแบบไบต์แทนที่จะเป็นตัวเลขแอสกี
- ไม่มีการใส่ช่องว่างในส่วนที่เก็บค่าความเข้มเทา และสามารถมีช่องว่างได้เพียงที่เดียว (โดยทั่วไปจะเป็นตัวขึ้นบรรทัดใหม่)หลังค่าที่เก็บค่าความเข้มเทาสูงสุด
- ขนาดของไฟล์เล็กกว่าและทำงานในการอ่านและเขียนได้เร็วกว่า

ตัวอย่างโปรแกรมการอ่านไฟล์รูปภาพ PGM

```
typedef struct {
        int width, height;          /* image size */
        int linewidth;
        int bitcount;
        unsigned char *data;        /* image data */
} IMAGE;

IMAGE img;

void ReadPGM(char *path)
{
        int i,j,tokens,tpixels;
        short unsigned int pix;
        char junk[100];
        char separators[] = " \t\v";
        char cparams[4][10];
        int params[4];
        FILE *fin;
        fin = fopen(path,"rb");

        tokens =0;
        do{
         fscanf(fin,"%s",junk);
         if(junk[0] ==  '#' ){   /* this is a comment line */
                fgets(junk,99,fin);
                continue;
         }
         if (junk[0] != '#'){   /* this is not a comment */
                strcpy(cparams[tokens],junk);
                tokens ++;
         }
        }
        while (tokens < 4);

        for(i=1;i<4;i++){
         params[i] = atoi(cparams[i]);
        }
```

```
if (params[3] != 255) {
  exit(1);
}
img.width= params[1];
img.height = params[2];
img.linewidth = img.width;
if(img.linewidth & 0x0003) img.linewidth = (img.linewidth | 3) +1; //Check octet boundary
tpixels = img.linewidth*img.height;
img.data = (unsigned char*)malloc(sizeof(unsigned char) * tpixels);
int nread;

// Allocate memory for the required number of bytes.

unsigned char *data_tmp = (unsigned char*)malloc(sizeof(unsigned char) * tpixels);

img.bitcount = 8;
if(strcmp(cparams[0],"P2")==0)   // Ascii Graymap = P2
{
        int k=0;
        for(i=0;i<img.height;i++)
        {
                for(j=0;j<img.width;j++)
                {
                        fscanf(fin,"%d",&pix);
                        data_tmp[k++] = pix;
                }
                for(int kk=1;kk<=img.linewidth-img.width;kk++)
                {
                        data_tmp[k++] = 128;   // Padding  to fit octet boundary
                }
        }
}
else    // Raw graymap = P5
{
        unsigned char *ch = (unsigned char*)malloc(sizeof(unsigned char)*1);
        int k=0;
        for(i=0;i<img.height;i++)
        {
                for(j=0;j<img.width;j++)
                {
                        fread((void*)ch,sizeof(unsigned char),1,fin);
                        data_tmp[k++] = ch[0];
                }
                for(int kk=1;kk<=img.linewidth-img.width;kk++)
                {
                        data_tmp[k++] = 128;
                }
        }
        free(ch);


}

for(i=0;i<tpixels;i++)
        img.data[tpixels-i-1] = data_tmp[i];  //Reverse order
        //img.data[i] = data_tmp[i];
free(data_tmp);

fclose(fin);
}
```

รูปแบบไฟล์รูปภาพ RAW (ข้อมูลชุดที่ 2 และ 3)

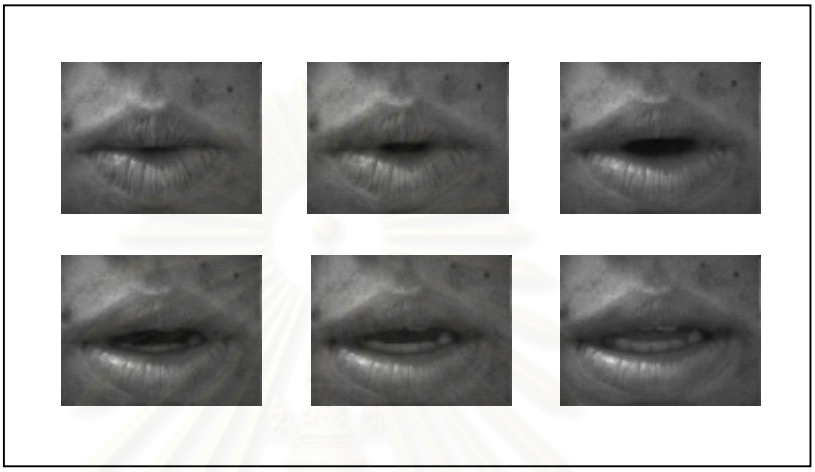
รูปแบบ RAW เป็นการเก็บข้อมูลรูปภาพในรูปแบบภาพเทา(grayscale) ในรูปแบบของไบนารีไฟล์ โดยไม่มีการเก็บข้อมูลรายละเอียดของภาพตรงส่วนหัวเลย โดยที่ลักษณะการเก็บค่ามีค่าอยู่ระหว่าง 0 ถึง 255 เริ่มต้นจากซ้ายบนของภาพ มีลำดับการเก็บเหมือนการอ่านภาษาอังกฤษคือจากซ้ายไปขวา และบนลงล่าง ค่า 0 คือ สีดำ และค่า 255 คือสีขาว ดังนั้นในการใช้งานตัวโปรแกรมจะต้องรู้ขนาดความกว้างและความสูงของภาพ

ตัวอย่างโปรแกรมการอ่านไฟล์รูปภาพ RAW

```
#define XSIZE       48
#define YSIZE       38

        void ReadRaw(char *path)
        {
                int i,j,tpixels;
                FILE *fin;
                fin = fopen(path,"rb");
                if(fin == NULL) return;
                img.width= XSIZE;
                img.height = YSIZE;
                img.linewidth = XSIZE;
                if(img.linewidth & 0x0003) img.linewidth = (img.linewidth | 3) +1;
                tpixels = img.linewidth*img.height;
                img.data = (unsigned char*)malloc(sizeof(unsigned char) * tpixels);
                img.bitcount = 8;

                int nread;
                unsigned char *data_tmp = (unsigned char*)malloc(sizeof(unsigned char) * tpixels);
                unsigned char *ch = (unsigned char*)malloc(sizeof(unsigned char)*1);
                int k=0;
                for(i=0;i<img.height;i++)
                {
                        for(j=0;j<img.width;j++)
                        {
                                fread((void*)ch,sizeof(unsigned char),1,fin);
                                data_tmp[k++] = ch[0];
                        }
                        for(int kk=1;kk<=img.linewidth-img.width;kk++)
                        {
                                data_tmp[k++] = 128;
                        }
                }
                free(ch);


                for(i=0;i<tpixels;i++)
                        img.data[tpixels-i-1] = data_tmp[i]; //Reverse order
                free(data_tmp);
                fclose(fin);
        }
```

ภาคผนวก ข
ตัวอย่างข้อมูลที่ใช้ในการทดลอง

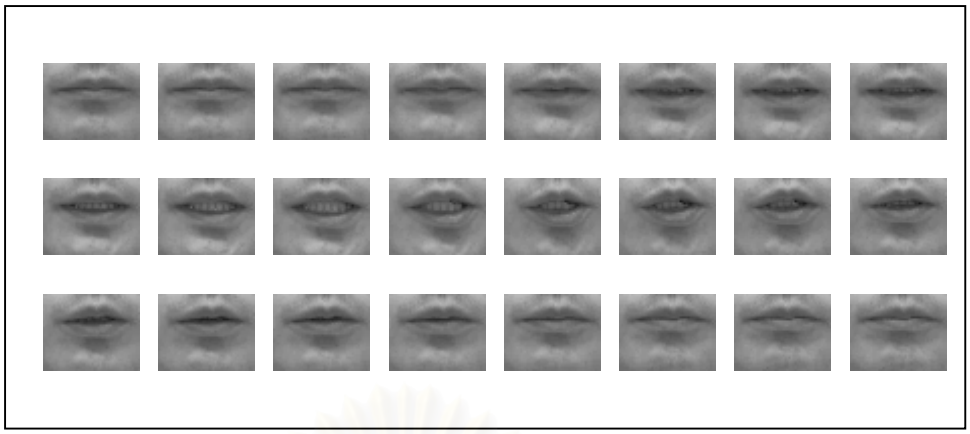
(1) ชุดข้อมูล Tulips1 [11][12] ซึ่งเป็นการพูด ตัวเลขภาษาอังกฤษ 1-4 โดยคน 12 คนโดยที่แต่ละคนพูดคนละ 2 ครั้ง และขนาดของภาพคือ 100x75



รูปที่ 1 แสดงลำดับภาพของการพูดตัวเลข 1 ในภาษาอังกฤษของผู้พูดคนหนึ่งใน 12 คน

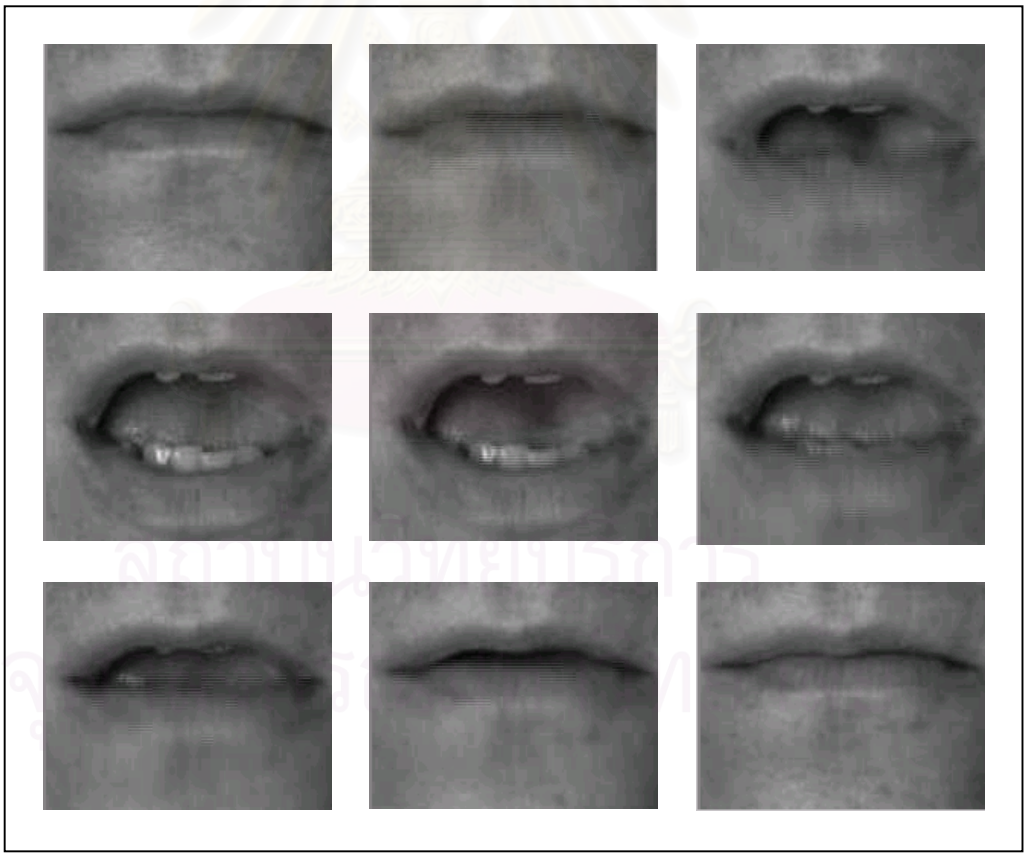
(2) ข้อมูลชุดที่ 2 เป็นชุดข้อมูลที่ได้จากงานวิจัย ของ K. Yu, X. Jiang, and H. Bunke [10] ซึ่งเป็นการพูดตัวเลขภาษาอังกฤษ 0-9 โดยคน 2 คน โดยแต่ละคนพูดคนละ 19 ครั้งต่อหนึ่งตัว และขนาดของภาพคือ 48x38

รูปที่ 2 แสดงลำดับภาพของการพูดตัวเลข 0 ในภาษาอังกฤษของผู้พูดคนที่ 1



รูปที่ 3 แสดงลำดับภาพของการพูดตัวเลข 0 ในภาษาอังกฤษของผู้พูดคนที่ 2

(3) ชุดข้อมูลชุดที่ 3 เป็นชุดข้อมูลจาก Computer Vision Lab, Computer Science Department,University of Central Florida, Orlando [19] ซึ่งเป็นการพูดตัวอักษรภาษาอังกฤษ 10 ตัวจาก A – J โดยคนเดียวและแต่ละตัวอักษรมีการพูด 18 ครั้งขนาดของภาพคือ 190x150



รูปที่ 3 แสดงลำดับภาพของการพูดตัวอักษร A

ภาคผนวก ค

ตัวอย่างโปรแกรมการคำนวณ PCA

```
Matrix X(NUM_OF_PIXELS,NUM_OF_PICS);    // Construct Matrix X
Matrix Xt(NUM_OF_PICS,NUM_OF_PIXELS);  // Construct Matrix Xt
Matrix SumXXt(NUM_OF_PIXELS,NUM_OF_PIXELS);   // Construct Matrix X*Xt

FOR(nCol=0;nCol<NUM_OF_PICS)
{
        ReadImage(g_img); // Read Image Data ;
        FOR(int i=0;i<g_img.width*g_img.height;i++)
        {
                X.cell(i,nCol) = g_img.data[i];    // Save data to Matrix X
        }


        FOR(i=0;i<g_img.width*g_img.height;i++)
        {
                Xt.cell(nCol,i) = g_img.data[i];  // Save data to Maxtrix Xt
        }
}

SumXXt = X*Xt;    // Compute X*Xt

Matrix P;
Vector LL;

SumXXt.factorEV(LL,P);   // Find Eigen Vectors and Eigen Values of X*Xt

for(i=LL.rows()-1 ; i >= 0;i--)  // Sort in Descending Order
{
        GetEigenValue(LL.cell(i));   // Save Eigen Value
}

// Generate only first 300 Eigen vectors
for(i=1523;i<P.columns();i++) // Sort in Descending Order
{
        for(int j=0;j<P.rows();j++)
        {
                GetEigenVector(P.cell(j,i));
        }
}
```

ภาคผนวก ง
ตัวอย่างโปรแกรมการคำนวณการแปลงฟูเรียร์

```
//*************** Compute FFT **************
printf("Compute FFT ...\n");
fftw_complex *in,*out;
fftw_plan p;

Matrix ImageFFT;
ImageFFT.setSize((XSIZE*YSIZE)/DIVIDER,NUM_OF_FFT);

for(k=0;k<(XSIZE*YSIZE)/DIVIDER;k++)
{
        in = (fftw_complex*)malloc(nNumFrameWord[i] * sizeof(fftw_complex));
        out= (fftw_complex*)malloc(nNumFrameWord[i] * sizeof(fftw_complex));
        // Prepare data for Fourier Transform
        for(j=0;j<nNumFrameWord[i];j++)
        {

                in[j].re = ImageFeatures.cell(k,j);
                in[j].im = 0;  // Imaginary Part  must be 0
        }
        p=fftw_create_plan(nNumFrameWord[i],FFTW_FORWARD,FFTW_ESTIMATE);
        fftw_one(p,in,out);  // Fourier Transform

#ifdef _SKIP_F0_
        for(int j=1;j<=NUM_OF_FFT;j++)
        {
        #ifdef _NORMALIZE_
                out[j].re = out[j].re/nNumFrameWord[i];      // Real Part
                out[j].im = out[j].im/nNumFrameWord[i];    // Imaginary Part
        #endif
                // Computer Magnitude
                imageFFT.cell(k,j-1) =  sqrt(out[j].re*out[j].re + out[j].im*out[j].im);
        }

#else

        for(int j=0;j<NUM_OF_FFT;j++)
        {
        #ifdef _NORMALIZE_
                out[j].re = out[j].re/nNumFrameWord[i];
                out[j].im = out[j].im/nNumFrameWord[i];
        #endif
                ImageFFT.cell(k,j) =  sqrt(out[j].re*out[j].re + out[j].im*out[j].im);
        }
#endif

        free(in);
        free(out);
        fftw_destroy_plan(p);
}


//************** End Compute FFT ***************
```

ประวัติผู้วิจัย

นายปกิต ศีลประชาวงศ์ เกิดเมื่อวันที่ 29 เดือนมิถุนายน พุทธศักราช 2515 ที่จังหวัดนราธิวาส ศึกษาระดับประถมศึกษาที่โรงเรียนเทศบาล1 ระดับมัธยมศึกษาตอนต้นที่โรงเรียนสุไหงโกลก ระดับมัธยมศึกษาตอนปลายที่โรงเรียนบดินทร์เดชา(สิงห์ สิงหเสนี) เข้าศึกษาต่อระดับปริญญาตรีที่คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีพุทธศักราช 2534 จบการศึกษาวิศวกรรมศาสตรบัณฑิต จากภาควิชาวิศวกรรมคอมพิวเตอร์ เมื่อปีพุทธศักราช 2538 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีพุทธศักราช 2541