การเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปาก

Accuracy Enhancement of Thai Voice-Command System Using Lip Reading

อิสมาแอล มะสาแม

Isamail Masamae

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Computer Engineering
Prince of Songkla University
2559

ชื่อวิทยานิพนธ์ การเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปาก
ผู้เขียน นายอิสมาแอล มะสาแม
สาขาวิชา วิศวกรรมคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

.....
(ดร. ปัญญยศ ไชยกาฬ)

คณะกรรมการสอบ

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร.มนตรี กาญจนะเดชะ)

.....กรรมการ
(ดร. ปัญญยศ ไชยกาฬ)

.....กรรมการ
(ดร.สมชัย หลิมศิโรรัตน์)

.....กรรมการ
(ดร.ปฏิมากร จันทร์พริ้ม)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์

.....
(รองศาสตราจารย์ ดร.ธีระพล ศรีชนะ)
คณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ .....
(ดร. ปัญญยศ ไชยกาฬ)
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลงชื่อ.....
(นายอิสมาแอล มะสาแม)
นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ.....

(นายอิสมาแอล มะสาแม)

นักศึกษา

ชื่อวิทยานิพนธ์ การเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปาก

ผู้เขียน นายอิสมาแอล มะสาแม

สาขาวิชา วิศวกรรมคอมพิวเตอร์

ปีการศึกษา 2558

บทคัดย่อ

วิทยานิพนธ์นี้เสนอเทคโนโลยีการอ่านริมฝีปากเพื่อเพิ่มความแม่นยำในการรู้จำเสียงด้วยเสียงพูดภาษาไทยที่ประยุกต์ใช้ในการควบคุมอุปกรณ์อิเล็กทรอนิกส์ภายในรถยนต์ คำที่ใช้ทดสอบเป็นคำสั้นๆภาษาไทย โดยมีการทดสอบกับตัวเลข (0-9) และคำสั่งในภาษาไทยไม่เกิน 3 พยางค์ จำนวน 10 คำสั่ง ระบบการรู้จำเสียงเริ่มต้นด้วยการสกัดคุณลักษณะเด่นโดยใช้ Mel-Frequency Ceptral Coefficients และใช้ฮิดเดนมาร์คอฟโมเดลในการรู้จำเสียง ระบบรู้จำริมฝีปากจะเริ่มต้นด้วยการหาใบหน้าด้วย Haar-Like Feature และใช้ Constrained Local Model เพื่อที่จะหาขอบเขตของริมฝีปากและใช้ฮิดเดนมาร์คอฟโมเดลในการรู้จำริมฝีปาก ส่วนระบบสุดท้ายเป็นระบบผสมผสานระหว่างการรู้จำเสียงและการรู้จำริมฝีปาก โดยทำการผสมผสานในระดับคุณลักษณะ (Feature Fusion) และการผสมผสานในระดับการตัดสินใจ (Decision Fusion) ระหว่างการอ่านริมฝีปากกับการรู้จำคำสั่งเสียงเพื่อเพิ่มความแม่นยำในการรู้จำเสียง จากการทดสอบผลที่ได้พบว่าการใช้ระบบรู้จำริมฝีปากสามารถเพิ่มประสิทธิภาพความแม่นยำของระบบรู้จำเสียงโดยเฉพาะอย่างยิ่งเมื่อภายในรถยนต์มีเสียงรบกวนสูง

Thesis Title	Accuracy Enhancement of Thai Voice-Command System Using Lip Reading
Author	Mr.Isamail Masamae
Major Program	Computer Engineering
Academic Year	2015

ABSTRACT

This thesis presents the use of lip-reading and Thai speech recognition to control electronic devices in a vehicle. Twenty Thai command words of no more than three syllables are supported by our system. The MFCC-DA features are extracted and a Hidden Markov Model classifier is utilized to recognize the speech. The face of the user is detected by means of a Haar-like feature and then the lip region is located by using a Constrained Local Model. The extracted feature from lip data is send to a Hidden Markov Model classifier. We tried combining the speech and lip data at both feature level and decision level. The results showed that the use of lip-reading can efficiently increase the accuracy of a speech recognition system especially when the noise in the car is at high level.

กิตติกรรมประกาศ

ในการดำเนินงานวิจัยและจัดทำวิทยานิพนธ์เล่มนี้ ผู้วิจัยขอขอบพระคุณ ดร.ปัญญยศ ไชยกาพ ที่ได้เสียสละเวลาในการให้คำปรึกษา ข้อเสนอแนะ แนวคิดและแนวทางในการทำงาน ทั้งยังให้กำลังใจและดูแลเอาใจใส่ข้าพเจ้าเป็นอย่างดี ตลอดจนการแนะนำสาหรับวางแผนการดำเนินงาน การแก้ไขปัญหาและตรวจทานวิทยานิพนธ์ที่ได้จัดทำขึ้นให้สำเร็จสมบูรณ์ในครั้งนี้

ขอขอบพระคุณ รองศาสตราจารย์ ดร.มนตรี กาญจนะเดชะ ดร.สมชัย หลิมศิโรรัตน์ และ ดร.ปฏิมากร จันทร์พริ้ม ที่สละเวลาในการตรวจทานแก้ไขข้อบกพร่องของวิทยานิพนธ์ ตรวจทานความถูกต้องของภาษา และให้แนวคิดต่าง ๆ ที่เป็นประโยชน์ในการวิจัย

ขอขอบคุณเพื่อน ๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ที่ร่วมเสนอความคิดเห็นคอยให้คำแนะนำและคอยให้ความช่วยเหลือกันตลอดมา

ขอขอบคุณทุกๆท่านที่มีส่วนร่วมในการแสดงความคิดเห็น ข้อเสนอแนะ ให้กำลังใจและช่วยเหลือผู้จัดทำนี้ให้สำเร็จลุล่วงไปได้ด้วยดี

อิสมาแอล มะสาแม

สารบัญ

รายการรูปภาพ

รายการตาราง

ตัวย่อและสัญลักษณ์

AVSR	=	Audio – Visual Speech Recognition
MFCC	=	Mel-Frequency Ceptral Coefficients
MFCC_D	=	MFCC เพิ่มค่าอนุพันธ์อันดับ 1
MFCC_DA	=	MFCC เพิ่มค่าอนุพันธ์อันดับ 1 และ อันดับ 2
HMM	=	Hidden Markov Model
SNR	=	Signal to Noise Ratio
ASM	=	Active Shape Model
ASM	=	Active Appearance Model
CLM	=	Constrained Local Model
ANN	=	Artificial Neural Network
KNN	=	K-Nearest Neighbor
PCA	=	Principal Component Analysis
FPS	=	Frame Per Second
WCR	=	Word Correct Rate

บทที่ 1

บทนำ

เนื้อหาในบทนี้เป็นส่วนที่ให้ข้อมูลเบื้องต้นเพื่อให้ผู้อ่านเข้าใจถึงที่มาและความสำคัญของวิทยานิพนธ์รวมถึงเอกสารและงานวิจัยที่เกี่ยวข้องเพื่อเป็นแนวทางในการทำวิจัย โดยมีรายละเอียดของเนื้อหาประกอบด้วยหัวข้อย่อยจำนวน 8 หัวข้อ ซึ่งประกอบด้วยหัวข้อดังต่อไปนี้

1.1 ความสำคัญและที่มาของวิทยานิพนธ์

1.2 งานวิจัยที่เกี่ยวข้อง

1.3 วัตถุประสงค์

1.4 ขอบเขตการวิจัย

1.5 ขั้นตอนและวิธีดำเนินการวิจัย

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.7 ทรัพยากรที่ใช้ในระบบ

1.8 ภาพรวมและหลักการของระบบ

1.1 ความสำคัญและที่มาของวิทยานิพนธ์

เทคโนโลยีการรู้จำเสียง(Audio Speech Recognition) มีการวิจัยอย่างกว้างขวางในปัจจุบัน แต่ปัญหาที่พบในการวิจัยเหล่านั้น คือ เสียงรบกวน (Noise) โดยในสภาพแวดล้อมที่มีเสียงรบกวนต่ำอัตราความถูกต้องของระบบนี้จะมีความถูกต้องสูงหรือมากกว่า 95 % [1] แต่หากในสภาพแวดล้อมที่มีเสียงรบกวนสูงอัตราความถูกต้องของระบบนี้จะถูกลดทอนลง

ปัจจุบันเทคโนโลยี การรู้จำเสียงถูกนำมาประยุกต์ใช้ในชีวิตประจำวันของมนุษย์มากขึ้น เช่น ระบบสั่งการด้วยเสียงพูดภายในรถยนต์ที่ถูกประยุกต์ขึ้นมาเพื่อความสะดวกสบายและความปลอดภัยของผู้ขับขี่รถยนต์ หากภายในรถยนต์ไม่มีเสียงรบกวนการใช้เสียงสั่งการจะมีความแม่นยำและมีประสิทธิภาพสูง แต่หากภายในรถยนต์มีเสียงรบกวนภายใน เช่น เสียงแอร์ เสียงเครื่องยนต์ ความเร็วของยานพาหนะ ผู้คนในรถกำลังพูด เสียงจากวิทยุ เป็นต้น และมีเสียงรบกวนจากภายนอก

เช่น เสียงเครื่องยนต์ของรถคันอื่น ซึ่งจะหลีกเลี่ยงสิ่งนี้ไม่ได้ในการใช้งานในชีวิตประจำวันของมนุษย์ เสียงรบกวนเหล่านี้อาจจะทำให้ผู้ขับขี่ไม่สามารถใช้เสียงสั่งการได้อย่างมีประสิทธิภาพ

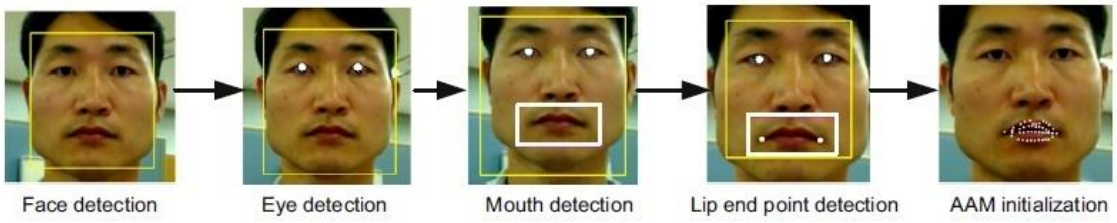
การติดต่อสื่อสารเป็นสิ่งที่เกิดขึ้นควบคู่มากับมนุษย์ เนื่องจากมนุษย์ต้องอยู่รวมกันเป็นกลุ่มเป็นก้อนโดยมนุษย์ใช้ภาษาเป็นสื่อในการส่งข้อมูล แลกเปลี่ยนข้อมูลซึ่งกันและกัน ซึ่งในภาษาที่มนุษย์ใช้ในการติดต่อสื่อสารกันนอกจากสัญญาณเสียงแล้วยังมีการติดต่อทางสัญลักษณ์ยกตัวอย่างเช่น การใช้สัญญาณควันไฟของชาวอินเดียแดง การแสดงสัญญาณมือสำหรับผู้พิการทางหู หรือยังสามารถมองใบหน้าของผู้พูดซึ่งอาจจะเป็นส่วนริมฝีปากเพื่อช่วยในการสื่อสารในสภาพแวดล้อมที่มีเสียงรบกวนสูง การใช้ริมฝีปากในการรับรู้คำพูดของมนุษย์ หรือเรียกว่าการอ่านริมฝีปาก (Lip Reading) สามารถเพิ่มประสิทธิภาพในการติดต่อสื่อสารของมนุษย์ได้มากขึ้นและยังสามารถไปประยุกต์ใช้กับผู้พิการทางหูเพื่อเพิ่มประสิทธิภาพในการติดต่อสื่อสารของบุคคลเหล่านั้น

จากปัญหาข้างต้นวิทยานิพนธ์นี้จึงมีวัตถุประสงค์เพื่อหาวิธีการเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปากที่สามารถนำมาประยุกต์ใช้งานภายในรถยนต์โดยผ่านทางกล้องเว็บแคม (Web Camera) และวิเคราะห์สัญญาณเสียงอย่างง่ายโดยทำการประมวลผลผ่านทางคอมพิวเตอร์แสดงผลออกมาในรูปของข้อความทางจอแสดงผล

1.2 งานวิจัยที่เกี่ยวข้อง

Jongju Shin[1] นำเสนอเกี่ยวกับระบบรู้จำเสียงแบบเรียลไทม์เพื่อแยกแยะคำในภาษาเกาหลี โดยที่ลำดับในการทำงานของระบบแบ่งออกเป็น 4 ขั้นตอน เริ่มต้นตรวจจับริมฝีปาก (ตรวจจับใบหน้า, ตรวจจับตา, ตรวจจับปาก และตรวจจับริมฝีปาก) ดังแสดงในรูปที่ 1.1 และใช้ Active Appearance Model (AAM) ในการติดตามขอบเขตของริมฝีปาก การติดตามริมฝีปากจะใช้ทั้งริมฝีปากบนและริมฝีปากล่างเพื่อเพิ่มความแม่นยำในการรู้จำ จะใช้งานการแบ่งคลาสแบบนิวรอลเน็ตเวิร์ค (NN) ในขั้นตอนการใช้งานริมฝีปาก (Lip Activation) ในขั้นตอนสุดท้ายใช้โมเดล HMM, ANN, KNN ในการรู้จำและเรียนรู้ระบบ โดยผลการทดลองระบบมีความถูกต้อง 97 %ในการทดสอบแบบขึ้นกับบุคคล (Person Dependent) และมีความถูกต้อง 46.50 % ในการทดสอบ

แบบไม่ขึ้นกับบุคคล (Person Independent) ยิ่งไปกว่านั้นหากมีการนำระบบนี้ไปรวมกับการรู้จำด้วยเสียงระบบจะมีความถูกต้องเพิ่มขึ้นสูงสุด 60 % ในสภาพแวดล้อมที่มีเสียงรบกวนสูง



รูปที่ 1.1 ขั้นตอนของการติดตามริมฝีปากในงานวิจัยของ Jongju Shin [1]

Navarathna [2] ได้นำเสนอเกี่ยวกับการใช้รูปภาพในการเพิ่มประสิทธิภาพความแม่นยำในการรู้จำเสียงภายในรถยนต์ เนื่องจากภายในรถยนต์มีเสียงรบกวนที่ดังมากและจะมีผลกระ ทบกับข้อมูลเสียงเท่านั้น ทำให้การรับคำสั่งภายในรถยนต์ไม่แม่นยำ แต่ข้อมูลรูปภาพจะไม่ได้รับผลกระทบจากสัญญาณเสียงรบกวน โดยจะใช้ข้อมูลจากรูปภาพริมฝีปากเพื่อเพิ่มศักยภาพในระบบ พวกเขาได้นำเสนอวิธีการอัลกอริทึม Viola-Jones ในการการติดตามใบหน้าและริมฝีปากภายในรถยนต์ดังรูปที่ 1.2 โดยใช้ข้อมูลวิดีโอจากฐานข้อมูล AVICAR พวกเขาแสดงให้เห็นว่าอัลกอริทึม Viola-Jones เป็นวิธีที่เหมาะสมในการติดตามและการหาตำแหน่งของริมฝีปากของผู้ขับขี่ถึงแม้ภายในรถยนต์จะมีความแปรปรวนของแสงที่มองเห็นก็ตาม

รูปที่ 1.2 ผลการทดสอบการติดตามใบหน้าและริมฝีปากภายในรถยนต์ [2]

Lee [3] ได้นำเสนอเกี่ยวกับระบบการรู้จำเสียงขนาดใหญ่ที่บันทึกในสภาพแวดล้อมภายในรถยนต์ ข้อมูลเสียงจะเก็บข้อมูลโดยการใช้ไมโครโฟนแบบอาร์เรย์จำนวน 8 ตัว ติดตั้งบนที่บังแดดและกล้องเว็บแคม 4 ตัว ติดตั้งบนแผงหน้ารถดังรูปที่ 1.3 คำสั่งที่ใช้ในการทดสอบระบบมี 4 หมวดหมู่ ได้แก่ ตัวเลข พยัญชนะ หมายเลขโทรศัพท์ และประโยคในภาษาอังกฤษ โดยผู้พูดจากหลากหลายภาษา ผู้หญิง 50 คน และผู้ชาย 50 คน พวกเขาได้ทดสอบในสภาพแวดล้อมที่มี

เสียงรบกวนต่างกันโดยจะแบ่งออกเป็น 5 สภาพเสียง คือ รถไม่เคลื่อนที่ รถเคลื่อนที่ด้วยความเร็ว 35 ไมล์ต่อชั่วโมง ตอนเปิดหน้าต่างและปิดหน้าต่างรถยนต์ และรถเคลื่อนที่ด้วยความเร็ว 55 ไมล์ต่อชั่วโมง ตอนเปิดหน้าต่างและปิดหน้าต่างรถยนต์

รูปที่ 1.3 การติดตั้งไมโครโฟนแบบอาร์เรย์ จำนวน 8 ตัว ติดตั้งบนที่บังแดด และกล้องเว็บแคม 4 ตัว ติดตั้งบนแผงหน้ารถ [3]

Potamianos [4] ได้นำเสนอการใช้ข้อมูลรูปภาพเพื่อปรับปรุงความแม่นยำของการรู้จำเสียง พวกเขาได้ตรวจสอบภาพและรู้จำเสียงในสองสภาพแวดล้อมในการทดสอบได้แก่ ก) สำนักงานทั่วไปที่ข้อมูลจะถูกบันทึกโดยคอมพิวเตอร์แบบพกพาที่ติดตั้งกล้องเว็บแคมทั่วไป ข) สภาพแวดล้อมภายในรถยนต์ที่มีการเก็บรวบรวมข้อมูลที่ระดับความเร็วของรถยนต์สามระดับ ดังรูปที่ 1.4 ผลการศึกษาที่ออกมานั้นพบว่าข้อมูลที่บันทึกรูปภาพในสภาพแวดล้อมที่สตูดิโอมีความถูกต้องมากแต่ในสภาพแวดล้อมจริง เช่น ออฟฟิศ ภายในรถยนต์ ความถูกต้องจะลดลงมาก แต่อย่างไรก็ตามการรู้จำรูปภาพริมฝีปากยังคงเป็นประโยชน์ต่อระบบการรู้จำเสียง

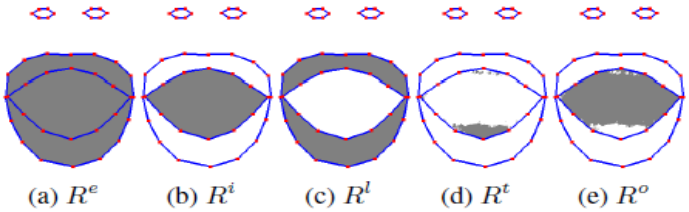
โดยงานวิจัยนี้จะมุ่งเน้นการแก้ปัญหาของเสียงรบกวน โดยใช้การผสมผสานข้อมูลของเสียงและรูปภาพเข้าด้วยกันในแบบการผสมผสานในระดับคุณลักษณะ(Feature Fusion) และการผสมผสานในระดับการตัดสินใจ (Decision Fusion) เพื่อให้ระบบนำความถูกต้องของทั้งสองระบบมาวิเคราะห์ร่วมกันเพื่อนำมาตัดสิน เมื่อไม่นานนี้งานวิจัยที่เกี่ยวข้องกับผสมผสานข้อมูลถูกนำมาประยุกต์ใช้ในงานการรู้จำคำสั่งเสียงและการอ่านริมฝีปากมากขึ้น

รูปที่ 1.4สภาพแวดล้อมในการทดสอบสำนักงานทั่วไปและภายใน
รถยนต์ที่มีการเคลื่อนที่ [4]

Glotin [5] ได้นำเสนอการปรับปรุงระบบการรู้จำเสียงและภาพ ภายใต้สภาพแวดล้อมไม่มีเสียงรบกวน และสภาพแวดล้อมที่มีเสียงรบกวนจากการใช้ข้อมูลภาพ นอกเหนือไปจากข้อมูลเสียงอย่างเดียว งานวิจัยนี้ใช้วิธีการผสมผสานแบบการตัดสินใจสำหรับข้อมูลภาพและเสียง โดยใช้ฮิดเดนมาร์คอฟโมเดลในการรู้จำ โดยเฉพาะอย่างยิ่งงานวิจัยนี้ได้แก้ปัญหาของการประมาณค่าน้ำหนักที่เหมาะสมในการผสมผสานสำหรับแต่ละส่วน โดยใช้สองเทคนิคที่แตกต่างกัน วิธีการแรก สกัดข้อมูลแบบอัตโนมัติที่จะปรับเปลี่ยนน้ำหนัก (ทั้งสภาพแวดล้อมที่ไม่มีเสียงรบกวนและมีเสียงรบกวน) ในขณะที่วิธีการที่สองเป็นวิธีการผสมผสานรูปแบบการจำแนกน้ำหนักที่กำหนดไว้ล่วงหน้าจะเพิ่มประสิทธิภาพในการลดความผิดพลาดของระบบ (สภาพแวดล้อมที่ไม่มีเสียงรบกวนเท่านั้น)

Lee [6] ได้นำเสนอการรวมระบบการรู้จำเสียงและภาพภายใต้สภาพแวดล้อมเสียงรบกวนต่าง ๆ จะเพิ่มประสิทธิภาพการทำงานในสามส่วนของระบบ ส่วนที่หนึ่งปรับปรุงประสิทธิภาพการทำงานของรูปภาพริมฝีปากและใช้ฮิดเดนมาร์คอฟในการรู้จำ ส่วนที่สองได้นำเสนอการเพิ่มประสิทธิภาพความแม่นยำของเสียง ส่วนที่สามผสมผสานระหว่างรูปภาพและเสียงเข้าด้วยกันเพื่อให้ระบบมีประสิทธิภาพมากขึ้น โดยใช้แบบจำลองโครงข่ายประสาทเทียมในการหาค่าน้ำหนักที่เหมาะสม พวกเขาแสดงให้เห็นประสิทธิภาพของการรู้จำโดยไม่ขึ้นอยู่กับผู้พูดผลการทดลองพบว่าระบบที่นำเสนอจะช่วยเพิ่มความทนทานต่อเสียงรบกวนมากกว่าเดิม

Saitoh [7] นำเสนอเกี่ยวกับการวิเคราะห์ประสิทธิภาพระบบการอ่านริมฝีปากเพื่อการใช้งานภาษาเริ่มต้นในหลากหลายภาษา ประยุกต์การสกัดค่าแบบ AAM ทั้งริมฝีปากด้านในและริมฝีปากด้านนอกริมฝีปากทำให้เกิดการตรวจจับฟันและภายในช่องปากดังรูปที่ 1.5 โดยใช้ 4 ภาษาในการรู้จำเสียงและบันทึก 20 คำต่อภาษา พวกเขานำเสนอการใช้งานของคุณลักษณะของเส้นวงโคจร (Trajectory Feature) จากคุณลักษณะสามอย่างของริมฝีปาก คือ พื้นที่ของริมฝีปากภายในอัตราส่วนของริมฝีปากภายใน และพื้นที่ของริมฝีปากภายนอก โดยผลการทดลองระบบมีความถูกต้อง 93.6 %



รูปที่ 1.5 คุณลักษณะเด่นทั้งหมดที่ถูกทดสอบในงานวิจัยของ Saitoh [7]

1.3 วัตถุประสงค์

1.3.1 เพื่อพัฒนาวิธีการอ่านริมฝีปาก (Lip reading) เพื่อเพิ่มความแม่นยำในการรู้จำของระบบสั่งการด้วยเสียงพูดภาษาไทย

1.3.2 เพื่อศึกษาวิธีการรู้จำการอ่านริมฝีปาก

1.3.3 เพื่อที่จะออกแบบซอฟต์แวร์สั่งการด้วยเสียงที่สามารถรู้จำการอ่านริมฝีปากในสภาพแวดล้อมที่มีเสียงรบกวนภายในรถยนต์อย่างมีประสิทธิภาพ

1.4 ขอบเขตการวิจัย

1.4.1 กล้องที่ใช้ในการทำวิจัยเป็น Webcam มีความละเอียดไม่ต่ำกว่า 640x480 พิกเซล

1.4.2 สภาพแสงที่ใช้ทดสอบเป็นสภาพแสงกลางวันภายในรถยนต์และไม่มีแสงภายนอกส่องกระทบหน้าคนขับขี่รถยนต์โดยตรง

1.4.3 บันทึกคำสั่งเสียงภายในรถยนต์ในสภาพแวดล้อมที่มีเสียงรบกวนหลายระดับ

1.4.4 คำที่ใช้ทดสอบเป็นคำสั้น ๆ ภาษาไทย โดยมีการทดสอบกับตัวเลข (0-9) และคำสั่งในภาษาไทยไม่เกิน 3 พยางค์ จำนวนไม่ต่ำกว่า 10 คำสั่ง เช่น เปิด ปิด ถัดไป ก่อนหน้า เป็นต้น

1.4.5 ติดตั้งกล้องบนพวงมาลัยรถยนต์และบันทึกภาพในขณะที่พวงมาลัยรถยนต์หยุดนิ่ง

1.5 ขั้นตอนและวิธีดำเนินการวิจัย

ขั้นตอนการดำเนินงานวิทยานิพนธ์แบ่งออกเป็น 12 ขั้นตอนโดยมีรายละเอียดขั้นตอนการดำเนินงาน ดังนี้

ขั้นที่1: ศึกษาแนวทางและวิธีการดำเนินงานวิจัย

ขั้นที่2: ศึกษาการใช้งานโอเพนซีวีคอมพิวเตอร์วิชชัน(Open Source Computer Vision) และจัดเตรียมกล้องวีดีโอที่ใช้ในการดำเนินงานวิจัย

ขั้นที่3: รวมรวบฐานข้อมูลวีดีโอริมฝีปากเพื่อใช้ในงานวิจัย

ขั้นที่4: ศึกษาและทดสอบอัลกอริทึมสำหรับการตรวจจับริมฝีปาก

ขั้นที่5: ออกแบบระบบสำหรับการประมวลผลภาพในการติดตามริมฝีปาก

ขั้นที่6: พัฒนาระบบหาพารามิเตอร์ที่เหมาะสมในการติดตามริมฝีปาก ทดสอบและตรวจสอบความถูกต้อง ในการติดตามริมฝีปาก

ขั้นที่7: ศึกษาการนำการเรียนรู้ของเครื่อง (Machine Learning) ที่ใช้งานในการรู้จำระบบ

ขั้นที่8: สร้างระบบสั่งการด้วยเสียง Thai Voice Command

ขั้นที่9: ผสมผสานระบบระหว่างรูปภาพกับเสียงเข้าด้วยกัน

ขั้นที่10: พัฒนาระบบทดสอบการทำงานความถูกต้องของระบบรวบรวมผลการทดสอบ

ขั้นที่11: ปรับปรุงและทดสอบระบบทั้งหมด

ขั้นที่12: สรุปผล จัดทำรายงานฉบับสมบูรณ์

1.6 ประโยชน์คาดว่าจะได้รับ

1.6.1 สามารถเพิ่มความแม่นยำในการรู้จำเสียง

1.6.2 ผสมผสานระหว่างการอ่านริมฝีปากกับการรู้จำคำสั่งเสียงเพื่อเพิ่มความแม่นยำในการรู้จำเสียง

1.6.3 สามารถใช้เสียงสั่งการในสภาพแวดล้อมที่มีเสียงรบกวนภายในรถยนต์ได้อย่างแม่นยำมากขึ้น

1.7 ทรัพยากรที่ใช้ในระบบ

ในงานวิจัยนี้เป็นการพัฒนาหาวิธีการเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปากที่สามารถนำมาประยุกต์ในงานภายในรถยนต์ ซึ่งมีซอฟต์แวร์และอุปกรณ์ที่ออกแบบและทดสอบดังนี้

1.7.1 Hardware

1. คอมพิวเตอร์ส่วนบุคคล มีหน่วยประมวลผล Intel Core i5 ความถี่ 2.66 GHz หน่วยความจำแรมขนาด 4 GB

2. คอมพิวเตอร์โน้ตบุค (Notebook) ที่มีหน่วยประมวลผล Intel Core i5 ความถี่ 2.66 GHz หน่วยความจำแรมขนาด 4 GB

3. กล้องวิดีโอเว็บแคม (Web Camera) ยี่ห้อ Oker รุ่น OE-177

1.7.2 Software

1.ระบบปฏิบัติการ Microsoft Windows 7 Ultimate (Service Pack 1)

2.โปรแกรม Microsoft Visual C++ Express Edition 2008

3. โปรแกรม Matlab 2011b

4. ไลบรารีจาก Open Source Computer Vision 2.3 (OpenCV)

1.8 ภาพรวมและหลักการของระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปาก

หลักการของระบบเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปากโดยทั่วไปจะประกอบไปด้วยขั้นตอนสำคัญดังนี้

1.8.1 โหลดรูปภาพเข้ามาในระบบ เป็นขั้นตอนแรกเพื่อจะนำรูปภาพมาประมวลผลในระบบตรวจสอบ

1.8.2 การประมวลผลเบื้องต้น (Preprocessing) เป็นขั้นตอนการประมวลผล เพื่อจัดเตรียมข้อมูลให้เหมาะสมในการตรวจจับริมฝีปาก

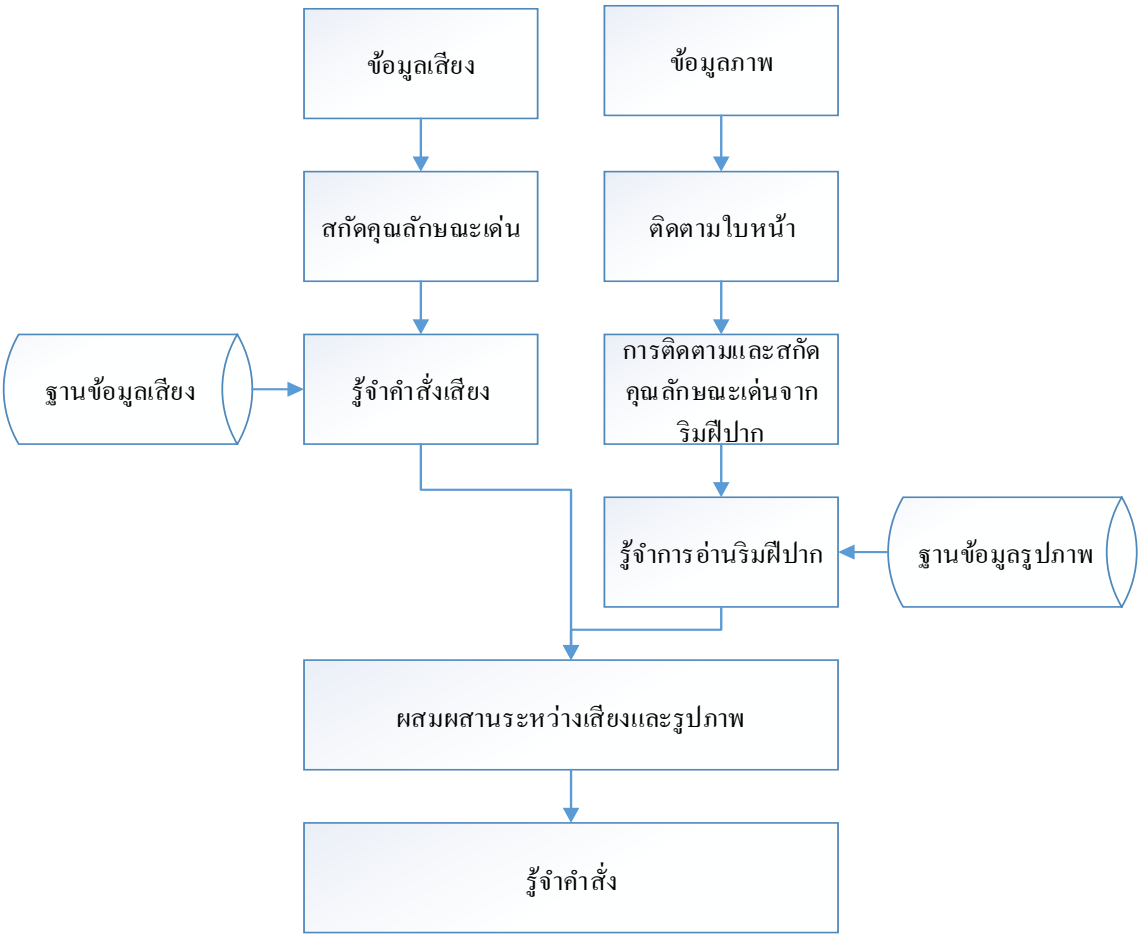
1.8.3 ติดตามและสกัดคุณลักษณะเด่น (Tracking and Feature Extraction) เป็นขั้นตอนการติดตามริมฝีปากและการสกัดเพื่อนำลักษณะเด่นมาประมวลผล

1.8.4 สร้างระบบรู้จำเสียงอย่างง่าย (Speech Recognition) เพื่อจะนำมาตรวจสอบความถูกต้องของระบบ

1.8.5 จำแนกรูปแบบของริมฝีปากของแต่ละบุคคล (Classification) เป็นขั้นตอนในการหารูปแบบใกล้เคียงของการรู้จำ

1.8.6 ผสมผสานระบบรู้จำระหว่างการรู้จำเสียงกับการรู้จำการอ่านริมฝีปากริมฝีปาก

1.8.7 การตัดสินใจว่าตรงกับข้อมูลอ้างอิงหรือไม่ (Decision making) เป็นกฎเกณฑ์ในการเลือกคำที่ต้องการแสดงคำว่าใกล้เคียงกับคำในฐานข้อมูลหรือไม่



รูปที่ 1.6 ภาพรวมและหลักการของระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปาก

บทที่2

ทฤษฎีและหลักการ

ในบทนี้เป็นการนำเสนอเนื้อหาที่ผู้วิจัยได้ทำการศึกษาและรวบรวมทฤษฎีเพื่อใช้ในงานวิจัยการเพิ่มประสิทธิภาพระบบรู้จำเสียงภายในรถยนต์โดยการใช้การอ่านริมฝีปากซึ่งมีหัวข้อการนำเสนอดังนี้

2.1 ระบบการรู้จำเสียง

2.2 ระบบการอ่านริมฝีปาก

2.3 การผสมผสานระบบการรู้จำเสียงและระบบการอ่านริมฝีปาก

2.4 การรู้จำด้วยเครื่องการเรียนรู้

2.5 สรุป

2.1 ระบบการรู้จำเสียง (Automatic Speech Recognition)

Automatic Speech Recognition (ASR) เป็นการฝึกสอนให้ระบบคอมพิวเตอร์สามารถวิเคราะห์และตัดสินใจได้ว่าเสียงพูดของผู้พูดนั้นเป็นคำพูดอะไร หรืออาจจะหมายถึงการแปลงให้เป็นข้อความที่ระบบคอมพิวเตอร์สามารถนำมาใช้ประมวลผลได้

2.1.1 ความรู้พื้นฐานเกี่ยวกับเสียง

ในทางฟิสิกส์เสียงเกิดจากแรงสั่นสะเทือนของแหล่งกำเนิดเสียงที่แพร่กระจายเป็นคลื่นกล เสียงสามารถเดินทางผ่านพาหะ เช่น อากาศ น้ำและของแข็ง เป็นต้น โดยแหล่งกำเนิดเสียงจะสร้างแรงสั่นสะเทือนและโอนถ่ายพลังงานผ่านอากาศกับวัตถุที่อยู่โดยรอบจนทำให้คลื่นเสียงเคลื่อนที่มาถึงหูเรา

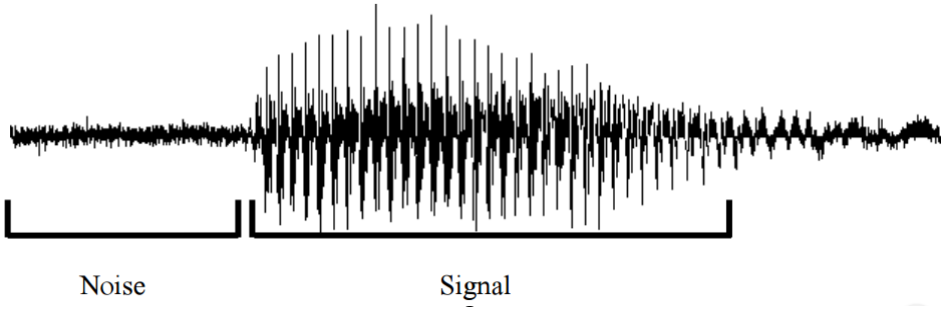
เสียงสามารถจำแนกได้เป็นสองประเภท คือ เสียงที่ต้องการ และเสียงรบกวน เสียงที่ต้องการเป็นเสียงที่ใช้ในติดต่อสื่อสารกันระหว่างแหล่งกำเนิดเสียงและผู้ฟัง แต่เสียงรบกวนทำให้ประสิทธิภาพความแม่นยำในการสื่อสารที่ต้องการถูกลดทอนลง

2.1.2 อัตราสัญญาณเสียงต่อสัญญาณรบกวน (Signal-to-Noise Ratio)

SNR (Signal-to-Noise Ratio) หมายถึงค่าความเข้มของสัญญาณ โดยใช้อัตราส่วนของสัญญาณระหว่างสัญญาณที่ต้องการกับสัญญาณรบกวนดังรูปที่ 2.1 และแสดงตามสมการที่ 2.1

$$SNR_{db} = 10\log_{10}[(\frac{A_{signal}}{A_{noise}})^2] \quad (2.1)$$

A_{signal} และ A_{noise} คือ ค่าแอมพลิจูดของสัญญาณเสียงที่ต้องการและสัญญาณเสียงรบกวน ตามลำดับ โดยจากสมการนี้ค่า SNR ที่ดีจะมีค่าสูง เมื่อแอมพลิจูดของสัญญาณที่ต้องการมีความสูงมากกว่าแอมพลิจูดของสัญญาณรบกวน แต่หากค่าแอมพลิจูดของสัญญาณที่ต้องการมีความสูงใกล้เคียงกับแอมพลิจูดของสัญญาณรบกวนแสดงว่าค่า SNR มีค่าต่ำและสัญญาณรบกวนสูง โดยที่ค่า SNR มีหน่วยเป็นเดซิเบล



รูปที่ 2.1 สัญญาณเสียงที่ต้องการกับสัญญาณเสียงรบกวน

ความหมายของค่าเบื้องต้น

SNR 5 เดซิเบล หรือต่ำกว่า แสดงว่าคุณภาพสัญญาณเสียงไม่ดีมีสัญญาณรบกวนมาก

SNR 10 เดซิเบล แสดงว่าคุณภาพสัญญาณเสียงพอใช้มีสัญญาณรบกวนพอใช้

SNR 20 เดซิเบล แสดงว่าคุณภาพสัญญาณเสียงดีมาก มีสัญญาณรบกวนเล็กน้อย

SNR มากกว่า 30 เดซิเบล หรือสูงกว่า แสดงว่าคุณภาพสัญญาณเสียงดีมากไม่มีสัญญาณรบกวน

2.1.3 หลักการของระบบรู้จำเสียง

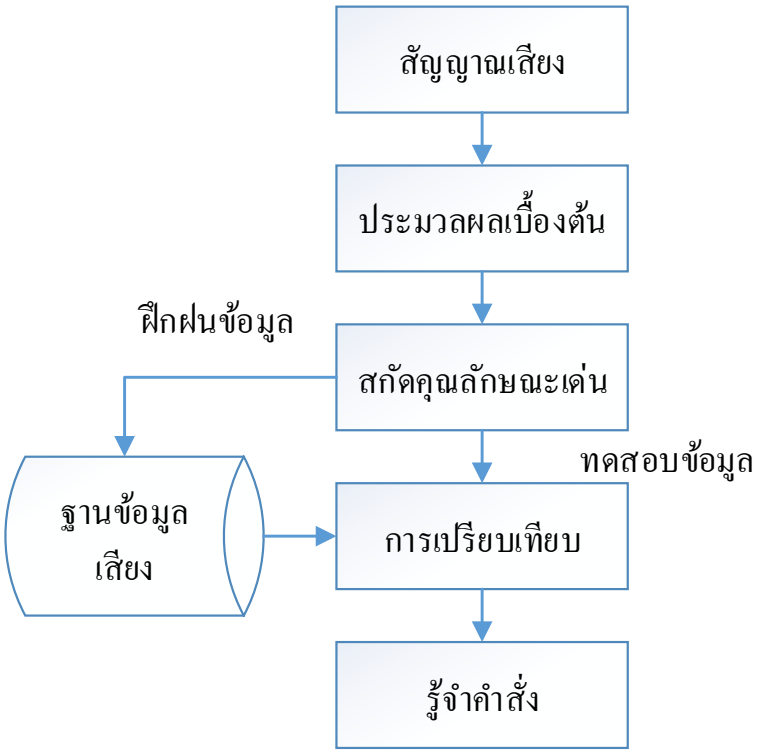
หลักการของระบบรู้จำเสียงโดยทั่วไปจะประกอบไปด้วยขั้นตอนสำคัญดังนี้

1. การประมวลผลเบื้องต้น เป็นขั้นตอนของการเตรียมข้อมูลเบื้องต้นเพื่อจัดรูปแบบของข้อมูลก่อนนำไปการประมวลผลในขั้นตอนต่อไป

2. การสกัดคุณลักษณะเด่น เป็นขั้นตอนของการสกัดข้อมูลเพื่อหาค่าคุณลักษณะเด่นของข้อมูลออกมา และลดทอนจำนวนข้อมูลเพื่อเพิ่มรวดเร็วในการประมวลผลในการรู้จำ

3. การจำแนกรูปแบบเพื่อการรู้จำเสียงพูด เป็นขั้นตอนในการจำแนกหารูปแบบของเสียงที่ต้องการตรวจสอบมาเทียบกับฐานข้อมูลเสียงอ้างอิง

4. การตัดสินใจเสียงพูด เป็นขั้นตอนในการประมวลผลลัพธ์ที่ได้จากการจำแนกรูปแบบมาตัดสินใจว่าเสียงที่ต้องการตรวจสอบเป็นเสียงคำสั่งใดในฐานข้อมูลที่อ้างอิง



รูปที่ 2.2 หลักการของระบบรู้จำเสียง

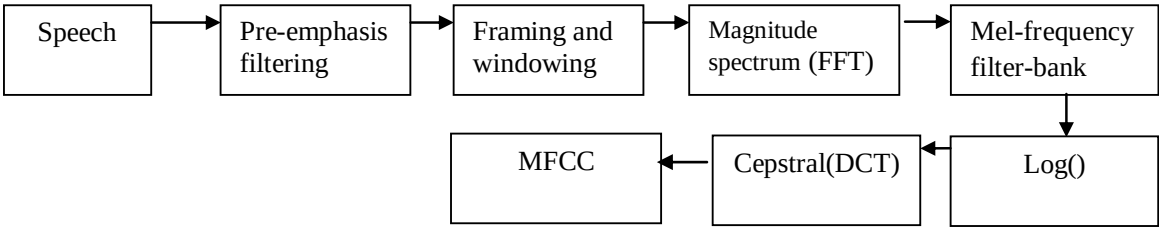
2.1.4 การประมวลผลเบื้องต้น

การประมวลเสียงเบื้องต้นเป็นขั้นตอนการจัดเตรียมข้อมูลดิบของสัญญาณเสียงพูดที่ได้จากการบันทึกเสียงผ่านผ่านการ์ดเสียง (Sound Card) ที่มีอัตราการสุ่มตัวอย่าง (Sampling Rate) 16 KHz และมีความละเอียดของข้อมูลขนาด 16 บิต มาผ่านขั้นตอนการประมวลผลสัญญาณเชิงตัวเลขเพื่อที่สามารถนำมาใช้งานในขั้นตอนต่อไป

2.1.5 การสกัดค่าคุณลักษณะเด่น

ในการรู้จำระบบเสียงไม่ได้นำเอาสัญญาณเสียงทั้งหมดไปใช้งานแต่จะพิจารณาเฉพาะค่าสำคัญของเสียงออกมาที่ถูกเรียกค่าสำคัญนี้ว่าการสกัดค่าคุณลักษณะเด่นของเสียง (Speech Feature Extraction) ซึ่งวัตถุประสงค์หลักของการสกัดค่าคุณลักษณะเด่น คือ การหาค่าที่จะใช้แทนสัญญาณเสียงด้วยกลุ่มตัวเลขที่ถูกลดทอนจำนวน แต่ต้องคงไว้ซึ่งลักษณะเด่นของเสียงด้วย

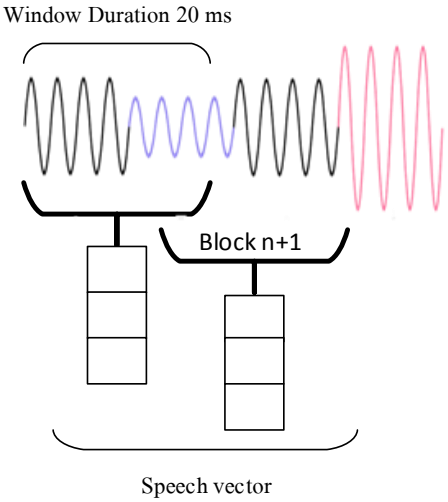
การสร้างระบบรู้จำเสียงพูดในงานวิจัยนี้ ได้เลือกใช้ค่าคุณลักษณะเด่น คือ สัมประสิทธิ์เคปสตรัมที่คำนวณบนสเกลเมล (MFCC) เป็นการหาค่าลักษณะเด่นโดยในการหาค่าเคปสตรัมจะมีการปรับสเกลของคลื่นความถี่ให้อยู่บนสเกลที่เหมาะสมสำหรับการตอบสนองของระบบการได้ยินของมนุษย์ ที่ถูกเรียกว่า สเกลเมล (Mel Scale) ซึ่งมีขั้นตอนการสกัดค่าลักษณะเด่นดังรูปที่ 2.3



รูปที่ 2.3 ขั้นตอนในการสกัดคุณลักษณะเด่นด้วยMFCC [8]

1. การเน้นสัญญาณเบื้องต้น (Pre-emphasis) เป็นการประมวลผลเสียงเบื้องต้นด้วยการนำสัญญาณเสียงผ่านวงจรกรองสัญญาณเพื่อลดทอนสัญญาณรบกวน

2. การแบ่งสัญญาณเป็นส่วนย่อย (Block to Frame) โดยใช้วินโดว์ (Window) ขนาด 20 มิลลิวินาทีและมีการใช้วินโดว์ Overlap กันขนาด 10 มิลลิวินาที ดังแสดงในรูปที่ 2.4



รูปที่ 2.4 การแบ่งสัญญาณเสียงเป็นส่วนย่อย

3. การวางกรอบแบบแฮมมิง (Hamming Window) เป็นขั้นตอนการวางกรอบแฮมมิงในสัญญาณที่ถูกแบ่งเป็นส่วนย่อยเพื่อเน้นสัญญาณบริเวณกลางเฟรมและลดความไม่ต่อเนื่องของสัญญาณบริเวณรอยต่อของเฟรม

4.การสกัดค่าคุณลักษณะเด่น Mel-Frequency Cepstral Coefficient โดยการคำนวณหาค่าสเปกตรัมพลังงานรวมทั้งปรับสเกลของสเปกตรัมให้อยู่ในรูปแบบของสเกลเมล โดยใช้การแปลงฟูเรียร์ผ่านสัญญาณเข้าไปยังตัวกรองแบบสามเหลี่ยมบนสเกลเมลตามสมการที่ 2.2 [9]

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \qquad (2.2)$$

เมื่อ f คือ ช่วงความถี่ที่ยังเป็นเชิงเส้น โดยในขั้นตอนนี้ผู้ใช้ได้กำหนดพารามิเตอร์สำหรับการคำนวณค่าลักษณะเด่นได้ดังนี้

-ระยะเวลาของเฟรม frame duration (ms) ค่าที่นิยมใช้อยู่ในช่วง 20-30 มิลลิวินาที

-ระยะห่างระหว่างเฟรม frame shift (ms))ค่าที่นิยมใช้อยู่ในช่วง 10-13 มิลลิวินาที

-จำนวนสัมประสิทธิ์สเปกตรัม ใช้ประมาณ 12-13ค่า

5.การปรับปรุงค่าลักษณะเด่น MFCC

ค่าลักษณะเด่นที่ได้จากการสกัดแบบ MFCC อย่างเดียวนั้นอาจจะไม่เพียงพอต่อการนำมารู้จำ จึงมีการเพิ่มลักษณะเด่นบางอย่างเข้าไป

ค่าอนุพันธ์อันดับหนึ่ง MFCC+D เป็นการนำค่าความแตกต่างของคุณลักษณะเด่นที่อยู่ติดกัน ซึ่งการเพิ่มค่าค่าอนุพันธ์อันดับหนึ่งทำให้ค่าลักษณะเด่นมีจำนวนมากขึ้นเป็น 2 เท่า (13+13 = 26 feature)

ค่าอนุพันธ์อันดับสอง MFCC+DA เป็นการนำค่าความแตกต่างของคุณลักษณะเด่นที่อยู่ติดกันของค่าอนุพันธ์อันดับหนึ่ง MFCC+D ซึ่งการเพิ่มค่าค่าอนุพันธ์อันดับสองทำให้ค่าลักษณะเด่นมีจำนวนมากขึ้นเป็น 3 เท่า

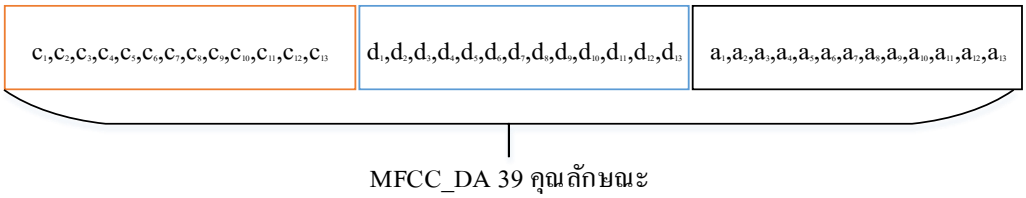
ตัวอย่างผลการสกัดคุณลักษณะเด่นและการปรับปรุงค่าลักษณะเด่น

MFCC = $[c_1,c_2, c_3,c_4 ,c_5,c_6 ,c_7,c_8 ,c_9,c_{10} ,c_{11},c_{12},c_{13}]^T$

ค่าอนุพันธ์อันดับหนึ่ง= $[d_1,d_2, d_3,d_4 ,d_5,d_6 ,d_7,d_8 ,d_9,d_{10} ,d_{11},d_{12},d_{13}]^T$

ค่าอนุพันธ์อันดับสอง= $[a_1,a_2, a_3,a_4 ,a_5,a_6 ,a_7,a_8 ,a_9,a_{10} ,a_{11},a_{12},a_{13}]^T$

จากนั้นจึงนำค่าอนุพันธ์ที่ได้มาจัดเรียงข้อมูลดังรูปที่2.5



รูปที่ 2.5 จำนวนค่าลักษณะในแบบ MFCC_DA

เมื่อสกัดข้อมูล ลักษณะโดยใช้ MFCC จำนวนค่าลักษณะจะมี 13 ค่า

MFCC = $[c_1 ,\ldots c_{13}]^T$

เมื่อคำนวณค่าอนุพันธ์อันดับหนึ่ง จำนวนค่าลักษณะเด่นเพิ่มขึ้นเป็น 26 ค่า

MFCC+D = $[c_1,\ldots c_{13}, d_1,\ldots d_{13}]^T$

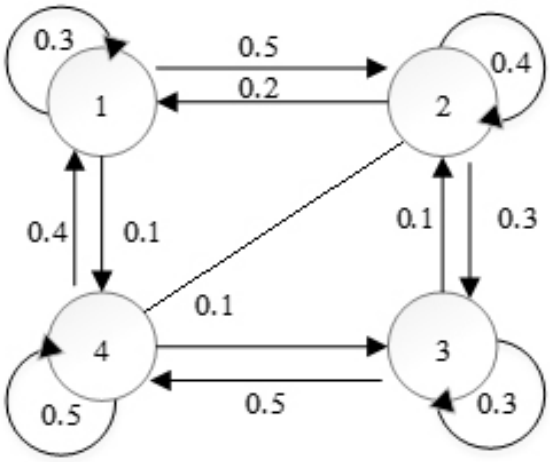
เมื่อคำนวณค่าอนุพันธ์อันดับหนึ่งและอันดับสอง จำนวนค่าลักษณะเด่นเพิ่มขึ้นเป็น 39 ค่า

MFCC+DA = $[c_1,\ldots c_{13}, d_1,\ldots d_{13} ,a_1,\ldots a_{13}]^T$

2.1.6 เทคนิคการรู้จำ

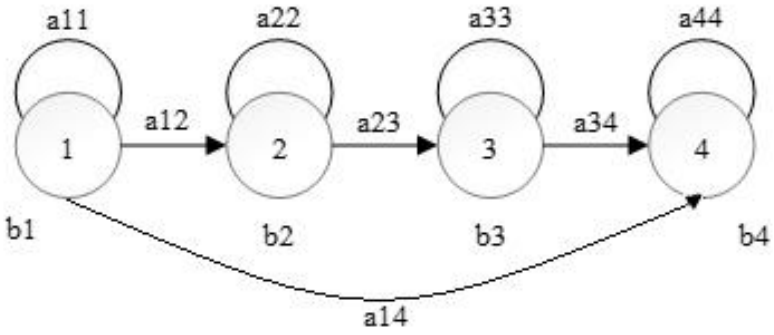
เทคนิควิธีการรู้จำเป็นกระบวนการทางคอมพิวเตอร์ที่ใช้ในการเปรียบเทียบข้อมูลที่เข้ามาเพื่อนำมาตัดสินใจในการรู้จำ ซึ่งในปัจจุบันเทคนิควิธีการรู้จำมีมากมายให้เลือกใช้งาน ดังนั้นวิธีการเลือกใช้งานเทคนิคจะขึ้นอยู่กับข้อกำหนดลักษณะของงาน ซึ่งเทคนิคการรู้จำที่เป็นที่นิยมใช้กันอย่างแพร่หลายในงานวิจัยต่าง ๆ คือ การรู้จำโดยใช้โครงข่ายประสาทเทียม[10], แบบจำลองฮิดเดนมาร์คอฟ [11] และซัพพอร์ตเวกเตอร์แมชชีน [12] ซึ่งในที่นี้จะได้อธิบายถึงหลักการเบื้องต้นของเทคนิคการรู้จำเพียงเทคนิคเดียว คือ แบบจำลองฮิดเดนมาร์คอฟ

ฮิดเดนมาร์คอฟโมเดล (Hidden Markov Model) พัฒนาขึ้นมาจากแบบจำลองมาร์คอฟ (Markov Models) ที่คิดค้นขึ้นโดย Andrei Andreevich Markov โดยที่แบบจำลองฮิดเดนมาร์คอฟเป็นเทคนิคของวิธีการเรียนรู้ทางสถิติ โดยหลักการทำงานของแบบจำลองฮิดเดนมาร์คอฟ เป็นการใช้ความน่าจะเป็นของลำดับของเหตุการณ์ที่สนใจมาใช้ในการอธิบายเกี่ยวกับสถานะในแบบจำลอง โดยที่ผู้ใช้จะไม่สามารถรู้ได้ว่าลำดับของสถานะคืออะไร แต่สามารถสังเกตเห็นได้แค่เพียงความน่าจะเป็นของผลลัพธ์ในแต่ละสถานะเท่านั้น ดังตัวอย่างที่แสดงในรูปที่ 2.6 ทุก ๆ สถานะจะถูกแสดงด้วยวงกลมและการเคลื่อนย้ายของสถานะจะถูกแสดงด้วยลูกศร โดยโครงสร้างของแบบจำลองฮิดเดนมาร์คอฟสามารถจำแนกออกเป็น 3 ชนิด คือ ก) Unconstrained Model สำหรับแบบจำลองนี้มีจุดเด่นที่ทุก ๆ สถานะสามารถที่จะเคลื่อนย้ายสถานะไปยังสถานะอื่น ๆ ได้ทุกสถานะ ข) Constrain Serial Model สำหรับแบบจำลองนี้จะเคลื่อนย้ายสถานะจากซ้ายไปขวาไม่สามารถการเคลื่อนย้ายย้อนกลับมาสถานะเดิมได้ ค) Constrain Parallel Model สำหรับแบบจำลองนี้จะมีการทำงานคล้ายกับแบบ Constrain serial model แต่การเคลื่อนย้ายสถานะจะเป็นแบบขนานทำให้มีความซับซ้อนมากขึ้น



รูปที่ 2.6 โครงสร้างแบบจำลองฮิดเดนมาร์คอฟโมเดล

แบบจำลองฮิดเดนมาร์คอฟจะมีจุดเด่นที่สามารถวิเคราะห์รายละเอียดของข้อมูลทางสถิติเกี่ยวกับลักษณะการเปลี่ยนแปลงตามเวลาได้มากกว่าการรู้จำแบบการจับคู่กับต้นแบบ (Template Matching) เนื่องจากข้อกำหนดลักษณะของงานทั้งด้านสัญญาณเสียงและภาพเคลื่อนไหวมีลักษณะการเปลี่ยนแปลงตามเวลา ดังนั้นการนำแบบจำลองฮิดเดนมาร์คอฟมาใช้งานในการรู้จำถือว่ามีความเหมาะสมมากที่สุด โดยจะใช้แบบจำลองฮิดเดนมาร์คอฟในแบบของ Constrain serial model ที่มีอินพุตที่ถูกเข้าทางโหนดที่ 1 และเคลื่อนย้ายสถานะไปยังโหนดถัดไปและเคลื่อนย้ายสถานะจนกระทั้งอินพุตออกทางโหนดสุดท้าย โดยการเคลื่อนย้ายสถานะของอินพุตจะเคลื่อนย้ายจากซ้ายไปขวาหรือวนลูปที่โหนดเดิมไม่มีการเคลื่อนย้ายย้อนกลับแสดงดังรูปที่ 2.7 เราจึงเรียกแบบจำลองฮิดเดนมาร์คอฟแบบนี้ว่า Left-to-Right Model



รูปที่ 2.7 โครงสร้างแบบจำลองฮิดเดนมาร์คอฟแบบ Left-to-Right

2.2 ระบบการอ่านริมฝีปาก

ในส่วนนี้จะกล่าวถึงทฤษฎีพื้นฐานต่าง ๆ ที่เกี่ยวข้องกับการอ่านริมฝีปากการค้นหาใบหน้าโดยใช้คุณลักษณะฮาร์ไลค์ (Haar-like Feature) ส่วนในด้านการติดตามริมฝีปากจะใช้หลักการทางสถิติในการวิเคราะห์เพื่อค้นหารูปร่างของริมฝีปาก

การอ่านริมฝีปาก คือ การที่ผู้ฟังพยายามอ่านและวิเคราะห์คำพูดโดยการสังเกตจากลักษณะการเคลื่อนไหวของริมฝีปากของผู้พูด เพื่อให้เข้าใจความหมายตรงกันเรื่องที่ผู้พูดกล่าวถึง การสังเกตการณ์การเคลื่อนไหวของริมฝีปากของผู้พูดนี้ เรียกว่าการอ่านคำพูด

2.2.1 ประมวลผลภาพ (Image processing)

การประมวลผลภาพ คือ การนำรูปภาพ (Image) หรือภาพวีดีทัศน์(Video) มาวิเคราะห์ และประมวลผลในด้านการประมวลผลสัญญาณเพื่อให้ได้ข้อมูลรูปภาพที่ต้องการโดยผ่านกระบวนการขั้นตอนต่าง ๆ เช่น การย่อและขยายรูปภาพ การแบ่งย่อยรูปภาพ การกำจัดสัญญาณรบกวนจากภาพ และการแปลงสีของรูปภาพเป็นต้น เพื่อนำมาประยุกต์สร้างเป็นระบบที่สามารถใช้ประโยชน์ในงานด้านต่าง ๆ เช่น การรู้จำใบหน้า การตรวจหาภาพเซลล์มะเร็งในทางการแพทย์ ระบบอ่านป้ายทะเบียนรถยนต์ และอีกมากมาย ๆ เป็นต้น

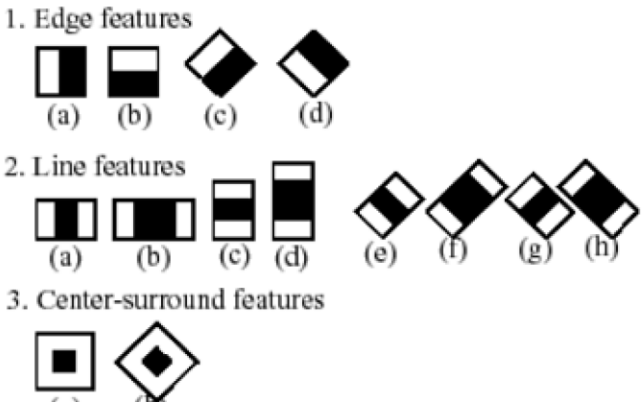
งานวิจัยชิ้นนี้ได้นำความรู้เรื่องประมวลผลภาพมาใช้ประโยชน์ในการวิเคราะห์ภาพเคลื่อนไหวจากกล้องเว็บแคมมาใช้งานในการประมวลผลภาพเบื้องต้นและเป็นพื้นฐานสำหรับประยุกต์ใช้ในเทคนิคต่าง ๆ ได้แก่ เทคนิคการตรวจจับใบหน้าและการค้นหาริมฝีปาก

2.2.2. การตรวจจับใบหน้าและการติดตามการเคลื่อนไหว (Detection and Tracking Motion)

ในส่วนนี้จะนำเสนอทฤษฎีในการหาตำแหน่งการเคลื่อนไหวของริมฝีปากและการสกัดคุณลักษณะเด่นที่นำมาใช้ในระบบด้านการประมวลผลภาพเพราะการหาตำแหน่งของการเคลื่อนไหวริมฝีปากเพื่อจะนำมาสกัดข้อมูลนั้นถือเป็นขั้นตอนที่สำคัญที่สุดในงานวิจัยนี้และเป็นตัวชี้วัดประสิทธิภาพความแม่นยำของระบบ โดยเริ่มต้นด้วยการตรวจจับใบหน้าด้วย Haar-Like Feature หลักจากนั้นค้นหาและติดตามตำแหน่งการเคลื่อนไหวริมฝีปาก ด้วยทฤษฎีที่มีความแม่นยำและนิยมใช้งานอย่างแพร่หลายในงานวิจัยด้านนี้ คือ ทฤษฎีแบบจำลองทางสถิติ 3 แบบ ได้แก่ Active Shape Model, Active Appearance Model และ Constrained Local Model

2.2.2.1 การตรวจจับใบหน้าด้วย Haar-Like Feature

Haar-Like Features ตามวิธีของ Viola และ Jones [13] เป็นการหาวัตถุที่ต้องการภายในรูปภาพโดยใช้ตัวกรองตามพื้นฐานของ Haar wavelet ลักษณะตัวกรองของ Haar-Like มีทั้งหมด 14 รูปแบบ ดังรูปที่ 2.8 โดยแต่ละรูปแบบจะมีลักษณะเป็นสี่เหลี่ยม แบ่งออกเป็นสีขาวและสีดำที่แสดงถึงผลต่างระหว่างพื้นที่ของวัตถุสามารถแบ่งเป็น 3 รูปแบบคุณลักษณะ ได้แก่ คุณลักษณะของขอบ คุณลักษณะของเส้น และคุณลักษณะของจุดตรงกลาง

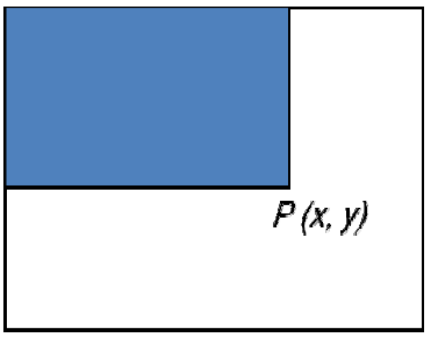


รูปที่ 2.8 รูปแบบของคุณลักษณะสำหรับการตรวจจับลักษณะแบบต่าง ๆ [13]

อัลกอริทึมนี้ได้นำเสนอวิธีการคำนวณหาผลรวมของค่าในทุก ๆ พิกเซลภายในรูปภาพเรียกว่า ”Integral Image” ดังแสดงในสมการที่ 2.3 และดังรูปที่ 2.9 ซึ่งช่วยเพิ่มประสิทธิภาพการคำนวณคุณลักษณะทำได้รวดเร็วมากขึ้น

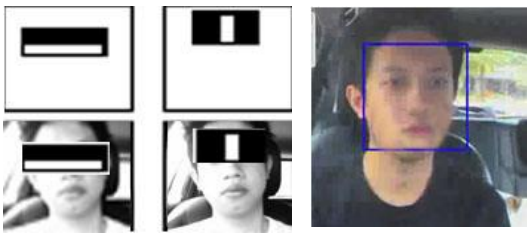
$$P(x,y)=\sum_{x'\le x,y'\le y} i(x',y') \tag{2.3}$$

เมื่อ $P(x,y)$ คือ Integral Image และ x',y' คือ จุดพิกเซลที่จะนำมาทำการค้นหาวัตถุที่ต้องการ และได้มีการประยุกต์ใช้อัลกอริทึม AdaBoost [13] เป็นวิธีการเรียนรู้ในการจำแนกวัตถุที่สนใจ



รูปที่ 2.9 การคำนวณแบบ Integral Image [13]

ในการสร้างฐานข้อมูลของ Haar-Like Feature จำเป็นต้องมีการฝึกสอนระบบเป็นจำนวนมาก โดยจะฝึกสอนระบบด้วยภาพที่มีวัตถุที่ต้องการ เรียกว่า Positive Image และจะฝึกสอนระบบด้วยภาพทั่วไปที่ไม่มีวัตถุที่ต้องการ เช่น พื้นหลัง เรียกว่า Negative Image บันทึกลงในฐานข้อมูล ส่วนในด้านการค้นหาวัตถุที่ต้องการในรูปภาพใหม่จะใช้ตัวกรองสี่เหลี่ยมในการหาผลต่างระหว่างพื้นที่ดังรูปที่ 2.10a) เป็นการใช้งานตัวกรองหาผลต่างระหว่างพื้นที่เพื่อค้นหารูปภาพใบหน้าที่ต้องการ ซึ่งผลการทดสอบการค้นหารูปภาพใบหน้าภายในรถยนต์แสดงดังรูปที่ 2.10 b)



(a) (b)

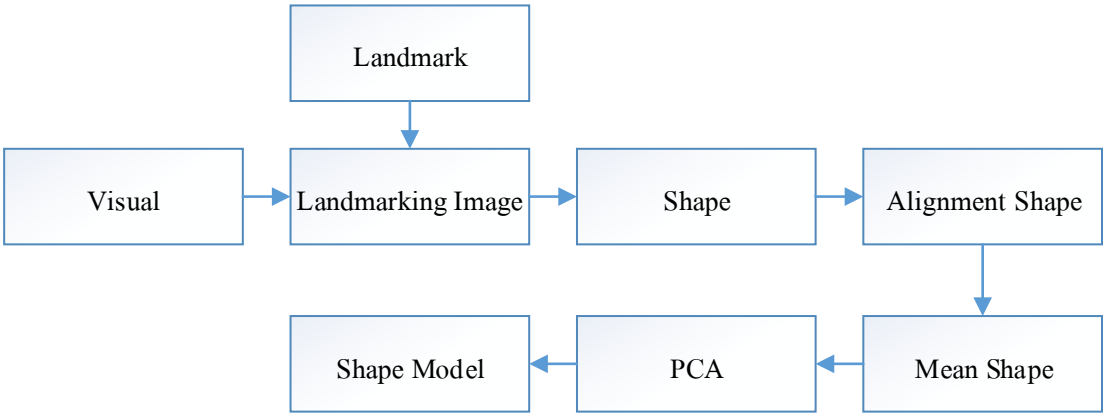
รูป 2.10 a) ตัวอย่างการใช้คุณลักษณะตรวจจับลักษณะต่าง ๆ
b) ผลการทดสอบภายในรถยนต์

ในการใช้งานในส่วนนี้ได้มีเรียกใช้ไลบรารี OpenCVhaartraining เทคนิคที่ได้รับการใช้กันอย่างแพร่หลายในการวิเคราะห์ภาพของใบหน้าและภาพทางการแพทย์

2.2.2.2 แบบจำลองรูปร่างโดยวิธีการวิเคราะห์ด้วยข้อมูลทางสถิติแบบ ASM

Active Shape Models (ASM) พัฒนาโดย Tim Cootes [14] เป็นแบบจำลองทางสถิติที่ฝึกสอนรูปร่างของวัตถุที่ต้องการ โดยใช้ Landmark ในการนำมาวิเคราะห์เพื่อปรับเปลี่ยนรูปร่างให้เหมาะสมเข้ากับรูปร่างวัตถุที่ต้องการในรูปภาพใหม่ เทคนิคนี้มีความรวดเร็วและแม่นยำสูงในการประมวลผลจึงถูกนำมาใช้งานอย่างแพร่หลายในด้านต่าง ๆ เช่นการค้นหา

รูปร่างใบหน้าการหารูปร่างของฝ่ามือ และการหารูปร่างของปอด หัวใจในทางการแพทย์ เป็นต้น โดยขั้นตอนการสร้างแบบจำลอง ASM เป็นดังรูปที่ 2.11



รูปที่ 2.11 ขั้นตอนการสร้างแบบจำลองรูปร่าง ASM [14]

เริ่มต้นจากการเตรียมชุดข้อมูลฝึกโดยการออกแบบโครงสร้างรูปร่างเรียกว่า Landmark ซึ่งประกอบด้วยจุดที่ระบุตำแหน่งที่มีความสำคัญบนรูปภาพและนำค่า x และ y ของโครงสร้างรูปร่างมาเรียงต่อกันในรูปแบบเวกเตอร์ดังสมการที่ 2.4

$$x = (x_1, y_1, \ldots, x_n, y_n)^T \tag{2.4}$$

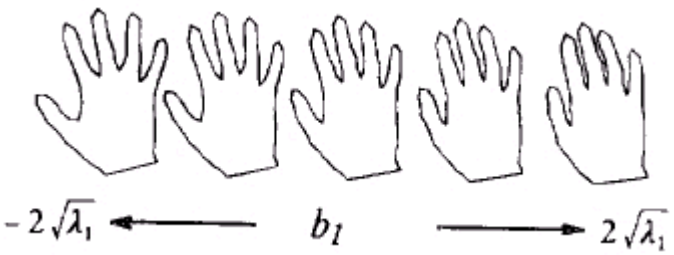
เมื่อ x เป็นเวกเตอร์ของชุดข้อมูลรูปร่างทั้งหมดหลังจากนั้นทำการจัดรูปร่างโดยการย้ายเวกเตอร์ของรูปร่างตัวอย่างเพื่อให้เข้าใกล้เวกเตอร์ของรูปร่างเฉลี่ย โดยทำให้เหลือเพียงแค่ความแตกต่างที่แท้จริงเท่านั้น ซึ่งวิธีการจัดรูปร่างแบบนี้ทำได้โดยการหาค่าความแตกต่างน้อยที่สุดดังสมการที่ 2.5 ซึ่งปรับค่าโดยใช้สมการ Transformation

$$E = \left(x_1 - M(s,\theta)[x_2]\right) - t)^T W \left(x_1 - M(s,\theta)[x_2]\right) - t) \tag{2.5}$$

เมื่อ $M(s,\theta)$ คือ การหมุนโดย θ การย่อและขยายขนาดโดย s ส่วน W คือ เมทริกน้ำหนักของทุก ๆ จุดขั้นต่อมาใช้สมการที่ 2.6 Transformation Parameters เพื่อทำการหาค่าเฉลี่ยของรูปร่างใหม่และทำซ้ำจนกว่าจะได้ค่าเฉลี่ยของรูปร่างคงที่

$$\begin{pmatrix} X_2 & -Y_2 & W & 0 \\ Y_2 & X_2 & 0 & W \\ Z & 0 & X_2 & Y_2 \\ 0 & Z & -Y_2 & X_2 \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ t_x \\ t_y \end{pmatrix} = \begin{pmatrix} X_1 \\ Y_1 \\ C_1 \\ C_2 \end{pmatrix} \tag{2.6}$$

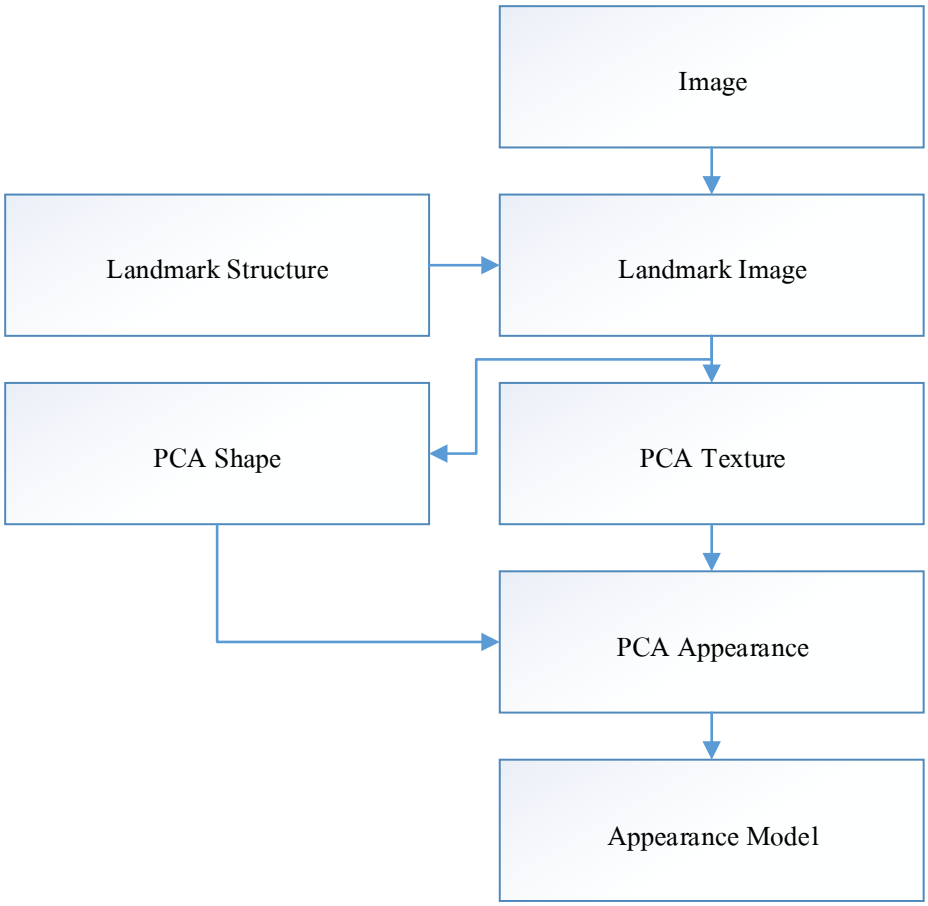
ขั้นตอนต่อไป คือ การลบเวกเตอร์ด้วยค่าเฉลี่ยของรูปร่างจากนั้นทำกระบวนการPCA [15] เพื่อหาค่าขององค์ประกอบหลักของรูปร่าง โดยแบบจำลองรูปร่างนี้มีคุณสมบัติที่สามารถแสดงให้เห็นความเปลี่ยนแปลงรูปร่างของข้อมูลที่นำมาฝึกเมื่อพารามิเตอร์มีการเปลี่ยนแปลง ดังรูปที่ 2.12 เมื่อ λ คือ ค่าของ Eigenvalue



รูปที่ 2.12 ความเปลี่ยนแปลงของรูปร่างของมือเมื่อพารามิเตอร์มีการเปลี่ยนแปลง [14]

2.2.2.3 แบบจำลองพื้นผิวโดยวิธีการวิเคราะห์ด้วยข้อมูลทางสถิติแบบ AAM

Active Appearance Model (AAM) พัฒนาโดย Tim Cootes [16] คือแบบจำลองทางสถิติที่ถูกพัฒนาขึ้นมากจากแบบจำลอง ASM ที่มีการผสมผสานแบบจำลองทางสถิติของรูปร่างและพื้นผิวเข้าด้วยกัน โดยทำการฝึกสอนรูปร่างและพื้นผิวของวัตถุที่ต้องการ โดยใช้ Landmark ของรูปร่าง และพื้นผิวของรูปภาพ ในการนำมาวิเคราะห์เพื่อปรับเปลี่ยนรูปร่างและพื้นผิวให้เหมาะสมเข้ากับวัตถุที่ต้องการในรูปภาพใหม่ ทำให้แบบจำลอง AAM สามารถค้นคืนการปรับเปลี่ยนรูปร่างและพื้นผิวของวัตถุให้พอดีในรูปใหม่ได้ โดยขั้นตอนการสร้างแบบจำลอง AAM เป็นดังรูปที่ 2.13



รูปที่ 2.13 ขั้นตอนการสร้างแบบจำลอง AAM [16]

ซึ่งชุดข้อมูลที่จะนำมาสร้างแบบจำลองนี้ต้องมีการสร้างองค์ประกอบของรูปร่างดังสมการที่ 2.7 และองค์ประกอบพื้นผิวของจากรูปภาพที่จะนำมาฝึกสอน ดังสมการที่ 2.8 โดยในการสร้างแบบจำลอง AAM เริ่มต้นด้วยการหาค่าสถิติของรูปร่างในการสร้างแบบจำลองรูปร่างก่อนด้วยการใช้ PCA เพื่อหาค่าองค์ประกอบหลักที่ใช้อธิบายรูปร่าง

$$x = \overline{x} + P_s b_s \tag{2.7}$$

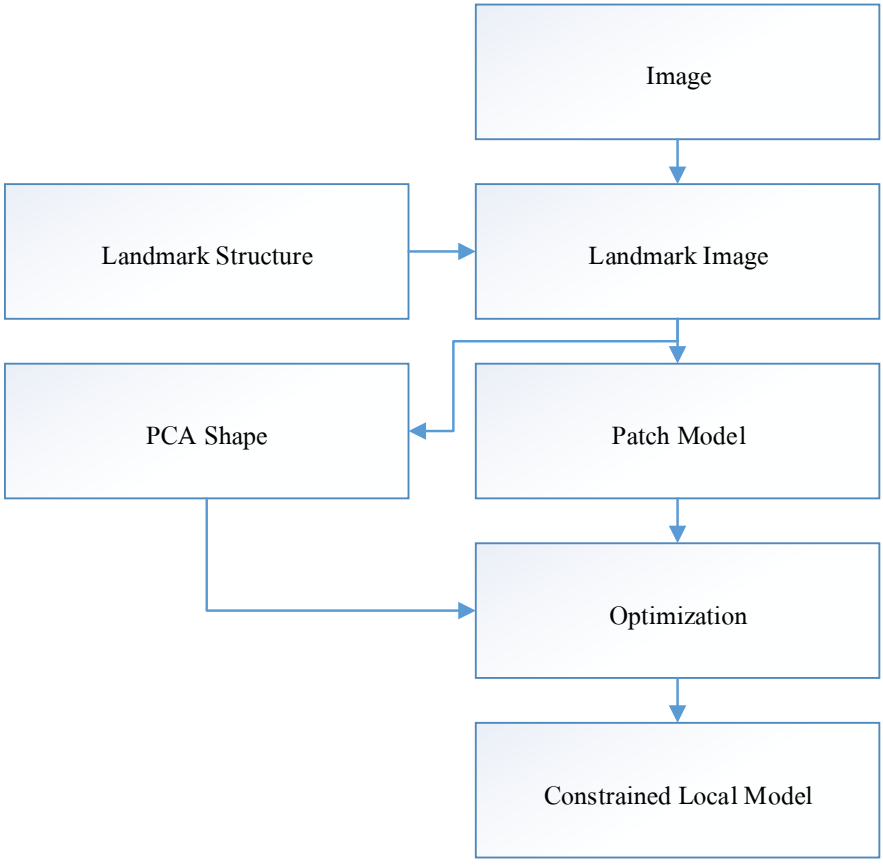
เมื่อ $\overline{x}$ คือ ค่าเฉลี่ยของรูปร่าง, b_s คือ ชุดข้อมูลพารามิเตอร์ของรูปร่าง ขั้นต่อมา คือ การสร้างแบบจำลองพื้นผิวโดยการใช้ PCA ชุดข้อมูลพื้นผิวนั้น

$$g = \overline{g} + P_g b_g \tag{2.8}$$

เมื่อ $\bar{g}$ คือ ค่าเฉลี่ยของเวกเตอร์รูปภาพ Gray Level, b_g คือ ชุดข้อมูลพารามิเตอร์ของรูปภาพ Gray Level หลังจากนั้นจะมีการใช้ PCA อีกครั้งในการรวมแบบจำลองรูปร่างและแบบจำลองรูปร่างพื้นผิวเข้าด้วยกัน

2.2.2.4 แบบจำลองพื้นผิวโดยวิธีการวิเคราะห์ด้วยข้อมูลทางสถิติแบบ CLM

Constrained Local Models (CLM) พัฒนาโดย TimCootes [17] คือ แบบจำลองทางสถิติโดยรวมแบบจำลองรูปร่างและพื้นผิวที่ถูกสร้างขึ้นมาจากชุดการฝึกอบรมจาก Landmark โดยวิธีการนี้จะมีความคล้ายกับ AAM อย่างไรก็ตาม วิธีการสุ่มตัวอย่างพื้นผิวมีความแตกต่างกัน เนื่องจาก CLM ใช้ตัวอย่างพื้นผิวที่แบ่งออกเป็นส่วนย่อยที่ถูกเรียกว่า แพทช์โมเดล (Patch model) เช่น ตา คิ้ว และปาก เป็นต้น ในการฝึกอบรมแบบจำลองนี้โดยขั้นตอนการสร้างแบบจำลอง CLM เป็นดังรูปที่ 2.14



รูปที่ 2.14ขั้นตอนการสร้างแบบจำลอง CLM [17]

ในการสร้างแบบจำลองพื้นผิวเริ่มต้นจากการเตรียมชุดข้อมูลฝึกโดยการออกแบบโครงสร้างรูปร่างเรียกว่า Landmark หลังจากนั้นสร้างแบบจำลองรูปร่างโดยการใช้ PCA และสร้างแบบจำลองพื้นผิวแบบ Local (Patch model) ดังรูปที่ 2.15 หลังจากนั้นทำการรวมแบบจำลองเข้าด้วยกัน

รูป 2.15 รูปร่างและแพทช์โมเดลของ CLM [17]

2.3 การผสมผสานระบบการรู้จำเสียงและระบบการอ่านริมฝีปาก

การผสมผสานข้อมูล (Data Fusion) คือ การนำข้อมูลจากหลากหลายข้อมูลที่ต่างกันตั้งแต่ 2 แหล่งกำเนิดขึ้นไปมาผสมผสานข้อมูลกันเพื่อที่ให้ได้ข้อมูลใหม่ที่มีความหลากหลายของข้อมูลยิ่งขึ้น ซึ่งวัตถุประสงค์หลักของการผสมผสานข้อมูล คือ เพื่อปรับปรุงและเพิ่มประสิทธิภาพของข้อมูลให้ดีขึ้น และจะสามารถนำข้อมูลที่ได้จากการผสมผสานข้อมูลดังกล่าวไปใช้งาน เช่น ช่วยเพิ่มความแม่นยำของการจำแนกตัวบุคคลโดยการรวมข้อมูลใบหน้าและลายนิ้วมือเข้าด้วยกัน เป็นต้น โดยการผสมผสานข้อมูลสามารถจำแนกออกเป็น 4 ระดับ คือ

1.ระดับสัญญาณ (Signal Fusion) เป็นการผสมผสานในระดับล่างโดยการนำสัญญาณจากแต่ละข้อมูลที่ต่างกันมารวมกันเพื่อที่จะได้สัญญาณใหม่ที่มีการปรับปรุงและเพิ่มประสิทธิภาพของสัญญาณดังกล่าว

2.ระดับพิกเซล (Pixel Fusion) เป็นการผสมผสานข้อมูลภาพพิกเซลต่อพิกเซลของข้อมูลรูปภาพที่ต่างกัน

3.ระดับคุณลักษณะ (Feature Fusion) ต้องมีการสกัดหาค่าคุณลักษณะเด่นและนำค่าคุณลักษณะเด่นของแต่ละระบบมารวมกันก่อนแล้วนำข้อมูลดังกล่าวมาจำแนก

4.ระดับตัดสินใจ (Decision Fusion) ต้องมีการสกัดคุณลักษณะเด่นแล้วนำไปจำแนกรูปแบบในแต่ละระบบก่อนหลังจากนั้นนำผลลัพธ์ที่ได้มารวมกันและทำการจำแนกอีกครั้ง

ซึ่งภายในงานวิจัยนี้ได้นำการผสมผสานข้อมูล แบบระดับคุณลักษณะและระดับตัดสินใจมาใช้งานในการผสมผสานข้อมูลระหว่างสัญญาณกับรูปภาพริมฝีปากเพื่อเพิ่มความแม่นยำในการรู้จำเสียง

2.4 การเรียนรู้ของเครื่อง (Machine Learning)

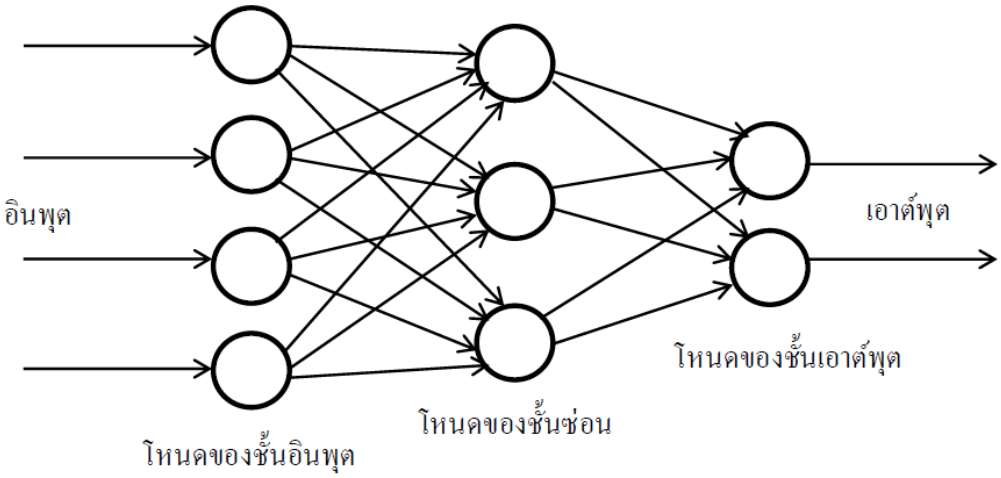
เทคนิคการเรียนรู้ของเครื่อง เป็นการฝึกสอนคอมพิวเตอร์ให้มีความชาญฉลาดในการแยกแยะวัตถุที่สนใจได้ เช่น การสร้างให้คอมพิวเตอร์สามารถแยกแยะรูปภาพใบหน้าของมนุษย์ แยกแยะรอยนิ้วมือของมนุษย์ เป็นต้น โดยทำการป้อนข้อมูลฝึกสอนให้คอมพิวเตอร์เรียนรู้เพื่อที่จะให้คอมพิวเตอร์สามารถวิเคราะห์แยกแยะรูปแบบจากสมมติฐาน และป้อนข้อมูลทดสอบสำหรับทดสอบหาค่าความแม่นยำในการแยกแยะวัตถุที่สนใจ ซึ่งในที่นี้จะอธิบายถึงหลักการเบื้องต้นของการเรียนรู้ของเครื่องแบบโครงข่ายประสาทเทียมที่จะใช้ในการหาค่าน้ำหนักที่นำมาใช้ในงานวิจัยนี้

โครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นแบบจำลองทางคอมพิวเตอร์ที่ถูกสร้างขึ้นจากการลอกเลียนของโครงสร้างและการทำงานของสมองมนุษย์ ในที่นี้เลือกใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Feedforward Backpropagation Neural Network) ดังรูปที่ 2.16 โดยโครงสร้างของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับแบ่งออกเป็น 3 ส่วนหลักดังนี้ คือ

1.จำนวนชั้นของโครงข่ายประสาทเทียมจะประกอบด้วย 3 ชั้นหลัก คือ ชั้นอินพุต (Input layer) ชั้นเอาต์พุต (Output layer) และชั้นซ่อน (Hidden layer)

2. การเชื่อมต่อระหว่างชั้นต่าง ๆ จะเชื่อมต่อกันเป็นระบบเครือข่าย โดยที่ชั้นอินพุตแต่ละโหนดจะเชื่อมต่อกับทุก ๆ โหนดในชั้นซ่อนระดับแรก และทุก ๆ โหนดในชั้นซ่อนระดับแรกจะเชื่อมต่อกับโหนดในชั้นซ่อนระดับถัดไป จนกระทั่งทุก ๆ โหนดในชั้นซ่อนระดับสุดท้ายจะเชื่อมต่อกับชั้นเอาต์พุต

3. การทำงานของชั้นต่าง ๆ ในชั้นอินพุตมีหน้าที่รับข้อมูลเพื่อส่งไปยัง ทุก ๆ โหนดในชั้นซ่อนเพื่อนำไปประมวลผล ส่วนชั้นเอาต์พุตมีหน้าที่รับข้อมูลที่ผ่านการประมวลผลจากชั้นซ่อนระดับสุดท้ายเพื่อนำข้อมูลไปในขั้นตอนต่อไป



รูปที่ 2.16 สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ [10]

บทที่ 3
ระเบียบวิธีวิจัย

ในบทนี้จะได้กล่าวถึงขั้นตอนในการทำงานของระบบและการนำเอาทฤษฎีต่าง ๆ ที่ได้กล่าวไปแล้วในบทที่ 2 มาประยุกต์ใช้ในงานวิจัยดังนี้ ซึ่งมีหัวข้อการนำเสนอ ดังนี้

3.1 ภาพรวมของระบบ

3.2 แนวคิดของงานวิจัย

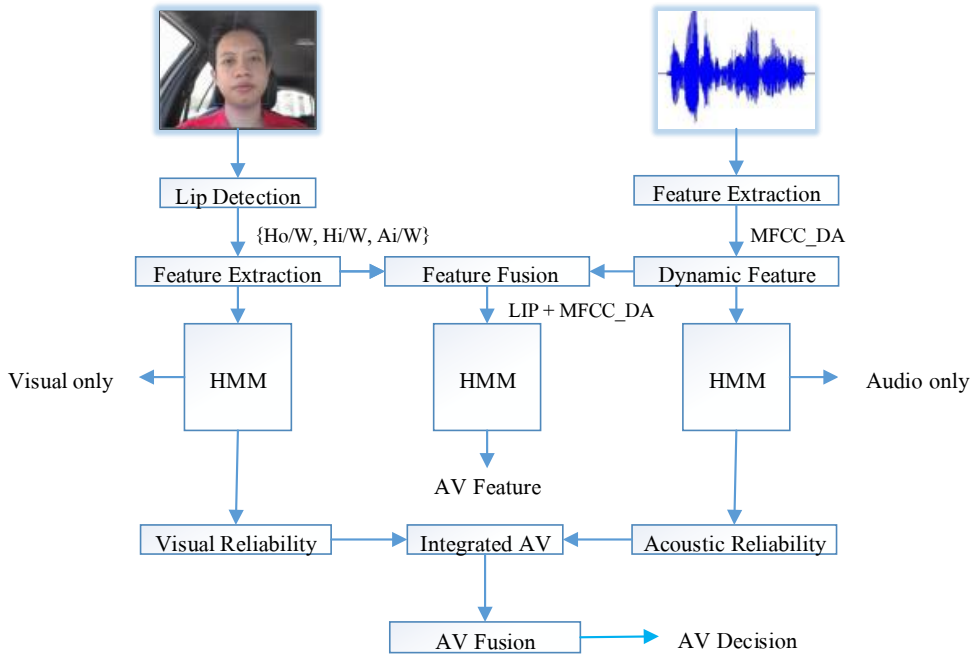
3.3 วิธีการรู้จำเสียง

3.4 วิธีการอ่านริมฝีปาก

3.5 วิธีการผสมผสานข้อมูล

3.1 ภาพรวมของระบบ

ในส่วนนี้จะกล่าวถึงการออกแบบระบบที่ใช้ในงานวิจัยนี้โดยเริ่มต้นด้วยการติดตั้งกล้องและไมโครโฟนภายในสภาพแวดล้อมภายในรถยนต์ งานวิจัยจะถูกแบ่งออกเป็นสามส่วน คือ เสียง รูปภาพ และการผสมผสานระหว่างข้อมูลเสียงกับรูปภาพโดยภาพรวมของระบบแสดงดังรูปที่ 3.1



รูปที่ 3.1 ภาพรวมของงานวิจัยการเพิ่มความแม่นยำในการรู้จำเสียงโดยการอ่านริมฝีปาก

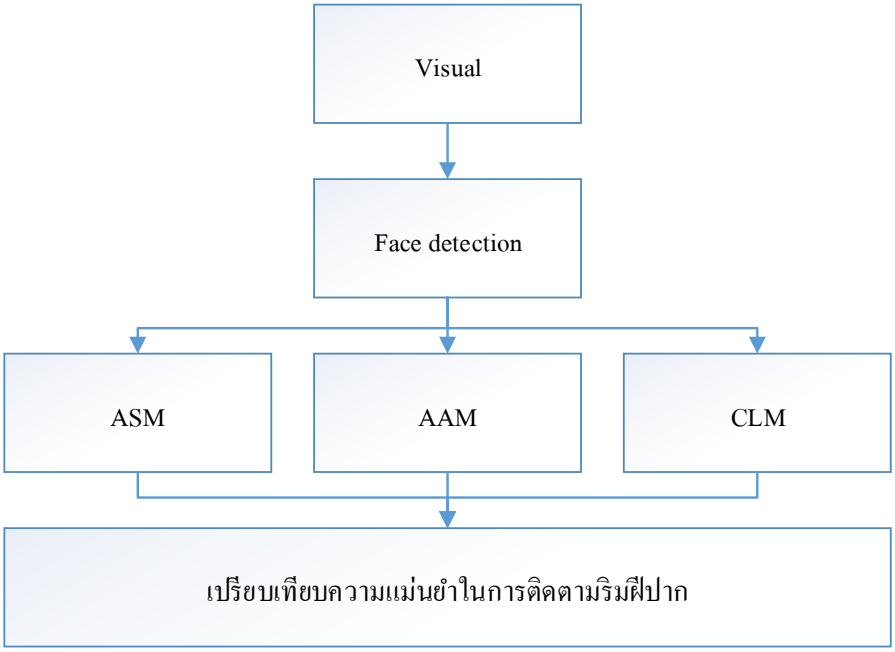
โดยระบบเสียงจะเริ่มต้นจากการรับเสียงผู้ใช้จากไมโครโฟน และบันทึกไว้ในรูปแบบไฟล์ .WAV เพื่อส่งไปยังขั้นตอนการสกัดคุณลักษณะเด่นและการรู้จำเสียง ส่วนระบบรูปภาพจะเริ่มต้นจากการรับภาพผู้ใช้ด้วยกล้อง และบันทึกไว้ในรูปแบบไฟล์ .AVI เพื่อส่งไปยังขั้นตอนการประมวลผลและรู้จำริมฝีปาก ส่วนสุดท้ายจะเป็นการผสมผสานทั้งสองระบบเข้าด้วยกัน ในแบบการผสมผสานด้านการรวมข้อมูล (Feature Fusion) และ การผสมผสานด้านการตัดสินใจ (Decision Fusion)

3.2. การดำเนินงานวิจัย

ในงานวิจัยนี้มีวัตถุประสงค์หลักในการเพิ่มความแม่นยำในการรู้จำเสียงโดยงานวิจัยนี้มีการเปรียบเทียบการรู้จำเสียงอย่างเดียว การอ่านริมฝีปาก และการผสมผสานระหว่างการรู้จำเสียงกับการอ่านริมฝีปากเข้าด้วยกัน นอกจากนี้เพื่อให้เห็นถึงความชัดเจนของประสิทธิภาพที่ได้จากแบบจำลอง ดังนั้นข้อมูลที่ใช้จึงมีทั้งแบบการรู้จำเสียงและการอ่านริมฝีปาก

3.2.1 การเปรียบเทียบแบบจำลองรูปร่างจาก ASM กับ AAM และ CLM

การทดลองนี้เป็นการทดลองใช้แบบจำลองรูปร่างโดยวิธีการวิเคราะห์ด้วยข้อมูลทางสถิติอย่าง AAM ASM และ CLM เพื่อเปรียบเทียบค้นหาความแม่นยำในการติดตามริมฝีปากว่าวิธีการใดให้ผลลัพธ์ดีที่สุด ดังรูปที่ 3.2



รูปที่ 3.2 แผนภาพกระบวนการเปรียบเทียบความแม่นยำการติดตามริมฝีปากโดยใช้แบบจำลองรูปร่างจาก ASM AAM และ CLM

1) ข้อมูลรูป (Visual) เป็นข้อมูลรูปถ่ายใบหน้าสำหรับทดสอบแบบจำลองรูปร่างข้อมูลรูปภาพใบหน้าสำหรับใช้ในการทดสอบระบบ ซึ่งสภาพแสงที่ใช้ทดสอบเป็นสภาพแสงกลางวันภายในรถยนต์และไม่มีแสงภายนอกส่องกระทบหน้าคนขับขี่รถยนต์โดยตรง

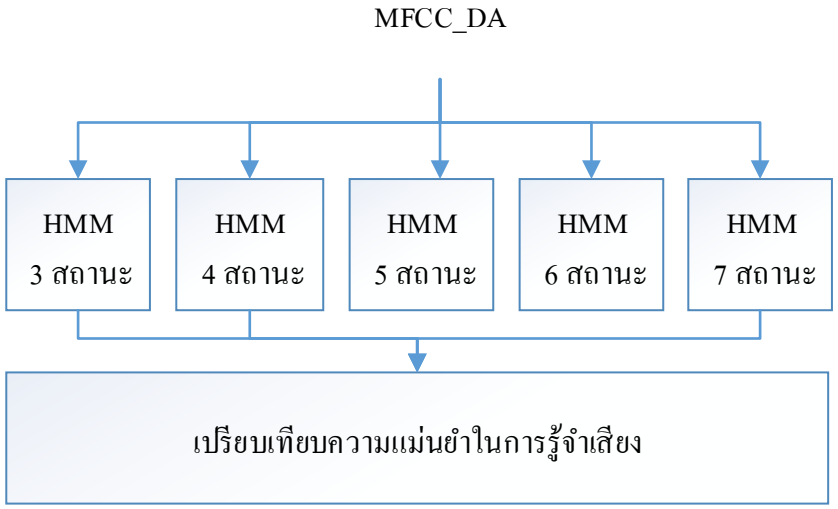
2) การค้นหาใบหน้า (Face Detection) เป็นระบบการตรวจหาใบหน้าของมนุษย์ วิธีที่นิยม คือ วิธีของ วีโอล่า โจนส์

3) ทดสอบแบบจำลองรูปร่าง ASM AAM และ CLM ขั้นตอนนี้ทำการติดตามรูปร่างของใบหน้าและริมฝีปากเพื่อจะนำรูปร่างริมฝีปากไปสกัดคุณลักษณะเด่น

4) การเปรียบเทียบแบบจำลองในขั้นตอนนี้เป็นการเปรียบเทียบแบบจำลองที่มีความแม่นยำในการติดตามริมฝีปากมากที่สุดมาใช้ในงานวิจัยนี้

3.2.2 การเปรียบเทียบจำนวนสถานะของฮิดเดนมาร์คอฟโมเดล

การทดลองนี้เป็นการทดลองหาจำนวนสถานะของฮิดเดนมาร์คอฟโมเดล ที่มีความแม่นยำมากที่สุดเพื่อนำมาใช้ในการรู้จำคำสั่งเสียง ดังรูปที่ 3.3



รูปที่ 3.3แผนภาพกระบวนการหาจำนวนสถานะที่เหมาะสม
ของฮิดเดนมาร์คอฟโมเดล (HMM)

การทดลองนี้ประกอบไปด้วย 3 ส่วน คือ

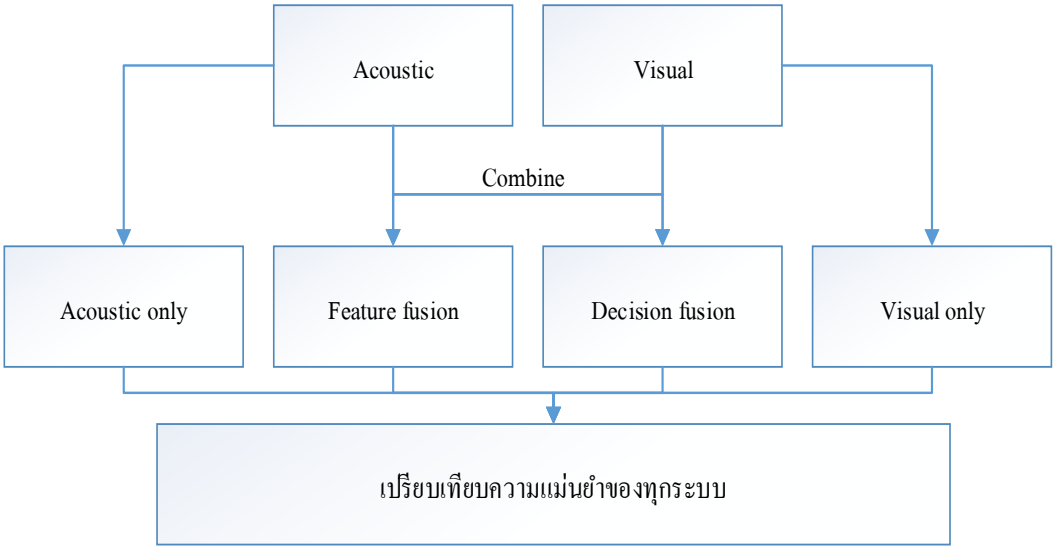
1) ข้อมูลเสียงและรูปภาพที่ถูกสกัดคุณลักษณะเด่นเพื่อนำมาหาจำนวนสถานะในการรู้จำ

2) ทดสอบด้วยจำนวนสถานะที่แตกต่างกันได้แก่ 3, 4, 5, 6 และ 7 สถานะ ตามลำดับด้วยแบบจำลองฮิดเดนมาร์คอฟโมเดล

3) การเปรียบเทียบจำนวนสถานะ ขั้นตอนนี้จะเปรียบเทียบหาจำนวนสถานะที่มีความแม่นยำในการรู้จำที่สุดมาใช้ในงานวิจัยนี้

3.2.3 การเปรียบเทียบความแม่นยำของแต่ละระบบ

การทดลองนี้เป็นการทดลองหาความแม่นยำในการรู้จำคำสั่งทั้ง 4 ระบบ ได้แก่ การรู้จำเสียงอย่างเดียว การอ่านริมฝีปากอย่างเดียว การผสมผสานในระดับคุณลักษณะ และการผสมผสานในระดับการตัดสินใจดังรูปที่ 3.4



รูปที่ 3.4 แผนภาพกระบวนการเปรียบเทียบความแม่นยำของแต่ละระบบ

การทดลองนี้ประกอบไปด้วย 6 ส่วน คือ

1) ข้อมูลเสียงและรูปภาพ เป็นการบันทึกข้อมูลเสียงและข้อมูลรูปภาพใบหน้าสำหรับใช้ในการทดสอบระบบซึ่งสภาพแสงที่ใช้ทดสอบเป็นสภาพแสงกลางวันภายในรถยนต์และไม่มีแสงภายนอกส่องกระทบหน้าคนขับขี่รถยนต์โดยตรงและบันทึกคำสั่งเสียงภายในรถยนต์ในสภาพแวดล้อมที่มีเสียงรบกวนหลายระดับ

2) การรู้จำเสียงอย่างเดียวเป็นระบบที่ใช้ข้อมูลเสียงอย่างเดียวเท่านั้นในการรู้จำคำสั่ง

3) การอ่านริมฝีปากอย่างเดียวเป็นระบบที่ใช้ข้อมูลรูปภาพริมฝีปากอย่างเดียวเท่านั้นในการรู้จำคำสั่ง

4) การผสมผสานในระดับคุณลักษณะเป็นระบบที่ใช้ข้อมูลเสียงและข้อมูลรูปภาพริมฝีปากมาผสมผสานในระดับคุณลักษณะในการรู้จำคำสั่ง

5) การผสมผสานในระดับการตัดสินเป็นระบบที่ใช้ข้อมูลเสียงและข้อมูลรูปภาพริมฝีปากมาผสมผสานในระดับการตัดสินใจในการรู้จำคำสั่ง

6) การเปรียบเทียบแต่ละระบบ ขั้นตอนนี้เป็นขั้นตอนที่เปรียบเทียบหาความแม่นยำในการรู้จำคำสั่งทุก ๆ ระบบ

3.3 วิธีการรู้จำเสียง

ในส่วนการรู้จำเสียงของงานวิจัยนี้จะดึงคุณลักษณะเด่นของเสียงพูดด้วยอัลกอริทึมที่นิยมใช้งานอย่างแพร่หลายในปัจจุบัน คือ Mel-Frequency Cepstral Coefficients (MFCC_DA) และงานวิจัยหลาย ๆ งานได้นำเสนอใช้งานฮิดเดนมาร์คอฟโมเดลในการรู้จำเสียงเนื่องจากอัลกอริทึมนี้มีประสิทธิภาพมากที่สุดในงานรู้จำเสียง ดังนั้นงานวิจัยนี้จะประยุกต์ใช้แบบจำลองฮิดเดนมาร์คอฟแบบซ้ายไปขวาจำนวน 5 สถานะ ตามที่แสดงผลในรูปที่ 3.5 ที่ได้นำเสนอระบบโดยรวมของการรู้จำเสียง

รูปที่ 3.5 ภาพรวมของระบบรู้จำเสียงอย่างเดียว

3.3.1 หลักเกณฑ์ในการบันทึกเสียง

การเก็บข้อมูลเสียงที่ใช้ในการสอนและทดสอบระบบเริ่มจากระบบรับเสียงผู้ใช้จากไมโครโฟน และบันทึกไฟล์เสียงในรูปแบบ Waveform Audio File Format (WAV) โดยมีอัตราการสุ่มตัวอย่าง (Sampling Rate) ที่ 16,000 เฮิรตซ์ ขนาด 16 บิต บันทึกเสียงภายใต้สภาพแวดล้อมภายในรถยนต์ โดยเสียงที่ใช้ในการสอนระบบจะบันทึกเสียงภายในโรงรถที่ไม่มีสัญญาณรบกวนจากภายนอก ส่วนเสียงที่ใช้ในการทดสอบจะแบ่งออกเป็น 2 ส่วน คือ สร้างเสียงรบกวนเข้าไปในระดับ 20, 10, และ 5 เดซิเบล ตามลำดับ และบันทึกเสียงสภาพแวดล้อมภายในรถยนต์ในขณะที่รถยนต์กำลังเคลื่อนที่ด้วยความเร็ว 20, 40, 60 และ 80-100 กิโลเมตรต่อชั่วโมง ตามลำดับ

3.3.2 คำสั่งที่ใช้ทดสอบ

เสียงคำสั่งภาษาไทยจำนวนคำสั่งละ 20 ชุดเพื่อฝึกฝนระบบ และอีก 20 ชุดในการทดสอบระบบ คำสั่งที่ใช้ในการทดสอบระบบ มีทั้งหมด 20 คำสั่งดังตารางที่ 3.1

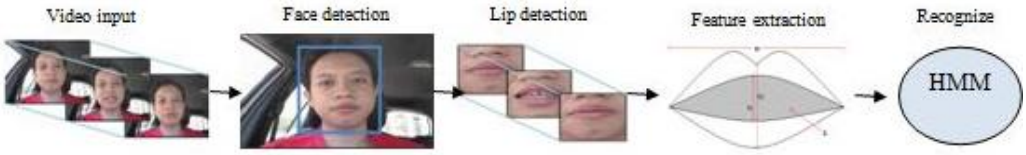
ตารางที่ 3.1 คำสั่งภาษาไทยที่ใช้ฝึกฝนและทดสอบ

คำสั่งภาษาไทย			
หนึ่ง	หก	เปิด	ก่อนหน้า
สอง	เจ็ด	ปิด	ซีดี
สาม	แปด	เล่น	เอฟ-เอ็ม
สี่	เก้า	หยุด	เอ-เอ็ม
ห้า	ศูนย์	ถัดไป	วิทยุ

ในงานวิจัยนี้จะทดลองระบบสั่งการด้วยเสียงพูดภาษาไทย คำที่ใช้ทดสอบเป็นคำสั้น ๆ ภาษาไทย โดยมีการทดสอบกับตัวเลข (0-9) และคำสั่งในภาษาไทยไม่เกิน 3 พยางค์ จำนวนไม่ต่ำกว่า 10 คำสั่ง เช่น เปิด ปิด ถัดไป ก่อนหน้า เป็นต้น

3.4 วิธีการอ่านริมฝีปาก

ในส่วนนี้จะนำเสนอวิธีการในการหาตำแหน่งการเคลื่อนไหวของริมฝีปากและการสกัดคุณลักษณะเด่นที่นำมาใช้ในระบบด้านรูปภาพเพราะการหาตำแหน่งของการเคลื่อนไหวริมฝีปากเพื่อจะนำมาสกัดข้อมูลนั้นถือเป็นขั้นตอนที่สำคัญที่สุดในงานวิจัยนี้และเป็นตัวชี้วัดประสิทธิภาพความแม่นยำของระบบตามที่แสดงผลดังรูปที่ 3.6 ได้นำเสนอระบบโดยรวมของการรู้จำรูปภาพริมฝีปาก

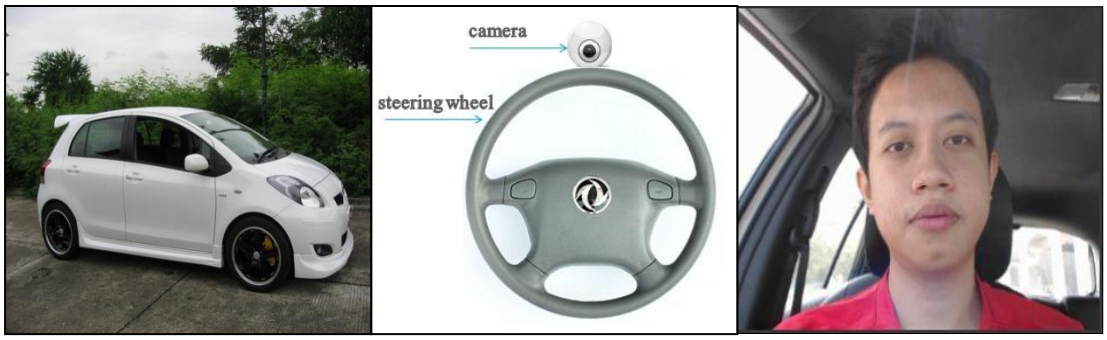


รูปที่ 3.6 ระบบโดยรวมของการรู้จำรูปภาพริมฝีปาก

โดยเริ่มต้นจากการนำรูปภาพวิดีโอที่พูดภายในรถยนต์ในสภาพหน้าตรงระบบจะนำรูปภาพวีดีโอมาวิเคราะห์หาใบหน้าด้วยอัลกอริทึม Haar-Like Features และงานวิจัยนี้จะเลือกใช้ Constrained Local Model (CLM) ในการติดตามและหาตำแหน่งของริมฝีปากภายนอกและภายในหลังจากนั้นนำรูปร่างของริมฝีปากในแต่ละเฟรมมาสกัดคุณลักษณะเด่น

3.4.1 การติดตั้งกล้องและบันทึกรูปภาพใบหน้าภายในรถยนต์

งานวิจัยนี้ได้ทดลองกับรถ โตโยต้า ยาริส (YARIS 1500 ซีซี) โดยมีการออกแบบการติดตั้งกล้องเว็บแคมข้างบนพวงมาลัยรถยนต์ของตามที่แสดงดังรูปที่ 3.7 และถ่ายรูปหน้าตรงเท่านั้นในสภาวะที่รถยนต์ไม่มีการเคลื่อนที่ และรถยนต์มีการเคลื่อนที่ด้วยความเร็ว 0, 20, 40, 60 และ 80 กิโลเมตรต่อชั่งโมง สภาพแสงที่ใช้ทดสอบเป็นสภาพแสงกลางวันภายในรถยนต์ และไม่มีแสงภายนอกส่องกระทบหน้าคนขับขี่รถยนต์โดยตรง



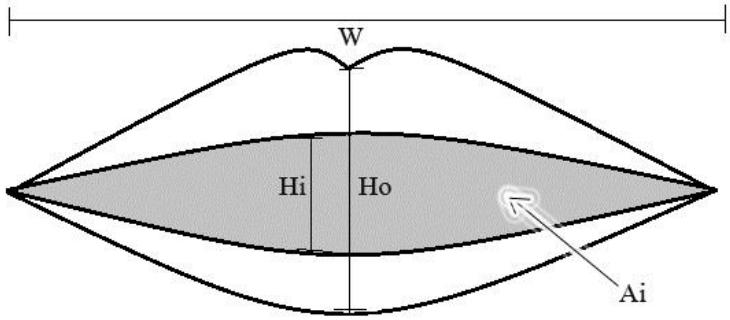
(ก) (ข) (ค)

รูปที่ 3.7 ก) รถยนต์ที่ใช้การทดสอบงานวิจัย ข) การติดตั้งกล้องบนพวงมาลัย
ค) ตัวอย่างรูปภาพใบหน้าในการติดตั้งกล้องบนพวงมาลัย

การเก็บข้อมูลรูปภาพใบหน้าที่ใช้ในการสอนและทดสอบระบบมีคำสั่งละ 20 ชุดเพื่อฝึกฝนระบบ และ อีก 20 ชุดในการทดสอบระบบภาพ โดยรูปภาพต้องมีความละเอียดไม่ต่ำกว่า 640x480 พิกเซลและมีอัตราเฟรมภาพ 30 เฟรมต่อวินาที

3.4.2 การหาค่าลักษณะเด่น

การหาค่าลักษณะเด่น (Features Extraction) จะกระทำหลังจากที่ได้รูปร่างริมฝีปาก โดยที่ในงานวิจัยนี้ จะใช้ความสูงของริมฝีปากภายนอก (H_o) ความสูงริมฝีปากภายใน (H_i) และพื้นที่ภายในปาก (A_i) ดังแสดงในรูปที่ 3.8



รูปที่ 3.8 ความสูง (H) และ ความกว้าง (W) และพื้นที่ภายในปาก(A)ของริมฝีปาก

ก่อนจะนำคุณลักษณะเด่นที่ได้มาใช้งานจำเป็นต้องมีการ Normalized Feature โดยมาหาอัตราส่วนระหว่างความสูงของริมฝีปากและความกว้างของริมฝีปากดังนั้นค่าลักษณะเด่นที่จะนำมาใช้ ดังสมการที่ 3.1

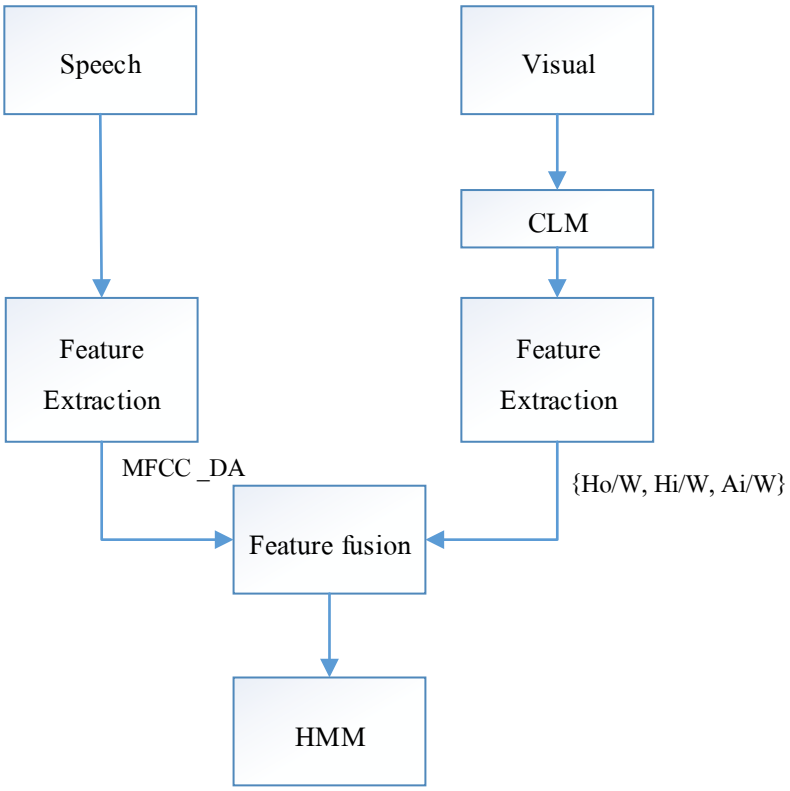
$$\{ \frac{H_o}{W}, \frac{H_i}{W} \text{และ } \frac{A_i}{W} \}^T \tag{3.1}$$

เมื่อ $\frac{H_o}{W}, \frac{H_i}{W}$ และ $\frac{A_i}{W}$ คือ ความสูงของริมฝีปากภายนอก ความสูงริมฝีปากภายในและพื้นที่ภายในปาก ที่มีการ Normalized Feature ด้วยความกว้างของริมฝีปาก ตามลำดับ

3.5 วิธีการผสมผสานข้อมูล

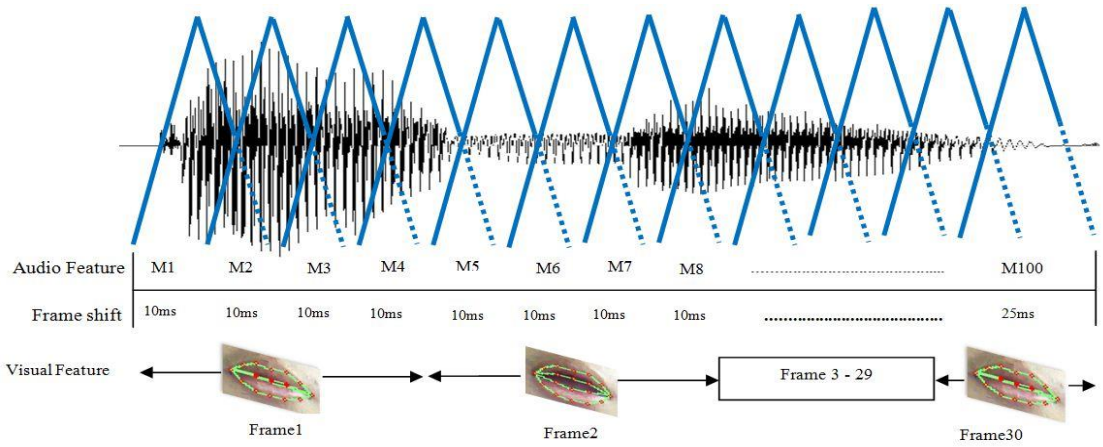
3.5.1 การผสมผสานในระดับคุณลักษณะ (Feature Fusion)

การผสมผสานในระดับคุณลักษณะจะกระทำหลังจากขั้นตอนที่สกัดข้อมูลคุณลักษณะเด่นจากเสียงและริมฝีปาก โดยระบบจะนำข้อมูลเมตริกคุณลักษณะเด่นของเสียงไปรวมกับเมตริกคุณลักษณะเด่นของริมฝีปากในแต่ละเมตริกของแต่ละคำสั่ง หลังจากนั้นนำเข้าไปฝึกฝนและทดสอบในฮิดเดนมาร์คอฟโมเดล ตามรูปที่ 3.9



รูปที่ 3.9 ระบบรวมของการผสมผสานในระดับคุณลักษณะ

การผสมผสานในระดับคุณลักษณะเริ่มต้นด้วยการสกัดข้อมูลโดยที่ค่า MFCC_DA นี้เก็บเป็นเวกเตอร์ เวกเตอร์หนึ่งแทนสัญญาณเสียงยาวประมาณ 25 มิลลิวินาที แต่ละเวกเตอร์แทนสัญญาณเสียง ที่ค่อย ๆ เลื่อนไปแบบคาบเกี่ยวกัน โดยงานวิจัยจะเลื่อนไปทีละ 10 มิลลิวินาที ดังรูปที่ 3.10

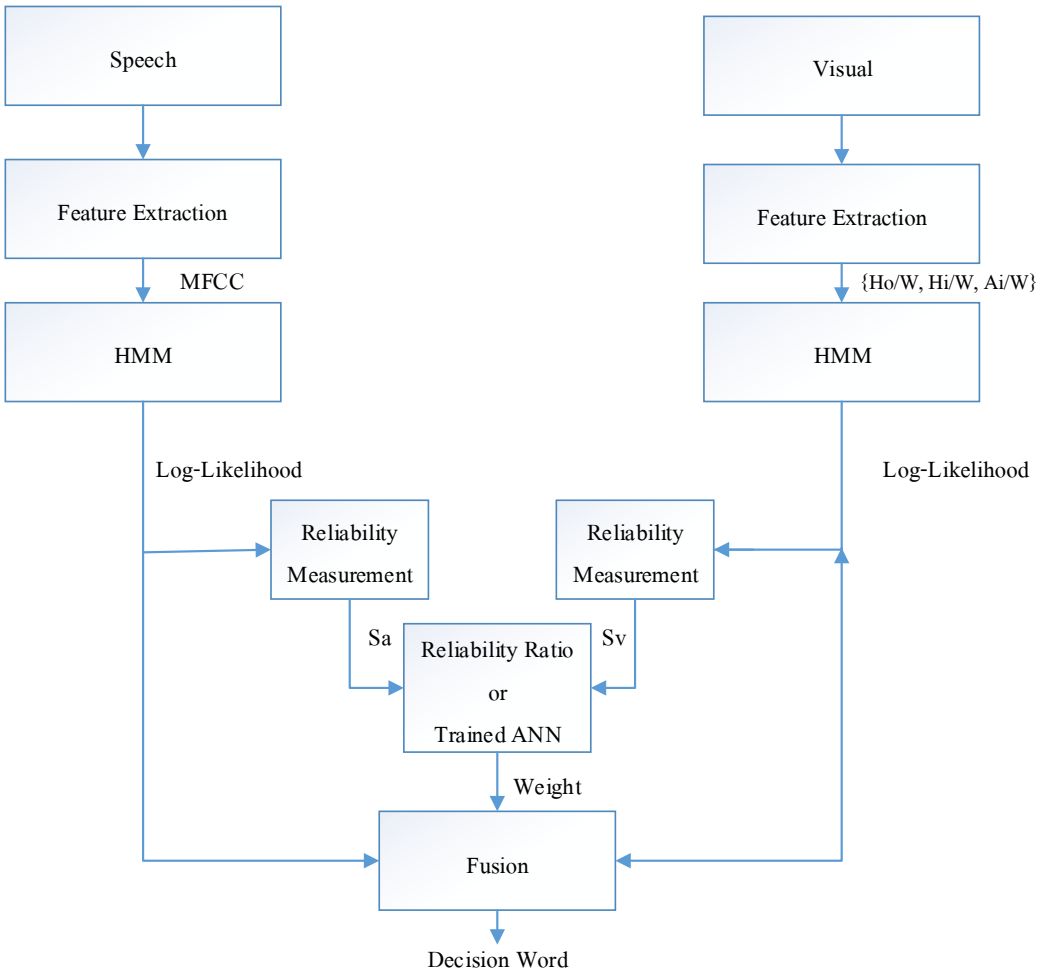


รูปที่ 3.10การผสมผสานข้อมูลเสียงและข้อมูลภาพเข้าด้วยกัน

ดังนั้นหากมีเสียงที่ยาว 1 วินาทีเข้ามาก็จะแทนด้วยเวกเตอร์จำนวน 100 อันส่วนข้อมูลรูปภาพ มีความเร็ว 30 เฟรมต่อวินาที ดังนั้นแต่ละเฟรมจะมีความยาว 33 มิลลิวินาทีต่อหนึ่งเฟรม โดยที่ข้อมูลเสียงที่ถูกสกัดด้วยMFCC-DA จะมี 39 คุณลักษณะจะนำมาผสมผสาน 3 คุณลักษณะจากการสกัดคุณลักษณะเด่นจากข้อมูลรูปภาพริมฝีปาก โดยระบบที่ใช้จะมีคุณลักษณะทั้งหมด 42 คุณลักษณะ เราจะฝึกฝนและทดสอบระบบนี้ด้วยแบบจำลองฮิดเดนมาร์คอฟแบบ 7 สถานะ

3.5.2 การผสมผสานในระดับการตัดสินใจ (Decision Fusion)

การผสมผสานในระดับการตัดสินใจ คือ การนำการตัดสินใจของทั้งสองระบบมาวิเคราะห์ร่วมกันเพื่อนำมาตัดสินดังรูปที่ 3.11 หลังจากการรู้จำโดยใช้ฮิดเดนมาคอฟโมเดลของเสียงและริมฝีปากจะได้ค่า Log-Likelihood มาเพื่อนำมาคำนวณหาค่าความน่าเชื่อถือ (Reliability)

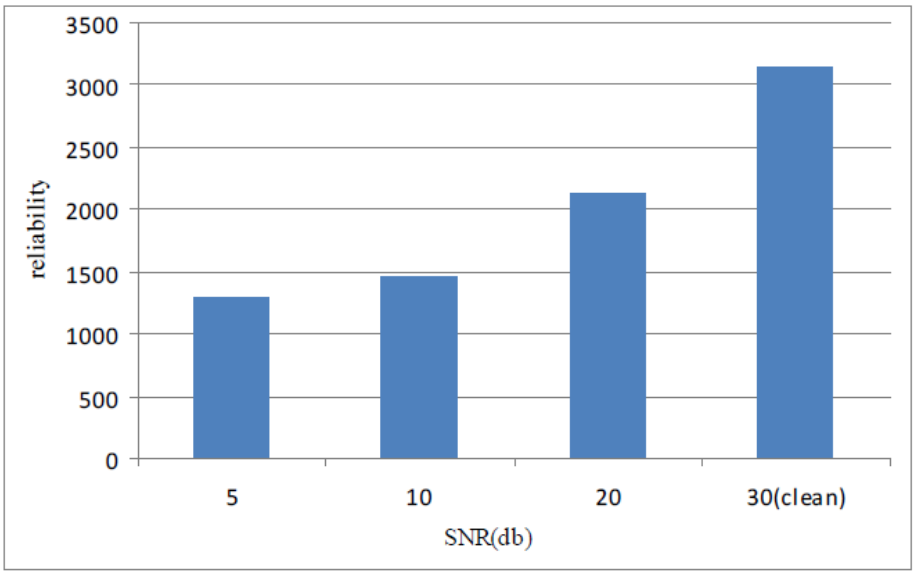


รูปที่ 3.11 ระบบรวมของการผสมผสานในระดับการตัดสินใจ

การหาค่าความน่าเชื่อถือที่มีประสิทธิภาพมากที่สุด ด้วยวิธีการค่าความแตกต่างของค่าสูงสุดของความน่าจะเป็น (MaxLog-Likelihood) จากสมการที่ 3.2 [18]

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(\max \log P(O \mid \lambda_a) - \log P(O \mid \lambda_v)) \tag{3.2}$$

$\log P(O \mid \lambda_a)$ และ $\log P(O \mid \lambda_v)$ คือ ค่าความน่าจะเป็น(Log-Likelihood) ของข้อมูลเสียง และข้อมูลรูปภาพ ตามลำดับโดยจากสมการนี้ได้ทดสอบทำให้เห็นว่าในสภาพแวดล้อมที่มีเสียงรบกวนน้อยค่าของ Reliability จะมีค่าสูงกว่าสภาพแวดล้อมที่มีเสียงรบกวนมาก ดังรูปที่ 3.12 ซึ่งเป็นการนำข้อมูลเสียงคำสั่ง “ห้า” ภาษาไทยมาหาค่าความน่าเชื่อถือในสภาพแวดล้อมที่สร้างเสียงรบกวนเข้าไป คือ Clean, 20, 10 และ 5 เดซิเบล ตามสมการที่ 3.2 พบว่าค่าความน่าเชื่อถือของเสียงในสภาพแวดล้อมที่ไม่มีเสียงรบกวน จะมีค่ามากว่าค่าความน่าเชื่อถือของเสียงในสภาพแวดล้อมที่มีเสียงรบกวน 20, 10 และ 5 เดซิเบล



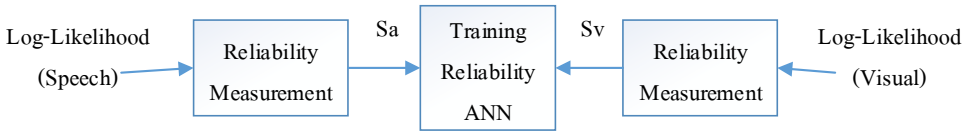
รูปที่ 3.12 หาค่าความน่าเชื่อถือด้วยวิธีการค่าความแตกต่างของค่าสูงสุดของความน่าจะเป็น (Max Log-Likelihood) ของคำสั่ง “ห้า”

หลังจากนั้นนำค่าความน่าเชื่อถือที่ได้มาทดลองหาน้ำหนัก (γ) ในสองกรณี: กรณีที่ 1 ทดสอบหาค่าน้ำหนักโดยหาค่าอัตราส่วนของความน่าเชื่อถือ (Reliability Ratio Base Method) จากสมการ 3.3[19]

$$\gamma = \frac{Sa}{Sa + Sv} \tag{3.3}$$

S_a และ S_v คือ ค่าที่ได้จากการหาค่าความน่าเชื่อถือของข้อมูลเสียงและรูปภาพตามลำดับ เมื่อเสียงที่ทดสอบมีเสียงรบกวนมากจะทำให้ได้ค่าความน่าเชื่อถือน้อย จะส่งผลกระทบให้ค่าน้ำหนักจากสมการมีค่าน้อย ในทางกลับกันหากเสียงที่ทดสอบมีเสียงรบกวนน้อยจะทำให้ได้ค่าความน่าเชื่อถือมากจะส่งผลกระทบให้ค่าน้ำหนักจากสมการมีค่ามาก

กรณีที่ 2 ทดสอบหาค่าน้ำหนักที่เหมาะสมด้วยโครงข่ายประสาทเทียม (Neural Network) โดยจะนำค่าความน่าเชื่อถือมาฝึกสอนระบบโดยใช้โครงข่ายประสาทเทียมเพื่อที่จะเรียนรู้ในการหาน้ำหนักที่เหมาะสมที่นำมาใช้งานในระบบดังรูปที่ 3.14



รูปที่ 3.13 ฝึกสอนระบบโดยใช้โครงข่ายประสาทเทียม

โดยจะนำค่าความน่าเชื่อถือในสภาวะ Clean, 20, 10 และ 5 เดซิเบล ให้เป้าหมายของน้ำหนัก (Target Weight) เป็น 1, 0.5, 0.25 และ 0 ตามลำดับหลังจากได้ค่าน้ำหนักก็จะนำค่ามาเข้าสมการที่ 3.4 [20] เพื่อมารู้จำคำสั่ง

$$Recognize = \arg\max\{\gamma \log P(O|\lambda_a) - (1-\gamma)\log P(O|\lambda v)\} \tag{3.4}$$

จากสมการหากข้อมูลเสียงอยู่ในสภาพแวดล้อมที่มีเสียงรบกวนน้อยค่าน้ำหนักของเสียงจะสูงเพื่อที่จะให้ระบบเลือกใช้งานการรู้จำเสียงมากกว่า ในทางตรงกันข้ามหากอยู่ในสภาพแวดล้อมที่มีเสียงรบกวนมากค่าน้ำหนักของการรู้จำเสียงจะน้อยเพื่อที่จะให้ระบบเลือกใช้การรู้จำริมฝีปากมากกว่าการรู้จำเสียง

บทที่ 4

ผลการทดลอง

ในบทนี้ได้ทำการทดลองและวิเคราะห์การทำงานของระบบการเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปาก โดยจะมีการแบ่งการทดสอบออกเป็น 3 ส่วนหลัก ๆ ได้แก่

4.1 ทดสอบการรู้จำเสียง

4.2 การอ่านริมฝีปาก

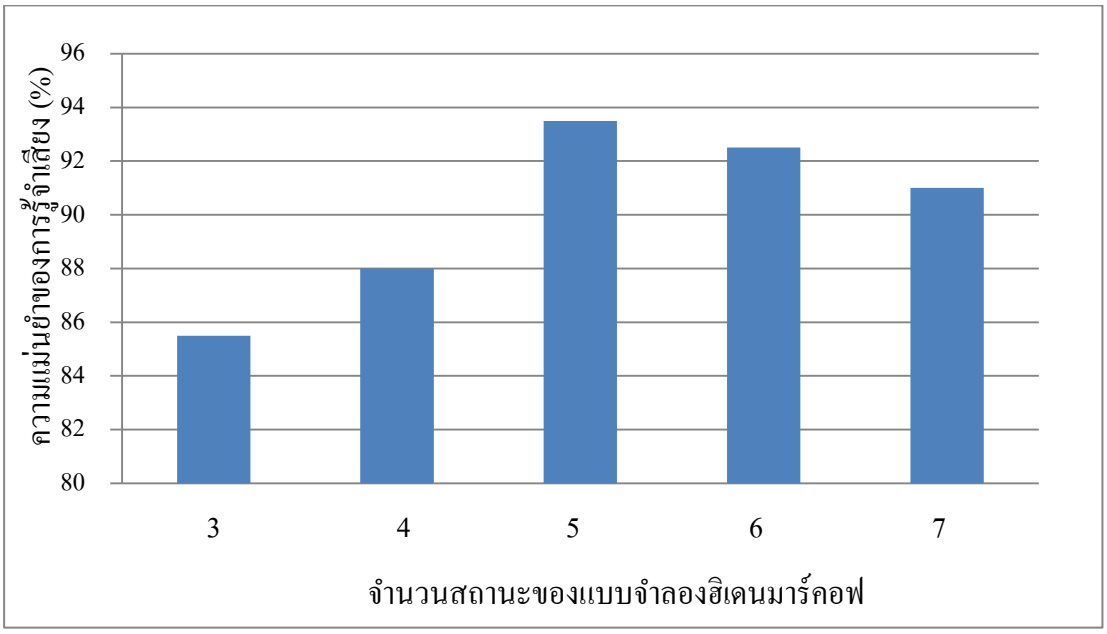
4.3 การผสมผสานระบบรู้จำเสียงและการอ่านริมฝีปากเข้าด้วยกัน

4.1 ทดสอบการรู้จำเสียง

ในการทดลองนี้จะทดสอบประสิทธิภาพเฉพาะการรู้จำเสียงอย่างเดียวเท่านั้นโดยจะเริ่มต้นด้วยการบันทึกเสียงภายในรถยนต์ในสภาพแวดล้อมที่ไม่มีเสียงรบกวน (บันทึกเสียงในโรงรถ) เพื่อนำมาฝึกสอนระบบ ส่วนการทดสอบจะเพิ่มเสียงรบกวนเข้าไปและบันทึกเสียงภายในรถยนต์ในสภาพแวดล้อมด้วยความเร็วต่าง ๆ

4.1.1 ผลการเปรียบเทียบจำนวนสถานะของฮิดเดนมาร์คอฟโมเดล

ในการทดลองนี้จะเป็นการคัดเลือกโครงสร้างฮิดเดนมาร์คอฟโมเดลที่มีประสิทธิภาพเพื่อคัดเลือกโครงสร้างที่เหมาะสมที่สุดสำหรับแบบจำลองภาษาไทย โดยใช้ค่าคุณลักษณะเด่น MFCC_DA การทดลองนี้เป็นการหาจำนวนสถานะของฮิดเดนมาร์คอฟโมเดลที่มีความแม่นยำมากที่สุดเพื่อที่จะนำมาใช้ในการรู้จำเสียง โดยได้เปรียบเทียบการใช้จำนวนสถานะ 3, 4, 5, 6 และ 7 สถานะโดยผลการเปรียบเทียบจำนวนสถานะเป็นดังรูปที่ 4.1

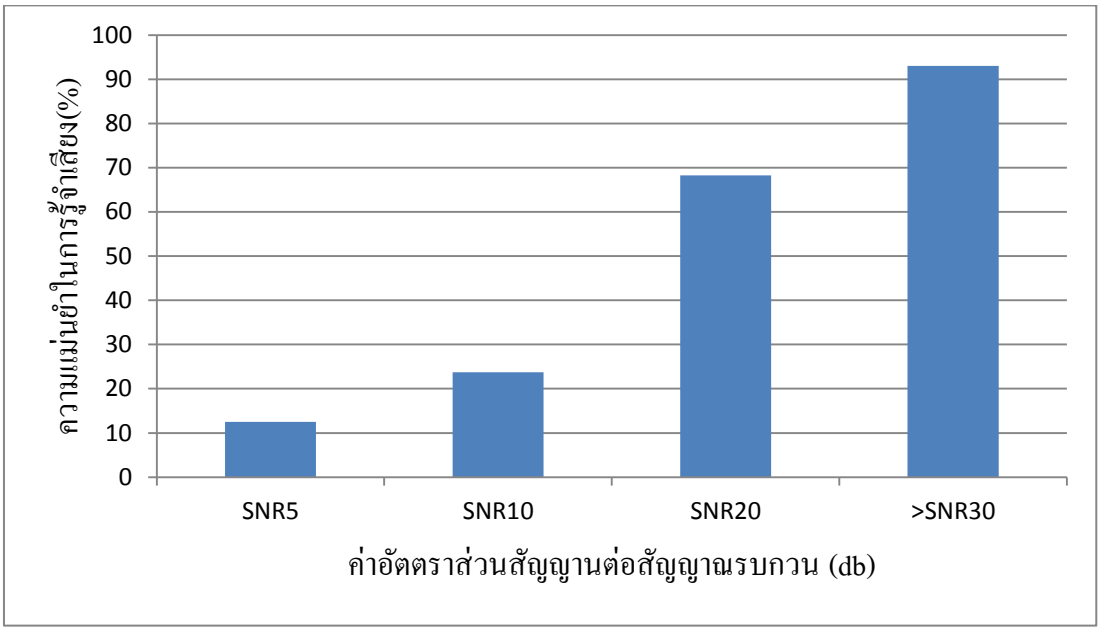


รูปที่ 4.1 ผลการทดลองนี้ได้เปรียบเทียบการใช้จำนวนสถานะต่าง ๆ กัน

จากผลการทดลอง พบว่าค่าจำนวนสถานะที่มีความแม่นยำมากที่สุด คือ 5 สถานะ ดังนั้นงานวิจัยนี้จึงกำหนดค่าสถานะเป็น 5 เป็นค่าที่เหมาะสมในการสร้างแบบจำลองข้อมูล

4.1.2 ผลการทดลองการรู้จำเสียงในสภาพแวดล้อมต่าง ๆ

การทดลองนี้เป็นหาความแม่นยำการรู้จำเสียงในสภาพแวดล้อมต่าง ๆ โดยสามารถแบ่งสภาพแวดล้อมของการทดสอบออกเป็น 3 สภาพแวดล้อม ได้แก่ สภาพแวดล้อมไม่มีเสียงรบกวน (บันทึกเสียงในโรงรถ) สภาพแวดล้อมที่ถูกเพิ่มเสียงรบกวนเข้าไปในระดับ 20, 10 และ 5 เดซิเบล ดังรูปที่ 4.2 และการทดลองสุดท้ายการรู้จำเสียงในสภาพแวดล้อมขณะขับรถจริงด้วยความเร็ว 0, 20, 40, 60 และ 80 กิโลเมตรต่อชั่วโมง

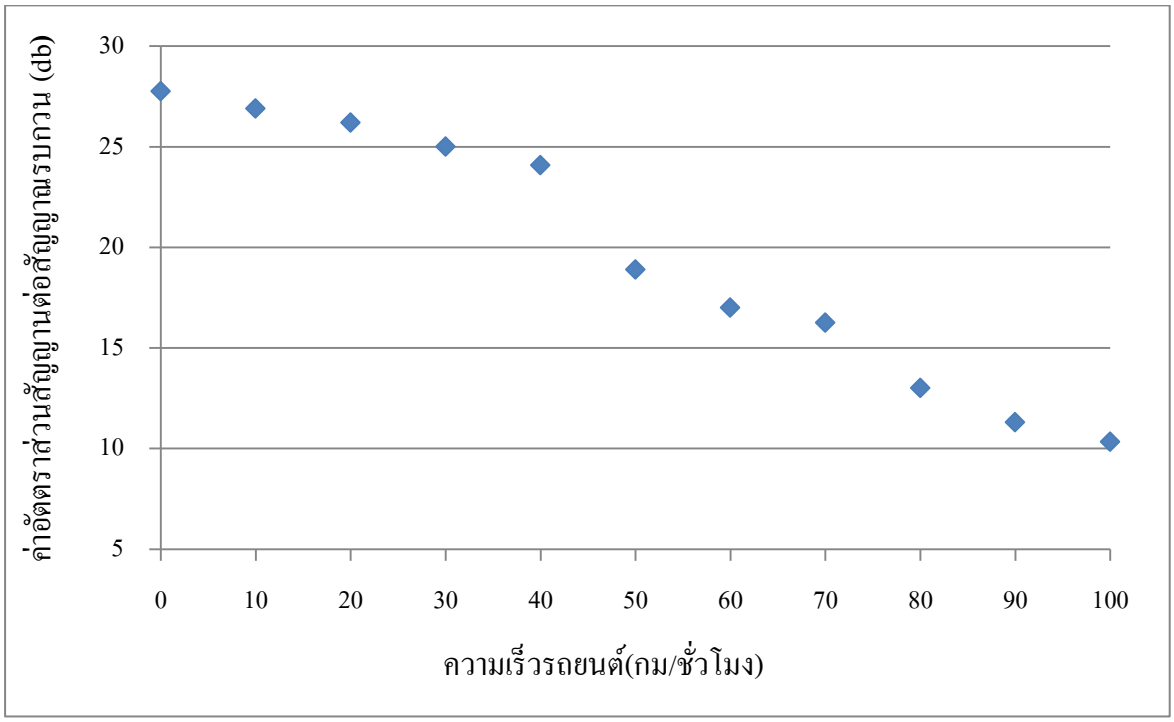


รูปที่ 4.2 ผลการทดลองความแม่นยำในการรู้จำเสียงในสภาพแวดล้อมที่มีเสียงรบกวน

จากผลการทดลอง พบว่าค่าเมื่อค่า SNR มีค่าน้อยหมายถึงสัญญาณเสียงนั้นมีเสียงรบกวนมากดังนั้นประสิทธิภาพการรู้จำเสียงจะถูกลดทอนลง ตัวอย่างเช่น สัญญาณเสียงที่ถูกสร้างเสียงรบกวนในระดับ SNR 5 เดซิเบล จะมีประสิทธิภาพความแม่นยำเพียง 13 % เท่านั้น ในทางกลับกันหากค่า SNR มีค่ามากหมายถึงสัญญาณเสียงนั้นมีสัญญาณรบกวนน้อย จะทำให้ประสิทธิภาพการรู้จำมีความแม่นยำมากขึ้น

4.1.3 ผลการทดลองหาอัตราส่วนของสัญญาณต่อสัญญาณรบกวนในสภาพแวดล้อมต่าง ๆ

ในส่วนนี้เราจะแสดงให้เห็นการทดลองเพื่อหาความทนทานของระบบรู้จำคำสั่งเสียงในสภาพแวดล้อมภายในรถยนต์มีเสียงรบกวนค่อนข้างมาก เช่น เสียงแอร์ เสียงเครื่องยนต์ ความเร็วของยานพาหนะ ดังรูปที่ 4.3 ที่มีการทดสอบหาค่า SNR ในสภาพแวดล้อมภายในรถยนต์ขณะการขับขี่ด้วยความเร็วค่าต่าง ๆ กัน



รูปที่ 4.3 ค่า SNR ในสภาพแวดล้อมภายในรถยนต์ขณะการขับขี่ด้วยความเร็วค่าต่าง ๆ กัน

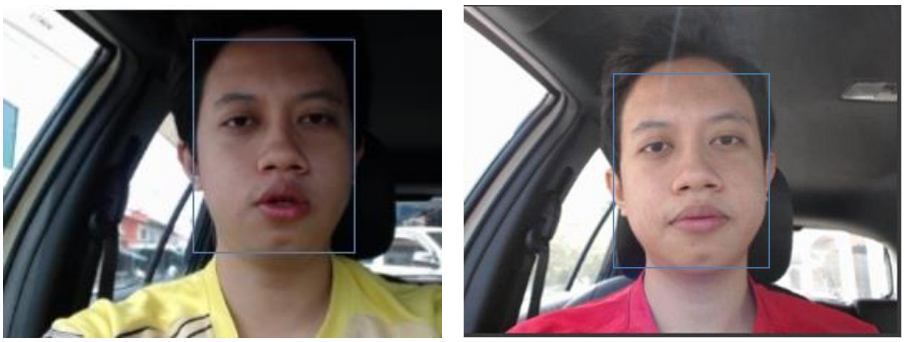
ในสภาพแวดล้อมภายในรถยนต์ขณะการขับขี่ด้วยความเร็ว 0, 20, 40, 60, 80 และ 100 กิโลเมตรต่อชั่วโมง จากกราฟแสดงให้เห็นระดับเสียงรบกวนของแต่ละความเร็ว โดยที่หากความเร็วรถต่ำค่า SNR จะมีค่ามากทำให้การรู้จำเสียงมีประสิทธิภาพสูง แต่ในตรงกันข้ามหากความเร็วรถสูงค่า SNR จะมีค่าน้อยทำให้การรู้จำเสียงมีประสิทธิภาพลดลงตามความเร็ว

4.2 ทดสอบการอ่านริมฝีปาก

ในการทดลองนี้จะทดสอบประสิทธิภาพเฉพาะการอ่านริมฝีปากอย่างเดียวเท่านั้นโดยจะเริ่มต้นด้วยการบันทึกวิดีโอภายในรถยนต์ในสภาพแวดล้อมควบคุมแสง เพื่อนำมาฝึกสอนระบบบันทึกเสียงภายในรถยนต์ในสภาพแวดล้อมด้วยความเร็วต่าง ๆ

4.2.1 ผลการทดสอบการค้นหาใบหน้า

การทดลองนี้ได้ทดสอบความแม่นยำการทดสอบการค้นใบหน้าโดยจะค้นหาใบหน้ารูปตรงเท่านั้นในสภาวะที่รถยนต์ไม่มีการเคลื่อนที่ สภาพแสงที่ใช้ทดสอบเป็นสภาพแสงกลางวันภายในรถยนต์และไม่มีแสงภายนอกส่องกระทบหน้าคนขับขี่รถยนต์โดยตรงตามที่แสดงดังรูปที่ 4.4



รูปที่ 4.4 ตัวอย่างผลลัพธ์การค้นหาใบหน้าตรงภายในสภาพแวดล้อมภายในรถยนต์

4.2.2 ผลการทดสอบการเปรียบเทียบแบบจำลองรูปร่าง

ความแม่นยำของการตรวจหาตำแหน่งปากนั้นมีผลต่อความแม่นยำในการรู้จำและความเร็วก็เป็นปัจจัยที่ต้องคำนึงถึงโดยผลการทดสอบระบบเป็นไปตามตารางที่ 4.1

ตาราง 4.1 ความแม่นยำและความเร็วในการติดตามริมฝีปาก

Lip Model	AAM	ASM	CLM
ความแม่นยำ	90%	75%	84%
ความเร็วในการประมวลผล	ช้า (1-2 fps)	เร็ว (35fps)	เร็ว (35fps)

แสดงให้เห็นว่า AAM มีความแม่นยำที่สุดแต่การประมวลผลมีความเชื่องช้าจึงไม่เหมาะสำหรับนำมาใช้งานในวิดีโอ ดังนั้นงานวิจัยนี้จึงเลือก CLM เพราะมีความแม่นยำและความเร็วในการประมวลผลสูงกว่าโมเดลอื่น ๆ

รูปที่ 4.5 ตัวอย่างผลลัพธ์การติดตามริมฝีปากโดยใช้ CLM

การอ่านริมฝีปากโดยใช้แบบจำลองฮิดเดนมาร์คอฟเป็นการนำข้อมูลริมฝีปากที่ถูกสกัดคุณลักษณะเด่นมาสร้างการฝึกสอนระบบจำนวน 20 แบบจำลอง เมื่อมีอินพุตเข้ามาในระบบจะหาแบบจำลองที่มีความน่าจะเป็นสูงสุดเพื่อเป็นคำตอบในการรู้จำของข้อมูลนั้น จากการทดสอบการอ่านริมฝีปากในสภาพแวดล้อมต่าง ๆ ภายในรถยนต์ที่ไม่มีการเคลื่อนที่จะมีการทดสอบคำสั่งละ 20 คำสั่ง โดยจะทดสอบคำสั่งละ 20 ครั้ง ดังตาราง 4.2

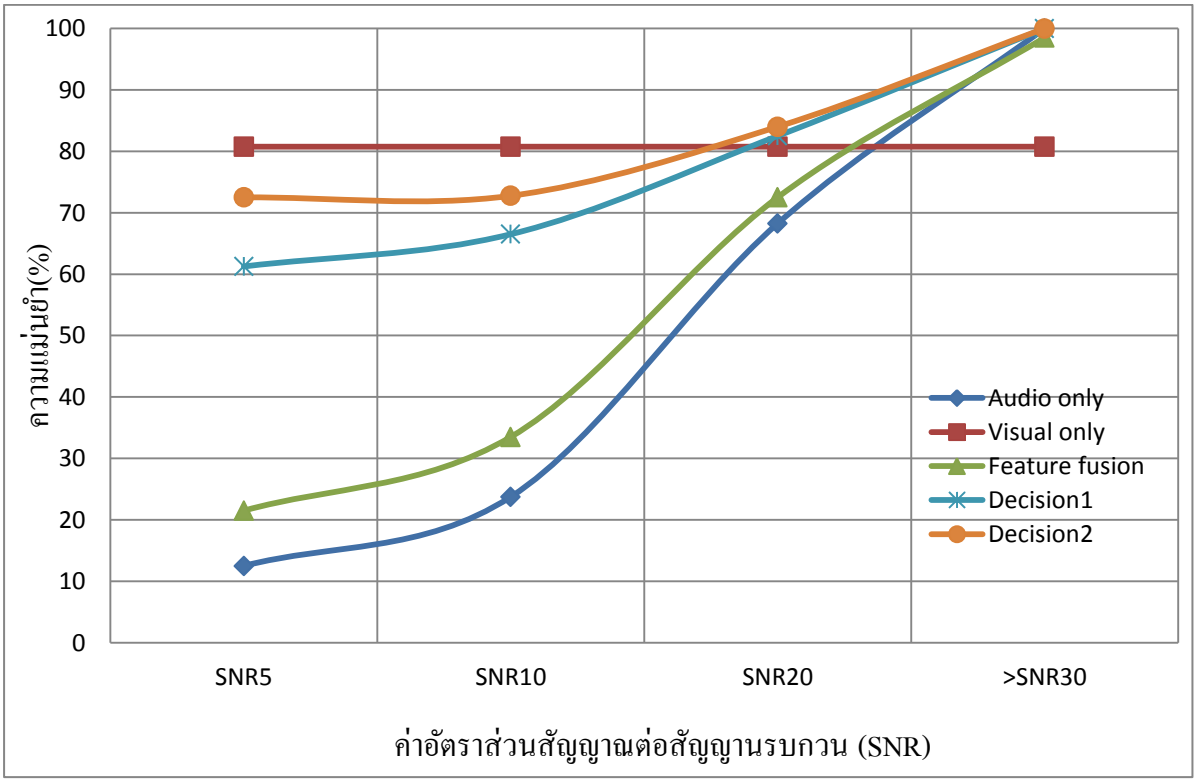
ตาราง 4.2 ความแม่นยำในการอ่านริมฝีปากในแต่ละคำสั่ง

คำสั่ง	ถูก(ครั้ง)	ผิด(ครั้ง)
ศูนย์	13	7
หนึ่ง	17	3
สอง	20	0
สาม	20	0
สี่	17	3
ห้า	19	1
หก	18	0
เจ็ด	12	8
แปด	13	7
เก้า	20	0
เปิด	12	8
ปิด	14	6
หยุด	15	5
เล่น	13	7
เอ-เอ็ม	16	4
เอฟ-เอ็ม	16	4
ก่อนหน้า	20	0
ถัดไป	14	6
ซีดี	14	6
วิทยุ	20	0
WCR	80.75	19.25

จากผลการทดลองพบว่าผลการอ่านริมฝีปากที่มีความแม่นยำอยู่ที่ 80.75 % ซึ่งระดับความแม่นยำในแต่ละคำสั่ง บางคำสั่ง ได้แก่ สอง สาม หก เก้า และ วิทยุ มีความถูกต้อง 100 % แต่คำสั่งเอ-เอ็มที่มีความแม่นยำน้อยที่สุด คือ 55 % เท่านั้น

4.3 ทดสอบการผสมผสานระหว่างระบบรู้จำเสียงกับการอ่านริมฝีปากเข้าด้วยกัน

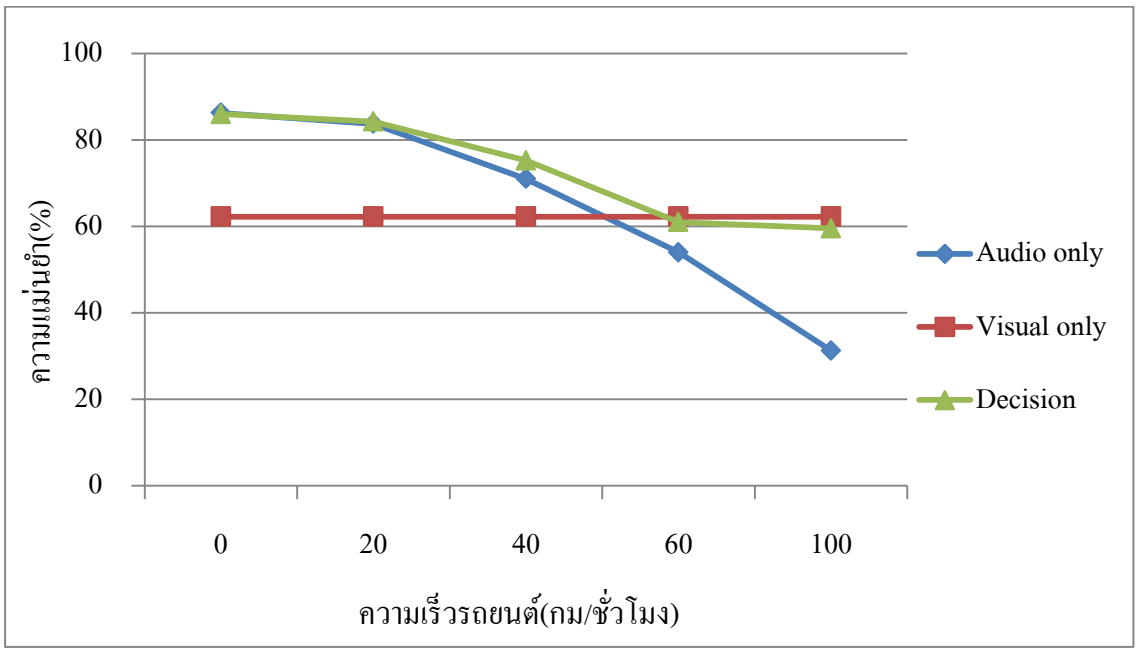
ในส่วนนี้เราได้แสดงให้เห็นการทดสอบในสภาพแวดล้อม 2 ส่วน ส่วนแรก จะบันทึกเสียงในสภาพแวดล้อมที่มีเสียงรบกวนน้อยสภาพแวดล้อมภายในรถยนต์ในโรงรถ และมีการสร้างเสียงรบกวนเข้าไปในระดับ 5, 10, 20 และ 30 (Clean) เดซิเบล เข้าไป ดังรูปที่ 4.6 โดยจะทดสอบระบบ 3 รูปแบบ ได้แก่ ก) การรู้จำเสียงอย่างเดียว ข) การอ่านริมฝีปากอย่างเดียว และ ค) การผสมผสานในระดับคุณลักษณะและการผสมผสานในระดับการตัดสินใจสองรูปแบบ



รูปที่ 4.6 ผลการทดลองระบบทั้งหมดในสภาพแวดล้อมที่มีเสียงรบกวนแบบ White Noise

ผลการทดลองปรากฏว่าความถูกต้องระบบรู้จำริมฝีปากอย่างเดียวมีความถูกต้อง 80.75 % ในทุกค่าของ SNR เนื่องจากว่าเสียงรบกวนของสัญญาณเสียงจะไม่มีผลกระทบกับสัญญาณรูปภาพ ส่วนความแม่นยำของการรู้จำเสียงอย่างเดียวนั้นในสภาพแวดล้อมที่มีเสียงรบกวนต่ำอัตราความถูกต้องของระบบนี้จะมีความถูกต้องสูงหรือมากกว่า 95 % แต่หากในสภาพแวดล้อมที่มีเสียงรบกวนสูงอัตราความถูกต้องของระบบก็จะถูกลดทอนลง ส่วนการผสมผสานในระดับคุณลักษณะนั้นอัตราความถูกต้องของระบบจะมีค่าสูงกว่าการรู้จำเสียงอย่างเดียว ส่วนการผสมผสานในระดับการตัดสินใจ อัตราความถูกต้องของระบบจะมีค่าสูงกว่าการรู้จำเสียงอย่างเดียวและการผสมผสานในระดับคุณลักษณะในทุก SNR และแบบการผสมผสานในระดับการตัดสินใจนั้น อัตราความถูกต้องของระบบจะมีค่าสูงกว่า การอ่านริมฝีปากอย่างเดียว ในช่วง SNR ประมาณ 17 เดซิเบล

ส่วนที่สองบันทึกเสียงในสภาพแวดล้อมขณะขับรถด้วยความเร็วในระดับต่าง ๆ เช่น 20, 40 ,60 และ 100 กิโลเมตรต่อชั่วโมงดังรูปที่ 4.7 โดยจะทดสอบระบบ 3 รูปแบบ ได้แก่ ก) การรู้จำเสียงอย่างเดียว ข) การอ่านริมฝีปากอย่างเดียว และ ค) การผสมผสานในระดับการตัดสินใจ



รูปที่ 4.7 ผลการทดลองในสภาพแวดล้อมที่มีเสียงรบกวนจากการเคลื่อนที่ของรถยนต์

ผลที่ได้พบว่าเมื่อรถยนต์มีความเร็วมากขึ้นประสิทธิภาพของการรู้จำเสียงอย่างเดียวจะถูกลดทอนลงเพราะเสียงรบกวนจากเครื่องยนต์ที่มากขึ้นตามความเร็ว ระบบรู้จำริมฝีปากอย่างเดียวมีความถูกต้อง 62.25 % ในสภาพแวดล้อมการขับรถยนต์จริง ดังนั้นหากรถยนต์มีการเคลื่อนที่การใช้การผสมผสานในระดับการตัดสินใจจะมีประสิทธิภาพความถูกต้องของระบบที่มากที่สุด

บทที่ 5
บทสรุปและข้อเสนอแนะ

ในบทนี้ได้กล่าวถึงการสรุปผลวิจัยที่ได้ดำเนินการสำหรับวิทยานิพนธ์นี้ รวมทั้งข้อเสนอแนะต่าง ๆ ที่จะเป็นประโยชน์ต่อการทำวิจัยด้านการออกแบบระบบร่วมระหว่างเสียงและรูปภาพหัวข้อการนำเสนอมีดังนี้

5.1 บทสรุป

5.2 สรุปผลการทดลองกระบวนการรู้จำเสียง

5.3 สรุปผลการทดลองกระบวนการอ่านริมฝีปาก

5.4 สรุปผลการทดลองกระบวนการผสมผสานสัญญาณเสียงและข้อมูลรูปภาพ

5.5 บทวิจารณ์และข้อเสนอแนะ

5.1 บทสรุป

งานวิจัยนี้ได้พัฒนากระบวนการเพิ่มความแม่นยำในระบบสั่งการด้วยเสียงพูดภาษาไทยด้วยการอ่านริมฝีปากที่สามารถนำมาประยุกต์ใช้ภายในรถยนต์ โดยผ่านกล้องเว็บแคม งานวิจัยนี้แบ่งออกเป็นสองส่วนคือ การรู้จำเสียง และการอ่านริมฝีปาก

โดยระบบการรู้จำเสียงจะจับสัญญาณเสียงของผู้ใช้งานเพื่อนำสัญญาณเสียงเหล่านั้นมาวิเคราะห์ใช้ในชุดคำสั่งที่สามารถนำไปประยุกต์ใช้ภายในรถยนต์ได้อย่างเหมาะสม กระบวนการรู้จำเสียงพูดสามารถแบ่งออกเป็น 2 ขั้นตอนสำคัญ ได้แก่ การฝึกฝนข้อมูลเสียง และการรู้จำเสียง ขั้นตอนการฝึกฝนระบบรู้จำเสียงเป็นการฝึกสอนระบบให้จดจำคำสั่งเสียงตามลักษณะที่กำหนดได้ โดยการป้อนข้อมูลเสียง และข้อมูลกำกับคำสั่งเสียง โดยใช้ HMM ในการสร้างแบบจำลองที่ใช้ในการรู้จำเสียงในแต่ละคำสั่ง

ส่วนระบบการอ่านริมฝีปากเริ่มต้นด้วยการค้นหาใบหน้าและติดตามการเคลื่อนไหวของริมฝีปากเพื่อสกัดข้อมูลคุณลักษณะเด่นออกมา ซึ่งผลการตรวจจับการเคลื่อนไหวของริมฝีปากสามารถแสดงข้อมูลความสูง พื้นที่ และความกว้างของริมฝีปากได้ ส่วนขั้นตอนการรู้จำการอ่านริมฝีปากใช้ HMM ในการสร้างแบบจำลองเพื่อใช้ควบคุมภายในรถยนต์

นอกจากนี้แล้วงานวิจัยนี้ได้ทดลองการผสมผสานระหว่างสัญญาณเสียง และสัญญาณรูปภาพเข้าด้วยกันเพื่อเพิ่มประสิทธิภาพของระบบรู้จำเสียง

5.2 สรุปผลการทดลองกระบวนการรู้จำสียง

การหาค่าความแม่นยำของการรู้จำคำสั่งเสียงอย่างเดียวได้ทดสอบในสภาพแวดล้อม 2 แบบ ส่วนแรก ทดสอบเสียงในสภาพแวดล้อมที่มีเสียงรบกวนน้อย (สภาพแวดล้อมภายในรถยนต์ในโรงรถ) และสร้างเสียงรบกวนเข้าไป ในระดับ 5, 10, 20 และ ไม่มีสียงรบกวนหรือมากกว่า 30 เดซิเบล ความแม่นยำของการรู้จำเสียงอย่างเดียวในสภาพแวดล้อมที่มีเสียงรบกวนต่ำอัตราความถูกต้องของระบบนี้จะมีความแม่นยำสูงสุดร้อยละ 93 แต่หากในสภาพแวดล้อมที่มีเสียงรบกวนสูงอัตราความแม่นยำของการู้จำก็จะถูกลดทอนลงไป

ส่วนที่สองได้ทดสอบการรู้จำคำสั่งเสียงในสภาพแวดล้อมภายในรถยนต์ที่มีเสียงรบกวนในแต่ละระดับความเร็วของรถยนต์ ได้แก่ 0, 20, 40, 60, 80, และ 100 กิโลเมตรต่อชั่วโมง โดยผลสรุปที่ได้รับ คือ หากความเร็วรถยนต์ต่ำค่าอัตราส่วนของสัญญาณต่อสัญญาณรบกวนจะมีมากทำให้การรู้จำเสียงมีความแม่นยำสูง แต่ในตรงกันข้ามหากความเร็วรถยนต์สูงค่าอัตราส่วนของสัญญาณต่อสัญญาณรบกวนจะมีค่าน้อยส่งผลให้ความแม่นยำของการรู้จำเสียงลดลงตามความเร็ว

5.3 สรุปผลการทดลองกระบวนการอ่านริมฝีปาก

การหาค่าความแม่นยำของการอ่านริมฝีปาก คำสั่งที่ใช้ในการทดสอบเป็นคำสั่งง่าย ๆ ที่ใช้ในเครื่องเล่นเพลงภายในรถยนต์ เช่น หยุด เล่น เปิด ปิด เป็นต้น โดยใช้กลุ่มตัวอย่างคำสั่งละ 20 ชุด เพื่อทำการฝึกสอนและทดสอบระบบโดยประสิทธิภาพของระบบการอ่านริมฝีปากอย่างเดียวมีความแม่นยำเฉลี่ยอยู่ที่ร้อยละ 80.75 ในทุก ๆ อัตราส่วนของสัญญาณต่อสัญญาณรบกวน เนื่องจากสัญญาณรบกวนของเสียงจะไม่มีผลกระทบต่อรูปภาพ

5.4 สรุปผลการทดลองกระบวนการผสมผสานสัญญาณเสียงและข้อมูลรูปภาพ

การผสมผสานระหว่างข้อมูลเสียงและข้อมูลรูปภาพถูกทดสอบ 2 แบบ คือ การผสมผสานในระดับคุณลักษณะ และการผสมผสานในระดับการตัดสินใจ ด้านการผสมผสานแบบระดับคุณลักษณะอัตราความแม่นยำของระบบจะมีค่าสูงกว่าการรู้จำเสียงอย่างเดียว ส่วนการผสมผสานในระดับการตัดสินใจอัตราความแม่นยำของระบบจะมีค่าสูงกว่าการรู้จำเสียงอย่างเดียวและการผสมผสานในระดับคุณลักษณะ และการผสมผสานในระดับการตัดสินใจจะมีอัตราความแม่นยำของระบบจะมีค่าสูงกว่าการอ่านริมฝีปากอย่างเดียวในช่วง SNR ประมาณ 17 เดซิเบล

5.5 บทวิจารณ์และข้อเสนอแนะ

5.5.1 การควบคุมสภาพแวดล้อมในการอ่านริมฝีปากเป็นสิ่งที่สำคัญมากแสงต้องมีความสว่างมากพอและแสงไม่กระทบหน้าผู้ขับโดยตรง เพื่อไม่ให้ภาพมีจุดอับแสงหรือเกิดเงา เพราะอาจจะทำให้การวิเคราะห์ผิดพลาดได้

5.5.2 เสียงรบกวนที่เกิดขึ้นในสภาพแวดล้อมขณะขับรถจริงไม่ได้มีแต่ความเร็วรถยนต์ แต่ยังมีเสียงรบกวนจากแหล่งกำเนิดอื่น เช่น เสียงเครื่องปรับอากาศภายในรถยนต์ เสียงรถยนต์รอบข้าง เสียงพูดภายในรถยนต์

5.5.3 แบบจำลองรูปร่างที่ใช้ในการติดตามการเคลื่อนไหวริมฝีปากที่ใช้ในงานวิจัยนี้เป็นแบบจำลองที่ถูกบันทึกและฝึกสอนระบบในสภาพแวดล้อมปกติ หากแบบจำลองถูกฝึกสอนในสภาพแวดล้อมภายในรถยนต์ ความแม่นยำในการติดตามการเคลื่อนไหวของริมฝีปากจะมีประสิทธิภาพมากขึ้น

5.5.4 ขอบเขตงานวิจัยนี้ระบบการอ่านริมฝีปากในขณะขับรถในเส้นทางตรงเท่านั้น แต่สภาพแวดล้อมในการขับจริง พวงมาลัยรถยนต์จะมีการเคลื่อนไหวค่อนข้างมากและเส้นทางที่ใช้มีลักษณะไม่เรียบ ดังนั้นทำให้ระบบระบบการอ่านริมฝีปากมีประสิทธิภาพที่ลดลง

5.5.5 ขอบเขตงานวิจัยนี้เป็นการทดสอบการใช้งานการรู้จำแบบขึ้นกับบุคคลดังนั้นหากเพิ่มจำนวนผู้ทดสอบระบบ จะสามารถเพิ่มประสิทธิภาพความแม่นยำของระบบในการใช้งานแบบไม่ขึ้นกับบุคคลได้

บรรณานุกรม

[1] J. Shin, J. Lee, and D. Kim, "Real-time lip reading system for isolated Korean word recognition," Pattern Recognition, Vol. 44, Issue. 3, Mar. 2011, pp.559-571.

[2] R. Navarathna, P. Lucey, D. Dean, C. Fookes, S. Sridharan, “Lip detection for audio-visual speech recognition in-car environment,” 10th International Conference on Information Science, Signal Processing and their Applications , 2010.

[3] B. Lee, M.H Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T. Huang, “AVICAR: Audio-Visual Speech Corpus in a Car Environment,” Annual Conference of the International Speech Communication Association - INTERSPEECH , 2004.

[4] G. Potamianos and C.Neti, “Audio-Visual Speech Recognition in Challenging Environments,” Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH, Geneva, Switzerland, Sept. 2003), pp. 1293-1296.

[5] H. Glotin, D. Vergyr, C. Neti, G. Potamianos, J. Luettin, “Weighting schemes for audio-visual fusion in speech recognition,” Proceedings. (ICASSP '01). 2001.

[6] J. Lee and C.H. Park, “Robust Audio-Visual Speech Recognition Based on Late Integration,” IEEE transactions on multimedia, vol. 10, no.5, august 2008.

[7] T. Saitoh, K. Morishita and R. Konishi, “Analysis of Efficient Lip Reading Method for Various Languages,” Pattern Recognition, 2008 . ICPR 2008. 19th International Conference on Tampa, FL, 8-11 Dec. 2008, pp 1-4.

[8] K. Kumar, “Delta-spectral cepstral coefficients for robust speech recognition,” 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 22-27 May 2011, pp. 4784 – 4787.

[9] D. O'Shaughnessy, “Speech communication: human and machine,” Addison-Wesley, 1987, pp. 150.

[10] H. Rowley, S. Baluja, and T. Kanade, “Neural networkbased face detection”, IEEE Patt.Anal. Mach. Intell., Vol. 20, pp. 22–38, 1998.

[11] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition”, Proceedings of the IEEE 77, Feb. 1989, pp. 257 – 285

[12] J. Han and M. Kamber., “Data Mining Concepts and Techniques,” Morgan Kaufmann Publishers, 2001.

[13] P. Viola and M. Jones, “Robust real-time face detection,” International Journal of Computer Vision, vol. 57, no. 2, 2004, pp. 137-154.

[14] T.F. Cootes, and C.J. Taylor, “Active Shape Models : their Training and Application,” Computer Vision and Image Understanding, vol. 61, no. 1, Jan. 1995, pp. 38-59.

[15] M. Turk and A. Pentland, “Face Recognition Using Eigenfaces,” IEEE International Conference on Computer Vision and Pattern Recognition, (1991), 586 – 591

[16] T.F. Cootes, G.J. Edwards, and C.J. Taylor, “Active Appearance Models,” Proc Fifth European Conf. Computer Vision, vol. 2, 1998, pp. 484- 498.

[17] D. Cristinacce and T. Cootes, “Automatic feature localization with constrained local models,” Pattern Recognition, vol 41, 2008, pp.3054-3067.

[18] T. W. Lewis and D. M. W. Powers, “Sensor fusion weighting measuresin audio-visual speech recognition,” in Proc. 27th Conf. AustralasianComputer Science, Dunedin, New Zealand, 2004, pp. 305– 314.

[19] A. Adjoudani and C. Benoît, , D. G. Stork and M. E. Hennecke, Eds.,“On the integration of auditory and visual parameters in an HMM based ASR,” in Speechreading by Humans and Machines: Models,Systems and Applications, ser. NATO ASI Series. Berlin, Germany:Springer, 1996, pp. 461–472.

[20] H. Liu, T. Fan and P. Wu “Audio-visual keyword Spotting Based on Adaptive Decision Fusion under noisy condition for human-robot interaction,” IEEE International Conference on robotics& Automation, 2014.

[21] E.Petajan , B.Bischoff, D,Bodoff, N. M.Brooke, “An improved automatic lipreading system to enhance speech recognition,” In proceeding of CHI ’88 Proceedings of the SIGCHI conference on Human factors in computing systems, 1988.

[22] ปกิต ศิลปะราวงศ์ , “การรู้จำริมฝีปากโดยใช้เทคนิควิเคราะห์สัญญาณแปรตามเวลาและนิวรอลเน็ตเวิร์ก,” จุฬาลงกรณ์มหาวิทยาลัย, 2543

[23] กิตติชัย วรรธนะจิตติกุล, “การอ่านริมฝีปากแบบอัตโนมัติโดยใช้แบบจำลองฮิดเดนมาร์คอฟ,” มหาวิทยาลัยชียงใหม่, 2551

ภาคผนวก ก. ตารางที่ ก-1 ประสิทธิภาพในการรู้จำคำสั่งเสียงอย่างเดียวในสภาพแวดล้อมที่ไม่มีเสียงรบกวน

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	100
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	100
ปิด	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	100
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	100
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	100
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	100
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	100
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	100

ตารางที่ ก-2 ประสิทธิภาพในการรู้จำคำสั่งเสียงอย่างเดียวในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 20 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	0	0	6	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4	7	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
2	0	0	16	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	80
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	15
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	2	7	0	0	0	0	0	0	0	0	0	11	0	0	10
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	1	0	0	95
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	100
ปิด	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	14	30
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	5	0	0	0	7	40
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	100
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	100
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	95
																				AVG	68.25

ตารางที่ ก-3 ประสิทธิภาพในการรู้จำคำสั่งเสียงอย่างเดียวในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 10 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	19	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	19	0	0	0
2	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	25
3	0	0	0	16	0	0	0	3	0	0	0	0	0	0	0	0	0	3	0	0	80
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
5	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	11	0	0	45
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0
7	0	0	0	0	0	0	1	18	0	0	0	0	0	0	0	0	0	0	0	2	90
8	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	17	0	0	35
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
เปิด	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	17	0	0	15
ปิด	0	0	0	0	0	0	0	0	0	0	0	4	0	1	0	0	0	0	0	15	20
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	14	0	0	30
เล่น	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
เอ-เอ็ม	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	8	0	2	0	0	0
เอฟ-เอ็ม	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	17	0	0	15
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
วิทยุ	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	4	20
																				AVG	23.75

ตารางที่ ก-4 ประสิทธิภาพในการรู้จำคำสั่งเสียงอย่างเดียวในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 5 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	4	2	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	1	1	0
2	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	1	0	0	0
4	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	2	0	0	0
5	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	5	0	0	0
7	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	2	90
8	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	18	0	0	10
9	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	12	0	0	0
เปิด	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	18	0	0	0
ปิด	0	0	0	0	0	0	0	1	0	0	0	13	0	0	0	0	0	0	0	6	65
หยุด	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	4	0	12	0
เล่น	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
เอ-เอ็ม	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
เอฟ-เอ็ม	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
ถัดไป	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	17	0	0	85
ซีดี	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
วิทยุ	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
																					12.5

ตารางที่ ก-5 ประสิทธิภาพในการรู้จำคำสั่งด้วยการอ่านริมฝีปากอย่างเดียวในสภาพแวดล้อมภายในรถยนต์

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	13	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	1	65
1	0	17	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	85
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	17	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	85
5	0	0	0	0	0	19	0	0	0	1	0	0	0	0	0	0	0	0	0	0	95
6	2	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	90
7	0	0	0	0	0	0	0	12	0	0	0	1	0	7	0	0	0	0	0	0	60
8	0	1	0	0	0	5	0	0	13	0	0	0	0	0	0	0	2	0	0	0	65
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	1	0	0	0	0	0	0	0	0	12	2	0	5	0	1	0	0	0	0	60
ปิด	0	0	0	0	0	2	0	0	0	0	3	14	0	0	0	0	0	1	0	0	70
หยุด	0	0	2	0	0	0	2	0	0	1	0	0	15	0	0	0	0	0	0	1	75
เล่น	0	0	0	1	0	1	0	0	0	0	0	0	0	13	0	0	0	5	0	0	65
เอ-เอ็ม	0	0	0	2	1	0	0	1	0	0	0	0	0	0	16	0	0	0	0	0	80
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	4	0	80
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	1	0	2	0	0	1	1	0	1	0	0	0	14	0	0	70
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	1	14	0	70
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	80.75

ตารางที่ ก-6 ประสิทธิภาพในการผสมผสานในระดับคุณลักษณะในสภาพแวดล้อมไม่มีเสียงรบกวน

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	1	0	0	95
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	100
ปิด	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	2	0	0	0	2	80
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	100
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	100
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	100
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	100
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	98.75

ตารางที่ ก-7 ประสิทธิภาพในการผสมผสานในระดับคุณลักษณะในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 20 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	18	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90
1	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95
2	0	0	16	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	80
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	20
5	0	0	0	2	0	17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	85
6	15	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	2	0	0	15
7	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	1	0	95
8	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	4	0	0	80
9	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	2	0	0	90
เปิด	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	100
ปิด	0	0	0	0	0	0	0	0	0	0	0	11	2	0	0	1	0	3	1	2	55
หยุด	1	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	1	0	0	90
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	17	0	15
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	15	0	3	0	0	10
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	7	0	65
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	4	0	0	80
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	85
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	72.5

ตารางที่ ก-8 ประสิทธิภาพในการผสมผสานในระดับคุณลักษณะในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 10 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	15	3	0	5
1	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	11	0	0	0	0	0	2	0	0	0	0	0	0	0	0	6	0	0	55
3	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	4	0
5	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	8	0	0	60
6	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	17	0	0	15
7	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	4	0	0	0	0	80
8	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	20	0	0	40
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
เปิด	0	0	0	0	0	0	0	0	0	0	6	2	0	0	0	0	0	12	0	0	30
ปิด	0	0	0	0	0	0	0	0	0	0	0	4	10	3	0	0	0	3	0	0	20
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	12	0	0	40
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	1	0	0	0
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	16	0	0	0	0	80
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	17	0	0	15
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	7	0	35
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	12	0	4	20
																				AVG	33.5

ตารางที่ ก-9 ประสิทธิภาพในการผสมผสานในระดับคุณลักษณะในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 5 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	5
1	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	15
3	0	0	0	9	0	3	0	0	0	0	0	0	0	0	0	0	0	8	0	0	45
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	4	0	0
5	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35
6	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	15	0	0	15
7	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	4	0	0	0	0	80
8	0	0	0	0	0	4	0	0	2	0	0	0	0	0	0	14	0	0	0	0	10
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
เปิด	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	18	0	0	10
ปิด	0	0	0	0	0	0	0	0	0	0	0	4	2	0	0	1	0	10	1	2	20
หยุด	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	10	6	2	0	0	50
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	4	0	20
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	1	4	20
																				AVG	15

ตารางที่ ก-10 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 1 ในสภาพแวดล้อมไม่มีเสียงรบกวน

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	100
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	100
ปิด	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	100
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	100
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	100
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	100
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	100
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	100

ตารางที่ ก-11 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 1 ในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 20 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	0	0	16	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	55
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	3	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	85
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	100
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	2	0	0	12	0	0	0	0	0	1	5	0	0	60
ปิด	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	2	90
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	5	0	0	0	0	0	3	2	0	0	10	0	0	15
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	95
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	17	0	0	0	0	85
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	13	0	65
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	82.5

ตารางที่ ก-12 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 1 ในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 10 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	9	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	45
1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	17	0
2	0	0	19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	95
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95
5	0	0	0	1	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95
6	1	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	4	0	0	0	0	3	0	0	0	0	0	0	0	13	0	0	0	15
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	9	0	0	6	0	0	0	0	0	2	3	0	0	30
ปิด	0	0	0	0	0	0	0	1	0	0	0	14	0	0	0	0	0	1	0	3	70
หยุด	0	0	1	1	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	10	40
เล่น	0	0	0	4	0	0	0	4	0	0	1	0	0	2	0	0	0	9	0	0	10
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	1	95
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	16	0	0	0	0	80
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	5	0	8	0	0	0
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	63.5

ตารางที่ ก-13 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 1 ในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 5 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	3	0	0	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	11	15
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	19	0
2	0	0	14	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	70
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	1	0	0	0	0	0	0	0	0	0	18	1	0	0
5	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	2	0	0	90
6	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	2	0	0	90
7	0	0	0	0	0	0	0	19	0	0	0	0	0	1	0	0	0	0	0	0	95
8	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	3	0	0	2	100
9	0	0	0	3	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	4	60
เปิด	0	0	0	0	0	0	0	7	0	0	5	0	0	0	0	0	1	3	0	0	25
ปิด	0	0	0	0	0	0	0	3	0	0	0	7	0	0	0	0	0	0	0	10	35
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	4	70
เล่น	0	0	0	0	0	0	0	6	0	0	0	0	1	3	0	0	0	10	0	0	15
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	1	95
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	15	0	0	0	0	75
ก่อนหน้า	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	17	0	0	0	85
ถัดไป	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	10	0	0	50
ซีดี	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	1	11	0	55
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	61.25

ตารางที่ ก-14 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 2 ในสภาพแวดล้อมไม่มีเสียงรบกวน

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	100
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	100
ปิด	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	100
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	100
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	100
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	100
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	100
ซีดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	100
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	100

ตารางที่ ก-15 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 2 ในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 20 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	0	0	13	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	75
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	95
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	3	0	0	9	0	0	0	0	0	2	5	0	0	45
ปิด	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	2	90
หยุด	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	100
เล่น	0	0	0	0	0	0	0	4	0	0	0	0	0	6	0	0	0	10	0	0	30
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	1	95
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	17	0	0	0	0	85
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	19	0	0	95
ซีดี	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	13	0	65
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	83.75

ตารางที่ ก-16 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 2 ในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 10 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	10	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	50
1	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	16	0
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
5	0	0	0	1	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95
6	11	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
7	0	0	0	0	0	0	0	18	0	0	1	0	0	1	0	0	0	0	0	0	90
8	0	0	0	3	0	0	0	0	4	0	0	0	0	0	0	0	13	0	0	0	100
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	100
เปิด	0	0	0	0	0	0	0	7	0	0	6	2	0	0	0	0	2	3	0	0	30
ปิด	0	0	0	0	0	0	0	0	0	0	1	16	0	0	0	0	0	2	0	1	80
หยุด	0	0	3	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	2	75
เล่น	0	0	0	1	0	0	0	0	0	0	1	0	0	8	0	0	0	10	0	0	40
เอ-เอ็ม	0	0	1	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	95
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	16	0	0	0	0	80
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	100
ถัดไป	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	15	0	0	75
ซีดี	0	0	0	0	1	0	0	5	0	0	0	0	0	0	0	0	0	5	9	0	45
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	72.75

ตารางที่ ก-17 ประสิทธิภาพในการผสมผสานในระดับตัดสินใจแบบที่ 2 ในสภาพแวดล้อมที่มีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน 5 เดซิเบล

คำสั่ง	0	1	2	3	4	5	6	7	8	9	เปิด	ปิด	หยุด	เล่น	เอ-เอ็ม	เอฟ-เอ็ม	ก่อนหน้า	ถัดไป	ซีดี	วิทยุ	WCR(%)
0	9	0	1	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	5	45
1	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	17	0
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	10	4	0	30
5	0	0	0	1	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	18	0	0	0	0	0	2	0	0	0	0	0	0	90
8	0	0	0	2	0	3	0	0	4	0	0	0	0	0	0	0	9	1	0	1	100
9	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	1	95
เปิด	0	0	0	1	0	0	0	6	0	0	4	5	0	0	0	0	0	3	0	0	20
ปิด	0	0	0	3	0	0	0	0	0	0	0	10	0	0	0	0	0	2	0	5	50
หยุด	0	0	1	0	0	0	1	0	0	0	0	0	16	0	0	0	0	0	0	2	80
เล่น	0	0	0	0	0	0	0	1	0	0	0	0	0	8	0	0	0	11	0	0	40
เอ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	1	95
เอฟ-เอ็ม	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	16	0	0	0	0	80
ก่อนหน้า	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	1	95
ถัดไป	0	0	0	0	0	1	0	7	0	0	1	0	0	0	0	0	0	11	0	0	55
ซีดี	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	16	0	80
วิทยุ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	100
																				AVG	72.5

ภาคผนวก ข

ผลงานตีพิมพ์เผยแพร่จากวิทยานิพนธ์

Isamail Masamae and Panyayot Chaikan, “Integrating Lip-Reading and Thai Speech to Control Electronic Devices in a Vehicle,” *IEEE 5th International Conference on System Engineering and Technology*, Aug. 10 - 11, UiTM, Shah Alam, Malaysia, pp.31-34**,** 2015

PROCEEDINGS

2015 IEEE International Conference on System Engineering and Technology (ICSET 2015)

Organizer
IEEE Control System Society Malaysia Chapter

Secretariat at
Faculty of Electrical Engineering
Universiti Teknologi MARA

Secretariat
Mohd Nasir Taib
Rozita Jailani
Ruhizan Liza Ahmad Shauri
Mohd Hezri Fazalul Rahiman
Nooritawati Md Tahir
Abd Manan Samad
Ramli Adnan
Hashimah Hashim
Megat Syahirul Amin Megat Ali
Mohd Farid Saaid
Ahmad Ihsan Mohd Yassin

ITB
Suwarno
Arief Syaichu-Rohman
Aciek Ida Wuryandari
Ary Setijadi Prihatmanto

Contact Information
ICSET 2015 Secretariat
Faculty of Electrical Engineering
Universiti Teknologi MARA
40450 Shah Alam
Malaysia

Email: *rozita@ieee.org*
adnanramli@yahoo.com

Tel: +603-55436083/5012
Fax: 603-55435077

UNIVERSITI TEKNOLOGI MARA

IEEE

CSS
IEEE CSS Chapter Malaysia

Integrating Lip-Reading and Thai Speech to Control Electronic Devices in a Vehicle

Isamail Masamae
Department of Computer Engineering,
Faculty of Engineering, Prince of Songkla University
Songkhla, Thailand.
5510120050@email.psu.ac.th

Panyayot Chaikan
Department of Computer Engineering,
Faculty of Engineering, Prince of Songkla University
Songkhla, Thailand.
panyayot@coe.psu.ac.th

***Abstract*—This paper presents the use of lip-reading and Thai speech to control electronic devices in a vehicle. The Viola-Jones algorithm detects the face of the driver and the constrained local model detects their mouth area before three lips features are extracted. Hidden Markov models are utilized to recognize speech and lip movement, with the lip movement recognizer offering better accuracy than the speech recognizer in a noisy environment. Three fusion methods are utilized to combine lip-movement and speech. We propose the use of vehicle speed for selecting the appropriate recognizer for different speech signal-to-noise ratios.**

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Controlling an in-vehicle electronic device by hand while driving is dangerous because it can distract the driver. For this reason, Automatic Speech Recognition (ASR) is one of the most popular methods for reducing a driver's distraction. However, the accuracy of ASR's recognition rate is considerably reduced in highly acoustic background noise, especially when the car is moving at high speed. To overcome this problem, some researchers have used lip movement data to improve speech recognition accuracy. Palecek and Chaloupka showed that the combination of lip and audio can improve speech recognition accuracy [1]. Navarathna et.al. demonstrated that the Viola-Jones algorithm is an appropriate method for detecting the driver's face and lips, and is also immune to the variability of light in an in-vehicle environment [2]. Shin et.al. used the active appearance model (AAM) and the Lucas-Kanade method for detecting the lip area from a human face [13]. Lee et.al. proposed adding an audio-visual English speech corpus in a car environment [3]. Liu et.al. proposed the mechanism of decision level fusion between the audio and visual features, and introduced a weighting scheme between each modality by means of reliability measures [4]. All of this research focuses on non-Thai commands.

Although Thai speech recognition research has a long history, the combination of Thai speech and visual data is much less reported. In this paper, we study the feasibility of using Thai visual-speech recognition for controlling in-vehicle electronics devices. Section II introduces our proposed audio-visual speech recognition system. Section III presents the fusion method for audio and visual data. Section IV describes our audio-visual database for training and testing the system. Experimental results are given in section V, and section VI concludes the paper. Future work is presented in section VII.

II. OUR PROPOSED AUDIO-VISUAL SPEECH RECOGNITION SYSTEM

Table 1 shows our choice of Thai speech commands designed for controlling the radio receiver and CD player in a car environment. A five-state hidden Markov model (HMM) is utilized to recognize input speech. Our system operates on isolated words in speaker dependent mode.

Table 1. Thai words supported by our visual- speech system.

Command	Meaning	Command	Meaning
หนึ่ง	One	เปิด	Open
สอง	Two	ปิด	Close
สาม	Three	เล่น	Play
สี่	Four	หยุด	Pause
ห้า	Five	ถัดไป	Next
หก	Six	ก่อนหน้า	Previous
เจ็ด	Seven	ซีดี	CD
แปด	Eight	เอฟเอ็ม	FM
เก้า	Nine	เอเอ็ม	AM
ศูนย์	Ten	วิทยุ	Radio

For lip recognition, a video camera is attached to the steering wheel to capture images of the driver's face, which is detected using the Viola-Jones algorithm [5], as shown in Figure 1. The location of the lips is obtained and the lips' features are extracted. A seven-state hidden Markov model is utilized to recognize lip movement, and three lip features are utilized:

$$\left\{\frac{H_i}{W}, \frac{H_m}{W}, \frac{N_s}{W}\right\},$$

where W = width of the mouth,

H_m = height of the mouth,

H_i = distance between the lower edge of the upper lip and the upper edge of the lower lip,

N_s = Number of pixels in the shaded area between the upper and the lower lip (see Figure 2).

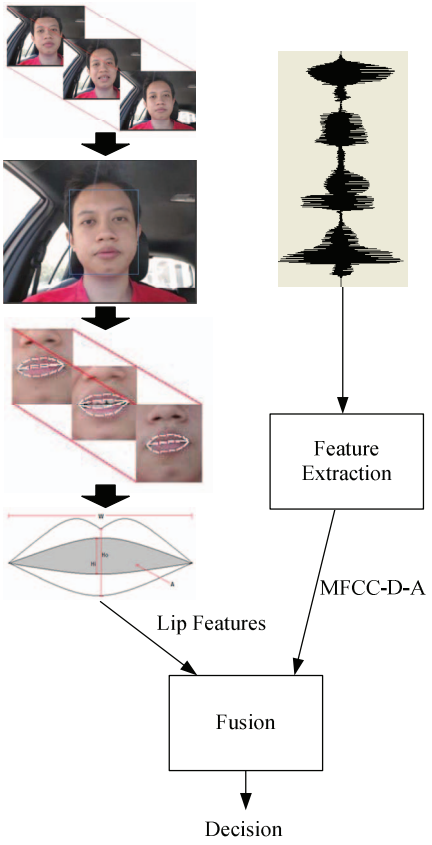


Figure 1. System Block diagram.

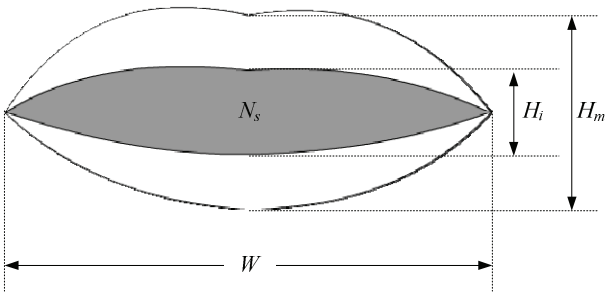


Figure 2. Lip features used in our system.

III. Audio-Visual Fusion Methods

Audio and lip features are combined in two ways to augment recognition accuracy: at the feature level and the decision level.

For feature level fusion, the Mel-frequency cepstral coefficients-delta-acceleration (MFCC-D-A) [14] data, consisting of 39 features extracted from the speech module, are combined with the lips' three features. These 42 features are used to train and test a seven-state HMM classifier. The input speech and lip movement video are recognized as a k^* class using:

$$k^* = \arg\max_i \{P(O_{A,V} \mid \lambda^i)\}, \quad (1)$$

where $O_{A,V}$ is the combined audio-visual observation feature, and $P(O_{A,V} \mid \lambda^i)$ stands for the likelihood obtained from the HMM of the i-th class.

For decision level fusion, the input speech and the lip-movement features are used to separately train an HMM classifier. The input speech and lip movement are classified together as c^* class using [4]:

$$c^* = \arg\max_i \{\gamma \log P(O \mid \lambda_A^i) + (1-\gamma) \log P(O \mid \lambda_V^i)\}, \quad (2)$$

where $P(O_A \mid \lambda_A^i)$ and $P(O_V \mid \lambda_V^i)$ stand for the likelihood obtained for the i-th class of the audio HMMs and video HMMs respectively. The weight γ specifies how much the audio effects the final decision, and is calculated using [8]

$$\gamma = \frac{S_A}{S_A + S_V}, \quad (3)$$

where S_A and S_V stand for the reliability of the outputs from the audio HMMs and video HMMs respectively. The reliability of each modality is calculated using [7]:

$$S = \frac{1}{N-1} \sum_{i=1}^{N} \{\max_j \log P(O \mid \lambda^j) - \log P(O \mid \lambda^i)\}, \quad (4)$$

where N is the total number of classes. Equation 4 assumes that the difference of the output of each HMM will be large in the case of low acoustic noise. For example, Figure 3 shows that the reliability of the command word "ห้า" is large when the voice is clean, and becomes much less so in more noisy voice data.

The overall recognition accuracy of decision level fusion was improved by replacing equation (3) for calculating the γ value by a neural network. The γ value is generated using S_A and S_V as the net's inputs. A feed-forward neural network was trained with the reliabilities obtained from the training data with different levels of noise. Speech data with signal-to-noise ratios (SNR) of 30, 20, 10, and 5 dB was trained with target γ values of 1.0, 0.5, 0.25 and 0 respectively.

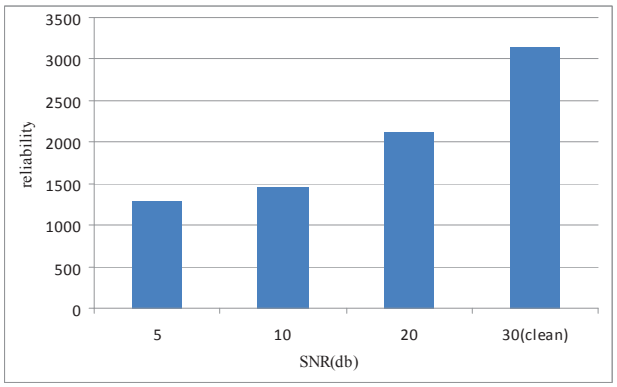


Figure 3. The reliability of the command word "ห้า"

at different SNR levels.

IV. Audio-Visual Database

Our database consists of two parts, for training and testing. Each part contains 20 command words, and each word comes with 20 speech examples and corresponding lip-movement videos. All audio-video data was collected in a controlled environment to ensure that the speech was clean (≥ 30 dB SNR). We collected the data in a car with its engine turned off, and all side and rear windows closed. A 640x480 pixels video camera, recording at 30 frames per second, was utilized to capture lip movement. No internal light source was used in the car, so the video was captured using only natural sunlight, and it was recorded during 11.00am-1.00pm to avoid sunlight being projected directly onto the driver's face.

Clean speech data in the database was used to train the audio HMM classifiers. To train the neural network for generating the γ value, white noise was added at 3 different levels; speech with SNR levels of 30 (clean), 20, 10, and 5 dB was obtained.

In order to test the system's tolerance for acoustic noise, white noise was added to the testing part of database at 3 levels in the same way as the testing phase; testing speech with SNR levels of 30, 20, 10, and 5 dB was derived.

V. Experimental Results

We chose the best algorithm for detecting the location of a car driver's lips from three methods implemented on a 2.5 GHz Core-i5 machine: active appearance model (AAM) [10], active shape model (ASM) [11], and constrained local model (CLM) [12]. The algorithms were tested on 200 video files, each of which contained only 10 video frames. The lip areas obtained from each algorithm were compared with a manually determined area, and the overlapping pixels between these two areas were counted to derive a correction percentages for each algorithm. The resulting values for AAM, ASM, and CLM were 90%, 75%, and 84% respectively. Although AAM has the best correction percentage, we did not choose it because the maximum video frame rate that it could support on our test machine was only 3 frames per second. However, since ASM and CLM support over 35 frames per second, we chose CLM as our lip detection algorithm.

We tested our Thai visual-speech command system in five configurations: (1) an *Audio-Only* recognizer using speech alone to recognize a command word, (2) a *Lip-Only* recognizer utilizing visual lip movements alone, (3) a *Feature-Fusion* recognizer which employed feature level fusion between the audio and video, (4) an *EQ3* recognizer using decision fusion along with equation (3) to calculate the γ weighting value, and (5) a *NN* recognizer which employed a neural network to obtain the γ weighting value.

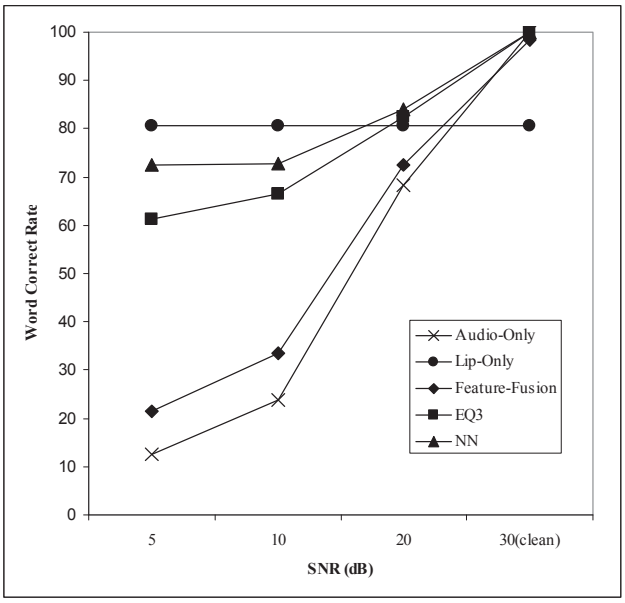


Figure 4. Recognition accuracy of 5 recognizers.

Figure 4 shows that the *Lip-Only* recognizer has a recognition accuracy of 80.75%. Since acoustic noise has no effect on the lip videos, the recognition of the lip features has a constant accuracy for all acoustic noise configurations. An *Audio-Only* gives very high accuracy in a clean configuration, but noisy speech, with SNR levels of 20, 10, and 5dB, causes the recognition accuracy to be considerably reduced, to 68.25, 23.75, and 12.50 percent respectively. The *Feature-Fusion, EQ3, and NN* recognizers give better accuracy than *Audio-Only* for all SNRs, but their accuracies are lower than *Lip-Only* for SNR levels less than or equal to 20 dB. The recognition accuracy of the *NN* recognizer is 84.00, 72.75, and 72.50 percent for SNR levels of 20, 10, and 5 dB respectively. This means that its performance is better than all the other fusion based recognizers, but is still less than the *Lip-Only* recognizer in a noisy environment.

The *NN* recognizer has better accuracy than the *Lip-Only* recognizer when the SNR is greater than 17 dB. However, the *Lip-Only* recognizer has superior accuracy when the SNR is less than 17 dB. Therefore, if we know the SNR level of the current input speech, then we can choose the appropriate recognizer to achieve the best overall recognition accuracy. Unfortunately, it is difficult to know the exact SNR value because the noise is highly dependent on the car's speed. However, we can approximately define a relation between the

car's speed and the speech's SNR. We collected speech data while a Toyota Yaris (1500 CC) was travelling at different speeds. Figure 5 shows that when the car's speed is greater than 60 km/h, the SNR is less than 17 dB. With this additional data, we added a car speed threshold, S_{th} to our system, set to 60 km/h. When the car speed is lower than S_{th}, the final decision comes from the *NN* recognizer, otherwise the final decision comes from the *Lip-Only* recognizer. The current speed of the car is obtained from an embedded board connected to the engine control unit (ECU) via its OBD-II port [9].

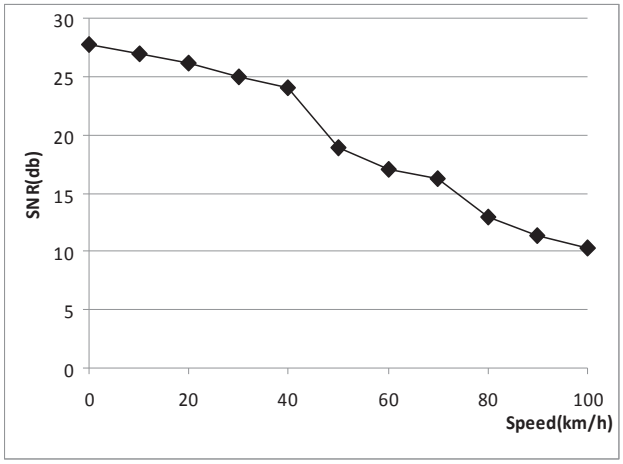


Figure 5. The signal-to-noise ratio of input speech obtained from a car travelling at different speeds.

VI. Discussion and Conclusion

We have demonstrated that the integration of audio and lip movement can improve car-based voice recognition accuracy for controlling electronic devices such as CD players and radios. Car speed information can help when choosing the appropriate recognizer at different levels of noises, but utilizing a speed threshold may not be the best method, and further study of this topic is required. The relationship between speed and acoustic SNR will vary between cars. Moreover, engine speed is not the only factor that affects car noise. A more effective noise detection mechanism is needed for recognizer selection.

VII. Future Work

We plan to test our system in real-world driving situations to better analyze the effect of acoustic noise on recognition accuracy. We will also examine the robustness of face detection and our lip detection algorithm in variable light.

Acknowledgment

The authors are grateful to Dr. Andrew Davison for his kind help in polishing the language of this paper.

References

[1] K. Palecek and J. Chaloupka "Audio-visual speech recognition in noisy audio environments," 36th International Conference on Telecommunications and Signal Processing (TSP), 2-4 July 2013.

[2] R.Navarathna, P. Lucey, D. Dean, C.Fookes, S.Sridharan, "Lip detection for audio-visual speech recognition in-car environment," 10th International Conference on Information Science, Signal Processing and their Applications ,2010

[3] B.Lee, M.H Johnson, C. Goudeseune, S.Kamdar, S.Borys, M.Liu, T.Huang, "AVICAR: Audio-Visual Speech Corpus in a Car Environment," Annual Conference of the International Speech Communication Association - INTERSPEECH , 2004

[4] H. Liu, T. Fan and P. Wu "Audio-visual keyword Spotting Based on Adaptive Decision Fusion under noisy condition for human-robot interaction," IEEE International Conference on robotics& Automation ,2014.

[5] P. Viola and M. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004.

[6] D.Cristinacce and T.Cootes, "Automatic feature localization with constrained local models," Pattern Recognition 41 ,(3054 – 3067),2008

[7] T. W. Lewis and D. M. W. Powers, "Sensor fusion weighting measuresin audio-visual speech recognition," in Proc. 27th Conf. AustralasianComputer Science, Dunedin, New Zealand, 2004, pp. 305–314.

[8] A. Adjoudani and C. Benoît, , D. G. Stork and M. E. Hennecke, Eds.,"On the integration of auditory and visual parameters in an HMM based ASR," in Speechreading by Humans and Machines: Models,Systems and Applications, ser. NATO ASI Series. Berlin, Germany:Springer, 1996, pp. 461–472.

[9] M. JinKim, J.WookJang, and Y.S, Yu., "A Study on In-Vehicle Diagnosis System using OBD-II with Navigation," IJCSNS International Journal of Computer Science and Network Security, vol.10 no.9, September 2010.

[10] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," Proc Fifth European Conf. Computer Vision, vol. 2, pp. 484-498, 1998.

[11] T.F. Cootes, and C.J. Taylor, "Active Shape Models : their Training and Application," Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38-59, Jan. 1995.

[12] D.Cristinacce and T.Cootes, "Automatic feature localization with constrained local models," Pattern Recognition, vol 41, pp.3054-3067,2008.

[13] J. Shin, J. Lee, and D. Kim, "Real-time lip reading system for isolated Korean word recognition," Pattern Recognition, Vol. 44, Issue. 3, pp.559-571, Mar. 2011.

[14] K. Nahar, H. Al-Muhtaseb, W. Al-Khatib, M. Elshafei, and M. Alghamdi, "Arabic Phonemes Transcription using Data Driven Approach," The International Arab Journal of Information Technology, Vol. 12, No.3, May. 2015.

ประวัติผู้เขียน

ชื่อ สกุล นายอิสมาแอล มะสาแม

รหัสประจำตัวนักศึกษา 5510120050

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วิศวกรรมศาสตรบัณฑิต (วิศวกรรมคอมพิวเตอร์)	มหาวิทยาลัยสงขลานครินทร์	2555

ทุนการศึกษา (ที่ได้รับในระหว่างการศึกษา)

ได้รับทุนบัณฑิตศึกษาวิศวกรรมศาสตร์ วิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์

การตีพิมพ์เผยแพร่ผลงาน

Isamail Masamae and Panyayot Chaikan, “Integrating Lip-Reading and Thai Speech to Control Electronic Devices in a Vehicle,” *2015 IEEE 5th International Conference on System Engineering and Technology*, Aug. 10 - 11, UiTM, Shah Alam, Malaysia, pp.31-34