# บทที่ 1

# บทนำ

## 1.1 ความเป็นมาและความสำคัญของปัญหา

มนุษย์มีวิธีในการติดต่อสื่อสารกันผ่านทางสื่อหลากหลายรูปแบบ ได้แก่ ข้อความ เสียง ภาพ หรือภาพเคลื่อนไหว ซึ่งมนุษย์ทำการสื่อสารกันผ่านทางการพูดมากที่สุด แต่ในบางครั้ง เสียงที่พูดออกมาเพื่อสื่อสารอาจเกิดความผิดพลาดได้ ทั้งที่เกิดอาจจากเสียงรบกวนในสภาพแวดล้อม และจากความผิดพลาดในการรับฟังของผู้ฟังเอง เพราะฉะนั้น ในบางครั้งมนุษย์อาจใช้การพิจารณารูปปากของผู้พูดเพื่อช่วยในการแยกแยะคำพูดเพื่อให้เกิดความถูกต้องในการรับข้อมูลเพิ่มขึ้น ซึ่งผลงานการวิจัยของ W. Sumby, และ I. Pollack[1] ได้กล่าวไว้ถึงการทดลองแยกแยะคำศัพท์ 64 คำ ในสภาวะที่ไม่มีเสียงรบกวน อัตราการแยกแยะคำพูดของผู้ฟังจะอยู่ที่ร้อยละ 80 ทั้งจากการแยกแยะด้วยเสียงเพียงอย่างเดียวและการแยกแยะโดยพิจารณาจากเสียงและรูปปากของผู้พูด แต่หากมีเสียงรบกวนที่ระดับ 30 dB อัตราการแยกแยะโดยพิจารณาจากเสียงและรูปปากของผู้พูดจะลดลงไปที่ร้อยละ 60 ในขณะที่อัตราการแยกแยะด้วยเสียงเพียงอย่างเดียวจะลดลงเหลือเพียงร้อยละ 20 เท่านั้น จึงสามารถกล่าวได้ว่า ความสามารถในการแยกแยะคำพูดของมนุษย์จะลดลงอย่างรวดเร็วหากมีเสียงรบกวนมากขึ้น แต่จะสามารถอาศัยการพิจารณารูปปากหรือการอ่านรูปปากเพื่อช่วยในการแยกแยะเสียงพูดหรือการรู้จำเสียงพูดได้ ซึ่งการเปลี่ยนแปลงของรูปปากในเสียงภาษาไทยตามหลักสรีรสัทศาสตร์(Articulatory Phonetics) จะเกิดจากความแตกต่างของเสียงสระและเสียงพยัญชนะบางตัว โดยลักษณะเฉพาะของเสียงสระที่แตกต่างกัน ก่อให้เกิดการเคลื่อนไหวของรูปปากที่แตกต่างกันไปด้วย จึงสามารถจำแนกลักษณะของรูปปากตามเสียงสระได้จากกันอย่างชัดเจน ในงานวิจัยนี้เราจะพิจารณาการสร้างแบบจำลองสำหรับการเคลื่อนไหวของรูปปากเพื่อจำลองการเคลื่อนไหวของตำแหน่งพิกัดของรูปปากตามเสียงสระที่ได้ป้อนเข้าไปโดยวิธี Backpropagation Neural Network โดยขอบเขตของงานวิจัยนี้จะครอบคลุมเฉพาะเสียงสระเดี่ยวแท้ในภาษาไทยเพียง 16 เสียงเท่านั้น และไม่พิจารณาความแตกต่างของรูปปากที่เกิดจากเสียงพยัญชนะต้นหรือตัวสะกดที่แตกต่างกัน รวมทั้งการเปลี่ยนแปลงอันเนื่องมาจากการยืดและหดตัวของกล้ามเนื้อรอบริมฝีปาก

## 1.2 วัตถุประสงค์ของการวิจัย

1. สามารถสร้างแบบจำลองการเคลื่อนไหวของรูปปากตามเสียงสระเสียงเดี่ยวได้โดยวิธีการ Neural Network
2. สามารถพัฒนาตัวอย่างโปรแกรมที่สร้างจากแบบจำลองเพื่อการสร้างภาพเคลื่อนไหวในการใช้งานจริงได้

## 1.3 ขอบเขตของการวิจัย

1. แบบจำลองที่สร้างขึ้นเป็นการพิจารณาการเคลื่อนไหวในตำแหน่งจุดพิกัดของรูปปากด้านนอกตามทิศทาง 2 มิติเท่านั้น โดยไม่พิจารณาถึงการหดหรือขยายของกล้ามเนื้อ หรือการเคลื่อนไหวของอวัยวะส่วนอื่นนอกเหนือไปจากส่วนริมฝีปาก
2. พิจารณาเฉพาะสระเสียงเดี่ยว 16 เสียงเท่านั้น และไม่พิจารณาถึงการเปลี่ยนแปลงที่เกิดจากพยัญชนะต้นหรือตัวสะกดที่แตกต่างกัน

## 1.4 ข้อจำกัดของการวิจัย

1. เนื่องจากวิธีในการเก็บข้อมูลจุดพิกัดรูปปากของผู้พูดที่เป็นตัวอย่างนั้นใช้การกำหนดจุดที่รอบริมฝีปากของผู้พูดก่อนแล้วจึงทำการถ่ายภาพวิดีโอในขณะทำการพูดคำตัวอย่าง ซึ่งอาจมีการสูญหายของข้อมูลไปได้เนื่องจากผลกระทบจากปริมาณแสงและการสะท้อนที่มีการเปลี่ยนแปลงตลอดเวลาในขณะที่ทำการพูดซึ่งต่างจากงานวิจัยที่ใช้การเก็บข้อมูลด้วยอุปกรณ์ประเภท Motion Capture ที่สามารถเก็บข้อมูลได้อย่างแม่นยำกว่า
2. การเก็บข้อมูลกระทำในแนว 2 มิติตามแกน XY ในขณะที่การเคลื่อนไหวของรูปปากมีการเคลื่อนไหวตามแนวแกน Z ซึ่งเป็นการเคลื่อนไหวแบบ 3 มิติด้วย ทำให้เกิดความผิดพลาดในทิศทางการเคลื่อนที่ของจุดพิกัดรอบริมฝีปาก

## 1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัยนี้คือ สามารถสร้างแบบจำลองสำหรับการเปลี่ยนแปลงของรูปปากตามเสียงสระเดี่ยวซึ่งสามารถเพิ่มประสิทธิภาพการรับสื่อข้อมูลให้กับผู้มีปัญหาในการรับฟังหรือการสื่อสารในพื้นที่ที่มีเสียงรบกวนมากได้ รวมทั้งการประยุกต์ใช้ในการสร้างภาพ Animation ได้อีกด้วย

## 1.6 วิธีดำเนินการวิจัย

1. การศึกษากระบวนวิธีในการสร้างแบบจำลองสำหรับการสร้างภาพเคลื่อนไหวของรูปปากด้วยวิธีการ Backpropagation Neural Network โดยอาศัยองค์ประกอบต่างๆของเสียงในภาษาต่างๆ
2. ศึกษาหลักการวิธีออกเสียงของคำตามหลักวิชาสรีรสัทศาสตร์
3. ศึกษาวิธีในการสร้างภาพเคลื่อนไหวตามหลักของกระบวนการสร้าง Animation
4. ทำการเก็บข้อมูลภาพเคลื่อนไหวของรูปปากตามเสียงสระจากผู้พูดแล้วทำการแปลงให้เป็นจุดพิกัด
5. ทำการสร้างระบบ Neural Network เพื่อสร้างแบบจำลองการเคลื่อนไหวของรูปปาก
6. ทำการสร้างแบบจำลองการเคลื่อนไหวของรูปปากตามเสียงสระ และพัฒนาตัวโปรแกรมทดสอบสร้างภาพเคลื่อนไหวเพื่อวัดประสิทธิภาพของแบบจำลองสำหรับการนำไปใช้งานจริง

## 1.7 ลำดับขั้นตอนในการเสนอผลการวิจัย

ในวิทยานิพนธ์เล่มนี้ ได้แบ่งเนื้อหาออกเป็น 5 บท โดยมีบทแรกเป็นบทนำ บทที่ 2 กล่าวถึงความรู้เบื้องต้นและทฤษฎีพื้นฐานที่ใช้ในงานวิจัย บทที่ 3 กล่าวถึงกระบวนการสร้างแบบจำลองการเคลื่อนไหวของรูปปาก ซึ่งได้แก่กระบวนการเก็บข้อมูล การกำหนดรูปแบบการเรียนรู้ของระบบ Neural Network เพื่อสร้างแบบจำลองการเคลื่อนไหวของรูปปากตามเสียงสระ

และการสร้างตัวโปรแกรมทดสอบสร้างภาพเคลื่อนไหว บทที่ 4 กล่าวถึงผลการวิจัยและการวิเคราะห์ผล และบทสุดท้าย จะเป็นการสรุปผลและข้อเสนอแนะของงานวิจัย

**1.8 เอกสารและงานวิจัยที่เกี่ยวข้อง**

Kalberer A.G. และ Luc Van Gool [2] ได้ทำการวิจัยเกี่ยวกับการสร้างภาพเคลื่อนไหวของใบหน้า โดยไม่ได้มีการใช้ข้อมูลเสียงแต่ทำการสร้างภาพโดยอาศัยรูปประโยคที่กำหนดเอาไว้ก่อนแล้ว ใช้วิธีการเก็บจุดพิกัดของรูปหน้าตำแหน่งต่างๆด้วย Triangular mesh model โดยมีลักษณะข้อมูลเป็นเวคเตอร์ 124 จุด หลักสำคัญของงานวิจัยนี้คือการใช้ Triangular mesh model มาจัดเก็บจุดพิกัดของรูปหน้าแบบ 3 มิติ ทั้งหมด 124 จุดแล้วนำ mesh model ของรูปหน้ามาสร้างเป็น Mask โดยอิงกับรูปแบบของตาราง Visemes หรือตารางรูปปากมาตรฐานสำหรับเป็นตัวแทนของเสียงพูดต่างๆ โดยอาศัยหลักการย่อ-ขยาย และการหมุนจุดเพื่อปรับขนาดของ Mask ให้พอดีกับใบหน้าของผู้พูด แต่งานวิจัยชิ้นนี้ก็ไม่ได้สรุปว่าสามารถให้ผลที่ถูกต้องสมบูรณ์ได้ เพราะมีการทดสอบข้อความเพียงแค่ประโยคเดียวและทำการทดลองกับผู้พูดเพียงแค่คนเดียว

Soonkye, Lee และ Dongsuk, Yook.[3] ได้ทำการวิจัยเรื่องการสร้างรูปปากจากข้อมูลเสียงในภาษาอังกฤษด้วยวิธี Hidden Markov Models แล้วทำการเปรียบเทียบผลที่ได้กับ Visemes มาตรฐาน ซึ่งได้ความถูกต้องที่ระดับร้อยละ 66.1 สำหรับการสร้างรูปปากจากข้อมูลเสียงโดยตรง และร้อยละ 70.3 ถ้าทำการแยกหน่วยเสียงก่อนแล้วจึงนำมาสร้างรูปปาก

Hong, Pengyu., Wen, Zhen., และ Huang, S., Thomas. [4] ได้ใช้ Time Delay Neural Network (TDNNs) ในการสร้างภาพเคลื่อนไหวของรูปปากจากสัญญาณเสียงที่ป้อนเข้ามา โดยทำการเก็บข้อมูลของรูปปากสำหรับเสียงต่างๆเป็นค่าเวคเตอร์แล้วทำการสร้างเป็น Mesh model และได้รูปแบบของ mesh model พื้นฐาน 4 รูปแบบ ส่วนข้อมูลเสียงจะถูกจัดกลุ่มด้วย Gaussian Mixture Model และ Multilayer Perceptron เพื่อให้มีความแม่นยำมากยิ่งขึ้น ซึ่งผลที่ได้จากงานวิจัยนี้ทำให้สามารถสร้างภาพเคลื่อนไหวของรูปหน้าจากการป้อนข้อมูลเสียงได้อย่างมีประสิทธิภาพและสมจริง

Kiyotsugu, Kakihara., Satoshi, Nakamura., และ Kiyohhiro Shikano.[5] ได้ทำการศึกษาเกี่ยวกับการใช้ Hidden Markov Models มาทำการสร้างรูปปากในลักษณะ 3 มิติ โดยใช้อุปกรณ์เก็บข้อมูลแบบ Motion Capture ที่เรียกว่า Optotrack มาทำการเก็บข้อมูลการเคลื่อนไหวของรูปปากตามเสียงภาษาญี่ปุ่น

Massaro, W., Dominic, Beskow,Jonas.,Cohen, M., Michael., and others. [6] งานวิจัยนี้ได้เสนอระบบประมวลผลภาพและเสียงสังเคราะห์ที่ได้จากคลื่นเสียงโดยตรง โดยวิธี Time Delay Neural Network และทำการเปรียบเทียบจากการทดลอง 2 ส่วน คือ การใช้คำสำหรับทดสอบ 400 คำโดยตัวอย่างผู้พูดเพียงคนเดียว และการใช้ตัวอย่างคำที่ไม่ได้มีการเตรียมไว้ล่วงหน้า โดยตัวอย่างผู้พูด 10 คน โดยวัดผลการทดสอบด้วยค่า root mean square error และค่า Correlation ของรูปปากที่สามารถสร้างออกมาได้ โดยมีค่า Correlation โดยเฉลี่ยเท่ากับ 0.84

R. R. Rao, T. Chen, และ R. M. Mersereau[7] ได้ทำการทดสอบเปรียบเทียบประสิทธิภาพของวิธีการที่แตกต่างกัน 3 วิธี ในการสร้างรูปปากที่เหมาะสม ได้แก่ การใช้ Classification-based strategy การใช้ Neural Networks และการใช้ Gaussian Mixture-Based Estimation ซึ่งผลที่ได้ปรากฏว่า การใช้ Gaussian Mixture-Based Estimation สามารถให้ผลได้ดีที่สุด รองลงมาคือ การใช้ Neural Network และการใช้ Classification-based strategy ตามลำดับ

Gutierrez,R.,Osuna. และคณะ[8] ได้ทำการวิจัยการสร้างรูปปากในลักษณะ 3 มิติ โดยพิจารณาถึงการเปลี่ยนแปลงของกล้ามเนื้อและการเคลื่อนไหวของส่วนกะโหลกและขากรรไกรในลักษณะของกายวิภาคศาสตร์ด้วย วิธีที่นำมาใช้คือ nearest-neighbor algorithm และ Karhunen-Loéve Transformation ในการคาดการณ์การเคลื่อนไหวของรูปหน้าจากข้อมูลของเสียง แล้วทำการสร้างภาพเคลื่อนไหวออกมาเป็น MPEG-4 ด้วยความถี่ภาพ 60 frames/s

ตารางที่ 1.1 รายการงานวิจัยที่เกี่ยวกับการสร้างรูปปากด้วยวิธีต่างๆ

| งานวิจัย | วิธีการที่ใช้ในการวิจัย |
|---|---|
| Face animation based on observed 3D speech dynamics. (Kalberer A.G. และ Luc Van Gool [2]) | Triangular Mesh Model |
| Audio-to-Visual Conversion Using Hidden Markov Models. (Soonkye, Lee และ Dongsuk, Yook.[3]) | Hidden Markov Models |
| Real-time Speech Driven Avatar with Constant Short Time Delay. (Hong, Pengyu., Wen, Zhen., และ Huang, S., Thomas. [4]) | Time Delay Neural Network |
| Speech – to – face movement synthesis based on speech – driven HMMs. (Kiyotsugu, Kakihara., Satoshi, Nakamura., และ Kiyohhiro Shikano.[5]) | Hidden Markov Models |
| Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. (Massaro, W., Dominic, Beskow,Jonas.,Cohen, M., Michael., and others. [6]) | Neural Network |
| Audio-to-visual conversion for multimedia communication. (R. R. Rao, T. Chen, และ R. M. Mersereau[7]) | Gaussian Mixture-Based Estimation ,Neural Network และ Ciassification-based strategy |
| Speech – driven Facial Animation with Realistic Dynamics. (Gutierrez,R.,Osuna. และคณะ[8]) | nearest-neighbor algorithm และ Karhunen-Loéve Transformation |