บทที่ 2

ความรู้เบื้องต้นและทฤษฎีพื้นฐานที่ใช้ในงานวิจัย

ในบทนี้จะกล่าวถึงความรู้เบื้องต้นและทฤษฎีพื้นฐานที่จำเป็นสำหรับการทำความเข้าใจเนื้อหาของวิทยานิพนธ์ฉบับนี้

2.1. สัทศาสตร์ : กลไกการเปล่งเสียง

สัทศาสตร์ คือการศึกษาถึงธรรมชาติของเสียงพูดของมนุษย์โดยทั่วไป ไม่ระบุเจาะจงไปว่าเสียงพูดนั้นๆเป็นเสียงในภาษาใด ซึ่งเป็นผลจากการศึกษาวิเคราะห์และพัฒนาเป็นองค์ความรู้ ดังนั้นเนื้อหาของวิชาสัทศาสตร์จึงมีมากมาย นักสัทศาสตร์จึงได้แบ่งวิชาสัทศาสตร์ออกเป็นสาขาต่างๆ ได้แก่

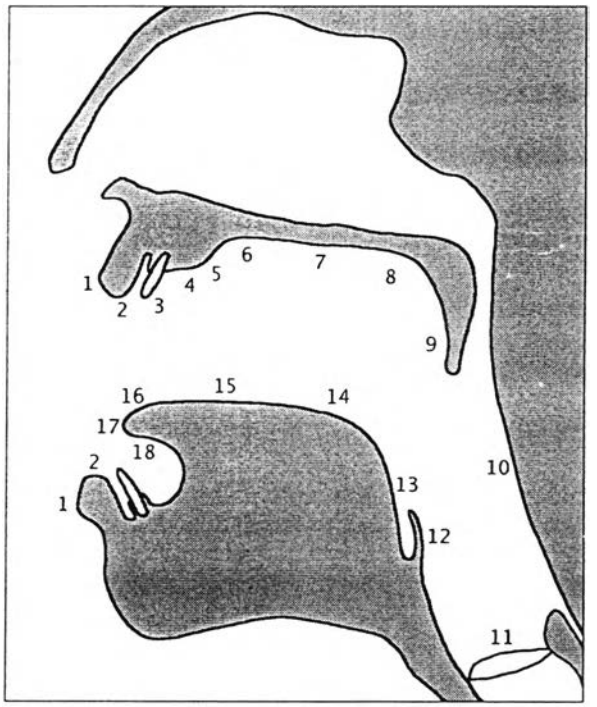
2.1.1. สรีรสัทศาสตร์ (Articulatory Phonetics) เป็นการศึกษากลไกกระแสลมที่ใช้ในการเปล่งเสียงพูด อวัยวะต่างๆที่ใช้ในการพูด กระบวนการออกเสียง กระบวนการเปล่งเสียง ตลอดจนเสียงประเภทต่างๆที่จำแนกตามการเกิดของอวัยวะที่ใช้ในการพูด

2.1.2. กลสัทศาสตร์ (Acoustic Phonetics) เป็นการศึกษาลักษณะทางกายภาพของคำพูด เช่น คลื่นเสียงลักษณะต่างๆ ความถี่ของคลื่นเสียง ความกังวานของเสียง เสียงรบกวนลักษณะต่างๆ เป็นต้น

2.1.3. โสตสัทศาสตร์ (Auditory Phonetics) เป็นการศึกษาเกี่ยวกับสรีระของหูการรับรู้คลื่นเสียงตลอดจนจิตวิทยาการรับรู้

สาขาของวิชาสัทศาสตร์ทั้งสามแสดงถึงวงจรของการพูดได้อย่างชัดเจน คือ เริ่มต้นด้วยกลไกการเคลื่อนไหวของอวัยวะในการพูดออกเสียงแบบต่างๆ ซึ่งการศึกษาในส่วนกลไกการเคลื่อนไหวเพื่อเกิดเสียงนี้เกี่ยวข้องกับหลักวิชาสรีรสัทศาสตร์ และเมื่อเสียงพูดที่เปล่งออกมาเคลื่อนผ่านอากาศไปสู่ผู้ฟังในลักษณะของคลื่นเสียง การศึกษาลักษณะของคลื่นเสียงก็จะเกี่ยวข้องกับหลักวิชากลสัทศาสตร์ เมื่อเสียงผ่านอากาศไปถึงผู้ฟัง คลื่นเสียงจะผ่านจากส่วนของหูชั้นนอกไปยังอวัยวะส่วนที่ใช้รับคลื่นเสียงเพื่อทำการแปรสัญญาณเสียงที่ได้รับเป็นข้อความหรือสารที่ส่งออกมา การศึกษากระบวนการส่วนนี้จะเกี่ยวข้องกับหลักวิชาโสตสัทศาสตร์

หลักของวิชาสรีรสัทศาสตร์นั้นจะเกี่ยวข้องกับการออกเสียง และกลไกการควบคุมอวัยวะที่ใช้ในการสร้างเสียง ซึ่ง อวัยวะที่ใช้ในการสร้างเสียงนั้น มีทั้งแบบที่สามารถมองเห็นได้จากภายนอก เช่น ริมฝีปาก ฟัน และแบบที่ไม่สามารถมองเห็นได้จากภายนอกแต่สามารถศึกษาได้จากภาพเอกซเรย์ เช่น เส้นเสียง ลิ้น



รูปที่ 2.1 ตำแหน่งเกิดเสียงในกระบวนการออกเสียง

1. ริมฝีปาก ด้านนอก (Exo-labial)
2. ริมฝีปาก ด้านใน (Endo-labial)
3. ฟัน (Dental)
4. ปุ่มเหงือก (Alveolar)
5. หลังปุ่มเหงือก (Post-alveolar)
6. หน้าเพดานแข็ง (Pre-palatal)
7. เพดานแข็ง (Palatal)
8. เพดานอ่อน (Velar)
9. ลิ้นไก่ (Uvular)
10. ช่องคอ (Pharyngeal)
11. เส้นเสียง (Glottal)
12. ลิ้นปิดกล่องเสียง (Epiglottal)
13. โคนลิ้นในช่องคอ (Radical)
14. ผนังลิ้นส่วนหลัง (Postero-dorsal)
15. ผนังลิ้นส่วนหน้า (Antero-dorsal)
16. ปลายลิ้น (Laminal)
17. ปลายสุดลิ้น (Apical)
18. ใต้ปลายสุดลิ้น (Sub-apical)

ที่มา : http://upload.wikimedia.org/wikipedia/commons/3/3e/Place_articulation.png

เนื่องจากงานวิจัยชิ้นนี้เน้นที่การสร้างแบบจำลองของรูปปากตามเสียงสระ ซึ่งรูปปากหรือริมฝีปากนั้นเป็นอวัยวะส่วนที่สามารถมองเห็นจากภายนอกได้ง่ายที่สุด ประกอบกับรูปแบบของเสียงนอกเหนือไปจากเสียงสระที่เกิดขึ้นจากการเคลื่อนไหวของอวัยวะส่วนต่างๆมีจำนวนมากและไม่มีความเกี่ยวข้องกับงานวิจัยของเรา จึงขอละเว้นรายละเอียดส่วนนี้ไป ซึ่งสามารถศึกษาเพิ่มเติมได้จากหนังสือของ อมร ทวีศักดิ์[8] และ พิณทิพย์ ทวยเจริญ[9]

เสียงพูดแต่ละคำที่เปล่งออกมา จะประกอบด้วยหน่วยของเสียงที่มีลักษณะต่างกันหลายๆส่วนประกอบกัน ซึ่งในทางสัทศาสตร์ได้ทำการแยกเสียงออกเป็นเสียงพยัญชนะและเสียงสระ โดยอาศัยกลไกการเปล่งเสียงพูดซึ่งประกอบด้วยกระบวนการ 3 ส่วน ได้แก่

1. กลไกกระแสลม คือกระบวนการบังคับให้เกิดการเคลื่อนไหวของกระแสลมจากจุดต่างๆที่เป็นแหล่งกำเนิดกระแสลมในการออกเสียง ได้แก่ กระแสลมจากปอด กระแสลมจากกล่องเสียง และกระแสลมจากเพดานอ่อน

2. กระบวนการเปล่งเสียง เพราะเพียงแค่การเกิดลมออกจากกลไกกระแสลมเพียงอย่างเดียว ไม่สามารถทำให้เกิดเสียงได้ จึงต้องมีการเปล่งเสียง ซึ่งอาศัยการเปิด-ปิด และการสั่นสะเทือนของเส้นเสียง ซึ่งการสั่นสะเทือนของเส้นเสียงก็เป็นองค์ประกอบสำคัญของการเกิดเสียงด้วยเช่นกัน

3. กระบวนการออกเสียง คือ กลไกที่ใช้อวัยวะต่างๆในช่องปาก มาควบคุมเสียงและกระแสลมที่เกิดจากกระบวนการก่อนหน้า เพื่อให้ได้เสียงในลักษณะต่างๆ ซึ่งรายละเอียดของกระบวนการนี้มีเนื้อหามากและไม่เกี่ยวข้องกับงานวิจัย จึงขออธิบายเพียงเพื่อให้เข้าใจกลไกการเกิดเสียงเท่านั้น

ซึ่งในทางสัทศาสตร์ ได้อธิบายเกี่ยวกับเสียงสระไว้ว่า หมายถึงเสียงที่อาศัยการเคลื่อนไหวของลิ้นประกอบกับรูปปากเท่านั้น และมีกลไกกระแสลมจากปอดเพียงอย่างเดียว ในขณะที่อธิบายเสียงพยัญชนะไว้ว่า หมายถึงเสียงที่เกิดจากการกระทบกันของอวัยวะในช่องปากไม่ว่าจะเป็นส่วนใดก็ตาม เช่น ฟัน ลิ้น เหงือก เพดานปาก ฯลฯ โดยไม่คำนึงว่ามีกลไกกระแสลมจากแหล่งใด

นักสัทศาสตร์ได้กำหนดเสียงสระมาตรฐานไว้ เพื่อเป็นเกณฑ์ในการอ้างอิงทางสัทศาสตร์ขึ้นมา โดย Daniel Jones นักสัทศาสตร์จากมหาวิทยาลัยลอนดอนได้กำหนดสระมาตรฐาน 2 ชุด คือ สระมาตรฐานชุดหลัก (primary cardinal vowels) และสระมาตรฐานชุดรอง (secondary

cardinal vowels) ทั้งนี้ เพื่อใช้สระมาตรฐานทั้ง 2 ชุดเป็นหลักในการบรรยายเสียงสระในภาษาต่างๆ โดยไม่ได้กำหนดให้เป็นเสียงสระในภาษาใดภาษาหนึ่งโดยเฉพาะ

Daniel Jones ได้กำหนดสระมาตรฐานโดยอาศัยรูปร่างของลิ้นและลักษณะการห่อปากเป็นเกณฑ์การจำแนก โดยแบ่งสระมาตรฐานหลักไว้ 8 เสียง และสระมาตรฐานรองอีก 8 เสียง ซึ่งเสียงสระของภาษาใดๆก็ตาม จะไม่ออกเสียงนอกเหนือไปจากเสียงสระมาตรฐานทั้ง 2 ชุด และเสียงผสมระหว่างสระมาตรฐานอย่างเด็ดขาด ซึ่งลักษณะของเสียงสระในภาษาไทยเองก็อยู่ในข่ายของเสียงสระมาตรฐานและเสียงที่ผสมกันระหว่างเสียงสระมาตรฐานด้วยเช่นกัน

ลักษณะของสระมาตรฐานชุดหลักทั้ง 8 แปดเสียง ได้แก่

1. [i] แทนลักษณะ สระสูง ใช้การเคลื่อนไหวของลิ้นส่วนหน้า ปากไม่ห่อ(เหยียดปากออก)

2. [e] แทนลักษณะ สระกึ่งสูง ใช้การเคลื่อนไหวของลิ้นส่วนหน้า ปากไม่ห่อ

3. [ɛ] แทนลักษณะ สระกึ่งต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหน้า ปากไม่ห่อ

4. [a] แทนลักษณะ สระต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหน้า ปากไม่ห่อ (ปากเปิดกว้างมาก)

5. [ɑ] แทนลักษณะสระต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหลัง ปากไม่ห่อ (ปากเปิดกว้างมาก)

6. [ɔ] แทนลักษณะสระกึ่งต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหลัง ปากห่อ

7. [o] แทนลักษณะสระกึ่งสูง ใช้การเคลื่อนไหวของลิ้นส่วนหลัง ปากห่อ

8. [u] แทนลักษณะสระสูง ใช้การเคลื่อนไหวของลิ้นส่วนหลัง ปากห่อมาก

ลักษณะของสระมาตรฐานชุดรอง 8 แปดเสียง ซึ่งกำหนดขึ้นเพื่อให้สามารถครอบคลุมเสียงสระในภาษาต่างๆส่วนใหญ่ได้ โดยยึดลักษณะของลิ้นในสระมาตรฐานชุดหลักเป็นเกณฑ์ แต่ปรับรูปปากให้อยู่ในลักษณะตรงกันข้าม

1. [y] แทนลักษณะสระสูง ใช้การเคลื่อนไหวของลิ้นส่วนหน้า คล้าย [i] แต่ปากห่อมาก

2. [ø] แทนลักษณะสระกึ่งสูง ใช้การเคลื่อนไหวของลิ้นส่วนหน้า คล้าย [e] แต่ปากห่อ

3. [ɶ] แทนลักษณะสระกึ่งต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหน้า คล้าย [ɛ] แต่ปากห่อ

4. [ɐ] แทนลักษณะสระต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหน้า คล้าย [a] แต่ปากห่อ

5. [ɒ] แทนลักษณะสระต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหลัง คล้าย [ɑ] แต่ปากห่อ

6. [ʌ] แทนลักษณะสระกึ่งต่ำ ใช้การเคลื่อนไหวของลิ้นส่วนหลัง คล้าย [ɔ] แต่ปากไม่ห่อ

7. [ɤ] แทนลักษณะสระกึ่งสูง ใช้การเคลื่อนไหวของลิ้นส่วนหลัง คล้าย [u] แต่ปากไม่ห่อ

8. [ɯ] แทนลักษณะสระสูง ใช้การเคลื่อนไหวของลิ้นส่วนหลัง คล้าย [e] แต่ปากไม่ห่อ

ซึ่งลักษณะของปากเมื่อออกเสียงจะมีลักษณะโดยรวมๆ แบ่งเป็น 3 แบบ ได้แก่

- Unrounded คือลักษณะที่มุมปากถูกดึงออกไป หรือปากเหยียดออก
- Neutral คือลักษณะปากที่อยู่ในท่าธรรมดา คือ ปากไม่ห่อกลมและไม่เหยียดออก
- Rounded คือลักษณะที่มุมปากถูดดึงเข้ามาหรือปากห่อกลม

แม้จะแบ่งโดยรวมๆได้ 3 รูปแบบ แต่เมื่อมีการเปล่งเสียงพูด ลักษณะของปากก็จะมีการเคลื่อนไหวในลักษณะที่ไม่ตายตัวขึ้นอยู่กับเสียงสระ บางเสียงอาจจะห่อปากมาก หรือห่อปากน้อย บางเสียงอาจมีการขยับปากเพียงแค่เปิดปากในลักษณะของ Neutral หรือเป็นแบบกึ่งๆก็ได้ เช่นเสียงสระผสมที่เกิดจากการผสมการออกเสียงของเสียงสระมาตรฐาน 2 เสียงขึ้นไปแบบกึ่งๆ ตัวอย่างของเสียงสระผสมในลักษณะนี้ได้แก่ เสียงสระเอียะ-เอีย หรือสระอัวะ-อัว เป็นต้น ซึ่งรูป

ปากที่เกิดจากเสียงสระผสมนี้จะมีการเคลื่อนไหวแบบกึ่งซึ่งแตกต่างไปจากการเคลื่อนไหวของสระเดี่ยวธรรมดา

2.2. Neural Network **หรือระบบเครือข่ายประสาทเทียม**

การศึกษาเรื่องโครงข่ายประสาทสมัยใหม่เริ่มต้นในศตวรรษที่ 19 ทำให้มีความตื่นตัวในการศึกษาการทำงานของสมองมนุษย์และพบว่าสมองมนุษย์ประกอบด้วยปมประสาท (Neuron) จำนวนมากแต่ละปมประสาทมีการเชื่อมต่อกันจำนวนมหาศาลทำให้การทำงานของสมองมนุษย์เป็นไปอย่างรวดเร็วมากเนื่องจากมีความไม่เป็นเชิงเส้นและมีการทำงานแบบขนาน จึงทำให้มีการสร้างแบบจำลองการทำงานของสมองมนุษย์แบบง่าย ๆ ขึ้น โครงข่ายประสาทเทียมจึงเป็นแบบจำลองการทำงานของสมองมนุษย์ที่ถูกนำมาประยุกต์สร้างเป็นวงจรอิเล็กทรอนิกส์หรือเป็นซอฟต์แวร์เพื่อการประมวลผลในงานต่าง ๆ

องค์ประกอบพื้นฐานของโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมประเภทต่าง ๆ มีองค์ประกอบที่เหมือนกัน 4 อย่างคือ

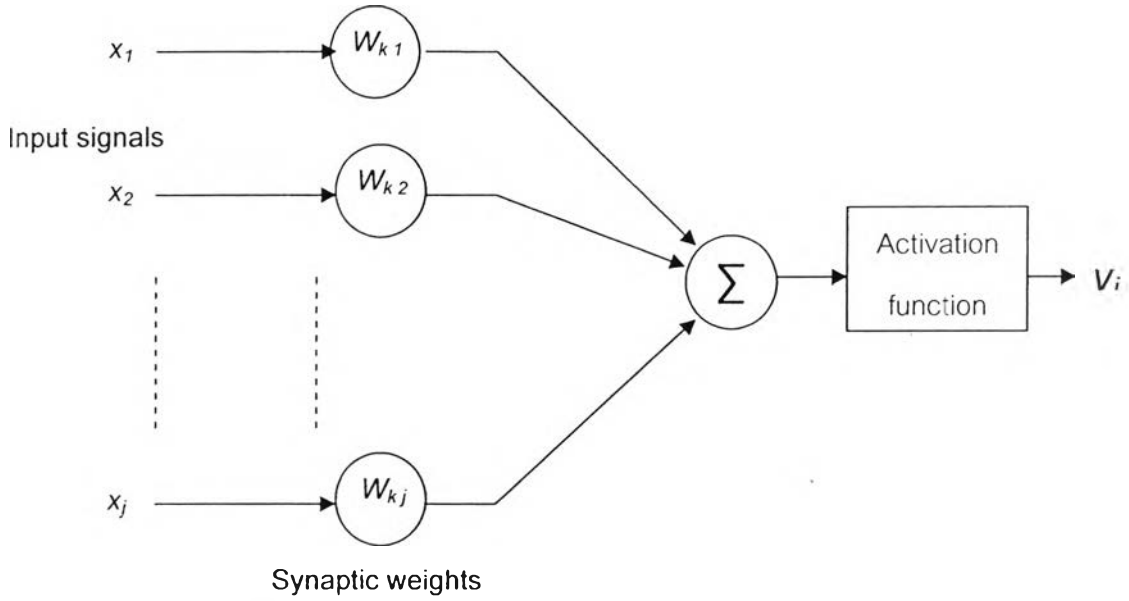
1. หน่วยประมวลผล (Processing Units)
2. การเชื่อมต่อ (Connections)
3. กระบวนการคำนวณ (Computing Procedure)
4. กระบวนการฝึกฝน (Training Procedure)

หน่วยประมวลผล (Processing Units)

หน่วยประมวลผลในโครงข่ายประสาทเทียมแบ่งออกเป็น 3 ส่วนคือ หน่วยข้อมูลเข้าทำหน้าที่รับข้อมูลจากภายนอก หน่วยซ่อนตัวทำหน้าที่แปลงข้อมูลภายในและ หน่วยข้อมูลออกทำหน้าที่ตัดสิน หรือควบคุมสัญญาณ

การเชื่อมต่อ (Connections)

หน่วยประมวลผลในโครงข่ายประสาทเทียมจะถูกจัดเรียงเป็นโครงสร้างต่างๆ โดยการเชื่อมต่อซึ่งมีค่ากำกับไว้ ค่าที่กำกับการเชื่อมต่อเรียกว่าน้ำหนักการเชื่อมต่อ



รูปที่ 2.2 แบบจำลองการทำงานที่ไม่เป็นเชิงเส้นของโนด

กระบวนการคำนวณ (Computing Procedure)

โนดเป็นหน่วยประมวลผลข้อมูลซึ่งเป็นพื้นฐานของโครงข่ายประสาทเทียม ซึ่งแบบจำลองของโนดมีองค์ประกอบดังนี้

1. การเชื่อมต่อซึ่งแต่ละการเชื่อมต่อจะมีคุณลักษณะคือน้ำหนักการเชื่อมต่อ สัญญาณเข้า x_j ของการเชื่อมต่อ j กับโนดที่ k จะถูกคูณด้วยน้ำหนักการเชื่อมต่อ w_{kj}
2. การบวกสำหรับการรวมสัญญาณเข้าที่ถูกคูณด้วยน้ำหนักการเชื่อมต่อ
3. ฟังก์ชันการกระตุ้นสำหรับจำกัดแอมปลิจูดของสัญญาณที่ออกจากโนด

กระบวนการฝึกฝน (Training Procedure)

การฝึกฝนโครงข่ายประสาทเทียมคือการปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อ หรือบางกรณีเป็นการปรับเปลี่ยนโครงสร้างของโครงข่ายประสาทเทียมซึ่งเป็นการเพิ่มหรือลบการเชื่อมต่อของโนด การปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อจะมีลักษณะทั่วไปมากกว่าการปรับเปลี่ยนโครงสร้างเพราะว่าการที่ค่าน้ำหนักการเชื่อมต่อเท่ากับศูนย์คือการลบการเชื่อมต่อนั้นออกจากโครงข่ายประสาทเทียม อย่างไรก็ตามการเปลี่ยนแปลงโครงสร้างของโครงข่ายประสาทเทียมจะเป็นการเพิ่มความเร็วในการเรียนรู้และเพิ่มความสามารถในการรู้จำรูปแบบทั่วไป

โครงข่ายประสาทเทียมมีโครงสร้างเป็นชั้นและมีความไม่เป็นเชิงเส้น การปรับเปลี่ยนค่าน้ำหนักการเชื่อมต่อทำโดยวิธีการวนซ้ำเช่น Gradient Descent การปรับเปลี่ยนค่า

น้ำหนักการเชื่อมต่อแต่ละครั้งจะต้องไม่ทำให้การเรียนรู้ที่ผ่านมาสูญเสียไป ค่าคงที่ที่ใช้ควบคุมขนาดของการปรับเปลี่ยนน้ำหนักการเชื่อมต่อเรียกว่าอัตราการเรียนรู้ (Learning Rate) การกำหนดค่าอัตราการเรียนรู้มีความสำคัญมาก ถ้ากำหนดค่าอัตราการเรียนรู้น้อยเกินไปจะทำให้การเรียนรู้นานมากและถ้ากำหนดค่าอัตราการเรียนรู้มากจะทำให้สูญเสียการเรียนรู้ที่ผ่านมา

โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP)

โครงข่ายประสาทเทียมแบบ MLP ประกอบด้วยชั้นข้อมูลเข้าสำหรับรับข้อมูล ชั้นซ่อนตัวและชั้นข้อมูลออกสำหรับการคำนวณ สัญญาณเข้าจะผ่านเข้าไปในโครงข่ายประสาทเทียมในทิศทางเดียวจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง โครงข่ายประสาทเทียมแบบ MLP ถูกประยุกต์ใช้สำหรับงานที่มีความซับซ้อนได้เป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบ Supervised และใช้ขั้นตอนวิธีการส่งค่าย้อนกลับสำหรับการฝึกฝน กระบวนการฝึกฝนแบบส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือการส่งผ่านไปข้างหน้า และการส่งผ่านย้อนกลับ สำหรับการส่งผ่านไปข้างหน้าข้อมูลจะผ่านโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และจะส่งผ่านจากชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิด คือผลต่างของผลตอบแท้จริง กับผลตอบเป้าหมายเกิดเป็นสัญญาณผิดพลาด ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงข้ามกับการเชื่อมต่อ ค่าน้ำหนักการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบแท้จริงเข้าใกล้ผลตอบเป้าหมาย

โครงข่ายประสาทเทียมแบบ MLP มีลักษณะเด่น 3 ประการคือ

1. แบบจำลองของแต่ละโนดมีความไม่เป็นเชิงเส้นดังกล่าว ต้องมีความราบเรียบคือสามารถหาอนุพันธ์ได้ทุกจุด ความไม่เป็นเชิงเส้นดังกล่าวถูกกำหนดจากฟังก์ชันซิกมอยด์

$$y_j = \frac{1}{1+\exp(-v_j)} \tag{2.1}$$

โดยที่ v_j เป็นผลรวมภายในของโนดที่ j และ y_j คือสัญญาณออกของโนดที่ j

2. โครงข่ายประสาทเทียมจะมีจำนวนชั้นซ่อนตัวที่มากกว่า 1 ชั้นได้ซึ่งไม่ใช่ชั้นข้อมูลเข้าและชั้นข้อมูลออก โนดในชั้นซ่อนตัวนี้จะทำให้โครงข่ายประสาทเทียมสามารถเรียนรู้งานที่มีความซับซ้อนได้ดีขึ้น

3. โครงข่ายประสาทเทียมแบบ MLP ที่มีจำนวนชั้นซ่อนตัว 2 ชั้นสัญญาณที่ในโครงข่ายประสาทเทียมแบบ MLP มี 2 ประเภท คือ Function Signal และ Error Signal ซึ่งมีรายละเอียดดังนี้

3.1 Function Signal เป็นสัญญาณเข้าที่มาจากโนดในชั้นก่อนหน้าและส่งผ่านไปข้างหน้าจากโนดหนึ่งไปสู่อีกโนดหนึ่ง

3.2 Error Signal เป็นสัญญาณที่เกิดขึ้นที่โนดในชั้นข้อมูลออกของโครงข่ายประสาทเทียมและถูกส่งผ่านย้อนกลับจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง

2.3 Audio-to-Visual Conversion

Audio-to-visual Conversion หรือ การแปลงข้อมูลเสียงเพื่อจำลองเป็นภาพ หมายถึง การสร้างภาพการเคลื่อนไหวของรูปหน้าในส่วนที่เกี่ยวเนื่องกับการพูด อันได้แก่ ริมฝีปาก ฟัน คาง และโครงหน้าส่วนล่าง ซึ่งในบางครั้งก็ได้ทำการจำลองทั้งใบหน้าด้วย โดยอาศัยข้อมูลหรือการจำแนกด้วยเสียง ซึ่งเป้าหมายของกระบวนการ Audio-to-Visual คือ การจำลองภาพการเคลื่อนไหวในขณะทำการพูดให้เหมือนจริงมากที่สุด เพื่อประโยชน์ในการสร้างภาพเคลื่อนไหวในการใช้งานในทางด้าน Animation เช่น Animation ในทางภาพยนตร์ Animation ที่ใช้สำหรับ Avatar หรือการประชุมทางไกล หรือการพัฒนาส่วน Interface สำหรับโปรแกรมบางอย่างหรืองานบริการที่ต้องการสวนเชื่อมต่อกับผู้ใช้ที่ให้ความเป็นกันเอง หรือเพื่อช่วยในการจำแนกการฟังสำหรับในพื้นที่ที่มีเสียงรบกวนมาก

แนวความคิดในการทำ Audio-to-Visual conversion หรือ Audio-Visual speech มาจากแนวความคิดที่ว่า ความสามารถในการแยกแยะเสียงด้วยการฟังของมนุษย์จะลดลงจากการเพิ่มขึ้นของเสียงรบกวนในสภาวะแวดล้อม แต่มนุษย์สามารถชดเชยได้ด้วยการสังเกตรูปปากของผู้พูดเพื่อแยกแยะสิ่งที่ผู้พูดต้องการจะสื่อหรือที่เรียกว่าการอ่านริมฝีปากนั้นเอง ซึ่งจากงานวิจัยของ [sumby W. and Pollack I.] ที่ได้ทดลองให้ผู้ฟังทำการรับฟังเสียงคำศัพท์ 64 คำ ที่ระดับเสียงรบกวน 0 dB ผลปรากฏว่า สามารถแยกแยะคำศัพท์ได้มากกว่าร้อยละ 80 ทั้งจากการรับฟังเพียงเสียงอย่างเดียว และการรับฟังเสียงประกอบกับการดูปากของผู้พูดด้วย แต่เมื่อเพิ่มเสียงรบกวนขึ้นเป็น 30 dB ปรากฏว่า การแยกแยะคำศัพท์ด้วยการฟังเสียงประกอบกับการดูปากของผู้พูดสามารถแยกแยะคำศัพท์ได้ที่ระดับร้อยละ 60 แต่การแยกแยะคำศัพท์ด้วยการฟังเพียงอย่างเดียวสามารถแยกแยะคำศัพท์ได้ต่ำกว่าร้อยละ 20 ซึ่งจากผลงานการวิจัยนี้ทำให้สามารถกล่าวได้ว่า การมองเห็นรูปปากของผู้พูดสามารถช่วยให้ประสิทธิภาพการรับรู้คำพูดของ

ผู้ฟังเพิ่มขึ้นได้ ถึงแม้ว่าจะมีในบางกรณีที่เสียงที่พูดออกมาจะมีความแตกต่างกันอย่างชัดเจน แต่ก็อาจจะให้รูปปากที่เหมือนกันได้ เช่น เสียง "บ" เสียง "ป" และ เสียง"พ" ซึ่งจะให้รูปปากที่คล้ายกันจนยากที่แยกแยะออกได้ด้วยการดูเพียงอย่างเดียว ทั้งนี้เพราะความสัมพันธ์ระหว่างเสียงพูดกับรูปปากเป็นความสัมพันธ์แบบ many-to-one

นอกจากนี้ ในการทำวิจัยทางด้าน Audio-Visual speech ก็ยังได้กำหนดรูปแบบของรูปปากเพื่อเป็นมาตรฐานสำหรับการทำวิจัยทางด้านนี้เช่นเดียวกับที่ในทางสัทศาสตร์ได้กำหนดสัญลักษณ์เสียงไว้ ซึ่งในทาง Audio-Visual speech ได้เรียกรูปปากแบบมาตรฐานนี้ว่า Visemes และกระบวนการเก็บข้อมูลสำหรับสร้างภาพของรูปปากได้แบ่งออกเป็น 2 วิธีใหญ่

1. การเก็บข้อมูลโดยใช้ภาพโดยตรง เป็นการบันทึกภาพลำดับการพูดในแต่ละเฟรมของการพูดคำ เพื่อนำมาจัดระดับแสง สี และเรียงลำดับคำใหม่ตามต้องการ ซึ่งจะสามารถได้ข้อมูลของภาพทั้งหมด โดยไม่ผิดเพี้ยนไปจากความเป็นจริงเลย แต่วิธีการนี้จะมีข้อด้อยในเรื่องของความตายตัวของจุด ตำแหน่ง ขนาดของรูปปาก รวมทั้งขนาดของข้อมูลที่ได้จะมีขนาดที่ใหญ่มาก

2. การเก็บข้อมูลในลักษณะของจุดพิกัด เป็นการเก็บข้อมูลเพื่อนำมาสร้างเป็นแบบจำลองแทนการใช้ข้อมูลภาพโดยตรง วิธีนี้มีความยืดหยุ่นมากกว่าวิธีแรก เพราะค่าที่ได้ออกมาจะเป็นเพียงค่าระบบพิกัดเพื่อนำไปสร้าง animation ในลักษณะของการสร้าง In-Between หรือช่วงเฟรมระหว่างจุดเริ่มต้นและจุดสุดท้ายต่อไป วิธีนี้มีข้อดีที่มีขนาดของข้อมูลที่น้อยกว่าวิธีแรกมากๆ และสามารถทำการประมวลผลได้เร็วกว่า ยืดหยุ่นกว่า แต่ก็มีข้อด้อยที่มีความผิดพลาดของข้อมูลที่สูงกว่าวิธีแรก และค่าที่ได้จากแบบจำลองมีเพียงค่าพิกัดของรูปปากเท่านั้น ส่วนองค์ประกอบอื่นๆทั้งพื้นผิว สี แสง เงา ต้องทำการสร้างเองในกระบวนการทำ animation

เป้าหมายของการพัฒนางานวิจัยทางด้าน Audio-Visual speech

- การวิจัยในเรื่องการับรู้และการสื่อสารของมนุษย์
- การพัฒนาเครื่องมือเพื่อช่วยในผู้บกพร่องทางการได้ยิน
- การพัฒนาระบบเชื่อมต่อกับผู้ใช้ที่เหมาะสมต่อการใช้งาน เช่น ในพื้นที่สาธารณะที่มีเสียงรบกวนจากสภาพแวดล้อม เช่น สถานีรถไฟ สนามบิน หรือศูนย์การค้า