บทคัดย่อ

งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาหาวิธีการรู้จำเสียงพูดคำไทยโดด ๆ แบบไม่ขึ้นกับผู้พูด โดยใช้วิธีการเปรียบเทียบ ข้อมูลสเปกโตแกรมของเสียงทดสอบ (test pattern) กับเสียงอ้างอิง (reference patterm) เพื่อหาระดับความเหมื่อนด้วยวิธีสห่สัมพันธ์ (Correlation) ผลการเปรียบเทียบได้เป็นข้อมูลที่แสดงระดับความเหมือนของเสียงทดสอบที่สัมพันธ์กับเสียงอ้างอิงอื่นๆ จากนั้นจึงนำข้อมูลดังกล่าวไปผ่านขบวนการตัดสินใจด้วยนิวรอลเน็ตเวอร์คที่ผ่านการฝึกสอนด้วยระดับความสัมพันธ์ของเสียงอ้างอิงกับเป้าหมายที่ต้องการ ทำให้ได้ผลการวิเคราห์ที่มีความถูกต้องสูง

Abstract

This research was aimed at finding methodes in speesh recognition of thai words in isolation fo independent speaker. The process of studing was comparing the spectrogram data of the voice test pattern with the referent pattern in order to find the levels of similarity through correlation. It is discovered that the data output is the data showing levels of similarity of the compared voices. In consequence, the data were bought to the decision process by using neural networks which had already been trained by levels of referent voices relation with target expecting. This has assured the reliability of the findings.

สารบัญ

บทที่ 1
บทนำ

หลายปีที่ผ่านมาเทคโนโลยีเข้าใจภาษามนษย์ของเครื่องคอมพิวเตอร์ได้ก้าวหน้ามาเป็นลำดับอย่างต่อเนื่องจากความก้าวหน้าทางเทคโนโลยีการสร้างวงจรรวมขนาดใหญ่ (VLSI) และความก้าวหน้าของศาสตร์การประมวลผลสัญญาณดิจิตอล จึงทำให้ในปัจจุบันนี้ได้มีอุปกรณ์ที่สามารถจะรับคำสั่งและทำงานตามคำสั่งที่เป็นเสียงพูดแทนจากการรับคำสั่งจากปุ่มฟังก์ชันการทำงานหรือจากอุปกรณ์ตัวตรวจจับชนิดต่าง แต่การที่คอมพิวเตอร์จะเข้าใจและทำงานตามคำสั่งที่เป็นเสียงพูดนั้น คอมพิวเตอร์จะต้องรับสัญญาณเสียงและทำการแปลงสัญญาณเสียงนั้นให้เป็นข้อมูลเชิงตัวเลขและทำการประมวลผลต่างๆมากมายหลายขั้นตอน แต่คอมพิวเตอร์ก็ไม่สามารถที่จะเข้าใจภาษาของมนุษย์ได้ทั้งหมดจะเข้าใจได้ในจำนวนที่น้อยไม่กี่คำสั่ง แต่คำสั่งที่ให้คอมพิวเตอร์เข้าใจนั้นก็เป็นคำสั่งที่เราใช้กันบ่อยๆ เช่น เปิด,ปิด ซ้ายและขวา เป็นต้นเพียงคำสั่งเหล่านี้ก็สามารถทำให้เรานำไปใช้งานต่างได้มากมาย

ระบบควบคุมด้วยเสียงพูดจากระยะไกลนั้นเป็นการประยุกต์การใช้งานของการรู้จำเสียงของคอมพิวเตอร์ร่วมกับอุปกรณ์สื่อสารและตัวควบคุมโดยคอมพิวเตอร์จะรับข้อมูลเสียงที่เป็นข้อมูลเชิงตัวเลขหรือที่เรียกว่าข้อมูลดิจิตอล (Digital) มาทำการวิเคราะห์คำนวณเพื่อให้สามารถจดจำเสียงพูดที่กำหนดให้เข้าใจความหมายแล้วจึงแปลความหมายของเสียพูดเหล่านั้นเป็นการกระทำโดยสั่งงานไปยังตัวควบคุมเพื่อควบคุมอุปกรณ์ ข้อมูลเสียงนั้นอาจจะเกิดจากการพูดผ่านระบบโทรศัพท์โดยเสียงเดินทางผ่านสายและชุมสายโทรศัพท์มายังเครื่องรับโทรศัพท์ซึ่งอยู่ที่บ้านที่มีจงจรรับโทรศัพท์อัตโนมัติแล้วจึงนำข้อมูลเสียงนี้ส่งให้คอมพิวเตอร์เพื่อการรู้จำเสียงพูดต่อไป อุปกรณ์ทางด้านโทรศัพท์และตัวควบคุมอุปกรณ์ไฟฟ้านั้นมีอยู่มากมายหลายชนิดในท้องตลาดและเป็นเทคโนโลยีที่เหมาะสมอยู่แต่ปัญหาใหญ่ของงานวิจัยนี้ก็คือโปรแกรมที่จะสามารถทำให้คอมพิวเตอร์เข้าใจความหมายของเสียงพูดนั้นเป็นเรื่องที่ยากต้องใช้ทฤษฎีการประมวลผลทางดิจิตอล (Digital Singnal Procession) ต่างๆ มากมายดังนั้นงานวิจัยนี้จึงมุ่งเน้นไปที่การพัฒนาวิธีการที่จะให้คอมพิวเตอร์รู้จำเสียงพูดเป็นหลัก

การวิเคราะห์เสียงของคอมพิวเตอร์โดยไม่ขึ้นกับผู้พูด เป็นการพัฒนาระบบคำสั่งของคอมพิวเตอร์ โดยไม่เจาะจงผู้ออกคำสั่ง เป็นการเปิดกว้างให้กับคนทั่วไปสามารถออกคำสั่งกับคอมพิวเตอร์ได้ เป็นจุดเริ่มต้นของการพัฒนาการสื่อสารระหว่างมนุษย์กับคอมพิวเตอร์ ดดยใช้เสียงคำพูดเป็นสื่อ ทำให้เกิดความสะดวกในกลุ่มสังคมที่ใช้ประโยชน์จากคอมพิวเตอร์ซึ่งประยุกต์ใช้งานได้มากมาย

<u>ระบบโทรศัพท์</u>

สัญญาณที่เครื่องชุมสายโทรศัพท์แจ้งสภาวะต่างในจะประกอบไปด้วย

1. สัญญาณให้หมุน (Dial tone) ใช้เพื่อแสดงให้ผู้เรียกหมุนหมายเลขผู้รับ เป็นสัญญาณเสียงต่อเนื่อง 400 Hz

2. สัญญาณไม่ว่าง (Busy tone) ใช้เพื่อเตือนผู้เรียกว่าผู้รับกำลังใช้งานอยู่ เป็นสัญญาณ 400 Hz 60 ครั้งต่อนาทีดัง 0.5 วินาที เงียบ 0.5 วินาที

3. สัญญาณกริ่งเรียก (Ringing tone Signal) เป็นสัญญาณที่ทางชุมสายโทรศัพท์ส่งมาเพื่อแสดงว่าได้ทำการต่อระบบการสื่อสารระหว่างผู้เรียกกับผู้รับ เป็นสัญญาณ 16 Hz ผสมสัญญาณ 400 Hz แบบ AM ส่ง 0.67-1.5 วินาที เงียบ 2-4 วินาที

4. สัญญาณเรียกกลับ (Ring back tone) ใช้เมื่อผู้เรียก เรียกมายังผู้รับเครื่องชุมสายโทรศัพท์ ดำเนินการต่อสำเร็จ แจ้งให้ผู้เรียกรู้ว่าการต่อสำเร็จ เป็นสัญญาณ 16 Hz ผสม 600 Hz แบบ AM ช่วงเวลาการส่งเช่นเดียวกับสัญญาณกริ่งเรียก

<u>สัญญาณในการติดต่อกันระหว่างเครื่องส่งและเครื่องรับโทรศัพท์</u>

เครื่องส่ง

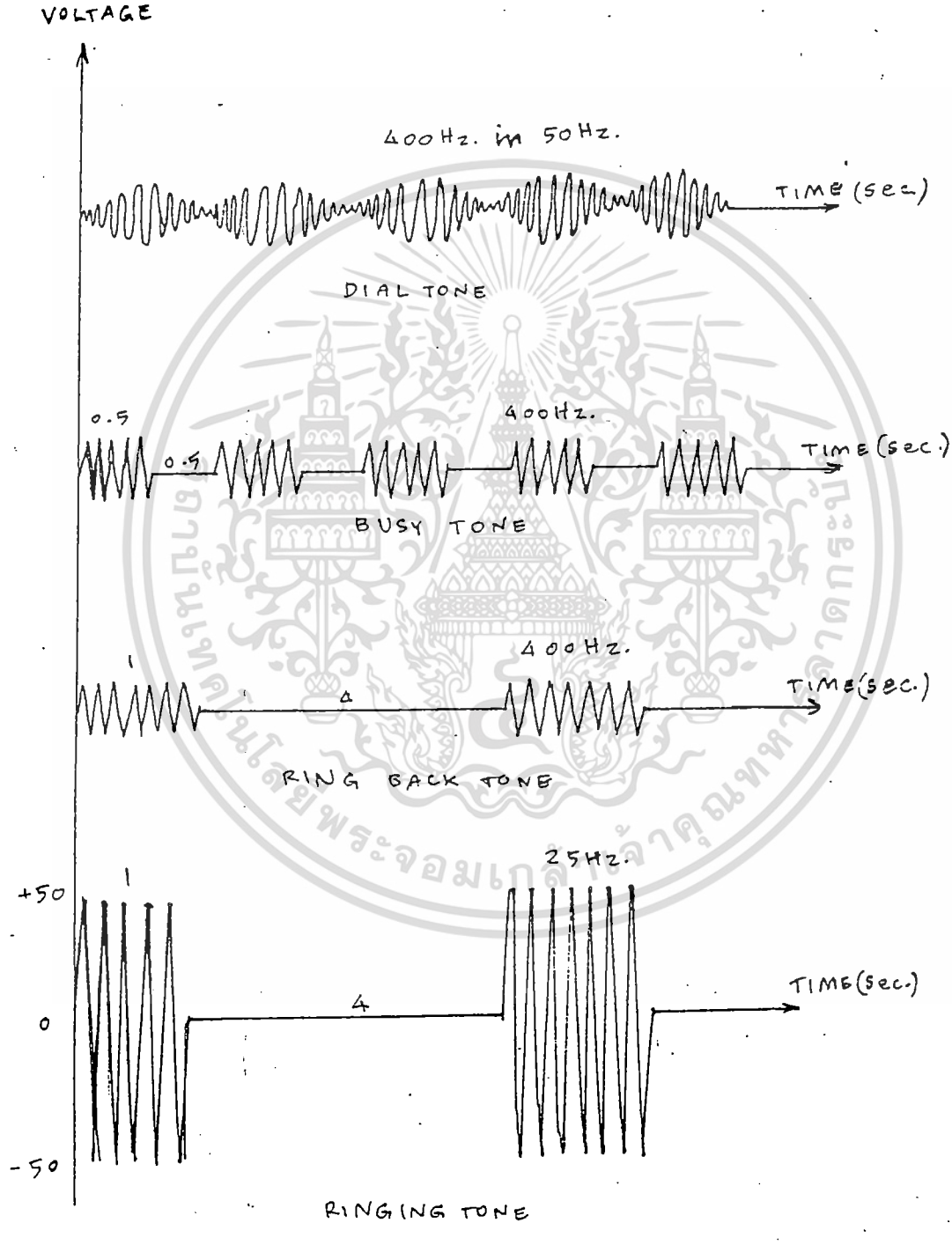
- ขณะที่ไม่ได้มีการยกหูโทรศัพท์ จะมีศักดาตกคร่อมสายโทรศัพท์เป็นสัญญาณกระแสตรง 48 โวลท์
- เมื่อผู้เรียกยกหูโทรศัพท์ ศักดาจะลดลงเหลือ 8 โวลท์ พร้อมทั้งมีสัญญาณให้หมุน ซึ่งเป็นสัญญาณกระแสสลับขนาด 250 มิลลิโวลท์ ความถี่ 400 Hz ผสมความถี่ประมาณ 50 Hz ซึ่งเมื่อกดรหัสสัญญาณความถี่แล้วสัญญาณให้หมุนจะหายไป
- กดรหัส (code) เบอร์โทรศัพท์ทั้งหมด 7 หลัก รหัสความถี่ที่ส่งจะเป็นสัญญาณผสมสองความถี่ เป็นความถี่สูงและต่ำผสมกัน แต่ละหมายเลขจะมี DTMF อยู่หนึ่งคู่
- ขณะที่รอการรับสาย จะมีสัญญาณตอบกลับ 2 แบบ เพื่อจะบอกว่าสายว่างหรือไม่ คือ สัญญาณเรียกกลับหรือสัญญาณสายไม่ว่างตามลำดับ
- เมื่อมีการรับสายแล้ว สัญญาณจะอยู่ที่ 8 โวลท์ โดยมีการกระเพื่อมตามลัษณะความถี่เสียง,ความดัง ของเสียงพูดตามสาย
- เมื่อวางหูโทรศัพท์เลิกการติดต่อ ขนาดศักดาจะกลับไปที่ 48 โวลท์ดังเดิม

เครื่องรับ

- ขณะที่ว่างหูจะมีศักดากระแตรงคร่อมสายอยู่ 48 โวลท์
- เมื่อสัญญาณกริ่งเรียกจะมีขนาดประมาณ 100 โวลท์ จังหวะ 1 วินาที หยุด 4 วินาที ซึ่งจะตรงกับสัญญาณเรียกกลับที่เครื่องส่ง

- จากนั้นเมื่อรับยกหูออกจากโทรศัพท์ ขนาดศักดา กระแสตรงจะเหลือ 8 โวท์ และมีการกระเพื่อมจากขนาดและความถี่ของเสียงพูด

- เมื่อว่างหูโทรศัพท์ขนาดศักดาจะกลับไปที่ 48 โวลท์ตามเดิม



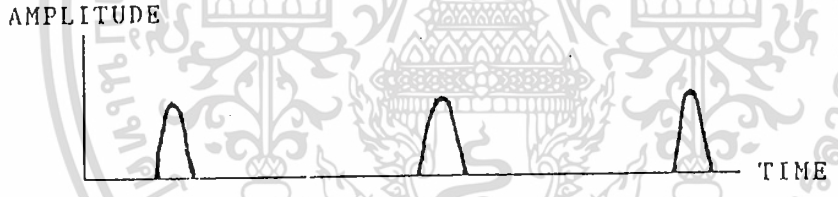
รูปที่ 1 สัญญาณพื้นฐานของระบบโทรศัพท์

บทที่ 2

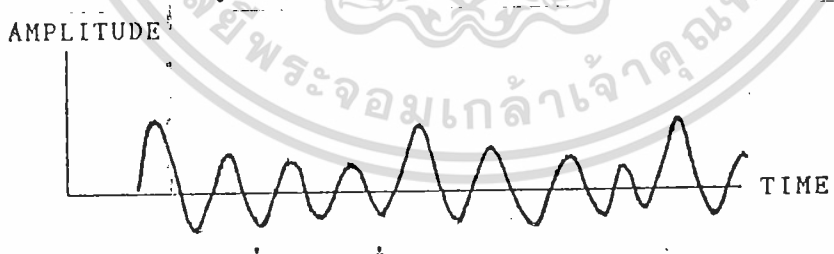
ระบบการสร้างเสียงและการได้ยินเสียงของมนุษย์

<u>ระบบกำเนิดเสียงพูดของมนุษย์</u>

เสียงเกิดจากการอัดหรือขยายตัวของอากาศ การสร้างเสียงของมนุษย์ ริ่มแรกมีลมจากปอดผ่านท่อลมมาสู่กล่องเสียง (GLOTTIS) ซึ่งมีลักาณะเป็นเส้นขึงอยู่ เมื่อพูดกล้ามเนื้อจะดึงเอ็นนี้ให้ตึงซึ่งมีผลให้เกิดการสั่นขึ้นเมื่อมีลมผ่านจะทำให้เกิด GLOTAL PULSE TRAIN ซึ่งมีความถี่ประมาณ 200-300 เฮิร์ตซ์ และมีความถี่เป็นความถี่มูลฐาน (FUNDAMENTAL FREQUENCY) แต่เสียงนี้จะไม่เป็นคำพูด จะต้องผ่านการกำธร (RESONANT) และ ฟิลเตอร์ (FILTER) ต่างๆ อีก สามารถออกเสียงโดยการเปิดปาก ให้ลิ้นและฟันอยู่สบายแล้วออก “ah” (อาห์) แต่เสียงที่ออกมานี้ไม่ใช้เสียงที่ออกจากกล่อง (GLOTTIS) โดยตรง เพราะได้ผ่านอวัยวะต่างๆ มาแล้วนั้นคือเสียงที่ผ่านการกำธร และ ฟิลเตอร์ มาแล้วทำให้เกิดฮาร์โมนิค (HARMONIC) ขึ้น ซึ่งปกติจะมี 2-4 ความถี่เรียกว่า FORMANT FREQUENCY ดังนั้นเสียงคนจึงมี ลักษณะเป็น PITCH (QUASI-PERIODIC) ซึ่งเป็นคาบซ้ำๆ ไปเรื่อยๆ ซึ่ง 1 PITCH ก็คือ 1 คาบของสัญญาณที่เกิดจากกล่องเสียง



รูปที่ 2 แสดง GLOTTAL PULSE



รูปที่ 3 แสดงเมื่อเกิด RESONANT

<u>ขั้นตอนการสร้างเสียง (Speech Production)</u>

ช่องทางเดินของเสียง (Vocal tract) จะเริ่มต้นจากช่องเปิดระหว่างเส้นเสียงใน กล่องเสียงไปสิ้นสุดที่ริมฝีปากส่วบประกอบของช่องทางเดินมีอยู่ 2 ส่วนใหญ่ๆ คือ ช่องคอ (Pharynx) และช่องปาก (Dral Cavity) ทั้งนี้เพราะว่าภายในช่องว่างภายในปากจะมีการเปลี่ยนแปลงขนาดตลอดเวลาที่เปลี่ยนเสียง และช่องจมูกก็ถูกควบคุมโดยเพดาอ่อน (Velum) ดังนั้น ถ้าช่องว่างต่างๆ เหล่า

นี้เปลี่ยนขนาดไป ก็จะทำให้ความถี่กำธรของเสียงเปลี่ยนไปด้วย โดยเฉลี่ยแล้วช่องทางเดินเสียงของผู้ชายจะมีความยาวประ 17 ซม. ส่วนผู้หญิงจะมีขนาดเล็กกว่าผู้ชายเล็กน้อย

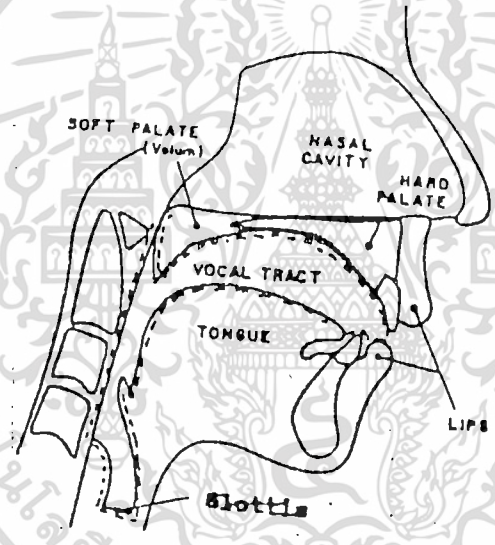
<u>อวัยวะที่ใช้ในการเปล่งเสียง (Articulation)</u>

อวัยวะที่ใช้ในการเปล่งเสียง แบ่งได้เป็น 3 พวกใหญ่ ๆ คือ

1. อวัยวะที่ใช้ในการสร้างลม คือ ส่วนที่ทำให้เกิดการเคลื่อนไหวของลม

2. อวัยวะส่วนที่เคลื่อนที่ได้ (Active Articulators) หมายถึงอวัยวะที่ติดกับกระดูกคางส่วนกลาง ได้แก่ ริมปากล่างและลิ้น

3. อวัยวะส่วนที่เคลื่อนที่ไม่ได้ (Passive Articulators) หมายถึงอวัยวะที่ติดกับกระดูกคางส่วนบนได้แก่ ริมฝีปากบน ฟันบน ปุ่มเหงือก เพดานอ่อน เพดานแข็ง ลิ้นไก่ และอวัยวะที่เป็นช่องว่างได้แก่ ช่องคอ ปาก ช่องจมูก



รูปที่ 4 ส่วประกอบของอวัยวะผลิตเสียง

<u>ลักษณะของเสียงพูด</u>

เสียงพูดสามารถจะแบ่งออกเป็นกลุ่มใหญ่ๆ ได้ 3 กลุ่ม ตามลักษณะการกระตุ้นของเสียงดังนี้

1. เสียงที่เกิดจากลำคอ (Voice Sound or Fricative) เป็นคำพูดที่เกิดจากลมผ่านกล่องเสียงซึ่งถูกปรับให้สั่นและสร้าง QUASI-PERIODIC PULSE ของอากาศออกมาและถูกดัดแปลง และ RESONANT โดยส่วนของ VOCAL TRACT (VOCAL TRACT คืออวัยวะที่ลมผ่านมาตั้งแต่กล่องเสียงจนถึงริมฝีปาก) เสียง /u/, /d/, /w/, /i/, /e/

2. เสียงเสียดสี (Unvoiced Sound or Fricative) เสียงพวกนี้ถูกสร้างโดยการรัดตัวของอวัยวะหรือบางจุดบริเวณ VOCAL TRACT และมีลมไหลผ่านช่วงที่รัดตัวนี้ด้วยความเร็วมากพอ

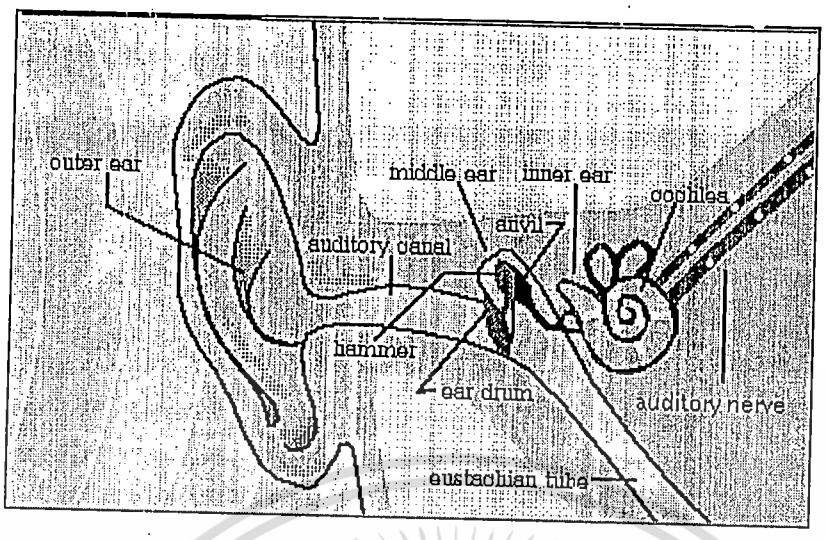
สมควร ผลคือทำให้มีการสร้าง BOARD-SPECTRUM NOISE ขึ้น เสียงนี้ใช้สัญญลักษณ์ ।ʃ। หรือ “sh”

3. เสียงอัด (Plosive Sound) เป็นผลมาจากการปิดช่องปากอย่างสนิท และสร้างความดันขึ้นข้างในด้านหลังส่วนที่ปิดนี้ แล้วเปิดมันอย่างรวดเร็ว เสียง PLOSIVE จะ ถูกสร้างขึ้น ใช้สัญญลัษณ์ ।ʧ ʃ।

<u>หูและการได้ยิน</u>

หู เป็นอวัยวะที่ใช้ในการรับเสียงทุกเสียง ประกอบด้วย 3 ส่วนได้แก่ หูส่วนนอก หูส่วนกลางและหูส่วนใน แรงดันของคลื่นเสียงที่เกิดขึ้นในธรรมชาติ จะเดินทางเข้ามาทางใบหู (Pinna) ช่องหู (Auditory) ซึ่งเป็นหูส่วนนอก ผ่านหูส่วนการที่มีโพรงอากาศประกอบด้วย Ear drum และกระดูก 3 ชิ้นได้แก่ Hammer ,Anvil และ Stirrup ทำหน้าที่ถ่ายทอดความสั่นสะเทือนสู่ของเหลวที่บรรจุอยู่ในอวัยวะรูปเปลือกหอยทาก (The snale-shaped cochlea) ซึ่งเป็นหูส่วนใน ของเหลวถ่ายทอดความสั่นสะเทือนเคลื่อนที่ไปตามความยาวของ Cochlea (The basilar membrane) ภายในเยื่อ Basilar มีเซลล์ขน (Hair cell) ที่มีคุณลักษณะการตอบสนองความถี่แตกต่างกันนับพันแผ่นอยู่ตั้งแต่ฐานถึงยอดของ Cochlea เมื่อเซลล์ขนเหล่านี้ถูกของเหลวทำให้ลักษณะเปลี่ยนไป จะเกิดเป็นสัญญาณอิมพัลส์ไฟฟ้า (Electrical impluses) กระตุ่นเป็นหลอดๆ ผ่านเส้นประสาทเสียง (Auditory nerve) เข้าสู่สมอง โดยเซลล์ขนจะมีคุณลักษณะการตอบสนองความถี่ต่ำที่ส่วนยอดของ Cochlea และส่วนฐานเซลล์จะมีคุณลักษณะการตอบสนองความถี่สูง

คุณลักษณะของเสียงที่มนุษย์สามารถแยกแยะได้ประกอบไปด้วย ปริมาณ (Volume), พิท (Pitch) และ โทน (Tone) ปริมาณความดังของเสียงขึ้นอยู่กับแอมปลิจูดหรือความเข้มของงเคลื่อนเสียง พิทเป็นความสัมพันธ์กับความถี่ของคลื่นเสียง หมายถึง จำนวนของคลื่นที่ผ่านจุดอ้างอิงต่อหนึ่งหน่วยเวลา เมื่อความถี่สูงขึ้นพิทก็จะมากขึ้น, โทนหรือ Quality ของเสียงมีคุณสมบัติที่ซับซ้อนมากกว่า วอลลุ่ม และ พิท Quality จะแปลไปตามจำนวนและชนิดของ Overtone หรือ ฮาร์โมนิคส์ (Combinations of frequencies) โดยทั่วไปมนุษย์สามารถได้ยินเสียงความถี่ตั้งแต่ 30 Hz-20KHz และช่วงแถบความถี่ของเสียงพูดที่ใช้ในระบบสื่อสาร ที่สื่อสารกันแล้วเข้าใจมีแบนวิทอยู่ในช่วง 300Hz-3.4KHz



รูปที่ 5 แสดงหูส่วนนอก,หูส่วนในและส่วนประกอบต่างของหู

<u>หลักการที่นำมาใช้ในการรู้จำเสียงพูด</u>

หลักการพื้นฐานอย่างหนึ่งที่ถือเป็นธรรมชาติของมนุษย์คือ การเปรียบเทียบซึ่งเป็นพฤติกรรมที่เป็นสัญชาติญาร การเปรียบเทียบเป็นการหาความแตกต่างระหว่างสองสิ่งอาจเป็นรูปธรรมหรือนามธรรมก็ได้ และสามารถเปรียบเทียบสิ่งที่เหมือนกันหรือไม่เหมือนกันก็ได้อาจเป็น รูปธรรมกับรูปธรรม,นามธรรมกับนามธรรม หรือณุปธรมมกับนามธรรม การเปรียบระหว่างของสองสิ่งจะต้องมีสิ่งหนึ่งเป็นบรรทัดฐานจึงจะบอกข้อแตกต่างของอีกสิ่งได้ ฉนั้น ผู้วิจัยจึงเลือกวิธีการบันทึกโครงสร้างทางความถี่ต่อเวลาของเสียงพูดจำนวนหนึ่ง แล้วนำมาเปรียบเทียบกับคำพูดของผู้ที่สั่งงานจริง (ใช้หลักการเปรียบเทียบลักษณะของข้อมูลเสียงทดสอบและข้อมูลอ้างอิง) ที่ใช้กันอย่างกว้างขวางในการวิจัยในสาขา การประมวลผลสัญญาณดิจิตอล หมวด Pattern recognition ก็คือวิธีการ Matching โดยยึดหลักการในแนวความคิดที่ง่าย แต่ได้ผลและหลีกเลี่ยงความซับซ้อน

งานวิจัยนี้ใช้การเปรียบเทียบข้อมูลด้วยสมการทางสถิติ โดยใช้สมการหาค่าสหสัมพันธ์ (Correlation) ระหว่างข้อมูลเสียงทดสอบกับข้อมูลเสียงอ้างอิง และทำไปตามขบวนการเพื่อให้ได้แพตเทอร์นที่แสดงระดับความสัมพันธ์ของเสียงทดสอบหนึ่งเสียงกับแบบทดสอบอ้างอิงของเสียงทั้งหมด ซึ่งก็คือคุณสมบัติเฉพาะตัวของเสียงนั้น จากนั้นจะใช้กรรมวิธีเลียบแบบโครงข่ายเซลล์สมองที่เรียกว่า โครงข่ายประสาทเทียม (Artificial Neural Network; ANN) โดยจะฝึกสอนเน็ทเวอร์คในลักษณะ Supervisor training เพื่อให้รู้จำ แพตเทอร์อ้างอิงในส่วนของ Training mode ตรวจสอบและตัดสินคุณลักษณะของเสียงกับแพตเทอร์นว่าเป็นเสียงใด โดยกำหนดให้รู้จำไม่เกิน 20 คำ สนับสนุนในงานควบคุมเสียงโดยกำหนดเสียงตัวอย่างที่ใช้คือ “ศูนย์”,”หนึ่ง”,”สอง”,”สาม”,”สี่”,”ห้า”,”หก”,”เจ็ด”,”แปด”,”เก้า”,”สิบ”,”เปิด”,”ปิด”,”หมุน”,”ยก”,

"วาง","ซ่อง","ซ้าย","ขวา","ไฟ" และใช้ขบวนการแปลงข้อมูลสัญญาณเสียงเชิงเวลา ไปเป็นสเปกโตรแกรมสามมิติในแกนของความถี่-เวลา-พลังงาน ขนาด Bandwidth 125Hz-4KHz ใช้สมการโควัรีเรชั่นในการทรานส์ฟอร์ม (Correlation transform) เพื่อลดขนาดข้อมูลและหาคุณลักษณะจำเพาะของเสียง จากนั้นจึงใช้นิวรัลเน็ตเวอร์ค (Neural Network) ช่วยในการวิเคราะห์คุณลักษณะจำเพาะของเสียงนั้นแล้วตัดสินใจว่าเป็นเสียงใด

บทที่ 3

การสร้างระบบการรู้จำเสียงพูดคำไทย

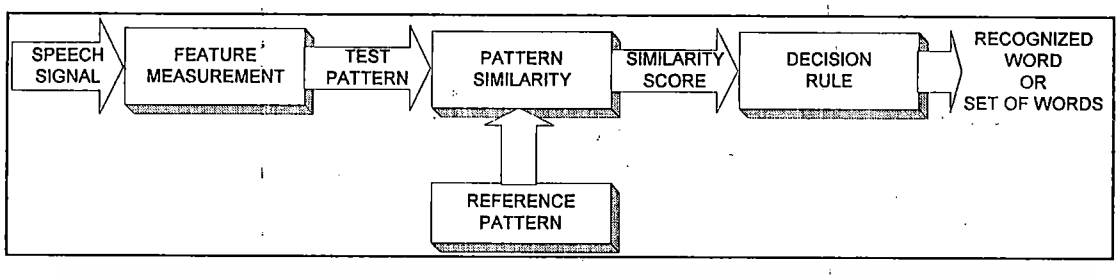
การวิเคราะห์เงื่อนไขการสร้างระบบการรู้จำที่ใช้ในการทดลอง

การสร้างระบบการรู้จำเสียงพูดคำไทยแบบไม่ขึ้นกับผู้พูดผู้วิจัยเริ่มต้นจากการศึกษาคุณลักษณะทั่วไป และเฉพาะตัวของเสียงคำพูด ระบบการได้ยิน การแปลความและวิเคราะห์ความหมาย ธรรมชาติการเปล่งเสียงของมนุษย์ ตลอดจนการศึกษาหาวิธีการวิเคราะห์เปรียบเทียบที่เหมาะสมเพื่อใช้ในระบบรู้จำที่จะสร้างขึ้น จากการศึกษาธรรมชาติการสื่อสารผ่านระบบเสียงและการรับฟังของมนุษย์ พอสรุปได้ว่าหากจะสร้างระบบรู้จำที่สามารถรู้จำเสียงแบบไม่ขึ้นกับผู้พูดได้ ระบบรู้จำจะต้องสามารถวิเคราะห์เปรียบเทียบข้อมูลเสียงได้โดยไม่ขึ้นกับความดัง ความถี่ และระยะเวลาของเสียงคำหนึ่งๆที่เปล่งออกมา ซึ่งต้องทำได้ดังนี้เป็นอย่างน้อย

ระบบการรู้จำที่สร้างขึ้นใช้ในการทดลอง

การรู้จำเสียงพูดนั้นจะเกี่ยวข้องกับการตัดสินใจระหว่าง แบบทดสอบ (Test pattern) กับแบบอ้างอิง (Reference pattern) หรือที่เรียกว่า Template ซึ่งจะเป็นการวิเคราะห์ถึงความสัมพันธ์ของตัวแปร (Parameter) ที่ใช้เป็นรูปแบบทั้งสอง เพื่อที่จะระบุว่าแบบทดสอบที่นำมาทดสอบนั้นมีความสัมพันธ์กับแบบอ้างอิงใดมากที่สุด จากนั้นจึงจะนำไปสู่การตัดสินใจต่อไป... (ระพีพัฒน์ เพ็ญศิริ 2538:16-18)

ภาพที่2



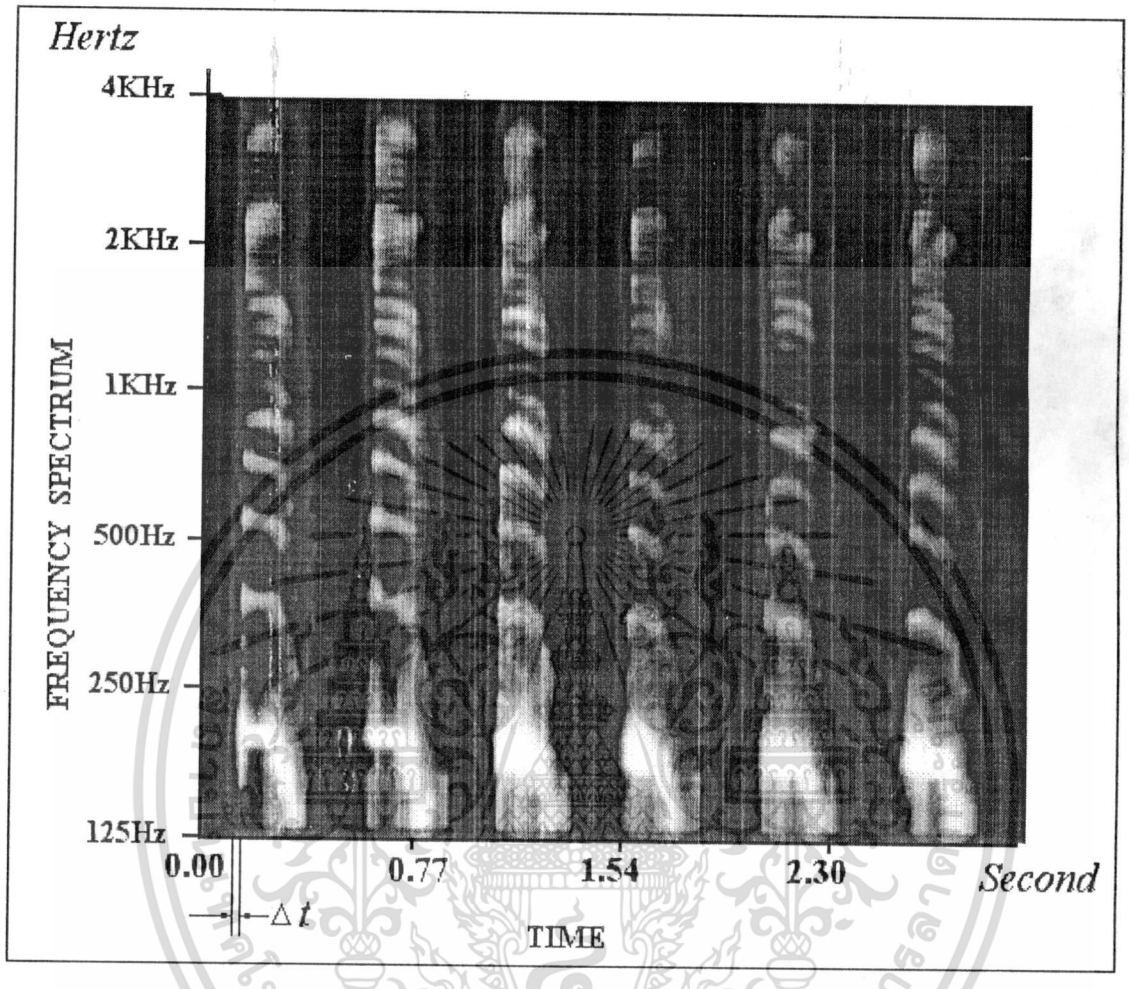
แสดงโครงสร้างระบบการรู้จำแบบ Isolated word recognition (Rabiner and Levinson,1981)

โครงสร้างระบบการรู้จำที่สร้างขึ้นเป็นแบบที่คล้ายกับ Isolated word recognition (Rabiner and Levinson, 1981) ดังภาพที่2 โดยในส่วนของ Feature measurement จะรวมเอา Pre-processing เข้าไปด้วยคือ A/D, FFT เพื่อสร้างเป็น Discrete ของสเปกตรัม ที่เป็นสเปกโตรแกรมของเสียงที่จะใช้ทดสอบ ส่วนของ Pattern similarity เป็นการเปรียบเทียบระหว่าง แพตเทอร์นทดสอบ กับ แพตเทอร์นอ้างอิง ซึ่งใช้วิธีการเปรียบเทียบด้วยสมการสหสัมพันธ์วิธีการนี้ทำให้ได้ผลลัพธ์เป็นคุณลักษณะจำเพาะตัวของเสียงทดสอบเทียบกับเสียงอ้างอิง และขนาดของข้อมูลลดลงอย่างมาก ซึ่งจะขอเรียกวิธีนี้ว่า โคร์รีเลชั่นทรานส์ฟอร์มและในส่วนของ Decision rule จะใช้การตัดสินใจด้วยนิวรัลเน็ตเวอร์คที่ผ่านการเทรนด้วยแพตเทอร์น อ้างอิงมาแล้วด้วยอัลกอลิทึ่มแบคพรอพาเกชั่น

การแก้ไขปัญหาการเปรียบเทียบเสียงที่มีความดัง,ระยะเวลาการเปล่งเสียงและความถี่ที่แตกต่างกัน

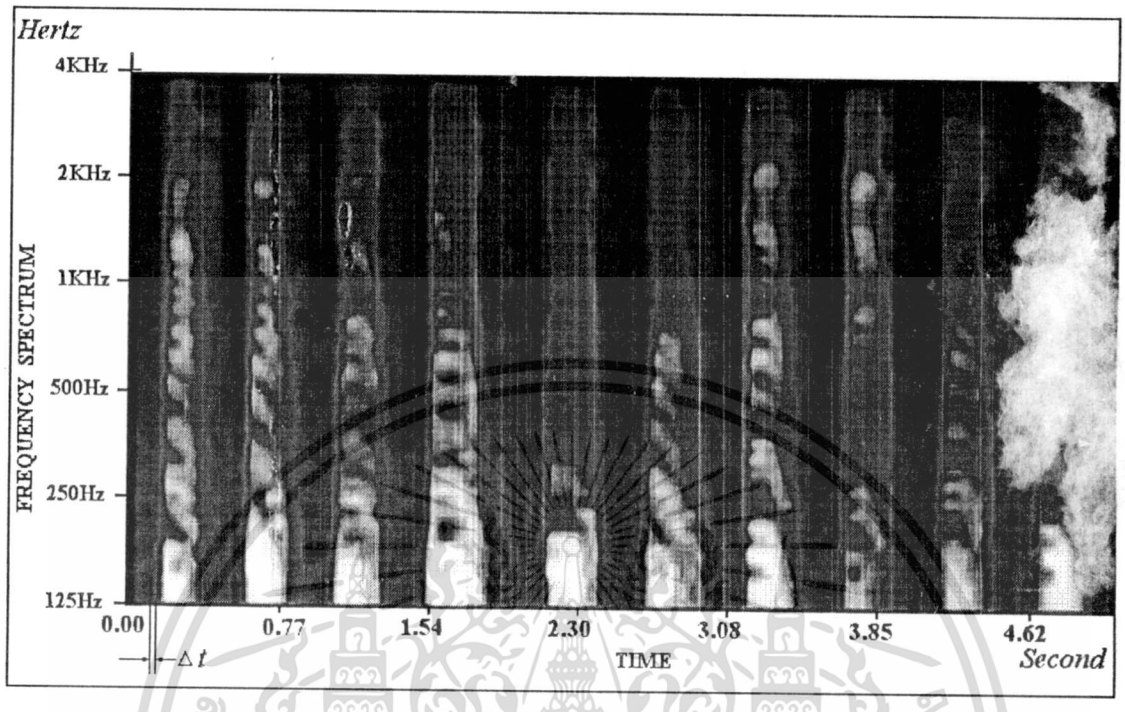
จากการศึกษาคุณลักษณะทางความถี่ของเสียงพูดตัวอย่างจากสเปกโตรแกรมพบว่า คำ คำเดียวกันที่พูดโดยผู้พูดคนละคน มีลักษณะของแถบความถี่แตกต่างกัน ทั้งนี้เนื่องจาก ลักษณะของช่องปาก, หลอดเสียง, และลักษณะการออกเสียง ของแต่ละคนมีความแตกต่างกัน รวมทั้งความดังของเสียงและระยะเวลาที่เปล่งที่แต่ละคนเปล่งออกมาก็ไม่เท่ากัน หรือแม้แต่คำคำเดียวกันที่ผู้พูดคนเดียวเปล่งออกมาแต่ละครั้ง ก็ไม่เหมือนกันเสียทีเดียวดังสังเกตุได้จากรูปที่3 และ 4

ภาพที่ 3



แสดงสเปกโตรแกรมของเสียง "เก้า" ที่เปล่งเสียงโดยผู้พูดคนเดียว 6 ครั้ง

ภาพที่ 4



แสดงสเปกโตรแกรมของเสียง "เก้า" ที่เปล่งเสียงโดยผู้พูด 10 คน

ในการวิเคราะห์ควรเลือกวิธีการเปรียบเทียบที่ทำได้โดยไม่ขึ้นกับตัวแปรทั้งสาม ได้แก่ ความดัง ความถี่และระยะเวลาของเสียง ซึ่งมีการเปลี่ยนแปลงอย่างอิสระต่อกัน งานวิจัยนี้ขอนำเสนอวิธีการแก้ปัญหาดังกล่าวดังต่อไปนี้

3.1 **ความดังของเสียง** มีวิธีการเปรียบเทียบทางสถิติวิธีหนึ่งที่น่าสนใจคือ วิธีการเปรียบเทียบแบบสหสัมพันธ์ (Correlation) ซึ่งเป็นการเปรียบเทียบลักษณะความสัมพันธ์ของการเปลี่ยนแปลงของข้อมูล โดยไม่คำนึงถึงระดับการเปลี่ยนแปลงของพลังงานเสียง คุณสมบัตินี้ตรงกับความต้องการเพราะสอดคล้องกับความเป็นจริง คือ ธรรมชาติการเปล่งเสียงของแต่ละคน มีระดับพลังงาน (ความดัง) ไม่เท่ากัน ดังนั้น ถ้าจะเปรียบเทียบสเปกตรัมของเสียงสองเสียง ซึ่งมีความดังแตกต่างกันด้วยวิธีปกติ จะต้องปรับเสียงทั้งสองให้มีสเกลความดังอยู่ในระดับเดียวกันเสียก่อน(แบบ Max-Min Scale) จึงทรานส์ฟอร์มไปเป็นสเปกตรัมของเสียงแล้วเปรียบเทียบ แต่ถ้าเปรียบเทียบความสัมพันธ์โดยการหาค่าสหสัมพันธ์ของข้อมูลแล้ว จะสามารถข้ามขั้นตอนดังกล่าวไปได้โดยยังคงให้ผลของการเปรียบเทียบเป็นเช่นเดิม ซึ่งงานวิจัยชิ้นนี้เลือกวิธีการแบบสหสัมพันธ์ ดังนั้นในส่วนของการหาคุณลักษณะเฉพาะตัวของเสียงจะใช้วิธีการเปรียบเทียบแบบสหสัมพันธ์โดยวิธีการสามารถทำได้โดยไม่คำนึงถึงระดับความดังของข้อมูลเสียง ซึ่งเป็นคุณสมบัติเฉพาะของวิธีการ

3.2 **ระยะเวลาที่เปล่งเสียงแต่ละพยางค์** การแก้ปัญหาที่เกิดจากระยะเวลาการเปล่งเสียงระหว่างเสียงทดสอบและเสียงอ้างอิงไม่เท่ากันนั้น อาศัยประโยชน์จากวิธีการเปรียบเทียบแบบสหสัมพันธ์ระหว่างแพตเทอร์นของสเปกตรัมเสียงทดสอบกับเสียงอ้างอิง โดยการเปรียบเทียบทำกันแบบเฟรมต่อเฟรม (ระยะเวลาการเปล่งเสียงแต่ละพยางค์ไม่เท่ากัน จำนวนเฟรมจึงไม่เท่ากัน) และหยุดเมื่อการเปรียบเทียบทำไปถึงเฟรมสุดท้ายของเสียงที่มีระยะเวลาการเปล่งเสียงที่น้อยที่สุด จากนั้นจึงหาค่าเฉลี่ยสหสัมพันธ์ของแต่ละเฟรมของเสียงทั้งสอง เหตุผลที่กำหนดจุดสิ้นสุดของการเปรียบเทียบดังกล่าวคือ

3.2.1 การเปล่งเสียงของเสียงเดียวกันแต่ละครั้งมีลักษณะคล้ายคลึงกัน และระยะเวลาการเปล่งเสียงใกล้เคียงกัน หากการเปรียบเทียบเฟรมขาดไปบ้างบางส่วน จะส่งผลกระทบน้อย เพราะผลลัพธ์ในแต่ละเฟรมยังต้องนำไปหาค่าเฉลี่ยสหสัมพันธ์รวมของเสียงที่เปรียบเทียบอีกครั้ง สำหรับกรณีเสียงทดสอบและเสียงอ้างอิงเป็นเสียงเดียวกัน แต่ระยะเวลาการเปล่งเสียงแตกต่างกันพอสมควร จะส่งผลกระทบมากขึ้น แต่ลักษณะความคล้ายคลึงกันยังคงมีอยู่จึงไม่กระทบกระเทือนต่อค่าเฉลี่ย ทำให้ผลการเปรียบเทียบยังพอเชื่อถือได้

3.2.2 กรณีเสียงทดสอบและเสียงอ้างอิงเป็นคนละเสียง แต่มีระยะเวลาการเปล่งเสียงใกล้เคียงกัน ผลการเปรียบเทียบก็จะชี้ชัดว่าระดับความสัมพันธ์ของข้อมูลต่ำเนื่องจากข้อมูลขาดความคล้ายคลึงกันตั้งแต่ต้น และหากระยะเวลาการเปล่งเสียงแตกต่างกันก็จะยิ่งส่งผลให้ระดับความสัมพันธ์ของข้อมูลต่ำลงไปอีก

3.3 **ความถี่** เสียงคำๆเดียวกันที่เปล่งโดยแต่ละคนมีความถี่มูลฐานไม่เท่ากัน การปรับความถี่เสียงใด ๆ ให้มีความถี่อยู่ในระดับเดียวกัน แล้วดำเนินการเปรียบเทียบ เป็นการบิดเบือนข้อมูลไปจากความเป็นจริง จึงหลีกเลี่ยง โดยอาศัยการนำแพตเทอร์นทดสอบไปเปรียบเทียบกับแพตเทอร์นอ้างอิงจำนวนมาก ซึ่งมีโอกาสพบกับแพตเทอร์นอ้างอิงที่ใกล้เคียงบ้าง แต่ยังคงมีการแบ่งกลุ่มชาย-หญิง เพื่อความสะดวกในการทดลองจะเลือกทำเฉพาะเสียงผู้ชายไทยทั่วไปก่อน ส่วนการดำเนินการกับข้อมูลกลุ่มเสียงอื่นจะแยกกัน แต่ยังคงใช้วิธีการเดียวกัน

<u>การนำนิวรัลเน็ตเวอร์คมาใช้ในระบบการรู้จำเสียงพูดคำไทย</u>

อัลกอลิที่มการเทรนนิ่ง แบบแบคพรอพาเกชั่นของนิวรัลเน็ตเวอร์ค (Backpropagation neural network;BPNN) เป็นการเรียนรู้แบบ Supervisor training ซึ่งต้องมีเป้าหมายที่ต้องการ เช่นเดียวกับการเรียนรู้ภาษาของมนุษย์ ต้องมีภาษาใดภาษาหนึ่งเป็นแม่แบบ และหากต้องการรู้ภาษาอื่นก็ต้องเริ่มเรียนรู้กันใหม่ เพราะภาษาไม่ใช้สัญชาติญาณ ด้วยเหตุนี้จึงเลือกนำเอาอัลกอลิที่มนี้มาใช้เฉพาะส่วนการตัดสิน คุณลักษณะจำเพาะของเสียงที่เป็นอินพุทว่าน่าจะเป็นเสียงใดเท่านั้น การจะนำ BPNN มาใช้รู้จำโดยตรงทั้งระบบ ยังติด

ปัญหาอยู่ที่ ความแปรปรวนของขนาดของแพตเทอร์นข้อมูลเสียง และขนาดของเน็ตเวอร์คก็มีขนาดใหญ่เกินไปจึงจำเป็นต้องปรับข้อมูลให้เหมาะสมเสียก่อนเพื่อให้เน็ตเวอร์คทำงานได้อย่างมีประสิทธิ์ภาพ

การสร้างแพตเทอร์นอ้างอิงจากสเปกโตรแกรมเสียงกลุ่มอ้างอิง

ข้อมูลเสียงที่ใช้ในการทดลองจะถูกแบ่งเป็น 2 ส่วนคือ ส่วนที่นำไปใช้เป็นข้อมูลอ้างอิง และส่วนที่นำไปใช้เป็นข้อมูลทดสอบ ข้อมูลเสียงที่ได้จากการ Quantizing และ Coding จากสัญญาณ อนาลอก ทุกชุดจะถูกทรานส์ฟอร์มไปเป็นสเปกโตรแกรม ความถี่-เวลา-พลังงาน (ธันวา ศรีประโมง 2537: 13-21) ได้เป็นกลุ่มสเปกโตรแกรมเสียงอ้างอิงและสเปกโตรแกรมเสียงทดสอบ ในการสร้างแพตเทอร์นอ้างอิงจะนำสเปกโตรแกรมเสียงกลุ่มอ้างอิงกลุ่มเดียวกันมาเทียบหาความสัมพันธ์ของข้อมูลด้วยกันเอง แบบเสียงหนึ่งเสียงนำไปเทียบกับเสียงทั้งหมด โดยใช้สมการสหสัมพันธ์(Correlative equation) จะได้ผลลัพธ์เป็นสัมประสิทธิ์สหสัมพันธ์ ของเสียงนั้นกับเสียงทั้งหมด จำนวนเท่ากับเสียงอ้างอิงที่นำไปเทียบ จากนั้นจะจัดเรียงสัมประสิทธิ์สหสัมพันธ์เสียใหม่ แบบมากไปน้อยโดยมีฟิลด์ชื่อประจำเสียงนั้นๆ เป็นตัวแปรตามเมื่อจัดเรียงเสร็จจะเลือกสัมประสิทธิ์สหสัมพันธ์ ที่มีค่ามากสุดออกมาประมาณ 5เปอร์เซนต์ ของจำนวนเสียงทั้งหมด แล้วนับคะแนนความถี่ในฟิลด์ชื่อประจำเสียงแต่ละชื่อที่ซ้ำกัน จะได้เป็นแพตเทอร์นอ้างอิง (หรือคุณสมบัติเฉพาะตัวของเสียงนั้น)ที่เทียบกับเสียงอ้างอิงทั้งหมด จำนวนของแพตเทอร์นอ้างอิงที่ได้จึงมีจำนวนเท่ากับจำนวนสเปกโตรแกรมเสียงอ้างอิงนั่นเอง

สหสัมพันธ์ (Correlation)

ค่าระดับความสัมพันธ์ของข้อมูล 2 ชุด เรียกว่า สัมประสิทธ์ของสหสัมพันธ์ (Correlative coefficient ;r) สามารถหาได้จากเทอมของ Z-Scores ดังนี้

$$r_{xy} = \sum_{i=1}^{n} \frac{Zx_i Zy_i}{n-1} \quad \text{.....(3.1)}$$

เมื่อ r_{xy} คือ สัมประสิทธ์ของสหสัมพันธ์ของข้อมูล X_i และ Y_i

Xi คือ ข้อมูลชุดที่1 (ทดสอบ)

Yi คือ ข้อมูลชุดที่2 (อ้างอิง)

n คือ จำนวนข้อมูลที่เป็นสมาชิกของ Xi ,Yi (โดยทั้ง 2 ชุดมีจำนวนเท่ากันคือ 128)

Zx_i และ Zy_i คือค่า Z-Scores เป็นระยะห่างระหว่างค่าข้อมูลกับค่า Mean ของชุดคำสั่งในเทอมของ Standard deviation Sx_i

โดย $$Zx_i = \frac{Xi - \overline{X}}{Sx} \quad \text{.....(3.2)}$$

และ $$Zy_i = \frac{Yi - \overline{Y}}{Sy} \quad \text{.....(3.3)}$$

เมื่อ $$Sx = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \overline{X})^2}{n-1}} \quad \text{.....(3.4)}$$

และ $$Sy = \sqrt{\frac{\sum_{i=1}^{n}(Yi - \overline{Y})^2}{n-1}} \quad \text{.....(3.5)}$$

จากสมการ จะได้ว่า ค่า rxy จะมีค่าอยู่ในระหว่าง -1 ถึง +1 พอตีความหมายได้ดังนี้คือ ถ้าค่า $|rxy| = 1$ หมายถึง ข้อมูลทั้ง 2 ชุด มีความสัมพันธ์กันทุกจุดข้อมูล ซึ่งในทางปฏิบัติแล้วมีโอกาสน้อยมาก

<u>เงื่อนไขการวิเคราะห์ระดับสัมประสิทธ์ของสหสัมพันธ์</u>

ผลการคำนวนหาระดับสัมประสิทธ์สหสัมพันธ์ได้ดังนี้ของข้อมูล2 ชุดพอวิเคราะห์ระดับของค่าระดับสัมประสิทธ์สหสัมพันธ์ได้ดังนี้

1.หากค่าสัมประสิทธ์สหสัมพันธ์เข้าใกล้ 1 อยู่ในช่วง 0.70 ถึง 0.90 ถือว่าค่าสหสัมพันธ์อยู่ในระดับสูง

2.หากค่าสัมประสิทธ์สหสัมพันธ์เข้าใกล้ 0.5 อยู่ในช่วง 0.30 ถึง 0.70 ถือว่าค่าสหสัมพันธ์อยู่ในระดับสูง

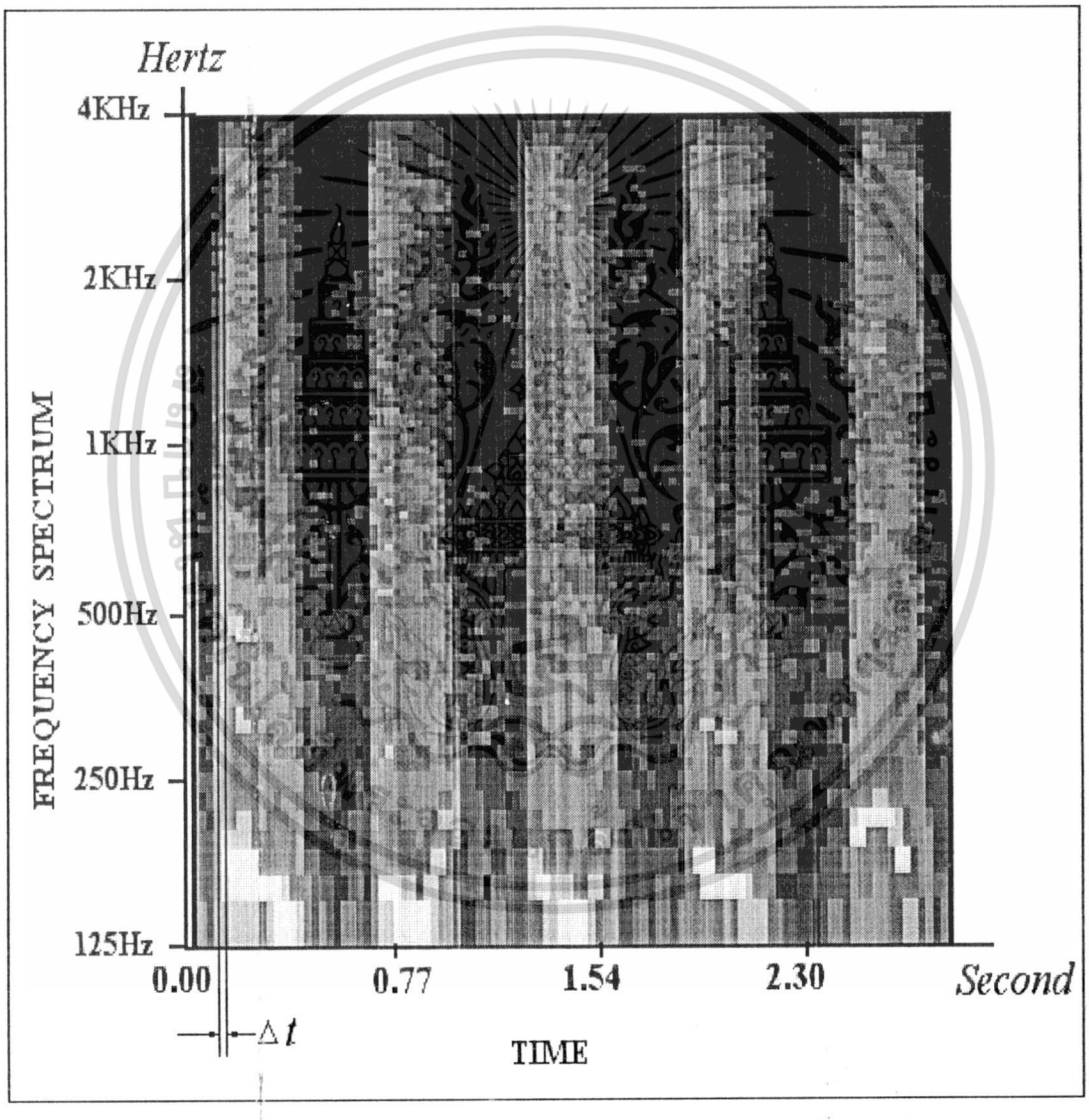
3.หากค่าสัมประสิทธ์สหสัมพันธ์เข้าใกล้ 0.0 อยู่ในช่วงต่ำกว่า 0.30 ถือว่าข้อมูลไม่มีความสัมพันธ์กันเลย

4.หากค่าสัมประสิทธ์สหสัมพันธ์มีค่าเป็นลบหมายถึงความสัมพันธ์ของข้อมูลทั้งสองชุด มีลักษณะตรงกันข้าม ถ้าเข้าใกล้ -1 แสดงว่า มีระดับความสัมพันธ์แบบตรงข้ามกัน ในระดับสูง

การเทียบหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างสเปกโตรแกรมเสียงสองชุด

สเปกโตรแกรมเสียง หมายถึง การนำสเปกตรัมของเสียงที่ได้จากการทรานส์ฟอร์มสัญญาณเสียงในเชิงเวลา ที่เวลา Δt ใด ๆ มาเรียงต่อกัน ในกรณีนี้เราจะสนใจ สเปกโตรแกรมเสียงนับตั้งแต่เริ่มเปล่งเสียงจนสิ้นสุดการเปล่งเสียง ซึ่งเสียงที่ใช้ทดสอบกำหนดเป็นเสียงคำไทยแบบ 1 คำ 1 พยางค์

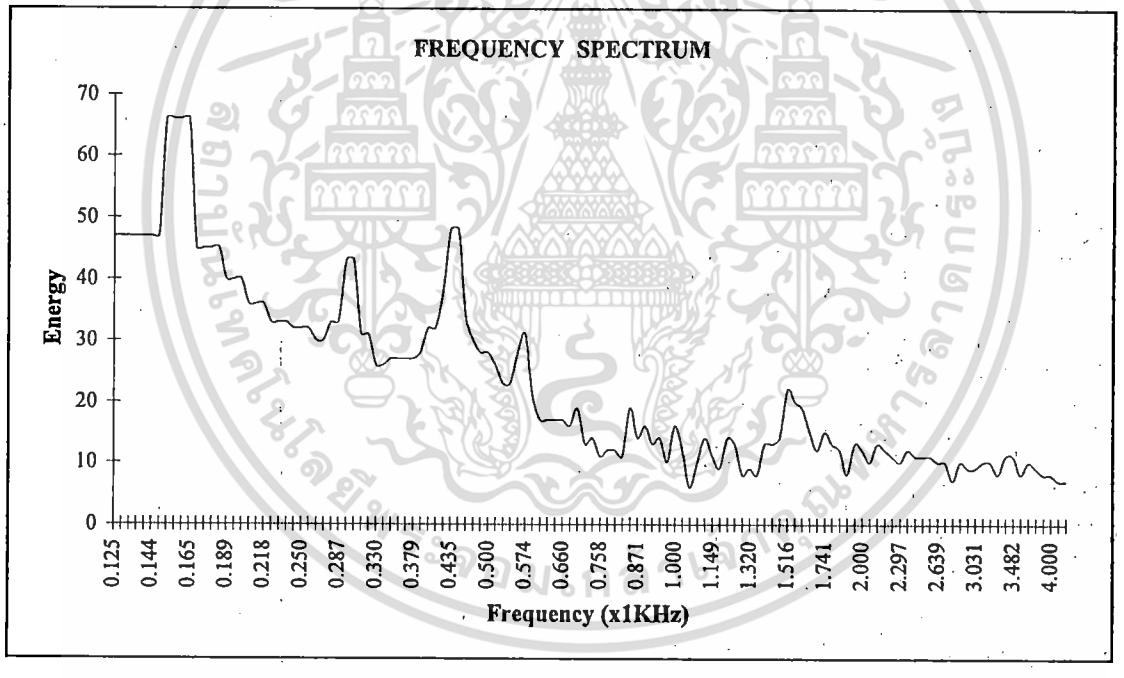
ภาพที่ 5



แสดงสเปกโตรแกรมของชุดข้อมูลเสียง "หนึ่ง" ถึง "ห้า" ที่เปล่งจากคนเดียวจากซ้ายไปขวาตามลำดับ

พิจารณาสเปกโตรแกรม 3 มิติ จากภาพที่ 5 พบว่า ชุดข้อมูลเสียงใด ๆ ประกอบด้วย Discrete ของระดับพลังงาน ของสเปกตรัมความถี่เสียง (แสดงเป็นแถบสี) มาเรียงต่อกัน โดยแต่ละ Discrete เป็นค่าเฉลี่ยในช่วงเวลา Δt แกนทางแนวนอน คือแกนเวลา โดยแสดงตั้งแต่เวลาที่ 0 ถึง 3.07 วินาที แต่ละ 1 ช่วงเวลา Δt = n แกนแนวตั้งเป็นแกนของแถบความถี่ ตั้งแต่ 125 Hz ไปจนถึง 4 KHz จำนวน 128 ค่า, จัดเรียงสเกลแถบความถี่แบบ Logalithm ส่วนระดับพลังงาน แสดงโดยระดับความเข้มของเฉดสีโดยเรียงเหลือง-ส้ม-แดง-น้ำตาล-เขียวขี้ม้า-เขียวออน่-เขียวแก่-น้ำเงินไปจนถึง ดำ จากระดับพลังงานสูงไปหาต่ำตามลำดับ จากภาพที่ 5 แสดงความเข้ม 16 ระดับ (โดยแปลงจากค่าที่คำนวณได้ผ่าน Fitting scale ที่ 0-16ระดับ แล้วนำมาแสดงผลโดยย้อมสีแทนระดับพลังงาน) เมื่อเปรียบเทียบระยะเวลาการเปล่งเสียงของชุดข้อมูลเสียงแต่ละชุด จะพบว่ามีเวลาการเปล่งเสียง แตกต่างกันไปบ้างตามแต่ผู้พูด ในช่วงเวลาส่วนย่อย (Discrete) อาจเขียนเป็นกราฟของแถบความถี่ได้ภาพที่ 6 ดังนี้

ภาพที่ 6



แสดงสเปกตรัมของแถบความถี่(เฟรม) ของเสียงพูด "หนึ่ง" ช่วงเวลาΔt ที่ 6

การเทียบชุดข้อมูลเสียงทั้งสองจะทำทีละเฟรม (หรือ Discrete หรือ Δt) แบบเฟรมต่อเฟรม โดยเฟรมที่1 ของสเปกโตรแกรมเสียงชุดที1 (ทดสอบ) ก็เปรียบเทียบกับเฟรมที่1 ของสเปกโตรแกรมเสียงชุดที่2 (อ้างอิง) ดังสมการ(3.6)

$$rxyp = \sum_{k=1}^{K} \frac{Zx_{ik}Zy_{jk}}{K} \quad \text{.....(3.6)}$$

เมื่อ $rxyp$ คือค่าสัมประสิทธิ์สหสัมพันธ์ของสเปกตรัมที่เป็นสมาชิกของ P

K คือจำนวนข้อมูลในแกนความถี่ (สำหรับการทดลองมี128ข้อมูล)

Zx_i , Zy_i คือ ค่า Z-Scores ได้จากสมการที่ (3.2)และ(3.3)

i , j คือ เฟรมที่เป็นสมาชิกของเสียงพยางค์ที่กำลังพิจารณาหาสัมประสิทธ์สหสัมพันธ์ กำหนดให้ i และ j มีค่าตั้งแต่ 1 ถึง P

และไปจนสิ้นสุดการพิจารณาที่เฟรมสุดท้ายของข้อมูลเสียงใดที่ถึงก่อน ดังสมการที่ (3.7)

$$P = I \text{ เมื่อ } J \geq I$$

$$P = J \quad \text{กรณีอื่น} \quad \text{.....(3.7)}$$

เมื่อ P คือ จำนวนเฟรมของข้อมูลเสียงพยางค์ 2ชุดที่นำมาเปรียบเทียบได้จากสมการที่ 3.7

I คือ จำนวนเฟรมของข้อมูลเสียงพยางค์ชุดที่หนึ่ง (ข้อมูลทดสอบ)

J คือ จำนวนเฟรมของข้อมูลเสียงพยางค์ชุดที่สอง (ข้อมูลอ้างอิง)

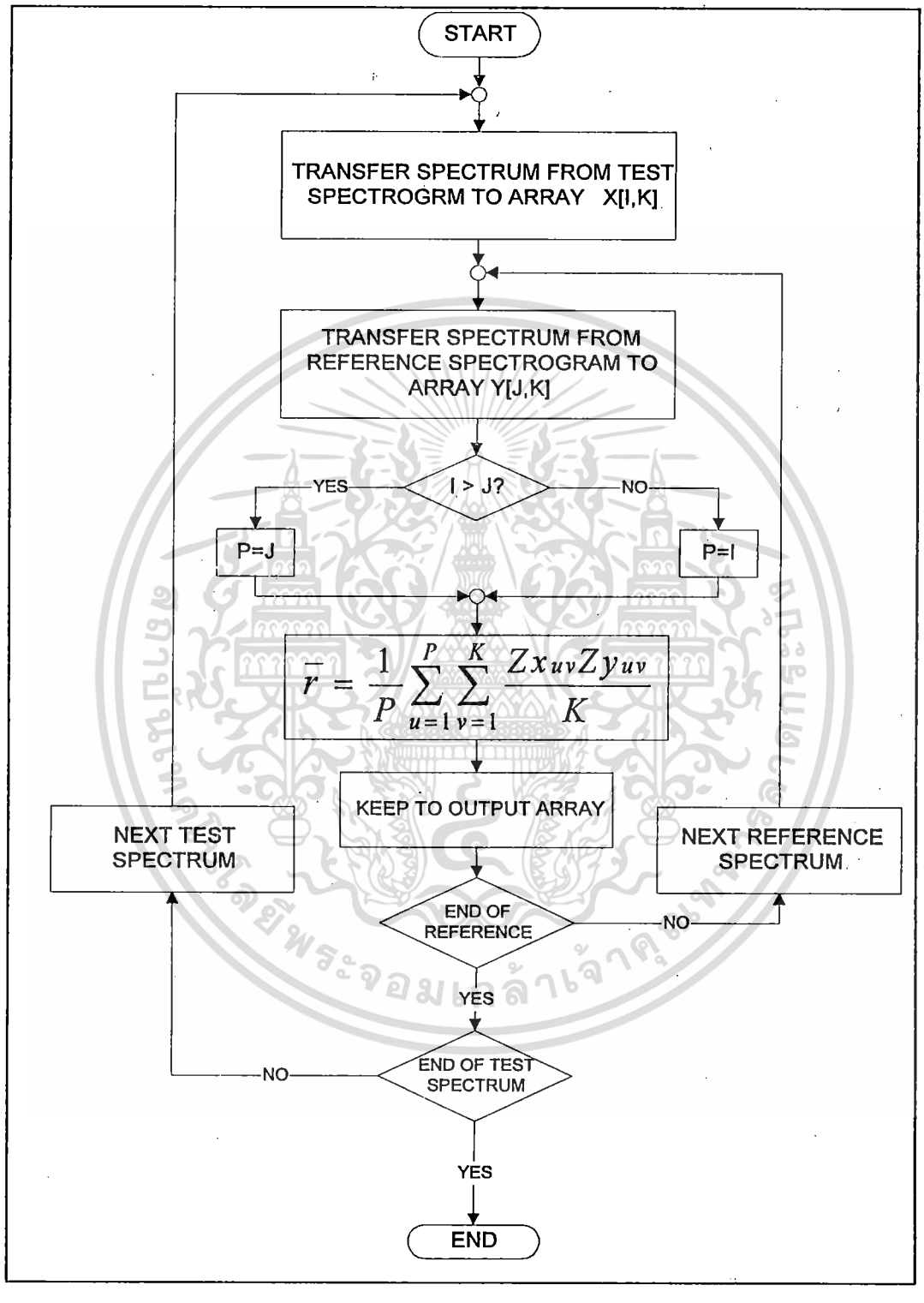
สาเหตุที่กำหนดเช่นนี้ มาจากข้อสังเกตุที่ว่า หากข้อมูลมีช่วงเวลาการเปล่งเสียงยาวนานออกไปมาก โอกาสที่จะเป็นเสียงเดียวกันก็น้อยลง หรือกรณีเป็นเสียงเดียวกันแต่มีช่วงเวลาการเปล่งเสียงยาวออกไปไม่มากนัก(1-2เฟรม) ก็ไม่จำเป็นต้องนำส่วนที่เกินมาเปรียบเทียบความสัมพันธ์เพราะเมื่อหาค่าเฉลี่ย(Mean)ของระดับสัมประสิทธิ์สหสัมพันธ์ ตามสมการที่(3.8)

$$R = \bar{r} = \frac{1}{P}\sum_{p=1}^{P} rxyp \quad \text{.....(3.8)}$$

เมื่อ R หรือ $\bar{r}$ คือค่า Mean ที่บอกถึงระดับความสัมพันธ์ของข้อมูลเสียงสองชุด

จะถูกหารด้วยจำนวน Discrete Pหรือ Δt ซึ่งหากเป็นเสียงเดียวกันแล้วค่าเฉลี่ยผลรวมที่ได้จะไม่ผิดพลาดมาก การหาค่าระดับของความสัมพันธ์ของข้อมูล แสดงเป็นโฟล์วชาร์ทได้ดังภาพที่7 ค่าระดับความสัมพันธ์ที่หาได้จะบ่งบอกถึงคุณสมบัติเฉพาะของเสียงนั้นโดยเทียบกับเสียงอ้างอิงทั้งหมด

ภาพที่ 7



แสดงผังงานการหาระดับความสัมพันธ์ระหว่าง สเปกโตรแกรมเสียงทดสอบ กับ เสียงอ้างอิง

โฟว์ชาร์ทดังภาพที่7 จะแสดงการหาจุดเริ่มต้นและจุดสิ้นสุดของชุดข้อมูลเสียงทั้งสอง โดยใช้สมการการหาค่าโคร์รีเลชั่นเฉลี่ย ของเสียงทดสอบและเสียงอ้างอิง จากสมการที่ 3.7และ 3.8 จะได้

$$\bar{r} = \frac{1}{P}\sum_{u=1}^{P}\sum_{v=1}^{K}\frac{Zx_{uv}Zy_{uv}}{K} \quad \text{.....(3.9)}$$

โดยกำหนด $\bar{r}$ คือ ค่า Mean ของโคร์รีเลชั่นเฉลี่ย ของเสียงทดสอบและเสียงอ้างอิง

K คือจำนวนของข้อมูลในแกนความถี่(ในการทดลองมีจำนวน 128ข้อมูลต่อ สเปกตรัม)

P คือจำนวนสเปกตรัมเฟรมที่สนใจที่จะนำมาหาค่าโคร์รีเลชั่นเฉลี่ย ของเสียงทดสอบและเสียงอ้างอิง โดยกำหนดให้มีจำนวนเท่ากับจำนวนสเปกตรัมของของพยางค์ที่เป็นเสียงทดสอบหรือเสียงอ้างอิงที่มีจำนวนน้อยที่สุด

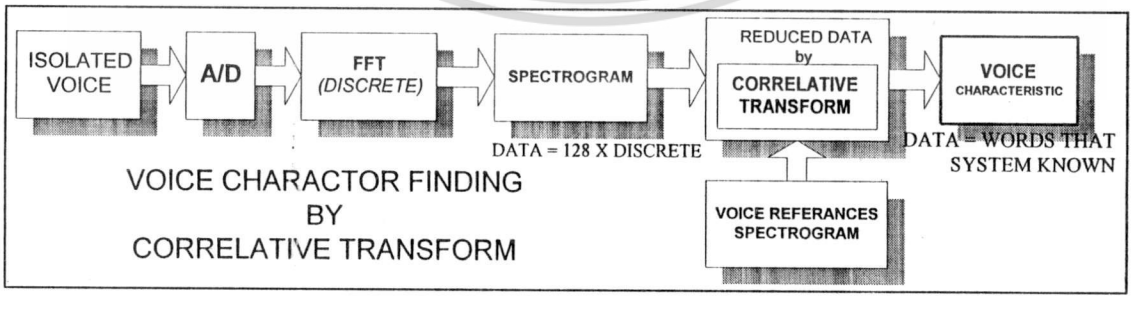
v คือ ข้อมูลที่เป็นสมาชิกของ K เริ่มตั้งแต่ 1 ถึง K

u คือ สเปกตรัมที่เป็นสมาชิกของเสียงหนึ่งพยางค์ในสเปกโตรแกรม เริ่มที่สเปกตรัมที่ 1 ถึง P

วิธีการหาคุณลักษณะจำเพาะของเสียง

การวิเคราะห์เสียงที่ใช้ในการทดลอง ใช้วิธีหาคุณลักษณะจำเพาะตัวของเสียงที่ต้องการทดสอบ โดยนำเสียงทดสอบ มาเทียบหาคุณลักษณะจำเพาะกับเสียงอ้างอิงว่า เสียงที่นำมาทดสอบมีคุณสมบัติใกล้เคียงเสียงอ้างอิงใด เพียงใด โดยใช้การทรานส์ฟอร์มแบบโคร์รีเลทีฟ (Correlative transform) ตามบล็อกไดอะแกรม ดังภาพที่8

ภาพที่ 8



แสดงบล็อกไดอะแกรมของวิธีการหาคุณลักษณะจำเพาะตัวของเสียงโดยใช้การทรานส์ฟอร์มแบบโคร์รีเลทีฟ

จากบล็อกไดอะแกรม ในภาพที่8 ชุดสัญญาณเสียงอนาลอกที่เป็นเสียงคำโดดที่ใช้เป็นตัวอย่าง จะถูกสุ่มเข้ามาโดยกำหนดอัตราสุ่มที่ 10KHz แล้วแปลงให้เป็นสัญญาณดิจิตอล ด้วยรายละเอียดของการแปลงข้อมูลขนาด 8 บิต ที่บล๊อก A/D หาสเปกโตรแกรมด้วยกรรมวิธีการแปลงฟูเรียร์อย่างเร็ว แล้วปรับ สเปกโตรแกรมเพื่อให้เหมาะสมสำหรับการวิเคราะห์เสียง (ธันวา ศรีประโมง,.2537:13-21)สเปกโตรแกรมที่ได้มีลักษณะเช่นเดียวกับภาพที่5 งานที่ทำต่อมาก็คือ การแยกคำเฉพาะเนื้อเสียงแต่ละพยางค์ที่ต้องการออกมาจากสเปกโตรแกรม แล้วนำมาหาคุณลักษณะจำเพาะของเสียงด้วยการทรานฟอร์มแบบสหสัมพันธ์ (Correlative transform) คุณลักษณะจำเพาะของเสียงที่หาได้นี้เป็นข้อมูลที่แสดงค่าระดับความสัมพันธ์กับสเปกโตรแกรมของเสียงอ้างอิง มีขนาดข้อมูลลดลงอย่างมาก ซึ่งจะนำไปใช้ในขบวนการวิเคราะห์เสียงในลำดับถัดไป

<u>การแยกคำออกจากสเปกโตรแกรม</u>

กระทำเพื่อแยกข้อมูลสเปกโตรแกรมเฉพาะส่วนที่เป็นพยางค์ออกมาจากแบคกราวด์นอยด์ (Background noise)เพื่อการทดสอบหรือ สร้างเป็นฐานข้อมูลอ้างอิง วิธีการแยกพยางค์หรือเสียงที่ต้องการวิเคราะห์ อาจกระทำได้หลายแบบดังนี้

1. แยกจากสัญญาณเสียงในขั้นต้นตอนที่เป็นอนาลอก โดยตรวจสอบระดับแอมปลิจูดของเสียงแล้วนำมาแยกเป็นส่วนๆตามระดับพลังงาน แล้วแปลงเป็นสเปกโตรแกรม

(ระพีพัฒน์ เพ็ญศิริ 2538:12-17)

2. แยกพยางค์ต่างๆหลายๆ ที่รวมมาในสเปกโตรแกรม ออกมา

(ธันวา ศรีประโมง 2537:24-25)

สำหรับงานวิจัยที่นำเสนอเลือกวิธีการที่สองเพราะสะดวกและรวดเร็วกว่า

เนื่องจากงานวิจัยมุ่งไปที่การวิเคราะห์เสียงคำโดด ๆ ประเภท 1คำ 1พยางค์ ซึ่งมีลักษณะพิเศษในการออกเสียงคือ การออกเสียงแต่ละคำมีช่องว่างค่อนข้างเด่นชัด ทำให้สามารถแยกคำออกมาได้โดยไม่ยากนัก หากสังเกตจากรูปสเปกโตร์แกรม จะเห็นขอบเขตของแต่ละเสียงได้ชัดเจน โดยช่องว่างระหว่างเสียงจะมีระดับพลังงานต่ำแสดงด้วยสีน้ำเงิน การแยกพยางค์ออกจากสเปกโตรแกรมอาศัยวิธีการหาผลรวมเฉลี่ยของระดับพลังงานในแถบความถี่เสียงแต่ละเฟรมทุกเฟรม แล้วหา เทรสโฮลที่ 20% จากค่าต่ำสุดไปถึงสูงสุด จากนั้นจึงพิจารณาในแต่ละสเปกตรัม(เฟรม) หากผล00รวมเฉลี่ยของระดับพลังงานเสียงในสเปกตรัมใด(เฟรม)มีค่ามากกว่าเทรสโฮล ก็จะดึงข้อมูลเฟรมไปเก็บไว้ที่บัฟเฟอร์ โดยถือเป็นเนื้อเสียงของพยางค์ที่จะนำไปพิจารณา วิธีการดังกล่าวเขียนเป็นสมการได้ดังนี้

$$E_{av(ti)} = \frac{\sum_{f=1}^{M} Ef}{M} \quad | \; ti = 1 \text{ to n} \qquad \text{.....(3.10)}$$

และ

$$T_{ref} = K(Eav_{max} - Eav_{min}) + Eav_{min} \qquad \text{.....(3.11)}$$

ข้อมูลแต่ละพยางค์จะถูกแยกออกมาเมื่อ

$$Eav_{(n)} > Tref \qquad \text{.....(3.12)}$$

กำหนด $Eav_{(n)}$ คือ ค่าระดับพลังงานเฉลี่ย ณ.ช่วงเวลา Δt ใด ๆ

Ef คือ ค่าระดับพลังงานของแต่ละความถี่ในสเปกตรัม โดย f คือลำดับข้อมูลของสเปกตรัมในแกนความถี่ มีลำดับตั้งแต่ 1 ถึง M

M คือ จำนวนข้อมูลในแกนความถี่ (การทดลองใช้ 128 ข้อมูล)

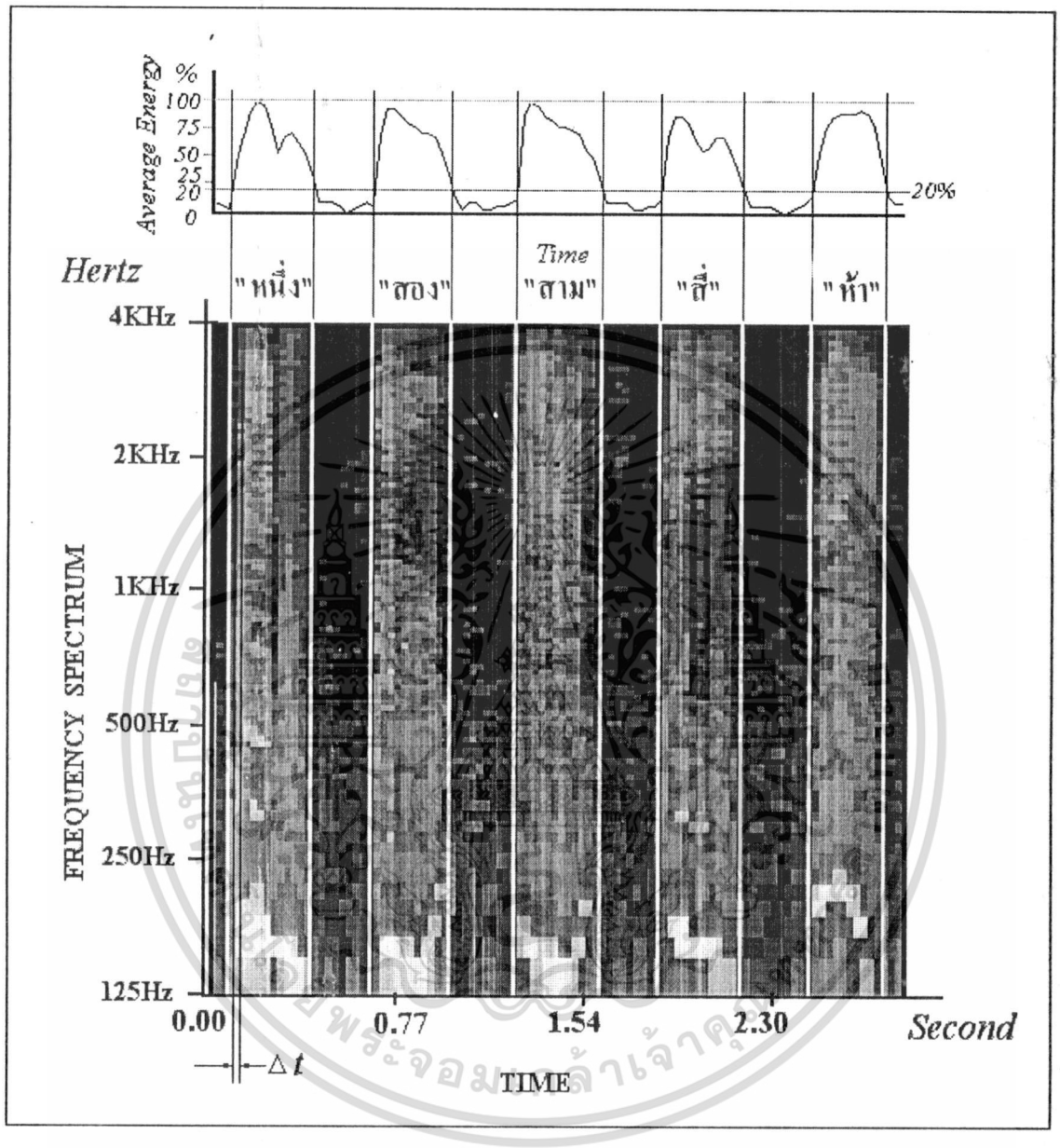
ti คือ เฟรมสเปกตรัมใด ๆ ที่กำลังสนใจ, ti มีค่าตั้งแต่ 1 ถึง n

n คือจำนวนสเปกตรัม ใดๆ(Δt) ทั้งหมดที่ประกอบเป็นสเปกโตรแกรม

$Tref$ คือ ค่าเทรสโฮล ที่กำหนดขึ้นเพื่อแยกพยางค์ออกมาพิจารณา

K คือ สัมประสิทธิ์ ที่ใช้กำหนดจุดแยกพยางค์ออกมาพิจารณา มีค่าตั้งแต่ 0.0-1.0 (จากการทดลองค่าที่ดีที่สุด คือ 0.2)

ภาพที่9



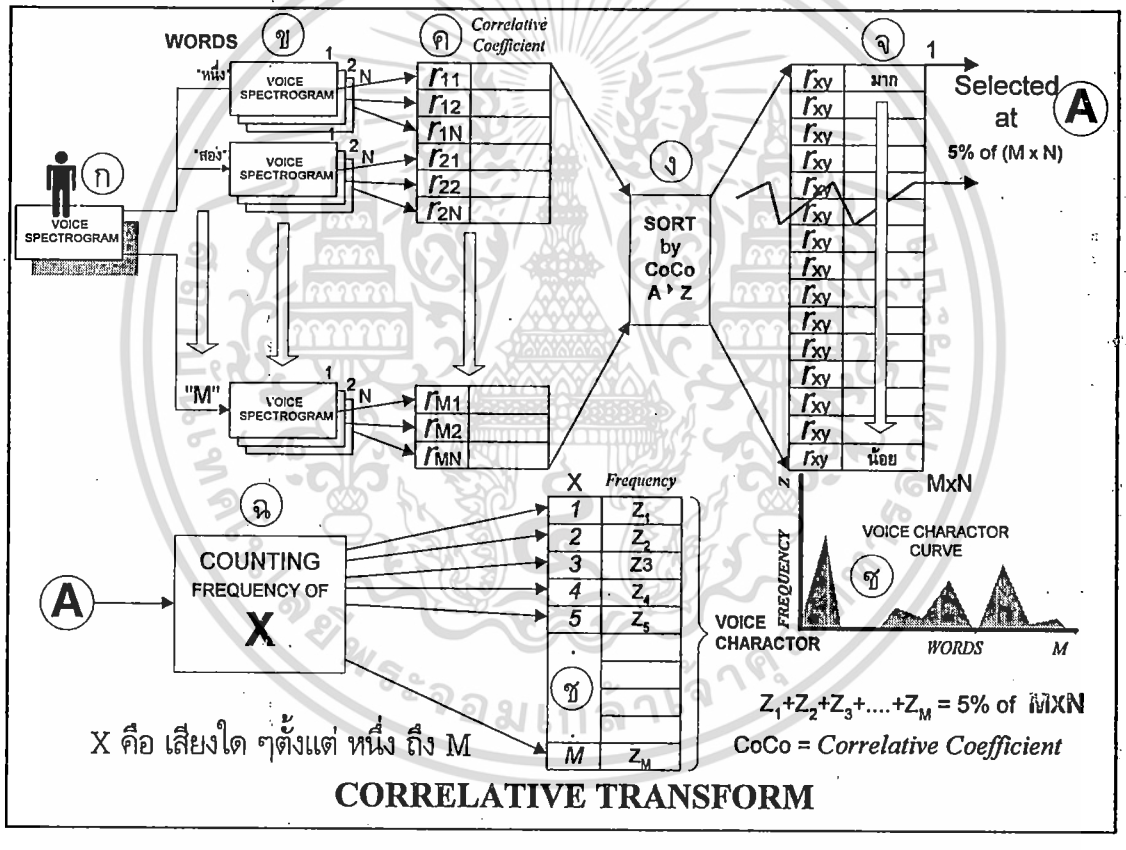
แสดงผลการแยกข้อมูลช่วงที่มีการเปล่งเสียงออกจากสเปกโตรแกรมโดยใช้สมการที่ 3.12

จากภาพที่9 กราฟด้านบนเป็นกราฟที่ได้จากสมการที่3.9 เช่นกันแต่นำมาปรับแสดงผลเป็นเปอร์เซ็นต์ จากการทดลองพบว่า ค่าเทรสโฮลที่เหมาะสมอยู่ที่ 20 % (สมการที่3.10)จากจุดต่ำสุดไปจุดสูงสุด การตัดคำแบบนี้มีความแม่นยำพอสมควร (มากกว่า 92%) โดยมีเงื่อนไขคือ ขณะบันทึกเสียง ต้องไม่มีเสียงสอดแทรกรบกวนเกินไปในช่องว่างระหว่างของพยางค์ จึงเพียงพอที่จะเลือกไปใช้แยกพยางค์เพื่อการทดลองได้

วิธีการทรานส์ฟอร์มสเปกโตรแกรมเพื่อหาคุณลักษณะจำเพาะและลดขนาดข้อมูล

โคร์รีเรทีฟทรานฟอร์ม(Correlative transform)ที่ใช้ในการทดลองเป็นการแปลงข้อมูลสเปกโตรแกรม ของเสียงที่ต้องการวิเคราะห์ให้เป็นข้อมูลคุณลักษณะจำเพาะประจำตัวของเสียงนั้น ใช้วิธีการเปรียบเทียบระดับความสัมพันธ์ของสหสัมพันธ์ ระหว่างสเปกโตรแกรมเสียงทดสอบกับเสียงอ้างอิง โดยคุณลักษณะประจำตัวของเสียงนั้นเป็นคุณสมบัติที่เทียบกับเสียงอ้างอิงเท่านั้น และจำเป็นต้องสร้างคุณลักษณะประจำตัวของเสียงอ้างอิงทุกเสียงก่อนเพื่อใช้ในขบวนการอ้างอิงเปรียบเทียบที่ (ข)

ภาพที่10



แสดงวิธีการทรานส์ฟอร์มแบบโครีเลทีฟที่สร้างขึ้น

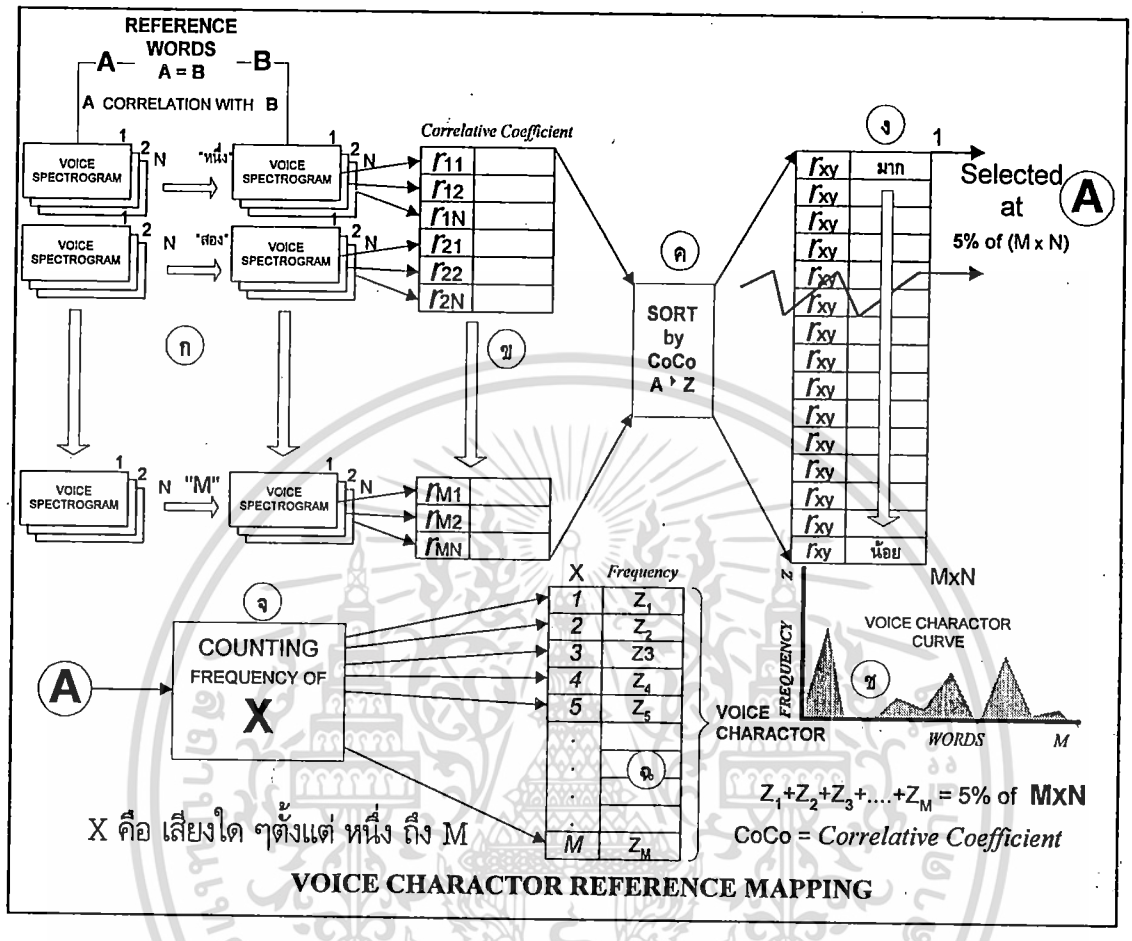
จากภาพที่10 (ก) คือสเปกโตรแกรมของเสียงที่ต้องการหาคุณลักษณะ (ข) คือกลุ่มสเปกโตรแกรมของเสียงที่ใช้สำหรับอ้างอิงซึ่งจะมีการคำนวณหาค่าระดับสหสัมพันธ์ของข้อมูลทั้งสองโดยใช้สมการที่3.6, 3.7 และ 3.8 โดยข้อมูลสเปกโตรแกรมเสียงที่ต้องการทดสอบ (ก) 1ชุดจะนำไปเทียบหาค่าสหสัมพันธ์กับข้อมูลสเปกโตรแกรม เสียงอ้างอิง (ข) ทุกชุดได้เป็นค่าสัมประสิทธิ์สหสัมพันธ์กับชุดสเปกโตรแกรมเสียงอ้างอิงทุกชุด (ค) ได้จำนวนเป็น r_{11} ถึง r_{MN} เมื่อ M คือจำนวนค่าใดๆ ที่ใช้ทดสอบ N คือจำนวนตัวอย่างของคำ

จากนั้น นำข้อมูล r_{11} ถึง r_{MN} ที่ได้จาก (ค) ไปจัดเรียงในบล็อก SORTS ที่ (ง) ซึ่งเป็นการจัดเรียงแบบมากไปน้อย ตามค่า Correlative coefficient ซึ่งมีชื่อของคำเป็นตัวแปรตาม ได้เป็นตาราง (จ) โดย r_{xy} คือตัวแปรตาม ที่แปรไปตามค่า Coefficient (มากไปน้อย) x คือ คำใดๆ ที่อาจเป็นไปได้ y คือ หมายเลขคำที่เท่าใดของคำหลายๆคำที่นำมาใช้ ที่อาจเป็นไปได้ จำนวนค่า Correlative coefficient ที่ได้นี้จะมีจำนวนเท่ากับ MxN การสรุปคุณลักษณะของเสียงทดสอบจะเลือกคำที่ได้จากการจัดเรียงมาจำนวน 5% ของจำนวนข้อมูลสัมประสิทธิ์สหสัมพันธ์ที่ได้ (MxN) ที่ (ฉ) จากนั้นจึงมานับความถี่ของคำที่ถูกเลือกมาความถี่ของแต่ละคำจะเป็นตัวเลขจากที่สุดได้ไม่เกิน N (ช) ซึ่งข้อมูลที่ได้ มีจำนวน M ข้อมูลนั่นคือ คุณลักษณะจำเพาะของเสียงนั้น ซึ่งสหสัมพันธ์กับข้อมูลอ้างอิงอาจเขียนเป็นกราฟได้ดังรูป(ช) ณ จุดนี้จะสามารถบอกความเป็นไปได้ของเสียงที่น่าจะเป็นได้แล้วในระดับหนึ่ง คือ คำ (X) ที่มีความถี่สูงที่สุดนั่นเอง (ช)

<u>การสร้างคุณลักษณะอ้างอิงมาตรฐาน</u>

การวิเคราะห์ในลักษณะดังกล่าว เป็นขั้นตอนที่นำเอาวิธีหาคุณลักษณะจำเพาะของเสียงมาประยุกต์เพื่อใช้สร้างคุณลักษณะจำเพาะของเสียงอ้างอิงและเปรียบเทียบกับคุณลักษณะจำเพาะของเสียงทดสอบ เพื่อให้ระบบสามารถบอกความเป็นไปได้ของเสียงทดสอบที่น่าจะเป็นได้อย่างถูกต้องมากขึ้น วิธีการสร้างคุณลักษณะจำเพาะของเสียงอ้างอิงทำได้โดยใช้สเปกโตรแกรมของเสียงอ้างอิงชุดเดียวกันนี้มาเทียบกันเอง เพื่อหาคุณลักษณะจำเพาะของเสียงอ้างอิงทุกเสียงที่สหสัมพันธ์กับเสียงตัวอย่างทุกเสียง แล้วเก็บไว้เป็นคุณลักษณะอ้างอิงมาตรฐาน มีวิธีการทรานฟอร์มแบบโคร์รีเลทีฟ แตกต่างกันตรงข้อมูลสเปกโตรแกรมที่นำมาเปรียบเทียบเป็นข้อมูลสเปกโตรแกรมชุดเดียวกันกับที่ใช้อ้างอิงด้วยกันเอง

ภาพที่11



แสดงไดอะแกรมการสร้างแบบอ้างอิงมาตรฐานจากสเปกโตรแกรมเสียงอ้างอิง

พิจารณาภาพที่11 จะพบว่า มีการคำนวณหาสัมประสิทธิ์สหสัมพันธ์ที่ Input ระหว่างชุดข้อมูล A และ B (ซึ่งเป็นชุดข้อมูลเดียวกัน) ถึง (MxN)x(MxN) ครั้ง เพื่อให้ได้คุณลักษณะจำเพาะของเสียงอ้างอิงที่ Output จำนวน MxN ชุด ใช้เป็นคุณลักษณะอ้างอิงมาตรฐาน ถึงแม้ว่าการสร้างแบบอ้างอิงมาตรฐานจะใช้เวลานานมากก็ตาม แต่เมื่อให้ระบบเรียนรู้ ชุดคำใดๆ ก็จะทำกับชุดตัวอย่างนั้นครั้งเดียวเท่านั้น

การแบ่งแบนด์ของสเปกโตรแกรมเพื่อเพิ่มความแม่นยำ

การแบ่งแบนด์ของสเปกโตรแกรม ผู้วิจัยตั้งใจไม่กล่าวถึงในลำดับแรก เพื่อหลีกเลี่ยงความสับสน แต่ขอกล่าวสรุปก่อนว่า วิธีการและขั้นตอนยังคงเป็นเช่นเดิม แต่มีการแบ่งพิจารณาเพิ่มขึ้นเป็นช่วง ๆ สามช่วง ดังจะกล่าวต่อไปนี้

เนื่องจากมีข้อมูลของสเปกตรัมในสเปกโตรแกรม ถึง 128 ข้อมูลต่อ 1 สเปกตรัม (1 สเปกตรัม = 1 ดีสคริตของสเปกโตรแกรม) แสดงค่าครอบคลุมแถบความถี่ตั้งแต่ 125 Hz ถึง 4 KHz ซึ่งยาวพอที่ทำให้เกิดข้อผิดพลาดของการเปรียบเทียบหาระดับความสัมพันธ์ระหว่างสเปกตรัมทดสอบกับสเปกตรัมอ้างอิงได้ ในกรณีได้ค่าความสัมประสิทธิ์ สหสัมพันธ์อยู่ในระดับกลาง ค่านี้อาจเกิดได้หลายกรณีเช่น

1. ค่าช่วงต้นของข้อมูลสัมพันธ์กันมากแต่ช่วงปลายสัมพันธ์กันน้อย
2. ค่าช่วงต้นของข้อมูลสัมพันธ์กันน้อย แต่ช่วงปลายสัมพันธ์กันมาก
3. ค่าช่วงกลางของข้อมูลสัมพันธ์กันมาก แต่ช่วงต้นกับปลายสัมพันธ์กันน้อย
4. ค่าความสัมพันธ์อยู่ในระดับกลางตลอด

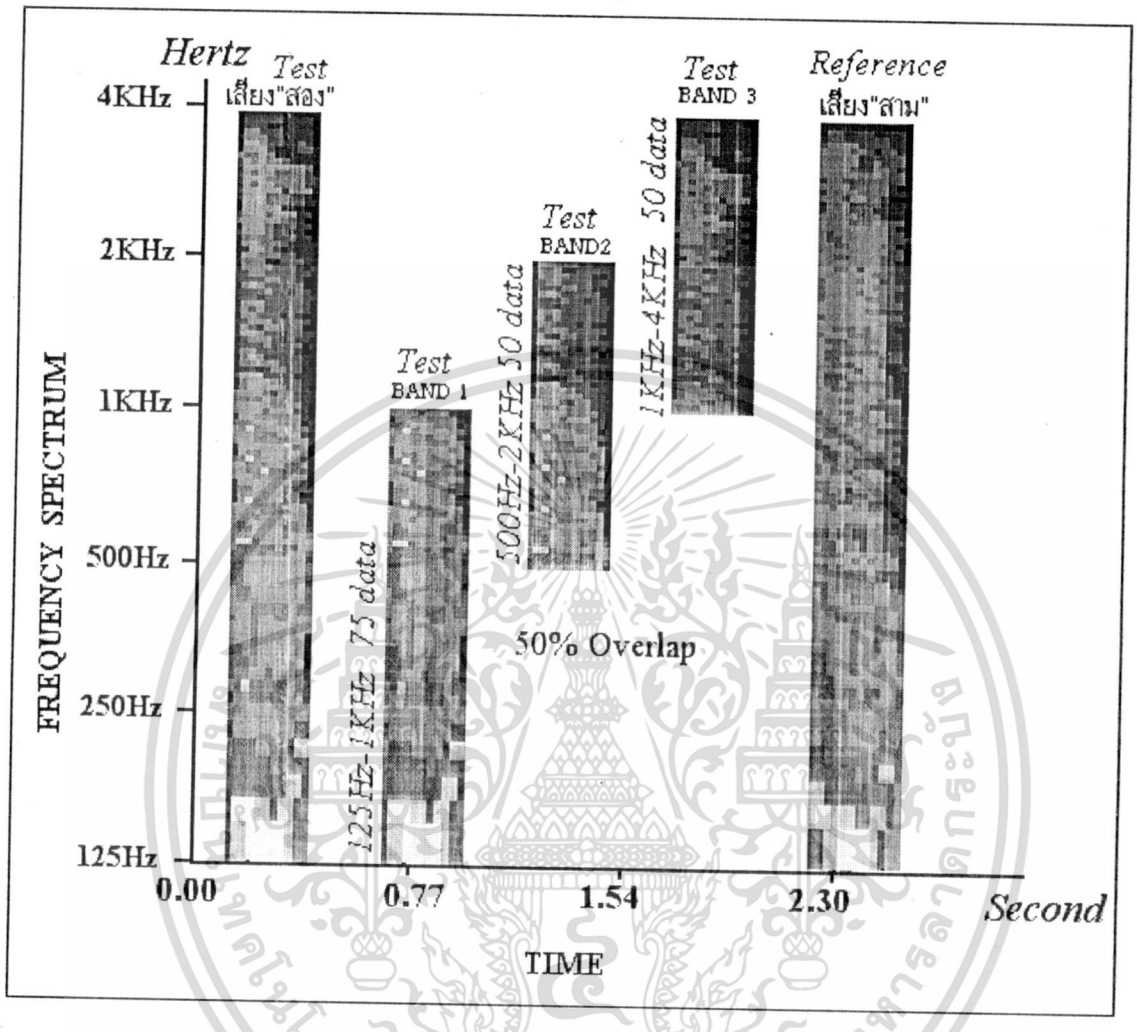
ดังนั้นจึงแก้ปัญหานี้โดยใช้วิธีแบ่งช่วงพิจารณาโดยแบ่งสเปกตรัมความถี่ของแต่ละช่วงเวลาออกเป็น 3 ย่าน คือ

ย่านแบนด์ต่ำ ครอบคลุมความถี่ 125 Hz - 1 KHz
รวม 75 ข้อมูล ในช่วงข้อมูลที่ 0-75ของข้อมูลในสเปกตรัม

ย่านแบนด์ต่ำ-กลางครอบคลุมความถี่ 500 Hz - 2 KHz
รวม 50 ข้อมูลในช่วงข้อมูลที่ 50 - 100ของข้อมูลในสเปกตรัม

ย่านแบนด์กลางครอบคลุมความถี่ 1 Khz - 4 KHz
รวม 50 ข้อมูล ในช่วงข้อมูลที่ 75 - 128 ของข้อมูลในสเปกตรัม

ภาพที่ 12



แสดงการตัดแบ่งสเปกโตรแกรมของเสียงหนึ่งออกเป็น 3 แบนด์ แบบซ้อนทับกัน 50% ของข้อมูล ซึ่งนำไปใช้เปรียบเทียบกับสเปกโตรแกรมของเสียงอ้างอิงต่าง ๆ แบบ แบนด์ต่อแบนด์

ซึ่งลักษณะของแบนด์จะมีลักษณะเหลื่อมซ้อนกัน 50% เพื่อลดข้อผิดพลาดที่เกิดขึ้นในช่วงรอยต่อของแบนด์ที่แบ่ง ดังนั้นไม่ว่าในขบวนการเตรียมข้อมูลแบบอ้างอิง หรือขบวนการทดสอบก็ตาม ดังที่กล่าวมาต้องดำเนินการทำ ทั้ง 3 แบนด์ หมายถึง ต้องทำตามวิธีการหาคุณลักษณะของเสียงดังกล่าว ถึง 3 ครั้ง ทำให้ได้ค่าแพตเทอร์นอ้างอิง 3 ชุด ข้อมูลที่นำมาทดสอบก็ทำเช่นเดียวกัน จากนั้นจึงหาค่าระดับความสัมพันธ์ระหว่างแบบทดสอบกับแบบอ้างอิงทั้ง 3 แบนด์ แบบแบนด์ต่อแบนด์ เมื่อได้ค่าระดับความสัมพันธ์ ระหว่างแบบดังกล่าวทั้ง 3 แบนด์ แล้วจึงนำมาหาค่าเฉลี่ยของความสัมพันธ์ทั้ง 3 แบนด์ อีกครั้งหนึ่งด้วยสมการ 3.13

$$CORR_{AVG} = \frac{1}{\mathrm{F}} \sum_{j=1}^{\mathrm{F}} \sum_{i=\mathrm{BSTRT}}^{\mathrm{BSTOP}} \frac{Zx_{ji} Z_{Yji}}{\mathrm{BSTOP}} \quad \text{.....(3.13)}$$

$CORR_{AVG}$ คือ ค่าสัมประสิทธิ์สหสัมพันธ์เฉลี่ยทั้งสามแบนด์

F คือจำนวนเฟรมของเสียง อาจเป็นจำนวนเฟรมของเสียงอ้างอิงหรือเสียงทดสอบก็ได้โดยพิจารณาจากจำนวนเฟรมของเสียงใดที่น้อยที่สุด โดย

$F = m$ เมื่อ $m-n \leq 0$, m คือ จำนวนเฟรมของพยางค์ที่ใช้ทดสอบ

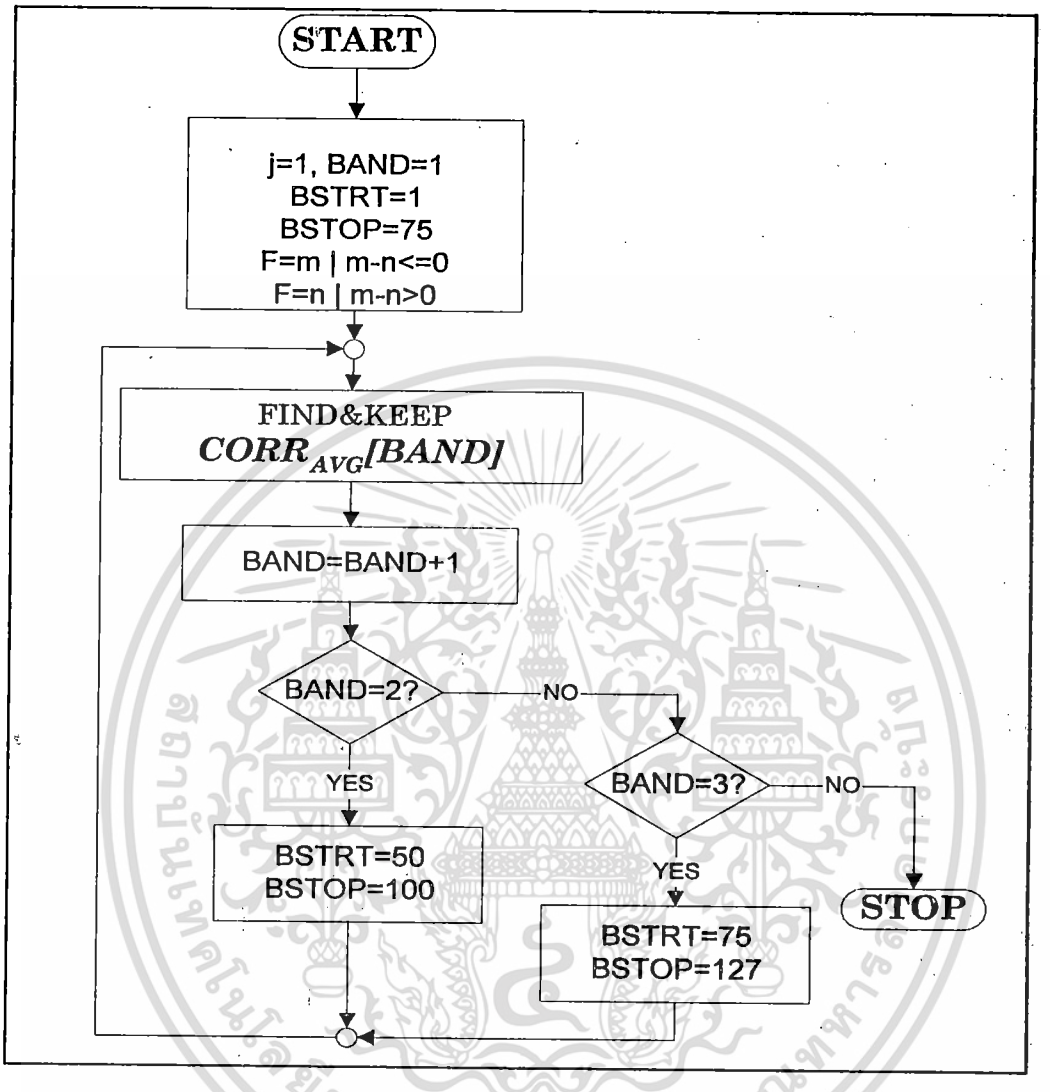
$F = n$ เมื่อ $m-n > 0$, n คือ จำนวนเฟรมของพยางค์อ้างอิง

BSTRT และ BSTOP คือตำแหน่งเริ่มต้นและสิ้นสุดของข้อมูลของแบนด์ที่กำลังพิจารณา

Zx_{ji} และ Zy_{ji} คือค่า Z-Score ของเฟรมทดสอบกับเฟรมอ้างอิง เฟรมที่ j

การหาค่าสัมประสิทธิ์สหสัมพันธ์เฉลี่ยทั้งสามแบนด์ พอแสดงเป็นโฟล์วชาร์ทได้ดังนี้

ภาพที่13



แสดงโฟล์วชาร์ทการหาค่าสัมประสิทธิ์สหสัมพันธ์เฉลี่ยทั้งสามแบนด์ ของสเปกโตรแกรมเสียงทดสอบกับเสียงอ้างอิง

หลังจากเสร็จสิ้นการหาค่าสัมประสิทธิ์สหสัมพันธ์เฉลี่ยทั้งสามแบนด์แล้ว จะนำค่า $CORR_{AVG}$ ที่ได้ใช้เป็นแบบอ้างอิงเพื่อใช้ในขบวนการต่อไป

บทที่4

นิวรัลเน็ตเวอร์ค (Neural Network)

ความรู้เบื้องต้นเกี่ยวกับนิวรัลเน็ตเวอร์ค

ในช่วงระยะเวลา 8-9 ปีที่ผ่านมา ในต่างประเทศมีการตื่นตัวในการวิจัยและพัฒนาเกี่ยวกับโครงข่ายประสาทเทียม (Artificial Neural Network:ANN) อย่างกว้างขวาง ทั้งทางด้านทฤษฎีและการประยุกต์ใช้งาน จนกระทั่งถึงปัจจุบันได้มีการประยุกต์นำนิวรัลเน็ตเวอร์ค มาใช้ในอุปกรณ์เครื่องใช้ต่างๆ มากขึ้น เช่น เครื่องมือหาปลา (Sonar) ที่มีความฉลาดมากขึ้น เช่น สามารถบอกได้มา ฝูงปลาที่กำลังตรวจจับอยู่นั้นเป็นปลาชนิดใด, จำนวนเท่าไร เครื่องโทรศัพท์แบบ (Voice phone) ที่สามารถหมุนหมายเลขที่จะเรียกให้อัตโนมัติเพียงยกหูแล้วพูดชื่อของผู้ที่จะติดต่อเท่านั้น เครื่องอ่านตัวอักษร (OCR) ที่สามารถเปลี่ยนภาพตัวอักษรให้เป็นรหัสตัวอักษรของคอมพิวเตอร์ ระบบนักบินอัตโนมัติ (Auto pilot aircraft) ระบบการคาดเดาอนาคตจากข้อมูลในอดีต (Forcasting, Prediction) ฯลฯ ซึ่งเครื่องมือและอุปกรณ์ต่างๆ ที่นำเอา ANN มาใช้ช่วยวิเคราะห์นั้น จะมีความฉลาดมากขึ้น และมีระบบความคิดที่มีการทำงานในลักษณะเดียวกับของมนุษย์

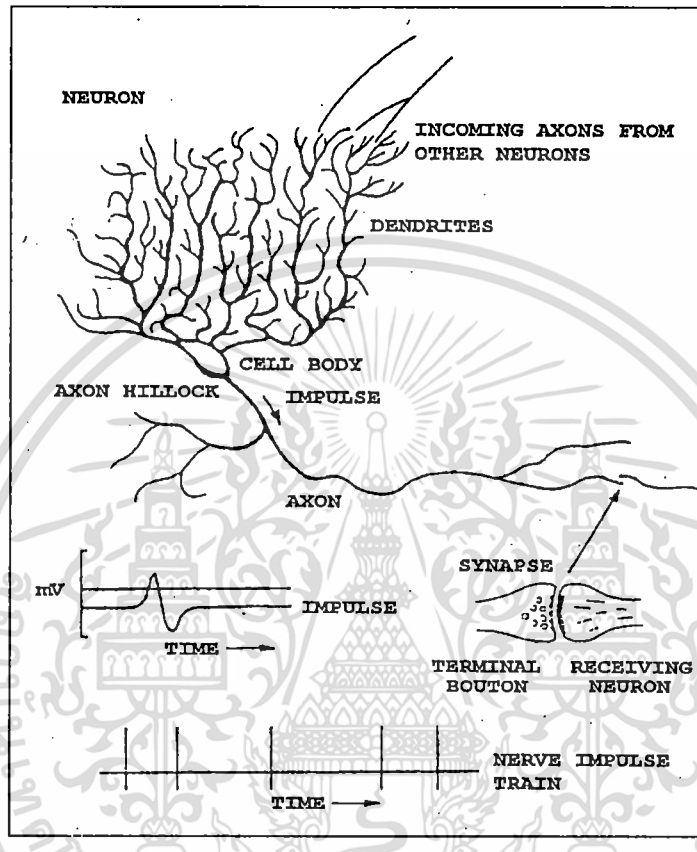
นิวรัลเน็ตเวอร์ค หมายถึงโครงข่ายใยประสาทที่เชื่อมต่อกันระหว่างเซลล์ประสาทจำนวนมากมายมหาศาลในสมอง มีความสามารถประมวลผลสูงบรรจุอยู่ในสมอง สมองชีวภาพที่เป็นจุดศูนย์กลางการควบคุมกิจกรรมของการดำเนินชีวิต การวิจัยสร้างโครงข่ายประสาทเทียม (Artificial Neural Network) มีแนวคิดเลียนแบบการทำงานของสมองชีวภาพ โดยเรียนรู้และศึกษาการทำงานของสมองชีวภาพเพื่อกำหนดแนวทางสำหรับการสร้างแบบจำลองขึ้นมา แล้วพยายามสมมติฐานลักษณะการทำงานให้เป็นโมเดลคณิตศาสตร์ที่มีลักษณะเดียวกันแล้วดำเนินการคำนวณโดยใช้คอมพิวเตอร์

นิวรัลเน็ตเวอร์คชีวภาพ

ระบบคิดคำนึงของมนุษย์ มีโครงสร้างพื้นฐานจากเซลล์สมอง ที่เรียกว่านิวรอล (Neurons) เรียงเป็นชั้นๆ อย่างซับซ้อน จำนวนมหาศาล ประมาณหมื่นล้าน (10^{11})นิวรอล และอาจมีจุดเชื่อมโยงส่งผานจุดเชื่อมโยงภายในถึงพันล้านล้านจุด (10^{15}) (Philip D. Wasserman,1989:12)แต่ละนิวรอลจะมีคุณลักษณะต่างๆ แตก

ต่างกันไป โดยมีการทำงานคล้ายกันคือ รับเข้า, ประมวลผล, ส่งออกสัญญาณไฟฟ้าเคมีผ่านไปยังนิวรอล ซึ่งจะส่งสื่อสารไปตามระบบของสมอง

ภาพที่14



แสดงโครงสร้างตัวอย่างของเซลล์ประสาทชีวภาพ

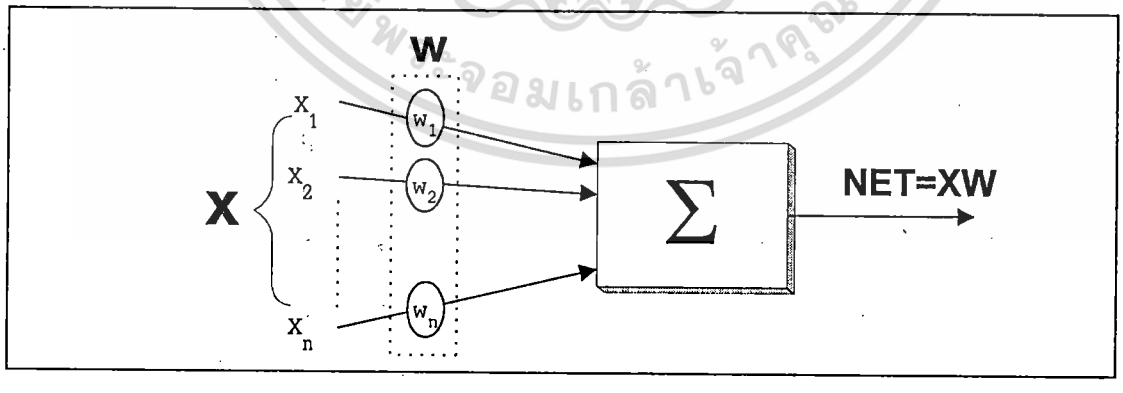
จากภาพที่14 ส่วนที่ขยายแยกออกไปจากตัวเซลล์ต่อไปยังเซลล์อื่น ๆ เรียกว่า เดนไดรท์ (Dendrites) ส่วนรับสัญญาณจากเซลล์อื่นเข้ามายังตัวเซลล์จะผ่านมาทางจุดเชื่อมต่อที่เรียกว่า ซินแนปส์ (Synapse) ซึ่งแอกซอน (Axon) จะเป็นตัวส่งสัญญาณเอาท์พุทออกไปยังนิวรอลอื่นจากผลการวิจัยพบว่า แต่ละนิวรอลจะเชื่อมต่อออกไปยังนิวรอลอื่นๆ ซึ่งแต่ละนิวรอลจะมีคุณสมบัติในการเพิ่มขยายหรือลดทอนความเข้มของสัญญาณบางสัญญาณที่เข้ามาทางเดนไดรท์ ของเซลล์ (ซึ่งมีแขนงมากมาย) อาจสามารถกระตุ้นตัวเซลล์ แต่บางสัญญาณก็อาจจะยับยั้งตัวเซลล์ เนื่องจากเซลล์ประสาท 1 เซลล์ มีเดนไดรท์มาก ฉนั้น สัญญาณกระตุ้นจากเดนไดรท์ ที่ รับเข้ามาจากเซลล์ประสาทอื่นๆ จะถูกนำมารวมกันที่ตัวเซลล์ประสาทที่เซลล์ประสาทจะมีค่าเทรสโฮล (Threshold) ค่าหนึ่งหากผลรวมของสัญญาณไฟฟ้าเคมี (Electrochemical) มีค่ามากกว่า เทรสโฮลเซลล์ เซลล์ประสาทก็จะส่งสัญญาณขนาดหนึ่งผ่านทางแอกซอนไปยังนิวรอลอื่นๆ

การจัดเรียงชั้น (Layer) และลักษณะการเชื่อมโยงระหว่างนิวรอลในสมองนั้นมีการจัดเรียงที่ซับซ้อน สอดคล้องกับหน้าที่การทำงานเฉพาะส่วน การเจริญเติบโตสิ่งแวดล้อมและการเรียนรู้ตลอดเวลา ซึ่งใช้เวลานานนับปี ดังนั้นจึงยากที่จะสร้างโมดุลขึ้นมาเพื่อเลียนแบบให้มีคุณลักษณะคล้ายสมองชีวภาพได้ทั้งหมด ผลงานที่ได้จากการทำวิจัยในปัจจุบันเป็นเพียงการจำลองและการเลียนแบบและการทำงานเฉพาะบางส่วนของโครงข่ายประสาท มาใช้เฉพาะกับงานใดงานหนึ่ง ซึ่งมีการวิจัยลักษณะของโครงข่ายแบบต่างๆ ขึ้นมา โดยแต่ละแบบจะเหมาะกับงานประเภทหนึ่งๆ เท่านั้น

โครงข่ายประสาทเทียม (Artificial Neural Network)

การออกแบบสร้างประสาทเทียมนั้นมีสมมติฐานขั้นแรกจากคุณสมบัติของระบบประสาทชีวภาพ ดังที่กล่าวมา กล่าวคือ ชุดรับสัญญาณข้อมูล อินพุทของเซลล์ประสาทหนึ่งได้จากสัญญาณเอาท์พุทของเซลล์ประสาทอื่นๆ ผ่านทางซินแนปและเดนไดรท์ ข้อมูลแต่ละค่าที่รับมาจะถูก จะลดขนาดด้วย ซินแนปติกส์ ซึ่งภายในประกอบด้วยสารเคมีประเภท K^+,Ca^{++},Na^+,Cl^- ซึ่งจะมีลักษณะทางความนำ พัลล์ (Pulse) สัญญาณไฟฟ้าเคมีที่แตกต่างกัน (James A. Freeman and David M.Skapura,1991:8-9) ด้วยเหตุนี้ โมดุลประสาทเทียมที่สร้างขึ้น จะต้องมีการถ่วงน้ำหนักให้กับโมดุลก่อน ที่จะนำเข้าสู่โมดุลประสาทเทียม ปริมาณของข้อมูลที่เข้าสู่นิวรอล จะถูกนำมารวมกัน และตัดสินใจด้วยระดับความสนใจของนิวรอล (Activation level) แล้วจะส่งเป็นเอาท์พุทออกที่แอกซอนไปยังนิวรอลอื่นๆ

ภาพที่15



แสดงไดอะแกรมของนิวรอลที่สร้างขึ้น (Artificial Neuron)

จากภาพที่15 แสดงถึงโมดุลที่สร้างขึ้นโดยแนวความคิดจากเซลล์สมองชีวภาพ สัญญาณอินพุท คือ $X_1, X_2, ..., X_n$ จะถูกป้อนเข้าไปยังนิวรอลที่สร้างขึ้น ซึ่งเปรียบเทียบได้กับสัญญาณที่ป้อนเข้ายัง ซินแนปของ

นิวรอลชีวภาพ สัญญาณอินพุทนี้จะนำไปคูณกับค่าเวจท์ (Weight: ค่าที่ใช้ถ่วงน้ำหนัก) W_1, W_2, ..., W_n ก่อนที่จะเข้าสู่บล็อกซัมเมชั่น (Σ:Summation) ซึ่งค่าถ่วงน้ำหนักนี้จะสอดคล้องกับค่าสเตรงท์(Strength) ของจุดต่อซินแนปชีวภาพแต่ละจุด (Single biological synaptic connection) บล็อกซัมเมชั่นนี้ก็จะทำหน้าที่สอดคล้องคล้ายกับตัวเซลล์สมองชีวภาพ ผลรวมทางคณิตศาสตร์ของอินพุตและเวจท์จะได้เป็นเอาท์พุท เราเรียกว่า เน็ต(NET) ซึ่งเราจะรวมกันในรูปของเวกเตอร์ได้ดังนี้

$$NET = X_1W_1 + X_2W_2 + \ldots\ldots\ldots + X_nW_n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.1)$$

จะได้

$$\mathbf{NET = XW} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.2)$$

ฟังก์ชั่นกระตุ้นความสนใจ (Activation Function)

เมื่อได้สัญญาณ NET แล้วขบวนการต่อมาที่นิวรอลต้องทำคือตัดสินใจ เราจึงต้องกำหนด ฟังก์ชั่นการตัดสินใจ เพื่อใช้เป็นระดับของการตัดสินใจให้กับนิวรอล เพื่อให้ได้สัญญาณเอาท์พุทของนิวรอล ออกมา ซึ่งเชื่อมต่อไปยังนิวรอลตัวอื่น ๆ เป็นโครงข่าย OUT ที่ได้อาจเป็น Simple linear function โดย

$$\mathbf{OUT = K[NET]} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.3)$$

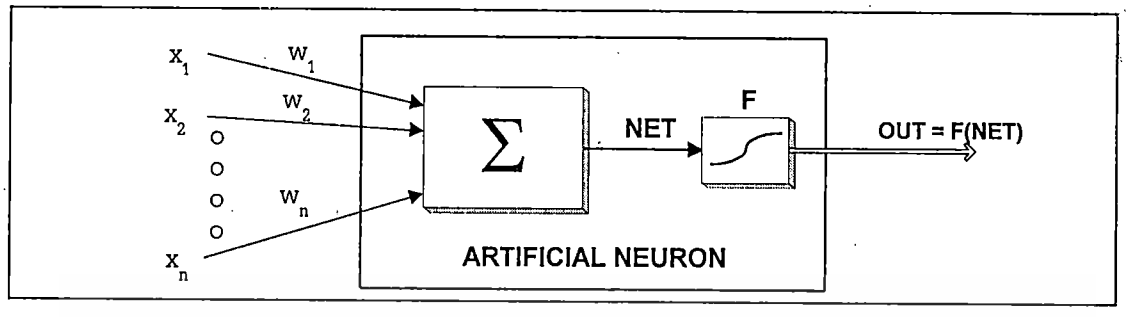
โดย K เป็นค่าคงที่ ที่เรียกว่า Threshold function
ตัวอย่างเช่น

OUT = 1 ถ้า NET > T

OUT = 0 เมื่อเป็นกรณีอื่น $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.4)$

และ T เป็นค่าเทรสโฮลคงที่ หรืออาจเป็น Function อื่น ๆ ที่เลียนแบบคุณสมบัติที่ไม่เป็นเชิงเส้นของเซลล์ประสาทชีวภาพได้อย่างใกล้เคียงกว่า และใช้เป็นฟังก์ชั่นให้กับโครงข่ายทั่วไปได้

ภาพที่16

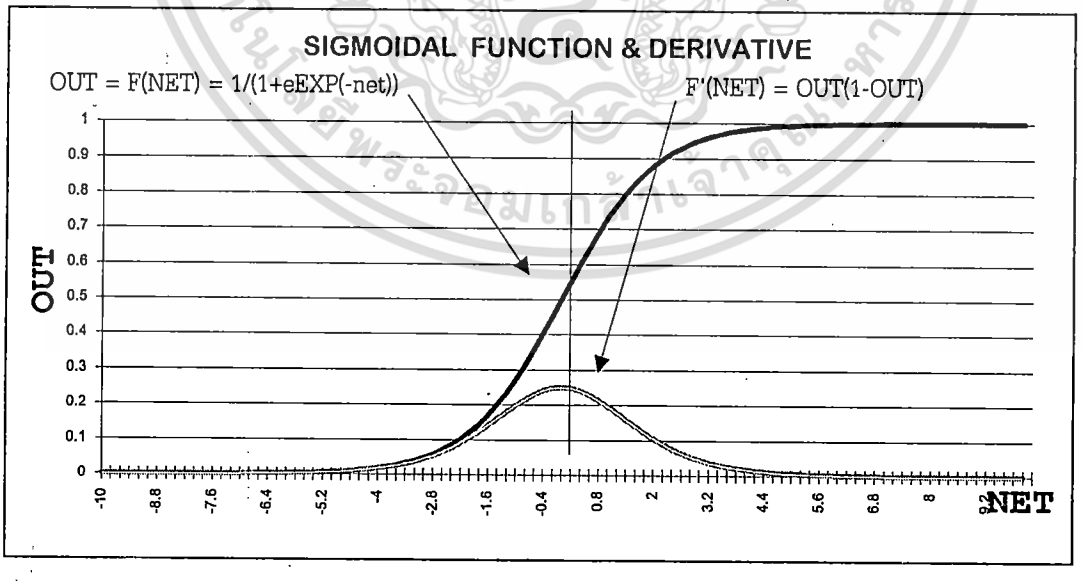


แสดงโมดุลนิวรอลที่สร้างขึ้นร่วมกับ Activation Function

ในภาพที่16 บล็อก F จะรับผลที่ได้จาก NET มาสร้างเป็นสัญญาณเอาท์พุทที่ OUT โดยขบวนการภายในบล็อก F จะบีบช่วงของ OUT ให้อยู่ในขอบเขตจำกัด ตามต้องการ ดังนั้น ค่า OUT จะมีค่าไม่ต่ำกว่าช่วงที่กำหนดโดยค่าของ NET เราเรียก บล็อก F นี้ว่า สแควชชิ่ง ฟังก์ชั่น (Squashing function) และโดยทั่วไปสแควชชิ่งฟังก์ชั่นที่ใช้เป็นแบบ ลอจิสติกฟังก์ชั่น หรือ "ซิกมอยด์"(Logistic function or "Singmoid") ซึ่งมีรูปร่างคล้ายตัว S โดยเขียนเป็นสมการคณิตศาสตร์ได้ดังนี้คือ

$$F(x) = \frac{1}{(1-e^{-x})} \qquad \text{.....(4.5)}$$

ภาพที่17

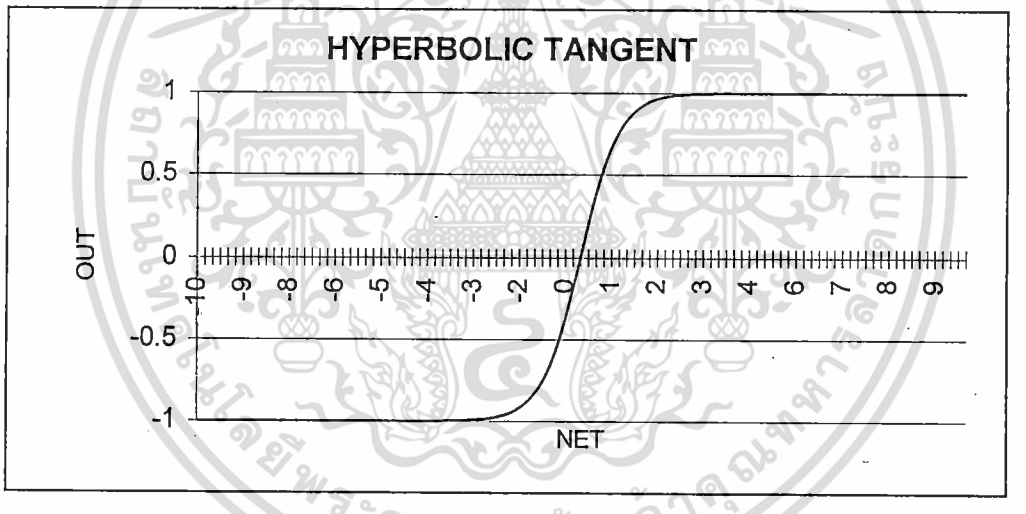


แสดงกราฟที่ได้จากสมการซิกมอยด์ลอจิสติกฟังก์ชั่น(Sigmoidal logistic function)

ลักษณะของเทรสโฮลฟังก์ชั่นมีลักษณะเป็น Non-linear function เช่น S-Curve เราจะได้ค่าเอาท์พุท ที่มีความไวต่อสัญญาณอินพุทที่มีขนาดเล็กๆ และเฉลี่ยต่อสัญญาณแรงๆ ซึ่งสัญญาณอ่อนๆ ไปทางบวกเพียงเล็กน้อยก็จะให้ OUT ใกล้เคียง "1" กระตุ้นหรือสัญญาณอ่อนๆ ทางลบเพียงเล็กน้อยก็จะทำให้ Output ใกล้เคียง "0" (ยับยั้ง) ขณะที่สัญญาณแรงๆ ทางบวกก็ยังคงให้ Output ใกล้เคียง "1" และสัญญาณทางลบแรงๆ ก็คงให้ Output ใกล้เคียง "0" เช่นกัน คุณลักษณะแบบนี้ เป็นแบบ NON-LINEAR GAIN ซึ่ง คลอสส์เบอร์ก (Grossberg,1973) พบว่า คุณลักษณะที่เป็น NON-LINEAR GAIN นี้สามารถแก้ปัญหา Noise-saturation dilemma ได้ และทำให้นิวรอลเทียมที่สร้างขึ้นสามารถทำงานกับขนาดของอินพุทได้กว้างมากขึ้น

ยังมีฟังก์ชั่นอื่นๆ อีกคือ ไฮเปอร์โบลิก แทนเจนท์ (Hyperbolic tangent) มันจะมีลักษณะคล้ายกับ Logistic function และนิยมใช้บ่อยๆ ในการสร้างโมดูลคณิตศาสตร์ การกระตุ้นเร้าความสนใจของเซลล์สมองเทียม ซึ่งมีคุณสมบัติคล้ายชีวภาพของเซลล์สมอง คือ OUT = Tanh(X)

ภาพที่18



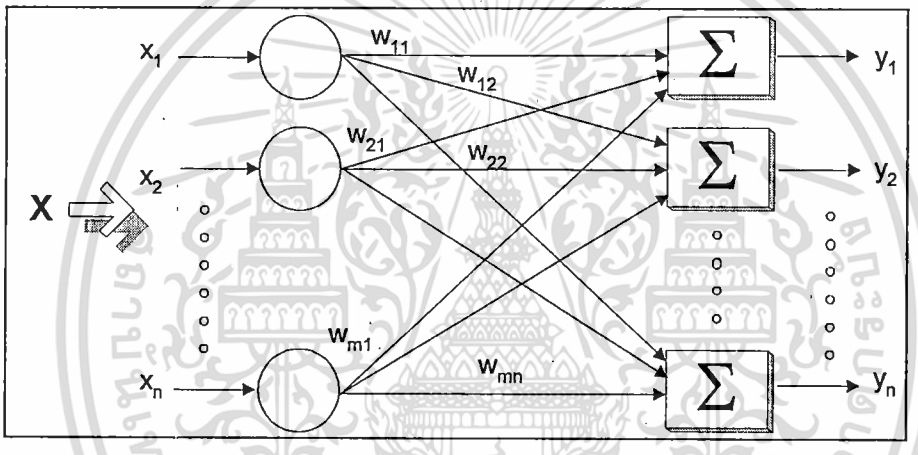
แสดง Hyperbolic Tangent Function

จากภาพที่18 ส่วนที่เหมือนกับ ซิกมอยด์ ลอจิสติก ฟังก์ชั่น คือมีลักษณะเป็น S แต่เนื่องจากมันจะมีความสมมาตรจึงให้ OUTPUT อยู่ระหว่าง "-1" ถึง "1" OUTPUT จะเป็น "0" เมื่อ NET เป็น "0" OUTPUT เข้าใกล้ "1" เมื่ออินพุทไปทางบวกและเข้าใกล้ "-1" เมื่อ อินพุทมีทิศทางไปทางลบ

โครงข่ายประสาทเทียมแบบชั้นเดียว (Single Layer Artificial Neural Network)

ที่กล่าวมาจนถึงจุดนี้ เป็นการกล่าวถึงหลักการและเหตุผลในการสร้างเซลล์ประสาทเทียมเพียง 1 เซลล์ โดยใช้แนวความคิดจากเซลล์ประสาทชีวภาพ การจะนำเซลล์ประสาทเทียมมาใช้งานได้นั้น ต้องใช้เซลล์ประสาทเทียมที่มีคุณลักษณะต่างๆ กัน (ค่า Weight จะทำให้คุณสมบัติของเซลล์ประสาทเทียมแต่ละเซลล์มีคุณลักษณะเปลี่ยนไป) มาเชื่อมโยงเป็นโครงข่าย ในลักษณะเดียวกับเซลล์สมองชีวภาพเสียก่อน ซึ่งลักษณะการเชื่อมโยงก็จะมีหลายแบบหลายหลักการซึ่งจะได้กล่าวต่อไป

ภาพที่19



แสดงลักษณะโครงข่ายประสาทเทียมแบบชั้นเดียว (Sigle-Layer Neural Network)

จากภาพที่19 เป็นโครงข่ายประสาทเทียมแบบชั้นเดียว ที่ประกอบด้วยเซลล์ประสาทเทียมง่ายๆ หลายๆชุด ความสามารถในการคำนวณของโครงข่ายประสาทเทียมได้มาจากลักษณะการเชื่อมต่อเป็นโครงข่ายประสาทเทียมโครงข่ายง่ายๆ เป็นกลุ่มโมดุลประสาทเทียมที่เชื่อมต่อกัน เป็นชั้น ๆ (Layer) ในภาพที่19 เป็นโครงข่ายประสาทเทียม แบบชั้นเดียว (Single layer) ที่ประกอบด้วยเอาท์พุทเลเยอร์ (กลุ่มของบล็อกซัมเมชั่นที่อยู่ทางขวามือ) และอินพุทเลเยอร์ (วงกลมทางซ้ายมือ) โดยไม่พิจารณาอินพุทเลเยอร์ว่าเป็น นิวรอล เลเยอร์ เนื่องจากอินพุทเลเยอร์จะทำหน้าที่เชื่อมต่ออินพุทที่รับมาและส่งออกไปให้ยังแต่ละอินพุทนิวรอลเลเยอร์ (ในที่นี้คือ Output layer) ในชั้นถัดไป โดยแต่ละอินพุทจะถูกคูณโดยค่าเวจท์เฉพาะแต่ละอินพุท โครงข่ายประสาทเทียมที่สร้างขึ้นในขั้นแรกไม่ซับซ้อน โดยแต่ละนิวรอลจะได้เอาท์พุทจาก

OUT = Logistic Function คูณ (ผลรวมของ Input **X** กับ **W**eight)

หรือ OUT = F(NET) .....(4.6)

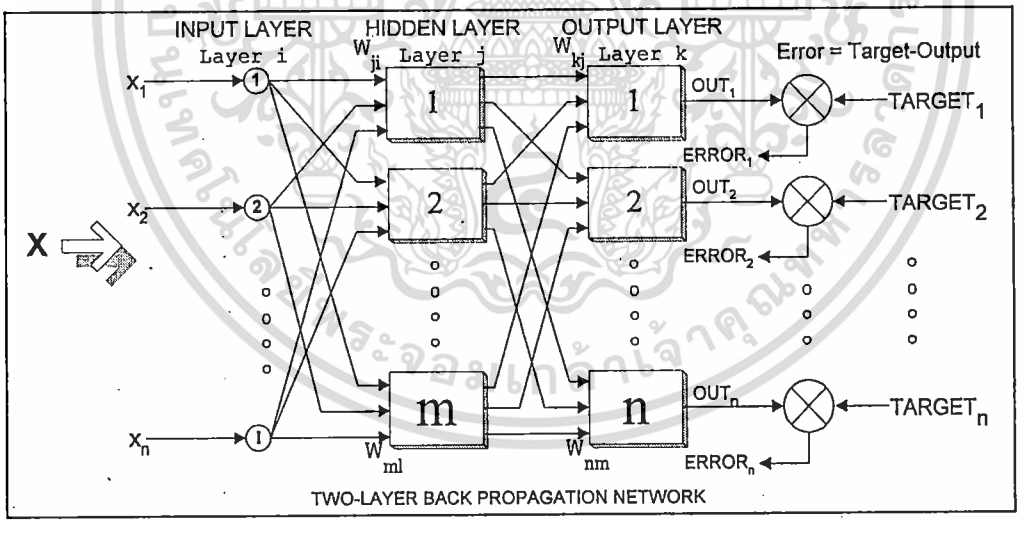
อย่างไรก็ดีลักษณะการเชื่อมโยงระหว่างโครงข่ายไม่ได้มีแบบเดียว การเชื่อมโยงระหว่างเลเยอร์อาจมีการเชื่อมโยงย้อนกลับมาที่อินพุทเลเยอร์อีก ซึ่งโครงข่ายประสาทชีวภาพก็มีลักษณะดังกล่าวเช่นกัน

สำหรับค่า Weight ในภาพที่19 มีวิธีการพิจารณาในรูปของ เวจท์เมตริก (Weight matrix) ซึ่งหากโครงข่ายมีหลายชั้น จะช่วยให้ระบุค่าเวจท์ ได้ง่ายขึ้น และเพื่อหลีกเลี่ยงความสับสนจะกำหนดเป็น ไดเมนชั่น (Dimensions) ของเมตริก โดยให้ m แทนจำนวนแถว หรือจำนวนของอินพุทและ n แทนจำนวนของนิวรอลที่สร้างขึ้น ตัวอย่างเช่น เวจท์ ที่เชื่อมระหว่างอินพุท ตัวที่ 4 กับนิวรอลตัวที่ 2 คือ $W_{4,2}$

<u>โครงข่ายประสาทเทียมแบบหลายชั้น</u> (Multilayer Artificial Neural Network)

โครงข่ายที่ซับซ้อนจะความสามารถในการคำนวณที่ดีขึ้นมันจะเป็นโครงข่ายที่มีโครงสร้างเป็นจินตนาการที่น่าเป็นไปได้โดยการจัดการเชื่อมโยงนิวรอลมีโครงสร้างเป็นชั้นๆ คล้ายส่วนหนึ่งของสมอง และมีการพัฒนาอัลกอลิที่มเกี่ยวกับการฝึกสอนให้โครงข่ายแบบหลายชั้นทำงานได้ตามความต้องการ แล้วเมื่อไม่นานมานี้ โครงข่ายแบบหลายชั้น อาจจะสร้างจาก กลุ่มของโครงข่ายแบบชั้นเดียวเอาท์พุทของ Layer หนึ่ง จะใช้เป็นอินพุทของ Layer ถัดไป ในภาพที่20 แสดงเน็ตเวอร์คที่มีการเชื่อมต่อแบบสองชั้น

ภาพที่20



แสดงไดอแกรมของ Backpropagation Neural Network แบบสองชั้น

ภาพที่20 แสดงโครงข่ายประสาทเทียมแบบหลายชั้น ที่ต่อเชื่อมโยงแบบเต็มชั้น ในโครงข่ายแบบหลายชั้นมีการเรียกชื่อชั้นต่างๆ ดังนี้ คือ ชั้นที่ต่อโดยตรงกับอินพุท เรียกว่า อินพุทเลเยอร์(Input lyer)ชั้นนี้จะไม่มีการคำนวณ แต่จะทำหน้าที่ต่อเชื่อมข้อมูลไปยังชั้นถัดไป ชั้นที่อยู่ท้ายสุดทางขวามือเรียกว่า เอาท์พุทเลเยอร์ (Output layer) เป็นชั้นที่โครงข่ายจะให้ผลลัพท์ ส่วนชั้นที่อยู่ระหว่างอินพุทเลเยอร์ และเอาท์พุท

เลเยอร์ จะมีกี่ชั้นก็ตามจะเรียกว่า ฮิดเดนเลเยอร์ (Hidden layer) หากฮิดเดนเลเยอร์ มีหลาย ๆ ชั้นก็จะมีการตั้งชื่อเฉพาะลงไปให้กับแต่ละชั้น

ฟังก์ชันกระตุ้นความสนใจแบบไม่เป็นเชิงเส้น (The Nonlinear Activation Function)

การนำเอาท์พุทของเลเยอร์หนึ่ง มาเชื่อมกับอินพุทของเลเยอร์ชั้นถัดไป โดยผ่านฟังก์ชันกระตุ้นความสนใจแบบไม่เป็นเชิงเส้น จะทำให้โครงข่ายมีความสามารถในการคำนวณเพิ่มขึ้น (หากไม่ผ่านฟังก์ชันดังกล่าว ความสามารถการคำนวณจะไม่เพิ่มขึ้นและจะมีความสามารถไม่แตกต่างไปจากSingle layer network)

การฝึกสอนให้กับโครงข่ายประสาทเทียม (Training of Artificial Neural Network)

ค่าเวจท์ มีความสัมพันธ์กับอะไร ? เปลี่ยนแปลงอย่างไร ? เช่นเดียวกับเด็กที่คลอดออกมาก็มีสมองแล้วแต่สมองยังไม่เจริญเติบโตและยังไม่ได้รับการฝึกสอนและเรียนรู้ เด็กจึงไม่สามารถทำกิจกรรมใดๆ ด้วยตนเอง เว้นแต่กิจกรรมที่ธรรมชาติสร้างมาพร้อมกับการกำเนิดที่เรียกว่า "สัญชาติญาณ" ซึ่งธรรมชาติใส่คุณลักษณะบางอย่างให้เซลล์สมองบางส่วนตั้งแต่ทารกเจริญเติบโตอยู่ในครรภ์มารดา เช่น ระบบควบคุมการหายใจ, ความรู้สึก, การเรียกร้องเมื่อหิว, การตอบสนองต่อสิ่งเร้า ฯลฯ เด็กจะพัฒนาการเรียนรู้ไปตามขั้นตอน หลังจากนั้นสมองของเขาจะได้รับการฝึกสอน และเจริญเติบโตไปพร้อมๆ กัน เซลล์สมองจะได้รับการปรับคุณลักษณะสอดคล้องกับการฝึกสอน และจะเจริญเป็นโครงข่ายสอดคล้องกัน

โครงข่ายประสาทเทียมที่สร้างขึ้นมีลักษณะเช่นเดียวกัน คือ เมื่อสร้างเสร็จ แต่ละเซลล์ประสาทที่สร้างขึ้นมานั้น จะไม่มีคุณลักษณะใดเลย เนื่องจากยังไม่มีการกำหนดค่าเวจท์ที่เหมาะสมกับงานที่ต้องการให้กับมัน ค่าเวจท์ที่ให้กับโครงข่ายเพื่อให้โครงข่ายทำงานใดๆนั้น ในขั้นแรกนี้อาจเป็นค่าสุ่มใดๆ (Random weight) หรือให้เป็นไปตามสมการใดสมการหนึ่ง ซึ่งสอดคล้องกับเอาท์พุท (กรณีนี้ไม่ต้องปรับค่าเวจท์ ถ้าเทียบกับเด็กทารก อาจเทียบได้กับสัญชาติญาณ) กรณีให้โครงข่ายทำงานใดงานหนึ่งโดยเฉพาะ

วัตถุประสงค์ของการเทรนนิ่ง (Objective of Training)

เนื่องจากค่าเวจท์ที่ให้เป็นค่าสุ่มใดๆ โครงข่ายจึงไม่แสดงคุณลักษณะใดออกมา การฝึกสอน (Training) ให้โครงข่ายก็คือการปรับค่าเวจท์ทุก ๆ จุดให้สอดคล้องกับอินพุทหลายๆแบบ เพื่อให้ได้เอาท์พุทตามต้องการนั่นเอง การฝึกสอนโครงข่าย จะต้องบรรลุถึงขบวนการเข้าใจพื้นฐานเสียก่อน คือ การเรียนรู้ในโครงข่ายประสาทเทียมนั้นมีขีดจำกัด ปัญหาต่างๆ ผู้ใช้คงต้องแก้ให้มันก่อน แล้วนำผลนั้นไปอ้างอิงสำหรับการปรับปรุงค่าเวจท์ หลังจากปรับเวจท์จนได้ค่าผิดพลาดที่เอาท์พุทเทียบกับเป้าหมายน้อยลงเป็นที่พอใจแล้ว โครงข่ายประสาทเทียมนั้นก็พร้อมที่จะวิเคราะห์อินพุทและให้เอาท์พุทตามลักษณะตัวอย่างที่มันเคยเรียนรู้มา การเรียนรู้จะมีการปรับเวจท์หลายๆรอบ จนค่าเวจท์สอดคล้องกับตัวอย่างหลายๆ ตัวอย่าง และให้เอาท์พุทตามต้องการ พบว่าโครงข่ายได้ตัวอย่างสำหรับการเทรนนิ่งมากๆ โครงข่ายก็จะมีความแม่นยำสูงขึ้น แต่ก็จะใช้เวลาในการเทรนนิ่งเพิ่มขึ้นเช่นกัน หากพิจารณาต่อไปจะพบว่า โครงข่ายประสาทเทียมที่สร้างขึ้นจะมีพฤติกรรมคล้ายกับระบบการเรียนรู้ของมนุษย์มาก เป็นเพราะมีต้นแบบมาจากระบบประสาทชีวภาพนั่นเอง

การเทรนนิ่งแบบควบคุม (Supervised Training)

เทรนนิ่งอัลลกอลิที่มถูกจัดเป็น 2 ประเภท คือ แบบควบคุม (Supervised training) และแบบอิสระ (Unsupervised training) โดย การเทรนนิ่งแบบควบคุม จะต้องการคู่ของการเทรนนิ่งระหว่างอินพุทกับเป้าหมายที่ต้องการ ที่เรียกว่า เทรนนิ่งแพร์ (Training pairs) โครงข่ายจะถูกเทรนไปตามจำนวนของคู่ที่เทรนนิ่ง (จำนวนคู่ของ Input กับ Output ที่ต้องการให้โครงข่ายรู้จัก) เอาท์พุทที่คำนวณได้จากโครงข่ายจะถูกเปรียบเทียบกับความสอดคล้องกับเป้าหมาย ค่าผิดพลาดที่เกิดขึ้นจะถูกป้อนกลับไปยังโครงข่ายและเปลี่ยนแปลงค่าเวจท์ให้สอดคล้องกับอัลกอลิที่ม ที่ทำให้แนวโน้มของค่าผิดพลาดที่เกิดขึ้นระหว่างเอาท์พุทกับเป้าหมายโดยเฉลี่ยมีค่าลดต่ำลง ตัวอย่างการเทนนิ่งแบบนี้ ได้แก่ การเทรนนิ่งแบบแพร่กลับ (Back propagation)

การเทรนนิ่งแบบอิสระ (Unsupervised Training)

ถึงแม้ว่าอัลกอลิทึมแบบควบคุม (Supervised training) สามารถจะประยุกต์ใช้เพื่อปรับคุณลักษณะของโครงข่ายได้สำเร็จ แต่ก็ยังมีข้อวิจารณ์อยู่ คือ มันเป็นไปอย่างแบบชีวภาพไม่ได้ และยากที่จะเชื่อได้ว่า กลไกการเทรนนิ่งของสมองจะต้องการ การเปรียบเทียบระหว่างค่าที่ต้องการกับเอาท์พุทจริง โดยขบวนการป้อนกลับไปแก้ไขคุณลักษณะของโครงข่าย และถ้าสมมติว่า ถ้าสมองมีกลไกเช่นนี้ ต้องมีผู้หาเอาท์พุทที่ต้องการเพื่อนำมาเป็นเป้าหมายตลอดเวลา และจะเอามาจากที่ใด ? สรุป คือ ต้องมีผู้คิดเป้าหมายให้กับโครงข่ายก่อน โครงข่ายไม่สามารถคิดและปรับคุณลักษณะได้ก่อนด้วยตนเอง ในทางตรงกันข้ามหากพิจารณาทารกแรกเกิดสมองของเขาสามารถจัดระบบเองได้อย่างไร? การเทรนนิ่งแบบอิสระ (Unsupervise learning) ที่สร้างขึ้นคงยังห่างไกลความเป็นไปได้ ที่จะมีลักษณะการเทรนนิ่งแบบระบบของสมอง จนกระทั่งมีการพัฒนาการเทรนนิ่งแบบอิสระนี้ขึ้นราวปี 1984 โดยโคโฮเนน (Kohonen) และบุคคลอื่นๆ โดยได้เสนอแนวคิดที่เป็นการเทรนนิ่งแบบไม่ต้องการเป้าหมาย ไม่มีการตัดสินใจด้วยเหตุผลในอุดมคติมาก่อน ชุดของการเทรนนิ่ง จะมีเพียงอินพุท เวกเตอร์เท่านั้น เทรนนิ่งอัลกอลิทึมจะเปลี่ยนแปลงค่าเวจท์ของโครงข่าย เพื่อสร้างเอาท์พุทที่มีความมั่นคง ยกตัวอย่าง เช่น หากให้โครงข่ายรู้จำภาพหน้าคนหนึ่ง หากภาพหน้าคนคนนั้นเปลี่ยนแปลงไปเล็กน้อย (Image อาจมี Noise ร่วมอยู่บ้าง) โครงข่ายนั้นก็ยังสามารถบอกได้ว่า คนคนนั้นเป็นคนเดิมเป็นต้น การเทรนนิ่งจะไม่มีการตัดสินใจมาก่อน ไม่มีการกำหนดแบบเอาท์พุทมาก่อน (อาจกล่าวได้ว่าแบบเอาท์พุทจะถูกกำหนดโดยอินพุทเวกเตอร์นั่นเอง) ดังนั้น เอาท์พุทของโครงข่ายก็เช่นกัน ส่วนใหญ่จะถูกแปรรูปซึ่งจะเข้าใจได้ภายหลังขบวนการเทรนนิ่ง ดังนั้นจึงไม่สามารถแก้ปัญหาที่เคร่งครัดสำคัญได้ แต่มักนิยมใช้โครงข่ายแบบนี้กับงานง่ายๆ ประเภทการเปรียบเทียบเอกลักษณ์, รูปแบบที่สัมพันธ์กันระหว่างอินพุท-เอาท์พุท ที่ถูกกำหนดโดยโครงข่าย

วิธีการแก้ปัญหาการฝึกสอน (Traning Algorithm)

ส่วนใหญ่แล้วทุกวันนี้ การแก้ปัญหาฝึกสอนของโครงข่ายค่อยๆ พัฒนาก้าวหน้าขึ้นจากแนวความคิดของ ดี โอ เฮบบ์ (ปี 1961) เขาได้เสนอโมเดลของ การเทรนนิ่ง แบบอิสระ (Unsupervise training) ในแบบซินแนปติกส์ สเตรงท์ หรือ เวจท์ ซึ่งจะเพิ่มขึ้น ถ้าทั้งแหล่งกำเนิด (Input Source) และจุดหมายปลายทาง (Destination)ของนิวรอลได้รับการสนใจ กรณีนี้ถ้ามีการใช้งานทางเส้นนี้บ่อยๆก็จะทำให้ซินแนปติกส์สเตรงจ์ (หรือเวจท์) แข็งแรงขึ้น (เซลล์สมองที่ใช้งานมากบ่อยๆ ก็จะทำให้ ซินแนปส์ใหญ่ขึ้นการส่งผ่านข้อมูลพัลล์

ไฟฟ้าทำได้ดีขึ้น ทำให้มีประสิทธิภาพดีขึ้น เช่นสามารถคิดหรือจดจำได้เร็วและดีขึ้น) โครงข่ายประสาทเทียมนี้ใช้การเรียนรู้แบบเฮบเบน (Hebbian learning) จะเพิ่มค่าเวจท์ของโครงข่ายอย่างสอดคล้องกับผลคูณของระดับความสนใจของแหล่งกำเนิดและจุดหมายของนิวรอล ตามสมการดังนี้

$$W_{ij}(n+1) = W_{ij}(n) + \alpha OUT_i OUT_j \quad \text{.....(4.7)}$$

โดย $W_{ij}(n)$ คือค่าเวจท์ จากนิวรอล i ไปยังนิวรอล j ก่อนปรับปรุงค่า

$W_{ij}(n+1)$ คือค่าเวจท์ จากนิวรอล i ไปยังนิวรอล j หลังปรับปรุงค่า

α คือค่าคงที่ของการเรียนรู้ (Learning rate coefficient)

OUT_i คือเอ้าท์พุทของนิวรอล i และเป็นอินพุทของนิวรอล j

OUT_j คือเอ้าท์พุทของนิวรอล j

โครงข่ายที่มีลักษณะการเทรนนิ่งแบบ เฮบเบน นั้น เป็นผลมาจากการพัฒนามาแล้วกว่า 20-30 ปี โดยเฉพาะงานของโรเซนเบลทท์ (Rosenblatt:1962), วิโดรว์(Widrow:1959), วิโดรว์และฮอฟฟ์ (Widrow&Hoff:1960) และอีกหลายๆ คนที่พยายามพัฒนาระบบการเทรนนิ่งแบบควบคุม ที่สร้างโครงข่ายที่สามารถเรียนรู้แบบของอินพุทได้อย่างกว้างขวางและมีอัตราการเรียนรู้สูง ที่บรรลุผลได้จากหลักการพื้นฐานของการเรียนรู้หรือเทรนนิ่งให้กับโครงข่ายเช่น Perceptrons, Hopfieldnets, Backpropagation Networks และCounter propagation.

บรรณานุกรม

กฤษดา เรเยส ต้นแบบเครื่องสังเคราะห์เสียงพูดด้วยวิธีเข้ารหัสแบบลิเนียร์พรีดิกทีฟ วิทยานิพนธ์หลักสูตรปริญญาวิศวกรรมศาสตร์มหาบัณฑิต, ภาควิชาวิศวกรรมไฟฟ้า, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย, 2530.

พงษ์เทพ ธนกิจสุนทร, พีรพล แดงสุภา, ภัสสร ธัญญสิริ, ศรีรัตน์ ชูโชติถาวร การประมาณพันธะเชิงเส้น, โครงงานปีการศึกษา 2532, ภาควิชาวิศวกรรมคอมพิวเตอร์, คณะวิศวกรรมศาสตร์, สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2532.

ศุนย์ปัน โสมภีร์ ร.อ. ความก้าวหน้าใหม่ในเทคโนโลยีกรรมวิธีทางเสียงและการประยุกต์ใช้งาน, คอมพิวเตอร์ อิเลคทรอนิคส์ เวิลส์, นิตยสาร, ฉบับที่ 129, 2533, หน้า 156-159.

ทวี ประทุมทาน การตรวจรู้เสียงพูดภาษาไทย โดยใช้หน่วยพยางค์, วิทยานิพนธ์หลักสูตรปริญญาวิศวกรรมศาสตร์มหาบัณฑิต, ภาควิชาวิศวกรรมคอมพิวเตอร์, จุฬาลงกรณ์มหาวิทยาลัย, 2532.

ชัยศรี เอี่ยมอ่ำไพ, นพดล ทวีท่านุสิน การตรวจหาจุดเริ่มต้นและสิ้นสุดของคำโดย ๆ, โครงงานพิเศษหลักสูตรวิทยาศาสตร์บัณฑิต, ภาควิชาฟิสิกส์ประยุกต์, คณะครุศาสตร์อุตสาหกรรมและวิทยาศาสตร์, 2530.

Lawrence Rabiner. Digital Processing of Speech Signal, Prentice-Hall, 1978.