บทที่ 2

การศึกษาระบบประมวลผลเสียงพูด

2.1 การติดต่อสื่อสารโดยใช้เสียง

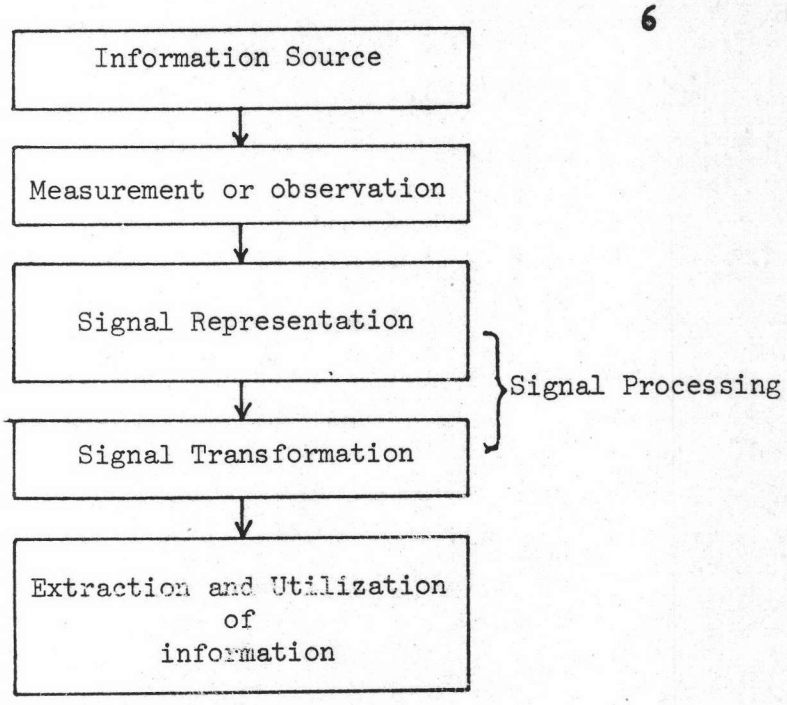
ในการติดต่อสื่อสารระหว่างมนุษย์นั้น ได้อาศัยการพูดสื่อความหมายระหว่างกันเป็นเวลานานมาแล้ว เราสามารถมองการพูดเป็นสัญญาณชนิดหนึ่งซึ่งพาข่าวสาร ขบวนการในการติดต่อสื่อสารโดยการพูด เริ่มจากการที่คนเราคิดข้อความซึ่งแทนบางสิ่งที่ต้องการจะให้ผู้อื่นเข้าใจ จากนั้นขบวนการพูดก็จะเกิดขึ้น ซึ่งสมองจะทำหน้าที่แปลข้อความนั้นเป็นสัญญาณควบคุมส่งมาทางระบบประสาท เพื่อออกคำสั่งให้อวัยวะที่ใช้ในการพูดทำงานไปตามลำดับขั้นของการพูด มนุษย์ได้มีการพัฒนาเทคโนโลยีไปอย่างมาก ทำให้มีการขยายขีดความสามารถในการติดต่อสื่อสารทางเสียง เสียงอาจถูกถ่ายทอดออกไปบันทึกไว้ หรือผ่านขบวนการต่าง ๆ ได้หลายแบบ โดยทั่ว ๆ ไป ระบบที่เกี่ยวกับเสียงจะมีความเกี่ยวพันในลักษณะคือ [3]

2.1.1 การรักษาข่าวสารซึ่งอยู่ในสัญญาณคำพูด

2.1.2 การแทนสัญญาณเสียงในอยู่ในรูปแบบที่เหมาะสมในการถ่ายทอด หรือบันทึกหรืออยู่ในรูปซึ่งง่ายต่อการดัดแปลงเพื่อประโยชน์อย่างใดอย่างหนึ่งโดยไม่ทำให้ข่าวสารที่ต้องการจะสื่อความนั้นเสียไป

2.2 การประมวลผลสัญญาณ

จุดมุ่งหมายของการประมวลผลสัญญาณ เพื่อที่นำข่าวสารซึ่งแฝงอยู่ในสัญญาณนั้นมาใช้ประโยชน์ ขบวนการที่ใช้ในการประมวลผลสัญญาณจะประกอบด้วย การนำสัญญาณมาผ่านขบวนการหนึ่งให้ได้สิ่งซึ่งแทนสัญญาณนั้น จากนั้นก็จะทำการเปลี่ยนแปลงให้อยู่ในรูปแบบซึ่งง่ายต่อการประมวลผลเพื่อดึงข่าวสาร และนำข่าวสารนั้นมาใช้ประโยชน์ในขั้นตอนสุดท้าย



รูปที่ 2.1 แสดงขั้นตอนโดยทั่วไปของระบบข่าวสาร

เนื่องจากเสียงเป็นสัญญาณชนิดหนึ่งซึ่งเป็นพาหนะของข่าวสาร ขบวนการซึ่งกระทำต่อสัญญาณเสียง จึงประกอบด้วยการแทนสัญญาณเสียงด้วยรูปแบบของข้อมูลอย่างหนึ่ง ซึ่งในขบวนการทางดิจิ ตอล สัญญาณเสียงจะถูกแทนด้วยข้อมูลทางดิจิ ตอลได้ใน 2 ลักษณะคือ [3]

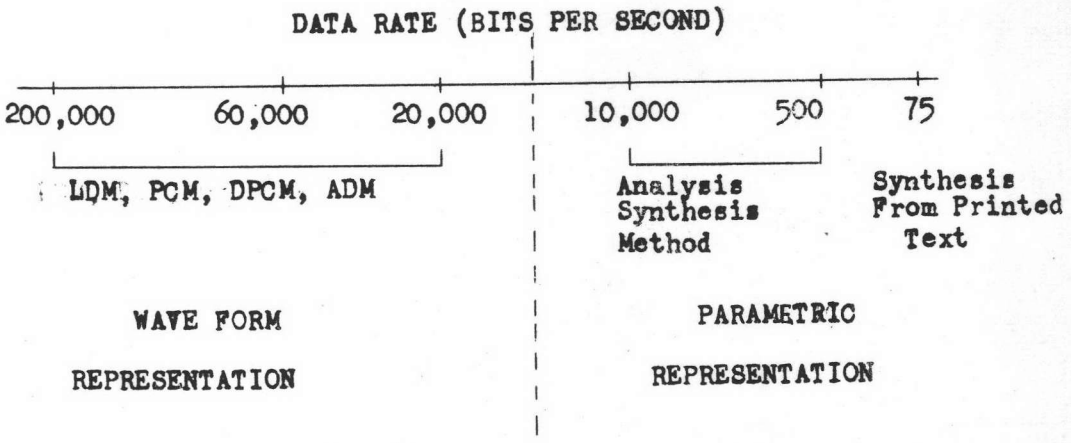
2.2.1 การแทนด้วยข้อมูลดิจิ ตอลตามลักษณะของรูปสัญญาณ วิธีการนี้รูปสัญญาณเสียงซึ่งเป็นสัญญาณอนาล็อกจะถูกสุ่มตัวอย่าง แล้วเปลี่ยนเป็นค่าทางดิจิ ตอลโดยตรง ค่าที่เก็บจะเป็นค่าตามรูปสัญญาณจริง

2.2.2 การแทนด้วยข้อมูลดิจิ ตอลโดยเก็บเฉพาะพารามิเตอร์ของสัญญาณ ในวิธีการนี้สัญญาณเสียงถูกผ่านขบวนการหนึ่ง ซึ่งจะให้ผลลัพท์เป็นค่าพารามิเตอร์ที่แทนการเกิดเสียงของคำ

การแทนด้วยข้อมูลดิจิ ตอลตามลักษณะของรูปสัญญาณ จำเป็นต้องสุ่มตัวอย่างอย่างละเอียด เพื่อให้ได้ข้อมูลที่แทนรูปสัญญาณได้ใกล้เคียงที่สุด วิธีการเหล่านี้ได้แก่ L D M (Linear delta modulation), P C M (pulse Code Modulation), A D M (Adaptive Delta Modulation), D P C M (Differential Pulse Code Modulation) เป็นต้น ส่วนการแทนด้วยข้อมูลดิจิ ตอล โดยเก็บเฉพาะ

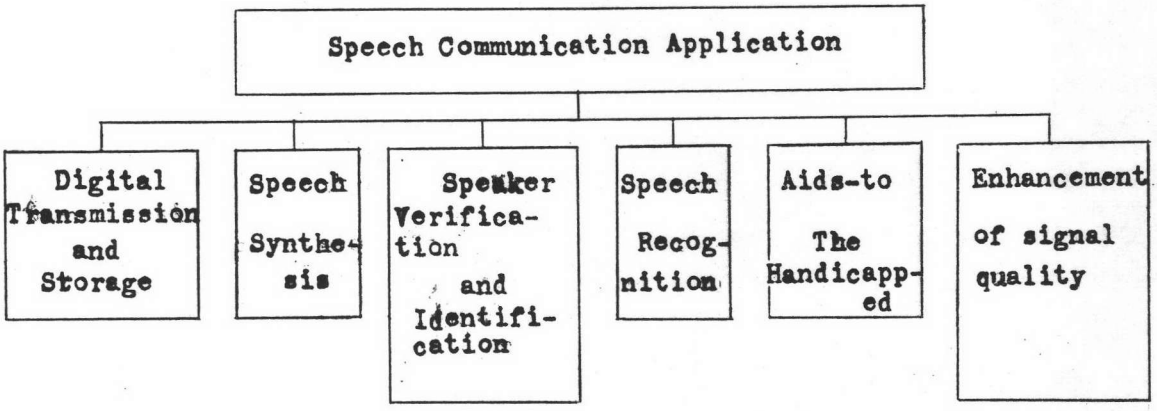
พารามิเตอร์ของสัญญาณ สามารถทำได้หลายลักษณะ เช่น การวิเคราะห์หาเสียงสระ เสียงพยัญชนะ อัตราการกระจายเสปคตรัม เป็นต้น

รูปที่ 2.2 แสดงการเปรียบเทียบความแตกต่างอัตราการแทนข้อมูลระหว่างทั้งสองวิธี



รูปที่ 2.2 แสดงขอบเขตของอัตราข้อมูลทางดิจิ ตอลที่ต้องใช้แทนสัญญาณคำพูด

การแปลงข้อมูลเสียงซึ่งถูกแทนเป็นข้อมูลดิจิ ตอลให้เป็นอีกลักษณะที่เหมาะสมตามขบวนการประมวลผลสัญญาณจะขึ้นกับการนำข่าวสารจากเสียงพูดไปใช้งาน การประยุกต์ไปใช้งานนี้ทำได้หลายแบบ ดังแสดงในรูปที่ 2.3



รูปที่ 2.3 แสดงการประยุกต์การประมวลผลเสียงพูดไปใช้งาน

2.3 ประวัติการค้นคว้าเกี่ยวกับระบบจดจำเสียง

การค้นคว้า พัฒนาเกี่ยวกับระบบจดจำเสียงนี้ได้กระทำมานานกว่า 30 ปีแล้ว และได้มีการสร้างระบบและทดลองต่าง ๆ มากมาย บางส่วนของการวิจัยเหล่านี้ ได้แก่ [4]

ในปี ค.ศ. 1952 Davis, Wiren และ Biddulph จาก Bell laboratory ได้สร้างระบบจดจำ ซึ่งสามารถจดจำเสียงพูดตัวเลขได้ 10 ตัว สามารถจำคำพูดจากผู้พูด 1 คน ได้ถูกต้อง 100% โดยผู้พูดผ่านโทรศัพท์ ความถูกต้องจะลดลงต่ำกว่า 50% ในกรณีที่เปลี่ยนคนพูด

ในปี ค.ศ. 1956 Wiren และ Stubbs จาก Northeastern University ได้สร้างเครื่องจดจำแยกประเภทของการออกเสียง เช่น คำที่เปล่งเสียง (voiced) ไม่เปล่งเสียง (unvoiced) คำที่มีการหยุดเสียง (stop) และลากเสียง (fricative) เสียงแหลม (acute) เสียงต่ำ (grave) เป็นต้น ระบบสามารถจดจำได้อย่างถูกต้อง 94% สำหรับผู้พูด 24 คน โดยจำเสียงสระ ในปีเดียวกัน Olson และ Belar จาก R C A ได้สร้างระบบสามารถจดจำคำ 1 พยางค์ ได้ 10 คำ มีความถูกต้อง 98% สำหรับผู้พูด 1 คน โดยผู้พูดจะต้องระวังการออกเสียงแต่ละพยางค์ในการพูดประโยค 1 ประโยค และจะต้องหยุดระหว่างคำ ระบบใช้วงจรกรองแถบความถี่ 8 วงจร และแบ่งการวิเคราะห์ออกเป็น 5 ช่วงเวลาต่อคำ ซึ่งจะได้แมทริก (Matrix) ขนาด 8 x 5 ใช้ในการวิเคราะห์ แต่ละสมาชิกของแมทริกจะมีค่า 1 หรือ 0 ขึ้นกับค่าพลังงานของแต่ละแถบความถี่จะมีค่ามากกว่าหรือน้อยกว่าระดับที่กำหนดวัด

ในปี ค.ศ. 1960 Danes และ Mathew ได้สร้างระบบจดจำเลข 10 จำนวน โดยใช้เครื่องวิเคราะห์เสปคตรัม 17 ช่องความถี่ จับสัญญาณและบันทึกข้อมูลลงบนเทปแม่เหล็กจากนั้นนำข้อมูลเข้าสู่คอมพิวเตอร์เพื่อทำการวิเคราะห์ด้านความถี่และเวลา แล้วเก็บบันทึกไว้เป็นแบบอ้างอิง คำพูดใหม่ที่เข้ามาจะถูกเปรียบเทียบโดยขบวนการครอสโคริเลชั่น (crosscorrelation process) กับแต่ละรูปแบบที่เก็บไว้ ซึ่งระบบจะเลือกรูปแบบที่คล้ายกันมากที่สุดกับรูปแบบของคำพูดใหม่ การเปรียบเทียบ

กระทำโดยใช้การนอมัลไลซ์ (normalization) และโดยไม่นอมัลไลซ์

ในปี ค.ศ. 1966 King และ Tunis ได้สร้างเครื่องจำคำ (word recognition) ใช้เครื่องวิเคราะห์เสปคตรัม 15 ช่อง และคอมพิวเตอร์ช่วยในการวิเคราะห์ โปรแกรมคอมพิวเตอร์ทำหน้าที่ตรวจจับยอดที่เกิดขึ้นในแต่ละแถบความถี่ โดยหาตำแหน่งที่เกิด สร้างเป็นแมทริกซ์ตามความถี่ – เวลา (frequency-time matric) ซึ่งสมาชิกจะมีค่าเป็น 1 ถ้าในแถบความถี่ ณ เวลานั้นเกิดยอดคลื่นขึ้น และเป็น 0 ถ้าไม่มียอด ผลลัพท์ให้ความผิดพลาด 0.2% เมื่อตัวอย่างเสียงเก็บจากผู้พูดคนเดียวกัน และผิดพลาดมากที่สุด 2.5% การสอนให้ระบบรู้จักจะใช้คำพูด 15 คำ แต่ละคำจะให้ตัวอย่าง 35 ถึง 50 ครั้ง การทดลองให้ผู้ทดลองพูดคำต่าง ๆ 90 ถึง 115 ครั้ง ผลลัพท์จะผิดพลาดมากที่สุดเมื่อตัวอย่างเสียงเก็บจากคนหนึ่ง และใช้ผู้พูดอีกคนหนึ่ง ความผิดพลาดจะเกิดขึ้นประมาณ 46% แต่เมื่อการเก็บตัวอย่างเสียงกระทำจากคนทั้ง 2 คน ความผิดพลาดจะมีเพียง 0.8%

ในปี ค.ศ. 1966 Fraipont ได้สร้างเครื่องจำคำ โดยใช้วงจรกรองแถบความถี่ 10 แถบ กรองความถี่ในช่วง 300 Hz ถึง 3000 Hz โดยจับเฉพาะแนวยอดคลื่น (envelope) และใช้วงจรกรองแถบความถี่อีก 2 วงจร สำหรับความถี่สูงกว่า 3000 Hz การสุ่มตัวอย่างสัญญาณใช้เวลา 500 msec การสุ่มคำพูดกระทำ 192 ครั้ง ข้อมูลถูกบันทึกลงบนเทปกระดาษแล้วป้อนเข้าสู่คอมพิวเตอร์ ซึ่งทำการวิเคราะห์โดยใช้ลิเนียร์ดิซิชั่น ไฮเปอร์เพลน (linear decision hyperplane) ซึ่งจะถูกปรับให้เหมาะสมระหว่างการสอนระบบ ระบบถูกสร้างให้จำตัวเลขได้ 10 ตัว และคำอื่น ๆ อีก 5 คำ ข้อมูลตัวอย่างเก็บจากคน 24 คน แต่ละคนจะพูดคำหนึ่งคำเพียงครั้งเดียว หลังจากสอนระบบเรียบร้อยแล้ว ระบบจะสามารถจดจำได้ถูกต้อง 99% สำหรับผู้พูดกลุ่มเดิม และถูกต้อง 87% เมื่อเปลี่ยนกลุ่มผู้พูด

2.4 ระบบจดจำคำพูด

ระบบจดจำคำพูดโดยทั่ว ๆ ไป จะใช้วิธีการเปรียบเทียบพารามิเตอร์หรือองค์ประกอบบางอย่างที่ใช้แทนเสียงพูดที่เข้ามาสู่ระบบกับลักษณะต้นแบบแต่ละอันที่ได้สร้าง

รูปที่ 2.4 แสดงขั้นตอนของระบบจดจำคำพูดที่ใช้กันทั่ว ๆ ไป สิ่งที่ต้องนำมาพิจารณาในการออกแบบระบบ ได้แก่ [3]

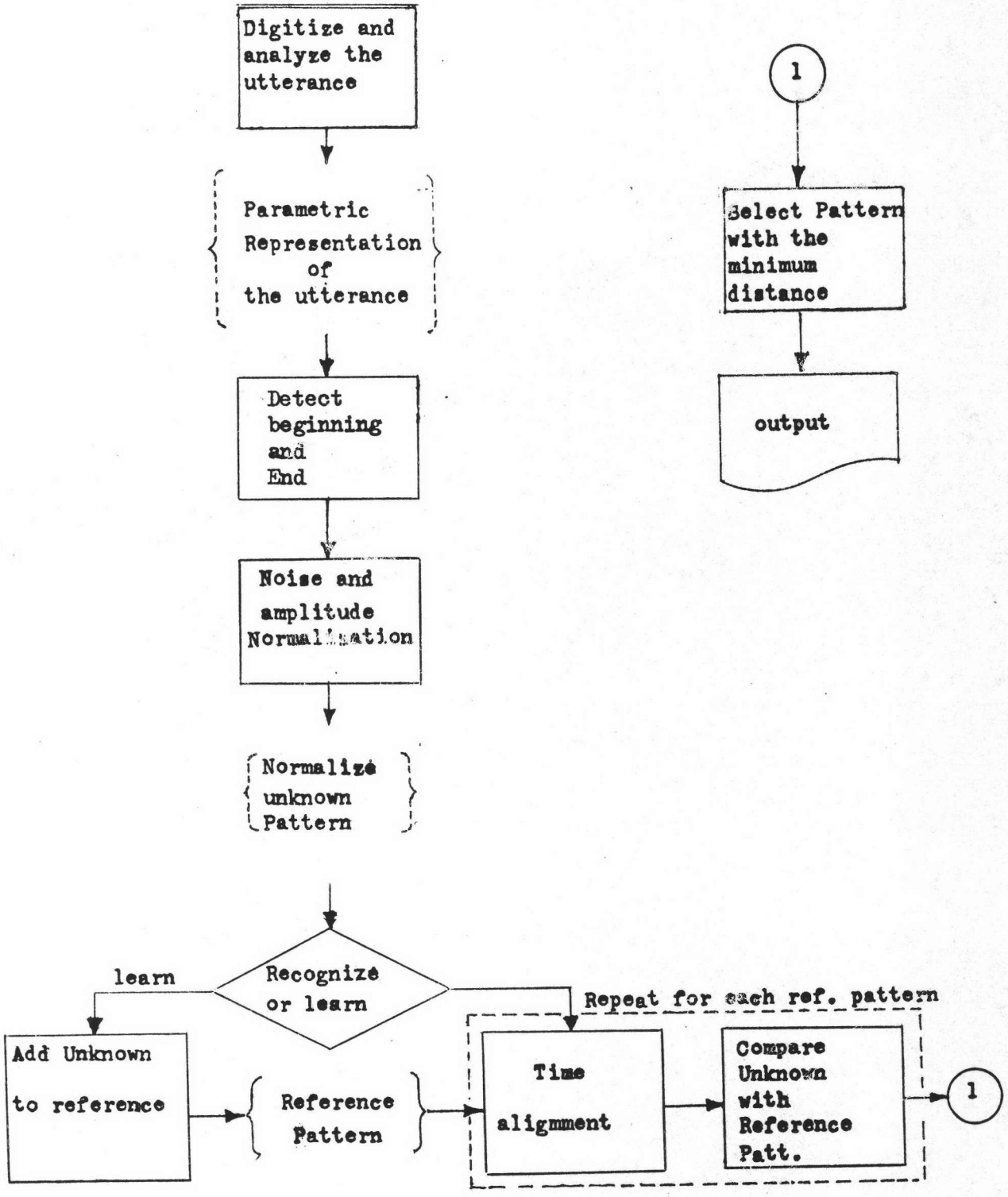
2.4.1 การนอมัลไลซ์ (nomalization) เนื่องจากในการพูดของคนเราจะมีการเปลี่ยนแปลงอยู่เสมอ ถึงแม้ว่าผู้พูดจะเป็นคนเดิม และใช้ไมโครโฟนอันเดิมก็ตาม สาเหตุอาจมาจากหลายทาง เช่น การเปลี่ยนแปลงอารมณ์ของผู้พูด หรือเสียงรบกวน เป็นต้น สิ่งเหล่านี้มีผลต่อขนาดสัญญาณ ระยะเวลาการออกเสียง และ S/N (Signal to Noise Ratio) เป็นต้น ดังนั้นก่อนที่จะทำการเปรียบเทียบสัญญาณที่ได้รับกับสิ่งที่เก็บไว้แล้ว บางครั้งจำเป็นต้องทำการนอมัลไลซ์สัญญาณก่อน เพื่อลดความแปรปรวนลง

2.4.2 การแทนสัญญาณในรูปของพารามิเตอร์ จากรูปที่ 2.2 ได้แสดงให้เห็นแล้วว่าการแทนข้อมูลตามรูปลักษณะของสัญญาณเสียงจำเป็นต้องใช้ข้อมูลจำนวนมาก การแทนในลักษณะของพารามิเตอร์จะช่วยให้ลดจำนวนข้อมูล และช่วยความจำที่ใช้เก็บข้อมูลลงได้อย่างมาก อย่างไรก็ดี การแทนด้วยพารามิเตอร์ จะทำให้เกิดความผิดพลาดเพิ่มขึ้น การพิจารณาหาพารามิเตอร์ที่เหมาะสมจะช่วยให้เกิดความผิดพลาดที่น้อยที่สุด

2.4.3 การสอนให้ระบบเรียนรู้ เนื่องจากระบบทั่วไปจำเป็นต้องมีแบบแผนอ้างอิง (reference pattern)เพื่อใช้ในการเปรียบเทียบและตัดสินใจเลือกความหมายที่เหมาะสมกับสัญญาณเสียงที่รับเข้ามา แบบแผนอ้างอิงเพียง 1 แบบแผน ไม่สามารถครอบคลุมความแปรปรวนในการออกเสียงของผู้พูด แม้จะเพียงคนเดียวได้ ในขณะที่ระบบตอบสนองผิดพลาดมากขึ้น ก็จำเป็นต้องเก็บตัวอย่างอ้างอิงเพิ่มขึ้นด้วย เพื่อลดความผิดพลาดให้น้อยลง

2.4.4 การเลือกขั้นตอนที่เหมาะสมในการเปรียบเทียบ เมื่อมีคำพูดมาให้เปรียบเทียบ 1 คำ จำเป็นต้องเปรียบเทียบคำพูดนั้นกับทุก ๆ รูปแบบที่เก็บไว้ เพื่อหารูปแบบ ที่เหมือนกัน เช่น ใช้วิธีการหาความแตกต่างที่น้อยที่สุด หรือรูปแบบอ้างอิงที่มีความสัมพันธ์ร่วมกันมากที่สุดกับเสียงที่เข้ามาใหม่ เป็นต้น ขั้นตอนที่เหมาะสมจะมีผลต่อความถูกต้องในการจดจำและเวลาที่ใช้ในการตอบสนองของระบบ เนื่องจาก

การเปรียบเทียบต้องอาศัยการคำนวณ และในกรณีที่มีรูปแบบอ้างอิงจำนวนมาก การเปรียบเทียบกับรูปแบบทั้งหมดจะใช้เวลา



รูปที่ 2.4 แสดงขั้นตอนของระบบจดจำคำพูด