บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื่องจากการค้นหาคำหลักในสัญญาณเสียงมีพื้นฐานมาจากระบบรู้จำเสียง ดังนั้นในส่วนแรกของบทนี้จะอธิบายถึงการรู้จำเสียง ความแตกต่างระหว่างการรู้จำเสียงกับการค้นหาคำหลักในสัญญาณเสียง และทฤษฎีที่ใช้ในวิทยานิพนธ์ ส่วนที่สองจะอธิบายถึงงานวิจัยอื่นที่เกี่ยวข้องกับการค้นหาคำหลักในสัญญาณเสียง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การรู้จำเสียง (Speech Recognition)

การรู้จำเสียง (Speech Recognition) คือ การประมวลผลโดยแปลงสัญญาณเสียงพูดให้อยู่ในรูปแบบของคำที่มีการเรียงลำดับต่อกันด้วยอัลกอริธึ่มที่ใช้ในโปรแกรมคอมพิวเตอร์ (Wikipedia, 2007) เช่น การป้อนเสียงที่เป็นตัวเลขของชั้นบนลิฟต์โดยสารอัจฉริยะ ผ่านไมโครโฟน เมื่อระบบรู้จำเสียงได้รับเสียงเข้าก็จะทำการประมวลผลให้ลิฟต์โดยสารขึ้นลงตามเลขชั้นที่ได้รับเข้ามา โดยตัวเลขชั้นต่าง ๆ จะต้องมีการสอน (Training) ให้ระบบรู้จำก่อน เพื่อให้คอมพิวเตอร์สามารถเรียนรู้เสียงพูดจากคน ซึ่งจะแตกต่างไปตามลักษณะของคน เช่น เพศ อายุ แล้วจดจำเอาไว้ใช้ในอนาคต เนื่องจากข้อจำกัดดังกล่าว จึงทำให้ระบบรู้จำเสียงในยุคแรก มีความสามารถแยกแยะคำได้อย่างถูกต้องก็ต่อเมื่อผู้พูดจะต้องมีการเว้นจังหวะคำพูด หรือพูดทีละคำ (Isolated Word) ต่อมาได้มีการพัฒนาระบบรู้จำเสียงให้สามารถรู้จำเสียงพูดแบบต่อเนื่อง (Continuous Speech) ได้

ประเภทของระบบรู้จำเสียง (Speech Recognition) สามารถแบ่งตามลักษณะการรู้จำเสียงได้ 3 ประเภท คือ

1. การรู้จำเสียงพูดจากสัญญาณเสียงพูดแบบคำโดด (Isolated speech) เป็นการฝึกฝนให้ระบบรู้จำเสียงพูดเป็นคำพูดสั้น ๆ หรือรู้จำคำสั่งเพียงไม่กี่คำสั่ง เพื่อให้ระบบสามารถโต้ตอบกับมนุษย์ได้รวดเร็ว เช่น

- ระบบที่ใช้เสียงในการระบุตัวบุคคล (ชัย วุฒิวิวัฒน์ชัย, สุทัศน์ แซ่ตั้ง และวารินทร์ อัจฉริยะกุลพร, 2543, น. 24-35) โดยระบบสามารถที่จะตรวจสอบได้ว่าเสียงของผู้นั้นเป็นเสียงของบุคคลใด
- การควบคุมการปฏิบัติการของแขนกลหุ่นยนต์คอมพิวเตอร์ด้วยคำสั่งเสียงภาษาไทย (ชมทิพ พรพนมชัย, ธรรมรัตน์ แสงโสภี และธีรวิชช์ วงษ์เสรี, 2548, น. 1-8) โดยการรู้จำเสียงมาใช้ควบคุมการปฏิบัติงานของแขนกลหุ่นยนต์ด้วยคำสั่งเสียงรวมทั้งหมด 6 คำสั่ง ได้แก่ ขึ้น ลง หยิบ ปล่อย ซ้าย และ ขวา
- ระบบ d-Ear Attendant (d-Ear Technologies, 2005) ซึ่งเป็นระบบตอบรับโทรศัพท์อัตโนมัติ เมื่อมีผู้โทรศัพท์เข้ามา ระบบจะให้ผู้นั้นบอกชื่อของผู้ที่ต้องการจะสนทนาด้วย จากนั้นระบบจะโอนสายไปยังผู้ที่ถูกระบุชื่อ เป็นต้น

2. การรู้จำเสียงพูดจากสัญญาณเสียงแบบพูดต่อเนื่อง (Continuous speech) เป็นฝึกฝนให้ระบบรู้จำมีความสามารถเพิ่มขึ้น โดยสามารถรู้จำคำทุกคำจากเสียงพูดแบบต่อเนื่องและทำการพิจาณาว่าสัญญาณเสียงที่เข้ามานั้นประกอบด้วยคำอะไรบ้าง เช่น

- โปรแกรมระบบ VoiceNotes (Lisa J. Stifelman, Barry Arons, Chris Schmandt and Eric A. Hulteen, 1993) สำหรับคอมพิวเตอร์มือถือ (Hand-Held Computer) ซึ่งโปรแกรมนี้จะช่วยอำนวยความสะดวกให้กับผู้ใช้งานที่อยู่สภาวะที่ไม่สามารถใช้ปากกา และกระดาษ เพื่อจดบันทึกข้อความ โดยผู้ใช้สามารถทำการจดบันทึกข้อความด้วยเสียงพูดแทน เช่น ผู้ที่กำลังขับขี่รถยนต์ เป็นต้น โดยก่อนที่โปรแกรมจะบันทึกข้อความเสียงนั้นลงในคอมพิวเตอร์มือถือ โปรแกรมจะทำการพิจารณาข้อความเสียงนั้น ก่อนว่าจะบันทึกลงหมวดหมู่ใด ตามหมวดหมู่ที่ผู้ใช้ได้กำหนดไว้ก่อนแล้ว
- ระบบรู้จำเสียงพูดภาษาไทยต่อเนื่องแบบเฉพาะบุคคลสำหรับการเข้าถึงอีเมล์ (สุนียา สัตยพานิช และ อัศนีย์ ก่อตระกูล, 2546) โดยการใช้เสียงพูดทั้งชายและหญิง สั่งงานอีเมล์ด้วยคำสั่งที่กำหนดขึ้นจำนวน 23 คำสั่ง เป็นต้น

3. การรู้จำเสียงพูดจากสัญญาณเสียงพูดแบบคำอุทาน (Spontaneous speech) เป็นการฝึกฝนระบบให้รู้จำคำจากเสียงพูดที่มีคำอุทาน เพื่อให้ระบบรู้จำสามารถรู้จำและเข้าใจเนื้อหาสำคัญที่อยู่ในสัญญาณเสียง เช่น ประโยคเสียงว่า “**เออ** ผมขอจองห้องพักหนึ่งห้อง” มีเนื้อหาคำ

สำคัญในเสียงคือ “ผมขอจองห้องพักหนึ่งห้อง” ส่วน “เออ” เป็นคำอุทาน ซึ่งไม่ใช่เนื้อหาสำคัญในสัญญาณเสียง ตัวอย่างเช่น

- ระบบสรุปข้อความในเสียงพูดแบบคำอุทาน (Sadaoki Furui, 2005) งานวิจัยนี้จะทำการถอดข้อความในสัญญาณเสียงพูดภาษาญี่ปุ่นแบบมีคำอุทานมาเป็นแฟ้มข้อความ ด้วยวิธีการสร้างโมเดลเสียง และโมเดลภาษาจากฐานข้อมูลคำอุทานภาษาญี่ปุ่น (Corpus of Spontaneous Japanese : CSJ) เป็นต้น

ความยากง่ายในการรู้จำเสียง (Speech Recognition) ยังมีปัจจัยประกอบอื่น ๆ เช่น อารมณ์และน้ำเสียงของผู้พูดในขณะนั้น ซึ่งจะทำให้คำคำเดียวมีการออกเสียงที่ต่างกันไปได้อีกด้วย อีกทั้งระบบรู้จำเสียงที่กล่าวข้างต้นทั้งหมดจำเป็นต้องอาศัยฐานข้อมูลเสียงขนาดใหญ่ เพื่อจะทำให้ระบบรู้จำเสียงให้ได้มากที่สุด ดังนั้นจำเป็นที่ต้องใช้เนื้อที่จัดเก็บ และใช้เวลามากในการสอน (Training) คอมพิวเตอร์ ให้รู้จำเสียง เนื่องจากประสิทธิภาพของระบบรู้จำจะขึ้นอยู่กับจำนวนของคำที่ต้องการให้ระบบรู้จำ ตัวอย่างเช่น ในงานการควบคุมการปฏิบัติการของแขนกลหุ่นยนต์คอมพิวเตอร์ด้วยคำสั่งเสียงภาษาไทย (ชมทิพ พรพนมชัย และคณะ, 2548) ถ้าต้องการควบคุมปฏิบัติงานของแขนกลหุ่นยนต์ให้มีประสิทธิภาพมากขึ้น เช่น ขึ้น 5 เซนติเมตร จำเป็นต้องสอนให้หุ่นยนต์รู้จำเพิ่มในเรื่องของระยะทางนอกเหนือจากคำสั่งที่กำหนดไว้เพียง 6 คำสั่ง

ดังนั้นเพื่อแก้ไขข้อจำกัดของระบบรู้จำที่จะต้องรู้จำเสียงจากฐานข้อมูลเสียงขนาดใหญ่และต้องใช้เวลาในการฝึกฝนคอมพิวเตอร์ผันแปรตามขนาดของฐานข้อมูลเสียง ในการวิจัยเกี่ยวกับค้นหาคำหลัก (Keyword Spotting) ในเสียงพูดนั้น จะอยู่บนพื้นฐานการฝึกฝนของระบบรู้จำเสียง โดยการสอนคอมพิวเตอร์ให้รู้จำเฉพาะคำหลักที่ต้องการเท่านั้น จึงทำให้ประหยัดเนื้อที่ในการจัดเก็บ และประหยัดเวลาในการสอน เนื่องจากระบบไม่จำเป็นเรียนรู้คำทุกคำ แต่ถ้าคำหลักที่ต้องการค้นหาไม่ได้ถูกสอนให้คอมพิวเตอร์รู้จัก ก็จะทำให้ระบบค้นหาคำหลักนั้นไม่พบ จำเป็นจะต้องสอนคอมพิวเตอร์ให้รู้จำคำหลักที่เพิ่มเข้ามาใหม่ ดังนั้นถ้าระบบค้นหาคำหลักสามารถสร้างคำหลักที่ไม่ได้ถูกสอนได้อัตโนมัติ ก็จะทำให้ประหยัดเวลาในการสอนคอมพิวเตอร์ทุกครั้งเมื่อมีคำหลักใหม่เข้ามา

สำหรับการแสดงขนาดของฐานข้อมูลเสียงตามจำนวนคำศัพท์ในฐานข้อมูลได้แสดงไว้ในตารางที่ 2.1

ตารางที่ 2.1

ตารางแสดงขนาดของฐานข้อมูลเสียงตามจำนวนคำศัพท์ในฐานข้อมูล

ขนาด	จำนวนคำศัพท์ในฐานข้อมูล
ขนาดเล็ก (small vocabulary)	10 คำ
ขนาดกลาง (medium vocabulary)	100 คำ
ขนาดใหญ่ (large vocabulary)	1,000 คำ
ขนาดใหญ่มาก (very-large vocabulary)	10,000 คำ

ที่มา : Andrew Hunt, 2004

ความสามารถของระบบรู้จำเสียง ในการที่จะรู้จำคำพูดว่าผู้พูดพูดอะไรนั้น และคำพูดแต่ละคำมีความแตกต่างกันอย่างไรนั้น ได้มีการนำเทคนิคหลายเทคนิคเข้ามาใช้ เช่น

- ใช้เทคนิคโมเดลของมาร์คอฟ (Hidden Markov Model : HMM)
- ใช้เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network)
- ใช้เทคนิคโดยให้คอมพิวเตอร์เรียนรู้ด้วยตนเอง (Machine Learning) เป็นต้น

นับเป็นเรื่องสำคัญมากในการที่จะจำลองเสียงพูดมาให้คอมพิวเตอร์รู้จำ เนื่องจากเสียงพูดเป็นสัญญาณที่มีความผันผวนตามเวลา ดังนั้น จึงจะต้องทำการกำหนดรูปแบบการของการจำลองเสียงพูดให้มีความสามารถที่จะแสดงให้เห็นถึงคุณสมบัติเด่นของเสียง โดยโมเดลนั้นจะต้องสามารถจำลองข้อมูลที่มีความผันผวนตามเวลา และใช้งานได้อย่างเหมาะสม โมเดลที่ว่านี้คือ โมเดลฮิดเด็นมาร์คอฟ (Hidden Markov Model : HMM)

2.1.1.1 โมเดลฮิดเด็นมาร์คอฟ (Hidden Markov Model : HMM)

โมเดลฮิดเด็นมาร์คอฟ สามารถนำไปใช้เพื่อจำลองคำ (Word) หรือหน่วยเสียง (Phoneme) ก็ได้ ขึ้นอยู่กับการใช้งานของระบบการรู้จำเสียงเช่น ถ้าระบบรู้จำเสียงมีข้อจำกัดคือให้รู้จำคำศัพท์ที่น้อยและรู้จำคำพูดที่ไม่ติดต่อกัน (Isolated Speech) การทำโมเดลเสียงเป็นคำเสียงจะดีกว่า เพราะจะให้ความแม่นยำมากกว่า ส่วนการทำโมเดลหน่วยเสียงจะทำได้ยากกว่าการจำลองคำ เพราะจะต้องทำการสร้างแบบหน่วยเสียงแต่หน่วยเสียงขึ้นมาก่อนที่จะนำไปรู้จำ

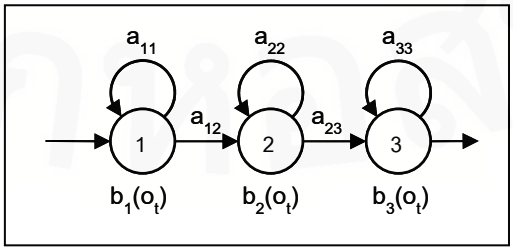
หน่วยเสียงนั้นๆ เมื่อรู้จำหน่วยเสียงได้แล้วก็จะต้องมีวิธีการสร้างคำจากหน่วยเสียงที่ได้รู้จำมา ว่าคำแต่ละคำประกอบไปด้วยหน่วยเสียงอะไรบ้าง ถ้ามีการรู้จำหน่วยเสียงหนึ่งผิดไปก็จะทำให้คำคำนั้นรู้จำผิดได้ แต่สำหรับการทำโมเดลคำเสียง ความแม่นยำในการรู้จำจะขึ้นตรงกับโมเดลคำเสียงที่ได้ฝึกฝนมา จึงไม่ต้องทำการสร้างคำจากหน่วยเสียง แต่ถ้าหากระบบที่มีคำศัพท์มาก (มากกว่า 1,000 คำ) และต้องการรู้จำเสียงพูดแบบต่อเนื่อง (Continuous Speech) แล้ว ระบบรู้จำที่ใช้โมเดลหน่วยเสียงจะมีประสิทธิภาพมากกว่า เพราะจำนวนโมเดลหน่วยเสียงที่ต้องการจะมีน้อยกว่าโมเดลที่สร้างจากคำเสียง และมีความเป็นไปได้สูงที่คำหนึ่งคำจะเปลี่ยนไปเมื่อใช้กับคำๆ อื่น ถ้าหากระบบมีการฝึกฝนที่ไม่ดีก็จะทำให้ระบบรู้จำมีความแม่นยำต่ำ แต่สำหรับหน่วยเสียงนั้นจะให้ความแม่นยำมากกว่า เนื่องจากมีเปลี่ยนแปลงน้อยกว่าเมื่อนำไปใช้กับคำ ๆ อื่น

ทฤษฎีโมเดลฮิดเด็นมาร์คอฟ คือ การเก็บรวบรวมของสถานะหลาย ๆ สถานะที่ถูกเชื่อมโยงโดยการเปลี่ยนสถานะในที่นี้คือ ลักษณะของสัญญาณเสียงหรือเครื่องหมายที่หมายถึงสัญญาณเสียงนั้น ณ เวลาหนึ่งๆ ส่วนการเปลี่ยนสถานะคือ การเลื่อนหรือเปลี่ยนแปลงของสัญญาณเสียงจากเวลาหนึ่งไปหาอีกเวลาหนึ่ง สำหรับการเปลี่ยนแปลงสถานะสามารถเปลี่ยนสถานะจากสถานะเดิมไปสถานะถัดไป (ข้างหน้า) หรือวนอยู่ที่สถานะเดิมเท่านั้น ไม่มีการย้อนกลับ จึงเรียกโมเดลฮิดเด็นมาร์คอฟแบบนี้ว่า Left-to-right model

โมเดลฮิดเด็นมาร์คอฟ มีค่าความน่าจะเป็นอยู่สองชนิดคือ ค่าความน่าจะเป็นที่จะมีการเปลี่ยนสถานะจากสถานะหนึ่งไปอีกสถานะหนึ่ง (Transition Probability) และ ทุก ๆ สถานะสามารถให้ผลลัพธ์ โดยใช้ค่าความน่าจะเป็นที่เรียกว่า ค่าความน่าจะเป็นของผลลัพธ์ (Output Symbol) เมื่อมีการเปลี่ยนสถานะเกิดขึ้น ผลลัพธ์ที่ว่านี้จะนำไปใช้ในการตัดสินใจว่าเสียงที่ได้ยินนั้นเป็นเสียงอะไร (Output Probability)

ภาพที่ 2.1

Left-to-right HMM



จากภาพที่ 2.1 เป็นตัวอย่างโมเดลฮิดเด็นมาร์คอฟรูปแบบมาตรฐานสำหรับแทนหน่วยเสียง 1 หน่วยเสียง ซึ่งสัญญาณเสียงที่ผ่านการสกัดลักษณะสำคัญของเสียงแล้วจะถูกป้อนเข้าทางซ้าย หรือ เข้าทางสถานะที่ 1 และออกทางสถานะสุดท้าย โดยที่ a_{ij} คือ ค่าความน่าจะเป็นที่จะมีการเปลี่ยนสถานะจากสถานะหนึ่งไปอีกสถานะหนึ่ง (Transition Probability) และ จะมีความน่าจะเป็นของผลลัพธ์คือ $b_i(o_t)$

2.1.1.2 **สัญลักษณ์แทนการออกเสียงพยัญชนะและสระ**

สัญลักษณ์การออกเสียง คือ สัญลักษณ์ที่ใช้แทนเสียงของหน่วยเสียง เช่น “k” แทน “ก” สำหรับสัญลักษณ์ของเสียงพยัญชนะและสระที่ใช้สร้างแบบจำลองเสียง (Acoustic Model) ในภาษาไทยสำหรับระบบรู้จำเสียงต่อเนื่อง มีจำนวนทั้งสิ้น 74 หน่วย สามารถแบ่งออกได้เป็น

1. สัญลักษณ์แทนการออกเสียงพยัญชนะต้นจำนวน 26 หน่วยเสียง และพยัญชนะท้าย จำนวน 12 หน่วยเสียง ดังแสดงไว้ในตารางที่ 2.2
2. สัญลักษณ์แทนการออกเสียงพยัญชนะควบกล้ำ จำนวน12 หน่วยเสียง ดังแสดงไว้ในตารางที่ 2.3
3. สัญลักษณ์แทนการออกเสียงพยัญชนะที่เป็นตัวสะกด จำนวน 12 หน่วยเสียง และสัญลักษณ์การออกเสียงสระประสม จำนวน 6 หน่วยเสียง รวมเป็น 24 หน่วยเสียง ดังแสดงไว้ในตาราง 2.4

ตารางที่ 2.2

สัญลักษณ์แทนเสียงพยัญชนะต้นและพยัญชนะท้าย

ในภาษาไทย (26 หน่วยเสียง และ 12 หน่วยเสียง)

พยัญชนะ	หน่วยเสียง		พยัญชนะ	หน่วยเสียง	
	พยัญชนะต้น	พยัญชนะท้าย		พยัญชนะต้น	พยัญชนะท้าย
ก	k	k^	บ	b	b^
ข, ค, ฆ	kh	k^	ป	p	p^
ง	ng	ng^	ผ, พ, ภ	ph	p^
จ	c	t^	ฝ, ฟ	f	p^
ฉ, ฌ, ช	ch	t^	ม	m	m^
ซ, ศ, ษ, ส	s	t^	ร	r	n^
ญ, ย	j	j^	ล, ฬ	l	n^
ฎ, ด	d	t^	ว	w	w^
ฏ, ต	t	t^	ห, ฮ	h	-
ฐ, ฑ, ฒ,ถ, ท, ธ	th	t^	อ	z	-
ณ, น	n	n^	หน่วยเสียงทับศัพท์	br, bl, fr, fl, dr	f^, s^, h^,l^

ที่มา : ภัชริภา คชสำโรง, ตรีภพ สรรเพชรนิยม, ศวิต กาสุริยะ, ณัฐนันท์ ทัดพิทักษ์กุล และ ชัย วุฒิวัฒน์ชัย, 2005

ตารางที่ 2.3

สัญลักษณ์แทนเสียงพยัญชนะควบกล้ำในภาษาไทย (12 หน่วยเสียง)

พยัญชนะควบ	หน่วยเสียง	พยัญชนะควบ	หน่วยเสียง
ปร	pr	กร	kr
ปล	pl	กล	kl
พร	phr	กว	kw
พล	phl	คว	khr
ดร	tr	คล	khl
ทร	thr	ขว	khw

ที่มา : ภัชริภา คชสำโรง, ตรีภพ สรรเพชรนิยม, ศวิต กาสุริยะ, ณัฐนันท์ ทัดพิทักษ์กุล และ ชัย วุฒิวัฒน์ชัย, 2005

ตารางที่ 2.4

สัญลักษณ์แทนเสียงสระในภาษาไทย (24 หน่วยเสียง)

ตำแหน่งลิ้น / ความสูงของลิ้น	หน้า (สั้น/ยาว)	กลาง (สั้น/ยาว)	หลัง (สั้น/ยาว)
สูง	i, ii (อิ, อี)	v, vv (อึ, อือ)	u, uu (อุ, อู)
กลาง	e, ee (เอะ, เอ)	q, qq (เออะ, เออ)	o, oo (โอะ, โอ)
ต่ำ	x, xx (แอะ, แอ)	a, aa (อะ, อา)	@, @@ (เอาะ, ออ)
สระผสม	ia, iia (เอียะ,เอีย)	va, vva (เอือะ, เอือ)	ua, uua (อัวะ, อัว)

ที่มา : ภัชริภา คชสำโรง, ตรีภพ สรรเพชรนิยม, ศวิต กาสุริยะ, ณัฐนันท์ ทัดพิทักษ์กุล และ ชัย วุฒิวัฒน์ชัย, 2005

2.1.1.3 คำ (Word) และหน่วยเสียง (Phoneme)

เสียงพูดประกอบด้วยคำ (Word) มากมายรวมกันเป็นประโยค (Sentence) โดยแต่ละคำ (Word) สามารถแยกย่อยออกมาได้หน่วยเล็ก ๆ เรียกว่า หน่วยเสียง (Phoneme) กล่าวคือ

หน่วยเสียง (Phoneme) คือ หน่วยที่เล็กที่สุดของเสียง เพราะฉะนั้นคำทุกคำก็คือการประสมกันของหน่วยเสียง (Phoneme) หลายหน่วยเสียง (Phoneme) มารวมกัน (ดังภาพที่ 2.2) ตัวอย่างดังภาพที่ 2.3 ประโยค “ฉันกินข้าว” ประกอบด้วยคำว่า “ฉัน” “กิน” “ข้าว” และคำว่า “ฉัน” ประกอบด้วยลำดับหน่วยเสียง (Phoneme) “ฉ” ตามด้วยสระ “อะ” และ “น” สามารถเขียนเป็นสัญลักษณ์การออกเสียงได้ดังนี้ “ch a n^” ลำดับต่อมาคำว่า “กิน” ประกอบด้วยลำดับหน่วยเสียง (phoneme) “ก” ตามด้วยสระ “อิ” และ “น” สามารถเขียนเป็นสัญลักษณ์การออกเสียงได้ดังนี้ คือ “k i n^” ลำดับสุดท้ายคำว่า “ข้าว” ประกอบด้วยลำดับหน่วยเสียง “k aa w^” จะเห็นได้ว่า คำว่า “ฉัน” และคำว่า “กิน” แม้ว่าจะเป็นคำคนละคำกัน แต่จะมีหน่วยเสียงสุดท้าย “น” (n^) เหมือนกัน ด้วยเหตุนี้จึงทำให้จำนวนของคำในพจนานุกรม (dictionary) มีจำนวนมากกว่าจำนวนหน่วยเสียง (phoneme)

ภาพที่ 2.2

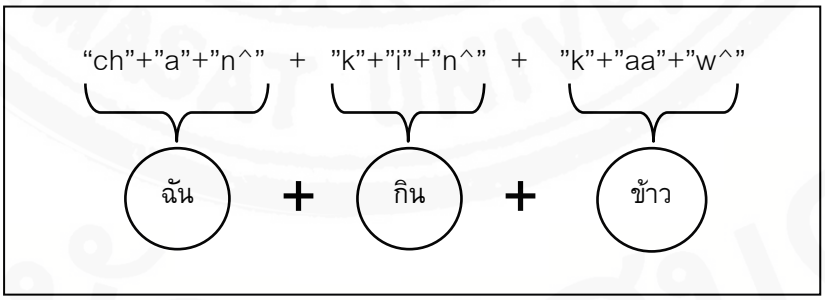
องค์ประกอบของประโยค และ คำ

ประโยค (Sentence) = คำ$_1$ + คำ$_2$ + … + คำ$_n$

คำ (Word) = หน่วยเสียง$_1$+หน่วยเสียง$_2$+…+หน่วยเสียง$_n$

ภาพที่ 2.3

ตัวอย่างประโยค “ฉันกินข้าว”



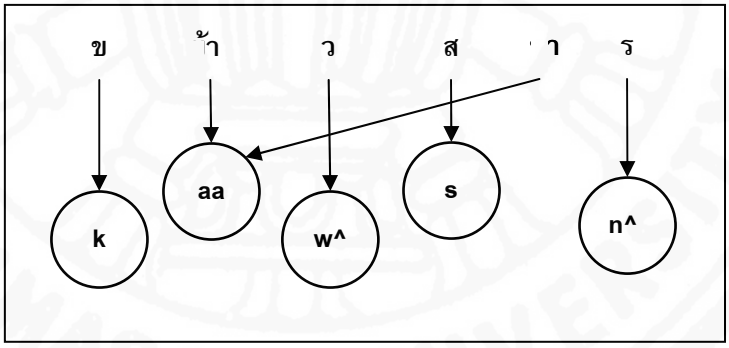
2.1.1.4 Monophone model และ Triphone model

Monophone model คือ การจำลองหน่วยเสียงของคำแต่ละหน่วยเสียงด้วยการใช้ HMM 1 โมเดลแทนหน่วยเสียง 1 หน่วย โดยไม่พิจารณาว่าหน่วยเสียงข้างๆ เป็นหน่วยเสียงอะไร

ดังนั้นสัญญาณเสียงของหน่วยเสียงที่เหมือนกันจะถูกนำเอามาฝึกสอน HMM ของหน่วยเสียงนั้นทั้งหมด ดังภาพที่ 2.4 แต่เนื่องจากเสียงของหน่วยเสียงเดียวกัน อาจจะแตกต่างกันได้หากอยู่ในตำแหน่งต่าง ๆ กันไป ขึ้นกับว่า หน่วยเสียงรอบข้างเป็นหน่วยเสียงอะไร ตัวอย่างเช่น เสียง "w^" ในคำว่า "k aa **w^**" (ข้าว) กับเสียง "w^" ในคำว่า "k aa **w^ s** aa n^" (ข้าวสาร) จะแตกต่างกัน เพราะเสียง "w^" ในคำว่า"ข้าวสาร" จะมีการควบเสียงกับเสียง "s" ในพยางค์ถัดไป ดังนั้นจึงมีการสร้างโมเดลหน่วยเสียงแยกกันในทุกกรณีที่หน่วยเสียงรอบข้างต่างกัน และเรียกโมเดลแบบนี้ว่า Context-dependent phone model (Chai Wutiwiwatchai, 2004) ตัวอย่าง Context-dependent phone model ที่เป็นที่นิยมคือ Triphone model ซึ่งแยกแยะหน่วยเสียงแต่ละตัวตามหน่วยเสียงข้างหน้าและข้างหลัง ดังภาพที่ 2.5 สัญลักษณ์ ”aa-w^+s" คือโมเดลหน่วยเสียง "w^" เฉพาะที่อยู่ระหว่าง "aa" กับ "s"

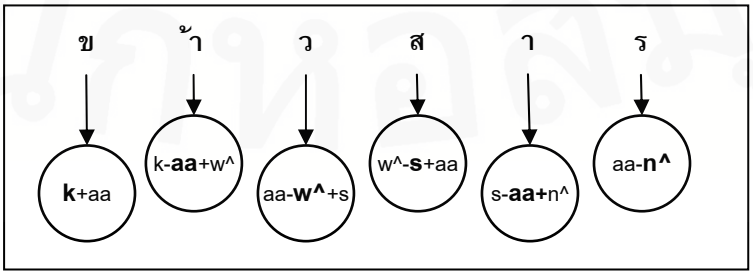
ภาพที่ 2.4

ตัวอย่าง Monophone model ของคำว่า “ข้าวสาร”



ภาพที่ 2.5

ตัวอย่างTriphone model ของคำว่า “ข้าวสาร”

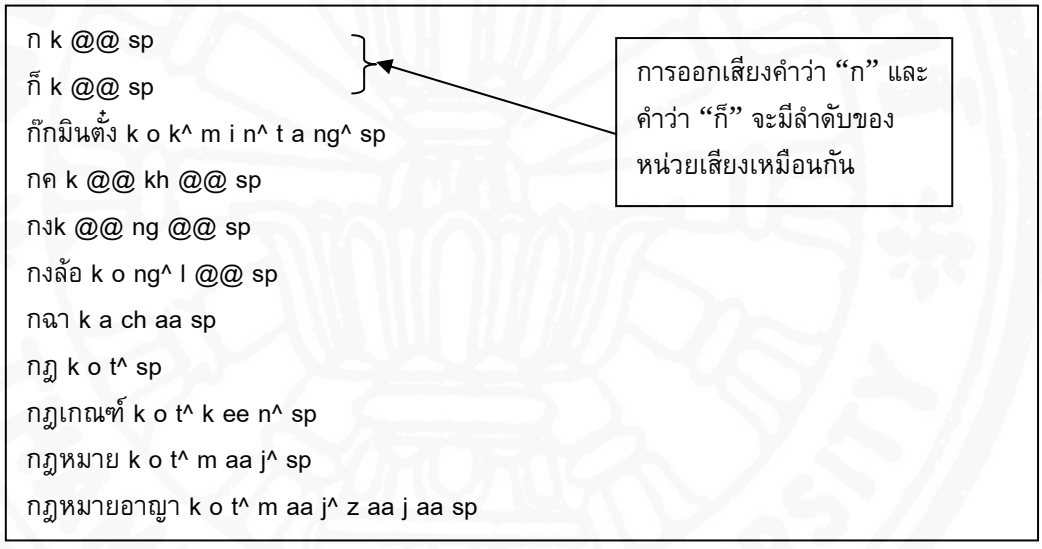


2.1.1.5 **พจนานุกรมการออกเสียง**

พจนานุกรมการออกเสียงมีหน้าที่บอกว่าคำแต่ละคำออกเสียงว่าอย่างไร ด้วยการแสดงลำดับหน่วยเสียง (Phoneme Sequence) ของคำ โดยไม่ได้คำนึงถึงวรรณยุกต์ เช่น คำว่า “ก” และ คำว่า “ก็” จะมีลำดับการออกเสียงเหมือนกัน คือ k @@ ส่วน “sp” คือหน่วยเสียงพิเศษที่ใช้กำหนดจุดสิ้นสุดของคำ (ดังภาพที่ 2.6)

ภาพที่ 2.6

ตัวอย่างพจนานุกรมการออกเสียง



2.1.1.6 **การใช้ HMM ฝึกฝนการรู้จำเสียง** (Speech Recognition)

เสียงพูดแบบต่อเนื่อง คือการนำหน่วยเสียงมาเรียงลำดับต่อกันให้เป็นประโยค ดังนั้นจึงต้องสอนให้คอมพิวเตอร์รู้ว่าหน่วยเสียงแต่ละหน่วยเสียงมีความน่าจะเป็นที่จะเกิดขึ้นเป็นประโยคที่หลากหลาย ได้อย่างไร เช่นประโยค “ฉัน กิน ข้าว” กับ “ฉัน กิน ก๋วยเตี๋ยว” เมื่อระบบทำการค้นหามาถึงคำว่า “กิน” ลำดับเสียงที่ต่อจากคำว่า “กิน” คือเสียงอะไร เป็นเสียง “ข” หรือ “ก” โดยตัดสินใจจากค่าความน่าจะเป็นรวม ณ ขณะนั้นว่าเส้นทางไหนมีค่ามากกว่า ก็จะเลือกเส้นทางนั้นเป็นเส้นทางที่จะทำการค้นหาต่อ การฝึกฝน HMM ให้รู้จำเสียงในวิทยานิพนธ์นี้ใช้วิธีการสอน

โดยมีผู้สอน (Supervised training) คือการฝึกฝนให้คอมพิวเตอร์รู้จำเสียงจากแฟ้มการออกเสียง (Transcription File) ว่าคำแต่ละคำประกอบไปด้วยหน่วยเสียงอะไรบ้าง

2.1.1.7 ขั้นตอนของระบบรู้จำเสียง

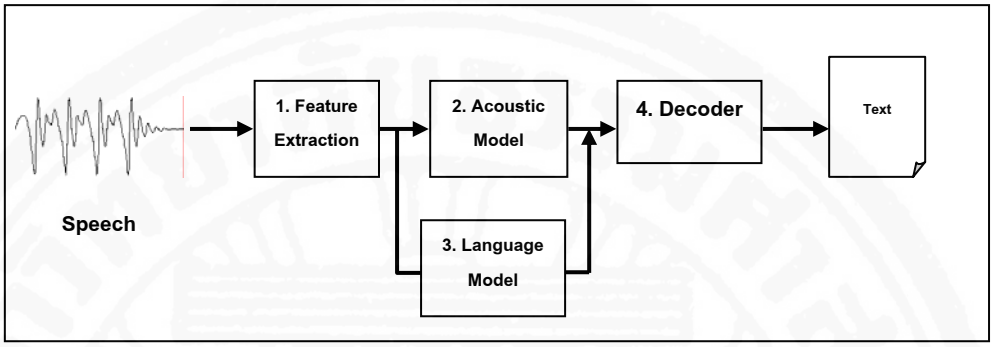
ระบบรู้จำ (Speech Recognition) โดยทั่วไป (ดังภาพที่ 2.7) ประกอบด้วย

1. การสกัดลักษณะสำคัญ (Feature Extraction) ของสัญญาณเสียงที่เข้ามาในระบบ
2. การกำหนดแบบจำลองเสียง (Acoustic Modeling) ทำหน้าที่วิเคราะห์ว่าสัญญาณที่เข้าในระบบเป็นเสียงอะไร
3. การกำหนดแบบจำลองภาษา (Language Modeling) ทำหน้าที่วิเคราะห์ความน่าจะเป็นในการเกิดขึ้นของคำจากลำดับของคำที่กำหนดไว้
4. ส่วนถอดรหัส (Decoder) มีหน้าที่ตีความสัญญาณเสียงเป็นข้อความ

ระบบรู้จำโดยทั่วไปจะมีการใช้ทั้งโมเดลเสียง (Acoustic Model) และ โมเดลภาษา (Language Model) ร่วมกัน หากใช้แต่ โมเดลเสียง (Acoustic Model) เพียงอย่างเดียว ก็จะทำให้ระบบรู้แต่เพียงว่าเสียงนั้นเป็นเสียงอะไร แต่ถ้าใช้โมเดลภาษา (Language Model) มาช่วย จะทำให้ ส่วนถอดรหัส (Decoder) สามารถวิเคราะห์ได้ว่าเสียงนั้นเป็นคำอะไร ซึ่งเป็นการลดอัตราการผิดพลาด (error rate) ที่จะเกิดขึ้น ตัวอย่าง เช่น สัญญาณเสียงของประโยค “ฉัน กิน ข้าว” ถ้าสัญญาณเสียงมีการออกเสียงคำว่า “กิน” ไม่ชัด แล้วใช้เพียงแต่โมเดลเสียง (Acoustic Model) อย่างเดียว ผลตีความจากส่วนถอดรหัส (Decoder) อาจจะได้เป็นคำว่า “กิ่ง” แทน แต่ถ้าใช้ โมเดลภาษา (Language Model) เข้ามาช่วยวิเคราะห์ลำดับของคำว่าคำก่อนหน้า คือ “ฉัน” และ คำที่ตามหลัง คือ คำว่า “ข้าว” ก็จะทำให้ความน่าจะเป็นของคำนั้นคือคำว่า “กิน” มากกว่า คำว่า “กิ่ง” เป็นต้น

ภาพที่ 2.7

ภาพรวมของระบบรู้จำ (Speech Recognition)

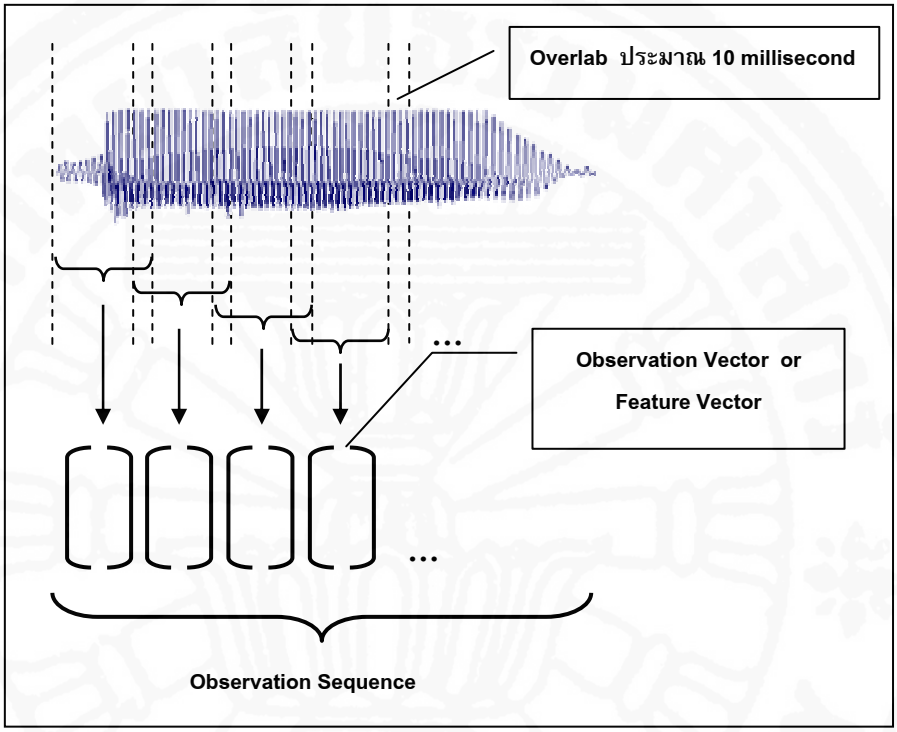


2.1.1.8 **การสกัดลักษณะสำคัญของเสียง** (Feature Extraction)

เนื่องจากเสียงพูดของคนแต่ละคนจะมีรูปแบบของสัญญาณเสียงไม่เหมือนกัน เพื่อให้ระบบรู้จำสามารถเรียนรู้ได้ว่าเสียงแต่ละเสียงนั้นมีความแตกต่างกันอย่างไร ต้องทำการสกัดค่าลักษณะสำคัญของเสียงก่อน ในขั้นตอนของการทำวิทยานิพนธ์นี้จะนำแฟ้มเสียงที่ได้จัดเตรียมไว้มาสกัดค่าลักษณะสำคัญ โดยการสกัดค่าลักษณะสำคัญของเสียงจะทำการสกัดแต่ละช่วงเวลา (ประมาณ 20 มิลลิวินาที) ต่อ 1 หน่วยเสียง ค่าที่สกัดได้แต่ละช่วงเวลา (แต่ละเวกเตอร์จะแทนสัญญาณเสียงแบบ overlap กันประมาณ 10 มิลลิวินาที) จะนำถูกเก็บไว้ในเวกเตอร์ค่าลักษณะสำคัญ (Observation Vector or Feature Vector) เพราะค่าลักษณะสำคัญของหน่วยเสียง 1 หน่วยเสียงคือการ เรียงลำดับต่อ ๆ กันของเวกเตอร์ค่าลักษณะสำคัญ (Observation Sequence) ดังภาพที่ 2.8

ภาพที่ 2.8

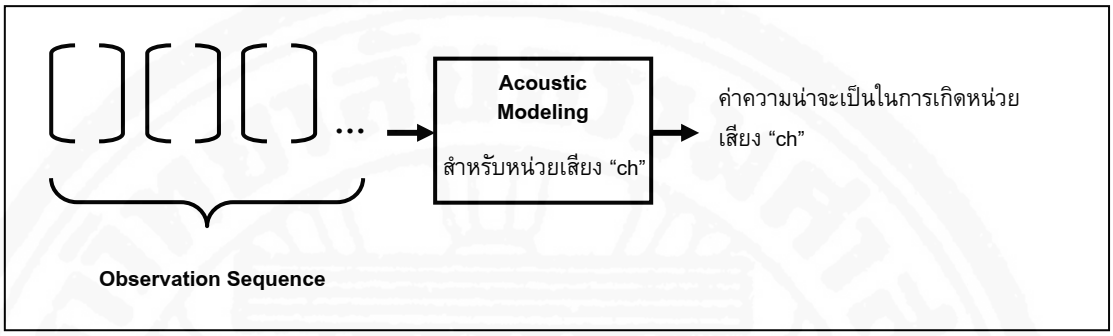
การสกัดค่าลักษณะสำคัญของหน่วยเสียง



2.1.1.9 **โมเดลเสียง** (Acoustic Model)

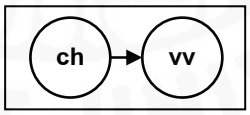
โมเดลเสียง (Acoustic Model) ดังภาพที่ 2.9 มีหน้าที่คำนวณหาความน่าจะเป็นของเสียงว่าเสียงที่เข้ามานั้นออกเสียงว่าอย่างไร โดยใช้โมเดลเสียง 1 โมเดลต่อ 1 หน่วยเสียง (1 Acoustic Model : 1 Phoneme) ในการคำนวณหาความน่าจะเป็นจะพิจารณาจากการเรียงลำดับของเวกเตอร์ลักษณะสำคัญ (Observation Sequence) ตัวอย่างเช่น เมื่อนำลำดับเวกเตอร์ลักษณะสำคัญ (Observation Sequence) ของเสียงที่ต้องการพิจารณาว่าเป็นเสียง “ch” (“ช”) เข้าสู่โมเดลหน่วยเสียง “ch” โมเดลนี้จะนำหน้าที่พิจารณาเสียงนั้นว่าเป็นเสียง “ch” หรือไม่

ภาพที่ 2.9
หน้าที่ของโมเดลเสียง

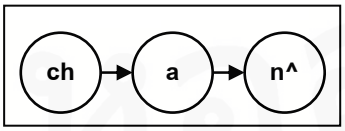


โมเดลคำ (Word Model) ของคำใด ๆ มีหน้าที่ บอกว่า คำ ๆ นั้น ออกเสียงอย่างไร โดยจะบอกเป็นลำดับของหน่วยเสียง (Phoneme Sequence) เช่นตัวอย่างคำว่า "ชื่อ" (ดังภาพที่ 2.10) จะมีลำดับการออกเสียงคือ "ch vv" และคำว่า "ฉัน" (ดังภาพที่ 2.11) จะมีลำดับของหน่วยเสียงคือ "ch a n^" วิธีสร้างโมเดลการออกเสียง สามารถทำได้ด้วยการสร้างพจนานุกรมการออกเสียง ซึ่งจะแสดงรายละเอียดของคำแต่ละคำว่ามีการออกเสียงอย่างไร (ลำดับของหน่วยเสียง) โดยมีทิศทางการออกเสียง (เปลี่ยนสถานะ) จากซ้ายไปขวา (Left-to-Right)

ภาพที่ 2.10
ตัวอย่างโมเดลการเสียงคำว่า “ชื่อ”



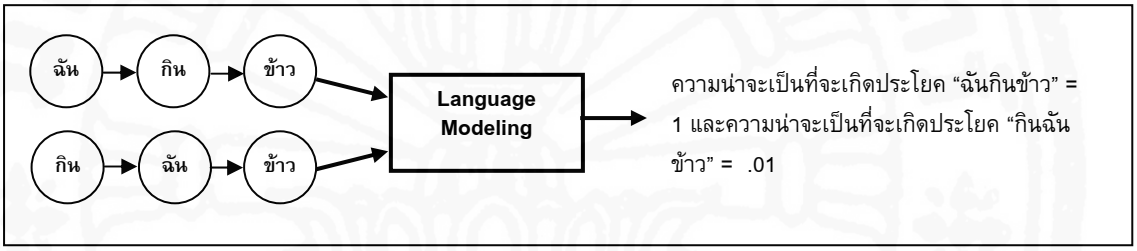
ภาพที่ 2.11
ตัวอย่างโมเดลการเสียงคำว่า “ฉัน”



2.1.1.10 **โมเดลภาษา** (Language Model)

โมเดลภาษา (Language Model) คือ โมเดลของภาษาที่จะบอกว่าคำแต่ละคำในประโยค มีลำดับของคำเป็นอย่างไร คำแต่ละคำสามารถตามด้วยคำใดบ้าง โดยพิจารณาจากค่าความน่าจะเป็นที่คำใด ๆ จะพูดต่อกัน เช่น ความน่าจะเป็นของการเกิดคำว่า "ฉัน กิน ข้าว" มีค่าความน่าจะเป็นมากกว่า "กิน ฉัน ข้าว" ดังภาพที่ 2.12

ภาพที่ 2.12
หน้าที่ของโมเดลภาษา



โมเดลภาษาสามารถสร้างได้ 2 วิธีคือ

1. Regular grammar โดยการเขียนกฎตาม EBNF (Steve Young et al., 2005, p. 165) ซึ่งได้กำหนดรูปแบบและความหมายของไวยากรณ์ดังแสดงไว้ในตารางที่ 2.5

ตารางที่ 2.5
ตารางแสดงเครื่องหมายและความหมาย เพื่อใช้ในการกำหนด Regular grammar ตามกฎ EBNF

Symbols	Meaning
|	denotes alternatives
[ ]	encloses options
{}	denotes zero or more repetitions
< >	denotes one or more repetitions
<< >>	denotes context-sensitive loop

ที่มา : Steve Young et al., 2005

ภาพที่ 2.13 เป็นตัวอย่างการใช้ Regular grammar เพื่อสร้างโมเดลภาษาของคำหลักในระบบลงทะเบียนนักศึกษา

ภาพที่ 2.13
ตัวอย่าง การใช้ Regular grammar เพื่อสร้างโมเดลภาษา
ของคำหลักในโดเมนระบบลงทะเบียนนักศึกษา

```
(silence (<ลง|ทะเบียน|ถอน|วิชา|เพิ่ม|ดร็อบ|บังคับ|เลือก|เสรี|เฉพาะ|หน่วยกิต|เทอม|
หลักสูตร|ภาค|ศึกษา>) silence)
```

2. N-gram เป็นวิธีที่ใช้แฟ้มข้อความ (text file) มาฝึกสอนให้ระบบรู้จำเรียนรู้ โดยแฟ้มข้อความที่จะนำมาใช้นั้น จะต้องมีการแบ่งประโยคออกเป็นลำดับของคำก่อน เช่น ประโยคตัวอย่าง “วิชาสอสอหนึ่งเจ็ดหนึ่งภาษาอังกฤษสอง” ดังภาพที่ 2.14

ภาพที่ 2.14
ตัวอย่าง ประโยคก่อนที่การแบ่งคำ

วิชาสอสอหนึ่งเจ็ดหนึ่งภาษาอังกฤษสอง

เมื่อนำมาประโยคตัวอย่างมาฝึกสอนจะต้องแบ่งประโยคออกเป็นคำ ดังนี้ “วิชา สอ สอ หนึ่ง เจ็ด หนึ่ง ภาษา อังกฤษ สอง” ดังภาพที่ 2.15

ภาพที่ 2.15
ตัวอย่าง ประโยคหลังจากที่มีการแบ่งคำแล้ว

วิชา สอ สอ หนึ่ง เจ็ด หนึ่ง ภาษา อังกฤษ สอง

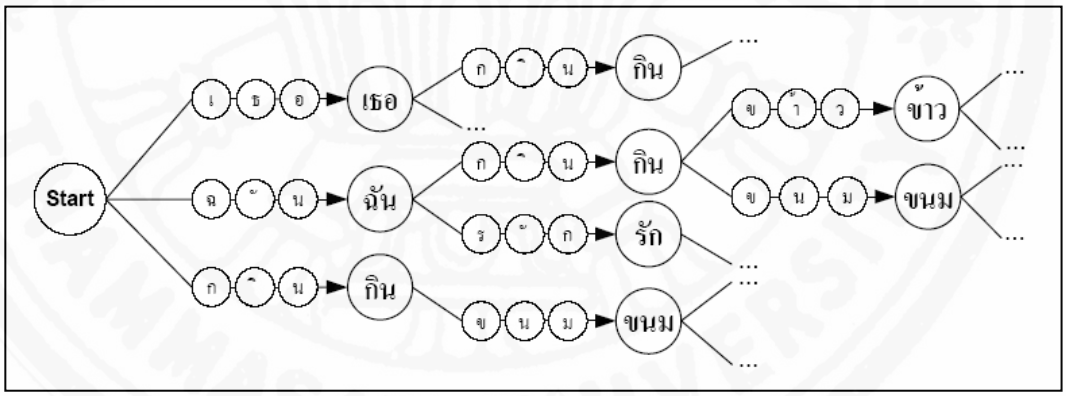
ในการแบ่งประโยคภาษาไทยออกเป็นคำ สามารถกระทำได้โดยวิธีด้วยมือ หรือใช้โปรแกรมการตัดคำภาษาไทย SWATH ซึ่งพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) มาช่วยการตัดแบ่งคำได้ แต่อย่างไรก็ดี โปรแกรมดังกล่าวยังมี

ข้อจำกัดเรื่องความถูกต้องในการตัดแบ่งคำ จึงทำให้ต้องทำการตรวจและแก้ไขด้วยมือหลังจากที่ได้ผลลัพธ์จากโปรแกรมดังกล่าวแล้ว

2.1.1.11 Word network

Word network ทำหน้าที่รวบรวมลำดับของคำ (Word Sequence) ที่เป็นไปได้ทั้งหมด (ดังภาพที่ 2.16) โดยในแต่ละเส้นทาง ของ Network จะประกอบด้วย ค่าความน่าจะเป็นของ Word sequence ซึ่งจะถูกคำนวณโดยโมเดลภาษา (Language Model) และความน่าจะเป็นของการเกิดสัญญาณเสียงนั้น เมื่อมีการกำหนด Word sequence ใด ๆ จะถูกคำนวณโดยโมเดลเสียง (Acoustic Model) ส่วนการค้นหาคำนั้นจะทำการค้นหาตามเส้นทางที่ให้ค่าความน่าจะเป็นรวมสูงที่สุด

ภาพที่ 2.16
ตัวอย่าง Word Network



2.1.2 การค้นหาคำหลัก (Keyword Spotting)

โดยธรรมชาติแล้วเสียงพูดของมนุษย์จะประกอบไปด้วยคำที่เป็นนัยสำคัญ และคำประกอบอื่น ๆ ที่ไม่นัยสำคัญ ดังนั้นระบบการค้นหาคำหลักในเสียงพูดจะต้องมีความแม่นยำในการพิจารณาเสียงพูดว่าคำใดเป็นคำที่มีนัยสำคัญ และคำใดไม่มีนัยสำคัญ ซึ่งในระบบการค้นหาคำหลักนี้จะมีกำหนดคำที่มีนัยสำคัญตามเรื่องที่สนใจ หรือโดเมนที่สนใจ ไว้เป็นคำหลัก (Keyword) เพื่อให้ระบบสามารถแยกเสียงคำหลักออกจากเสียงพูดได้

2.1.2.1 **โมเดลคำหลัก** (Keyword Model)

โมเดลคำหลัก คือ การจำลองหน่วยเสียงของคำหลักว่ามีลำดับหน่วยเสียงอย่างไร ซึ่งหน่วยเสียงที่นำมาสร้างเป็นโมเดลคำหลักจะนำมาจากโมเดลเสียงที่ได้ผ่านการฝึกฝนให้รู้จำเฉพาะคำที่เป็นคำหลักเท่านั้น

วิธีการที่ดีที่สุดทำให้ระบบค้นหาคำหลักให้สามารถแยกคำหลักออกจากเสียงพูด คือ ฝึกฝนให้ระบบการรู้จำคำทุกคำในคลังเสียงคำศัพท์เสียงพูดต่อเนื่องขนาดใหญ่ (Large Vocabulary Continuous Recognizer : LVCR) อย่างไรก็ตามการทำวิธีนี้จะทำให้มีค่าใช้จ่าย (Cost) ที่สูงมาก และเมื่อคำหลักไม่ปรากฏอยู่ในคลังเสียงดังกล่าวก็จะต้องฝึกฝนระบบให้รู้จำคำที่เพิ่มขึ้นใหม่ แต่มีอีกวิธีหนึ่งคือการกำหนดแบบจำลองของคำที่ไม่อยู่ในพจนานุกรม (out-of-vocabulary : oov) ด้วย แบบจำลองที่เรียกว่า โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model)

2.1.2.2 **โมเดลคำที่ไม่ใช่คำหลัก** (Non-keyword Model)

โมเดลคำที่ไม่ใช่ (Non-keyword Model) คือ การจำลองหน่วยเสียงของคำที่ไม่ใช่คำหลักว่ามีลำดับหน่วยเสียงอย่างไร โดยแบบจำลองนี้ทำหน้าที่พิจารณาเสียงพูดว่าคำใดเป็นคำที่ไม่ใช่คำหลัก ซึ่งโมเดลคำที่ไม่ใช่คำหลัก มีชื่อเรียกที่แตกต่างกันว่า “Background Model” หรือ “Filler Model” หรือ “Garbage Model “ แต่โมเดลทั้ง 3 แบบ มีจุดประสงค์และหน้าที่เหมือนกันดังที่ได้กล่าวมาแล้ว

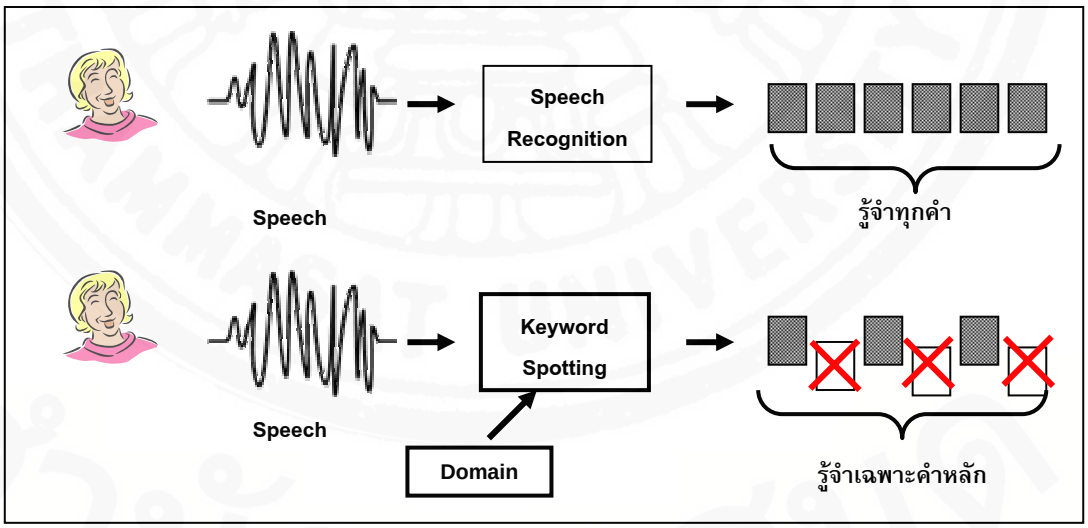
2.1.2.3 Confidence Measurement

ในการพิจารณาลำดับของคำที่ค้นหาได้ว่าเป็นคำใดนั้น มีเกณฑ์ที่ใช้พิจารณาที่เรียกว่า “Confidence Measurement” โดยจะนำที่เอาผลรวมค่าความน่าจะเป็นที่ได้จากโมเดลเสียง เมื่อเสียงพูดผ่านสถานะแต่ละสถานะ มาทำการเปรียบเทียบกับ Threshold ของคำนั้น ผลลัพธ์ที่ได้คือคำที่มีค่า Confidence Measurement ใกล้เคียงกับค่า Threshold ที่กำหนดไว้มากที่สุด

2.1.2.4 ความแตกต่างระหว่างการค้นหาคำหลัก (Keyword Spotting) กับ การรู้จำเสียง (Speech Recognition)

การรู้จำเสียงแบบต่อเนื่องต้องใช้การตีความจากสัญญาณเสียงที่สมบูรณ์ และผลที่ได้คือความสมบูรณ์ในการตีความประโยค โดยการรู้จำเสียงแบบต่อเนื่องนี้คอมพิวเตอร์จะต้องทำการเรียนรู้ทุกคำ ส่วนการค้นหาคำหลักต้องการเพียงผลลัพธ์ว่ามีคำหลักอยู่ในสัญญาณเสียงนั้นหรือไม่ การรู้จำก็จะรู้จำเพียงคำหลักที่อยู่ในโดเมน (Domain) ที่กำหนดไว้เท่านั้น ซึ่งส่วนนี้คือส่วนที่ทำให้การค้นหาคำหลักในเสียงพูดแบบต่อเนื่องแตกต่างจากการรู้จำเสียงแบบต่อเนื่อง จึงทำการค้นหาคำหลักทำได้ยากขึ้น เพราะคำหลักสามารถจะอยู่ ณ ตำแหน่งใดก็ได้ในสัญญาณเสียง (ดังภาพที่ 2.17) ดังนั้นระบบค้นหาคำหลักจะต้องมีความสามารถในการแยกสัญญาณเสียงคำหลักออกจากสัญญาณเสียงทั้งหมดให้ได้

ภาพที่ 2.17
ความแตกต่างระหว่างการค้นหาคำหลัก (Keyword Spotting)
กับการรู้จำเสียง (Speech Recognition)



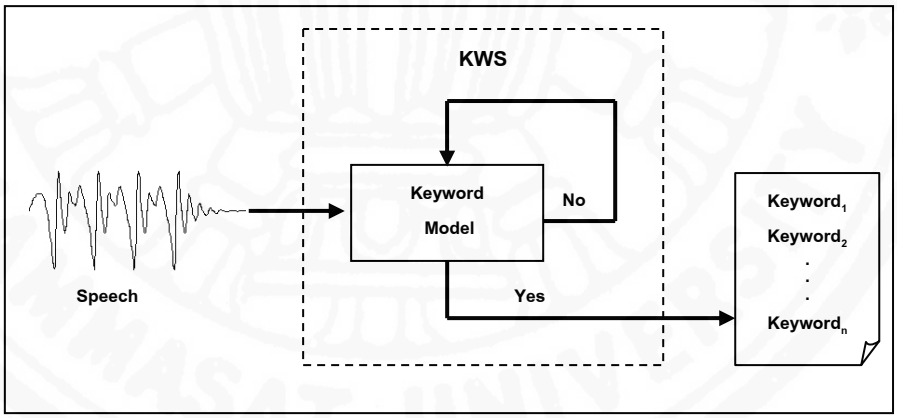
2.2 สรุปงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวกับการค้นหาคำหลักในเสียงพูด (Keyword Spotting System : KWS) สามารถแบ่งออกได้เป็น 2 กลุ่มหลัก คือ กลุ่มงานวิจัยที่สนใจเรื่องโมเดลคำหลัก (Keyword Model) และ กลุ่มงานวิจัยที่สนใจเรื่องโมเดลคำที่ไม่ใช่คำหลัก (Non-Keyword Model)

2.2.1 กลุ่มงานวิจัยที่สนใจเรื่องโมเดลคำหลัก (Keyword Model)

กลุ่มแรกคือ กลุ่มงานวิจัยที่ให้ความสนใจในการนำโมเดลคำหลัก (Keyword Model) มาค้นหาคำหลัก ดังภาพที่ 2.18

ภาพที่ 2.18
การค้นหาคำหลักโดยใช้ Keyword Model



โดยงานวิจัยในกลุ่มนี้จะมุ่งเน้นไปที่ประสิทธิภาพของโมเดลเสียง (Acoustic Model) ที่ได้ฝึกฝนมา และอัลกอริทึ่มที่ใช้ค้นหาคำหลัก งานวิจัยในกลุ่มนี้ (ดังตารางที่ 2.6) ได้แก่

J. Junkawitsch, L. Neubauer, H. Höge, and G. Ruske (1996) ใช้วิธีการค้นหาคำหลักด้วยโดยการแก้ไขอัลกอริธึ่มที่ใช้ค้นหาคำหลัก ด้วยการเก็บผลรวมค่าความน่าจะเป็นของสถานะ (state) HMM ในทุกเส้นทางที่คำหลักนั้นสามารถผ่านได้ และนำคำที่ค้นหาได้มาเปรียบเทียบกับค่า Threshold ของคำหลักคำนั้น ซึ่งค่า Threshold ได้มีการกำหนดไว้ก่อนล่วงหน้าแล้ว ว่าคำที่ค้นหามาได้คือคำหลักคำใด จากผลการทดลองในการค้นหาคำหลัก 25 คำ ได้ความถูกต้อง 94.4%

Marius-Calin Silaghi and Herve Bourlard (1999) ได้ใช้ค้นหาคำหลักด้วยการแก้ไขอัลกอริธึ่มที่ใช้ค้นหาคำหลักเช่นเดียวกัน โดยอัลกอริธึ่มจะเปรียบเทียบสัญญาณเสียงที่เข้ามากับโมเดลคำหลัก (Keyword Model) ที่สร้างขึ้นไว้ก่อน ซึ่งอัลกอริธึ่มนี้จะเปรียบเทียบสัญญาณเสียงกับโมเดลคำหลักทุกคำ แล้วพิจารณาว่าตรงกับโมเดลคำหลักคำใดมากที่สุด วิธีการนี้จะใช้หน่วยความจำมากถ้ามีโมเดลคำหลักที่ใช้เปรียบเทียบมาก ผลการทดลองสามารถค้นหาคำหลัก 100 คำได้ถูกต้อง 95%

Ya-Dong Wu and Bao-Long Liu (2003) ใช้วิธีที่คล้ายคลึงกันแต่จะใช้วิธีการสร้างเทมเพลต (Template) ของคำหลักของคำหลักแต่ละคำเป็นชุดไว้ก่อนล่วงหน้า จากนั้นจะนำคำหลักที่ค้นหาได้มาทำการเปรียบเทียบกับเทมเพลตของคำหลักนั้นว่าเหมือนกับชุดไหนมากที่สุด และได้ทำการทดลองค้นหาคำหลักจากเสียงพูดที่ขึ้นตรงกับผู้พูดจึงทำให้การค้นหาคำหลักทำได้รวดเร็วและมีความถูกต้องสูดสุดถึง 100% แต่วิธีการนี้ใช้เฉพาะกับเสียงพูดแบบเจาะจงผู้พูด (Speaker Dependent)

Ru Haibo, Chen Yining, Liu Jia and Liu Runsheng (2003) สามารถค้นหาคำหลักได้โดยไม่ขึ้นตรงกับประมวลคำศัพท์ (Vocabulary) ด้วยคุณสมบัติสองประการคือ ประการแรกคือ การสร้างแบบจำลองเสียง (Acoustic Modeling) และ สร้างเครือข่ายการค้นหา (Searching Network) ขึ้นในระหว่างที่ทำกระบวนการรู้จำ ประการที่สอง คัดเลือกโมเดลเสียงของคำต่าง ๆ ที่ดีที่สุด จากการฝึกฝนโมเดลเสียงด้วย Large vocabulary continuous speech recognizer (LVSCR) มากำหนดเป็นโมเดลคำหลัก เพื่อใช้เปรียบเทียบกับคำที่ค้นหาได้ จึงทำให้ไม่จำเป็นต้องใช้โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) ทำการค้นหาคำหลักในเสียงพูดที่มีคำหลักมากกว่า 1 ตัว แม่นยำถึง 95% แต่ถ้ามีการเพิ่มคำหลักใหม่ต้องทำการฝึกฝนระบบใหม่

Thomas Fang Zheng, Jing Li, Zhanjiang Song and Mingxing Xu (2004) ได้เสนอวิธีการค้นหาคำหลักจากโดเมน d-Ear Attendant ซึ่งระบบตอบรับโทรศัพท์อัตโนมัติ ซึ่งประกอบด้วย ขั้นตอน 2 ขั้นตอน คือ

1. ทำการสร้างลำดับของสัญลักษณ์ทางเสียง (Acoustic Symbols) ในระหว่างกระบวนการรู้จำเสียงต่อเนื่อง ว่าเสียงเริ่มต้นและเสียงสุดท้ายของคำคือเสียงใด แล้วจัดเก็บไว้เป็นเทมเพลต (Template)
2. ค้นหาคำหลักโดยทำการเปรียบเทียบหน่วยเสียงของสัญญาณเสียงที่ใช้ทดลองกับเทมเพลต (Template) ที่สร้างขึ้นในขั้นตอนแรก โดยไม่ต้องใช้โมเดลภาษา (Language Model)

วิธีการนี้จะใช้เวลาและหน่วยความจำน้อย เนื่องจากว่าอัลกอริธึ่มที่ใช้ค้นหาคำหลัก จะทำการค้นหาไปเปรียบเทียบหน่วยเสียงของแฟ้มเสียงไปตามลำดับของสัญลักษณ์ทางเสียง ซึ่งการค้นหาจะดูว่าเสียงเริ่มต้นตรงกับเทมเพลต (Template) ที่กำหนดไว้หรือไม่ ถ้าใช่ก็จะทำการค้นหาไปจนถึงเสียงสุดท้าย ผลการทดลองได้ความถูกต้องในการค้นหาคำหลัก 88.0% วิธีการนี้มีข้อจำกัดคือ ถ้าสัญญาณเสียงที่ใช้ทดสอบมีการออกเสียงคำที่เป็นคำหลักไม่ชัดเจน จะทำให้ค้นหาคำหลักนั้นไม่เจอ เนื่องจากโมเดลเสียง (Acoustic Model) จะให้ผลลัพธ์ของหน่วยเสียงไม่ตรงกับเทมเพลต (Tempate) ที่สร้างขึ้น

Shan-Ruei You et al. (2004) ใช้โมเดลเสียงภาษาจีน (Chinese Acoustic Model) กับ โมเดลเสียงภาษาอังกฤษ (English Acoustic Model) ร่วมกัน ในการค้นหา keyword ในภาษาจีนที่มีภาษาอังกฤษผสมอยู่ เพื่อแก้ปัญหาเรื่องชื่อของชาวไต้หวันที่มักจะมีชื่อที่ประกอบด้วยคำภาษาจีนและคำภาษาอังกฤษ จากผลการทดลองสามารถค้นหาคำได้ถูกต้องสูงสุดถึง 97% แต่ใช้ได้เฉพาะการค้นหาคำหลักที่เป็นแบบเจาะจงผู้พูด (Speaker Dependent) และสัญญาณเสียงพูดเป็นคำโดด (Isolated Word)

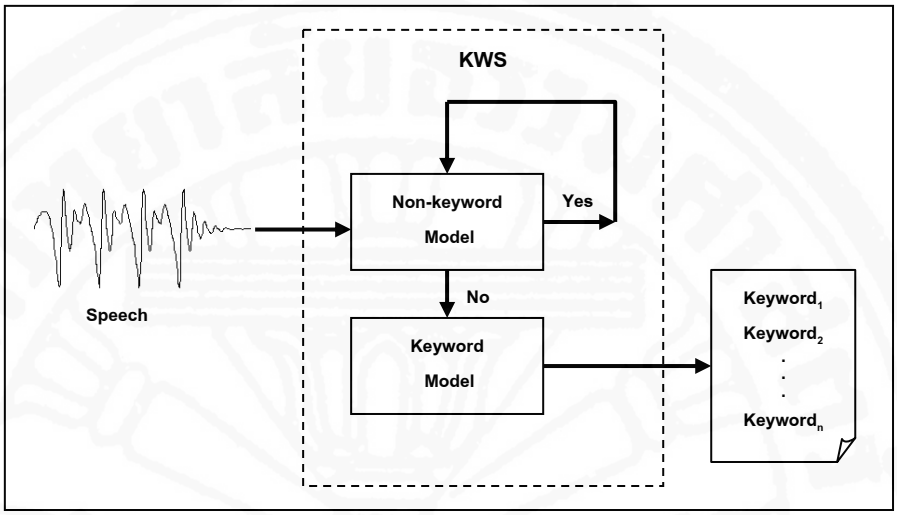
อย่างไรก็ตามงานวิจัยในกลุ่มนี้มีข้อจำกัดเรื่องของคำหลัก (keyword) คือจะต้องทำการฝึกฝนโมเดลเสียงของคำหลัก (Keyword Model) ทุกครั้ง เมื่อมีคำหลักใหม่เข้ามาในโดเมน จึงมีงานวิจัยอีกกลุ่มหนึ่งที่แก้ไขข้อจำกัดของงานวิจัยในกลุ่มแรกด้วยการใช้โมเดลคำที่ใช่คำหลัก (Non-keyword Model) เพื่อแยกคำที่ไม่ใช่คำหลักออกจากสัญญาณเสียง

2.3.2 กลุ่มงานวิจัยที่สนใจเรื่องโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model)

กลุ่มงานวิจัยที่ใช้โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) (ดังตารางที่ 2.7) นี้มีสมมติฐานว่าคำที่ไม่ใช่คำหลักมีความน่าจะเป็นที่จะเกิดคำใหม่ขึ้นในระบบ น้อยกว่าความน่าจะเป็นที่จะเกิดคำหลักคำใหม่ขึ้นในระบบ จึงใช้โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) นำมาช่วยแยกคำที่ไม่ใช่คำหลักออกจากสัญญาณเสียง และใช้โมเดลคำหลักเพื่อพิจารณาว่าคำนั้นเป็นคำหลักคำใด ดังภาพที่ 2.19

ภาพที่ 2.19

การค้นหาคำหลักโดยใช้ Non-keyword Model



งานวิจัยในกลุ่มนี้มีจุดประสงค์เพื่อแก้ปัญหาในการสร้างและฝึกฝนโมเดลคำหลักใหม่ทุกครั้งเมื่อมีคำหลักใหม่เพิ่มเข้ามาในระบบ โดย

Alan L. Higgins and Robert E. Wohlford (1985) ใช้วิธีการสร้างระบบค้นหาคำหลักแบบ dynamic time warping เพื่อทำการเปรียบเทียบ สัญญาณเสียงที่เข้ามากับเทมเพลต (Template) ของคำที่ไม่ใช่คำหลักที่ได้มีการกำหนดไว้ก่อน จากนั้นระบบจะทำการตัดสินใจว่าคำนั้นเป็นคำหลัก หรือไม่ใช่คำหลัก ถึงแม้ว่าวิธีการนี้จะสามารถค้นหาคำหลักได้ แต่สิ้นเปลืองหน่วยความจำมากในระหว่างกระบวนเปรียบเทียบสัญญาณเสียงที่เข้ามากับเทมเพลตที่กำหนดไว้

Richard C. Rose and Douglas B. Paul (1990) ด้วยการฝึกฝนโมเดลคำหลัก (Keyword Model) ให้รู้จำเพียงคำหลักที่ต้องการเท่านั้น และฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) จากฐานข้อมูลเสียงคำศัพท์ขนาดใหญ่ จากการทดลองการรู้จำคำหลักสำหรับการพูดสนทนาแบบมาตรฐาน (Standard Conversational) ได้ความถูกต้องถึง 82% สำหรับการค้นหาคำหลัก 20 คำ แต่วิธีการนี้จะต้องทำการแยกแฟ้มเสียงที่จะนำมาฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword) ไม่ให้มีคำหลักที่กำหนดไว้อยู่ในแฟ้มเสียงเหล่านั้นเสียงก่อน มิฉะนั้นจะทำระบบไม่สามารถแยกแยะสัญญาณเสียงที่ใช้ทดสอบได้ว่าเป็นคำหลักหรือคำที่ไม่ใช่คำหลัก ซึ่งงานดังกล่าวจะต้องใช้เวลาในการจัดเตรียมมาก เพื่อแก้ปัญหาในการที่จะต้องฝึกฝนโมเดลคำที่ไม่ใช่คำหลักด้วยการใช้แฟ้มเสียงจำนวนมาก และถ้ามีการเพิ่มคำที่ไม่ใช่คำหลักก็จะต้องทำการฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) ใหม่

K. M. Knill and S. J. Young (1995) ได้นำโมเดลภาษา (Language Model) ที่สร้างจากแฟ้มข้อความแบบ N-gram นำมาช่วยแยกคำที่ไม่ใช่คำหลักออกจากเสียง ซึ่งวิธีนี้ทำให้ไม่ต้องจัดเตรียมแฟ้มเสียงจำนวนมากเพื่อฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword model) แต่สำหรับภาษาที่ไม่มีการแบ่งคำหรือมีการแบ่งคำไม่ชัดเจน เช่น ในภาษาไทย (Sinaporn Suebvisai, Paisarn Charoenpornsawat, Alan Black, Monika Woszczyna and Tanja Schultz, 2005) ก็จะทำให้สิ้นเปลืองเวลาในการจัดเตรียมแฟ้มข้อความ จากการทดลองได้ความถูกต้อง 91.5%

Chen Yining, Liu Jing, Zhong Lin, Lin Jia and Liu Runsheng (2001) ได้ใช้วิธีการแยกคำที่ไม่ใช่คำหลัก (Non-keyword) ด้วยการใช้ไวยากรณ์ (Grammar) ร่วมกับโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) คือเมื่อมีสัญญาณเข้ามาจะตรวจสอบดูว่าคำที่อยู่ในสัญญาณเสียงตรงกับคำที่กำหนดไว้ในไวยากรณ์หรือไม่ ถ้าไม่ก็จะใช้โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) มาทำการตรวจสอบคำนั้นอีกครั้งหนึ่งว่าเป็นคำที่ตรงกับโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) หรือไม่ ถ้าไม่ตรงกันอีกแสดงว่าคำนั้นเป็นคำหลัก วิธีการนี้ทำให้สามารถเพิ่มรายการของคำที่ไม่ใช่คำหลักได้ด้วยการกำหนดไวยากรณ์เพิ่มเติม ทำให้ไม่ต้องทำการฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) เพิ่มเติมถ้ามีคำใหม่เพิ่มเติม แต่วิธีการใช้ไวยากรณ์ไม่สามารถแยกคำที่ไม่ใช่คำหลักได้ดีเท่ากับโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) จากการทดลองค้นหาคำหลักเฉพาะชื่อบุคคลที่เป็นภาษาอังกฤษ และภาษาจีนได้ความถูกต้อง 88.15% ส่วนการทดลองเมื่อทำการค้นหาคำหลักจากประโยค ได้ความถูกต้องเพียง 77.57%

Hamed Ketabdar, Jithendra Vepa, Samy Bengio and Herve Boulard (2006) ใช้วิธีการเปรียบเทียบสัญญาณเสียงว่าเป็นคำหลักหรือไม่ ด้วยการพิจารณาสัญญาณเสียงที่เข้ามาว่ามีลำดับของหน่วยเสียงอย่างไร แล้วทำการเปรียบเทียบกับโมเดลคำหลัก (Keyword Model) และ โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) แล้วทำการให้คะแนนกับคำนั้นว่าตรงกับโมเดลใดมากที่สุด สำหรับการให้คะแนนนั้น ระบบจะทำการเปรียบเทียบไปตามสถานะของโมเดลทั้งสอง แต่ละสถานะแล้วทำการให้คะแนน สำหรับการเปรียบเทียบไปตามสถานะของโมเดลทั้งสองนั้นจะต้องอาศัยความรู้ก่อนหน้าว่าโมเดลคำหลัก (Keyword Model) และ โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) มีลำดับของหน่วยเสียงอย่างไร จากผลการทดลองค้นหาคำหลัก 12 คำ ได้ความถูกต้อง 98.3% แต่วิธีการนี้สิ้นเปลืองหน่วยความจำมากเนื่องจากระบบจะต้องทำการจัดเก็บคะแนนของคำแต่ละคำที่เปรียบเทียบกับโมเดลทั้งสองไว้ก่อน เมื่อทำการค้นหาเสร็จ

แล้ว จึงนำคะแนนที่มาได้พิจารณาหาคะแนนสูงสุดของคำนั้นว่าเป็นคำใกล้เคียงกับเป็นคำหลักหรือคำที่ไม่ใช่คำหลักมากที่สุด

JianEn Liang, Meng Meng, XiaoRui Wang, Peng Ding and Bo Xu (2006) นำเสนอวิธีการค้นหาคำหลักในภาษาแมนดาริน โดยการเพิ่มคำหลัก (Keyword) เข้าในรายการคำหลัก (Keyword List) ที่จะมาฝึกฝนโมเดลคำหลัก (Keyword Model) ด้วยการพิจารณาจากความถี่ในการเกิดขึ้นของคำหลัก (Keyword) แล้วจึงนำมาฝึกฝนระบบค้นหาคำหลักด้วยวิธีการที่เรียกว่า MCE (Minimum Classification Error) และใช้โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) เพื่อกำจัดคำที่ไม่ใช่คำหลักร่วมด้วย ทำให้ผลการค้นหาคำหลักมีความแม่นยำขึ้น เมื่อทำการทดลองแล้วพบว่าความยาวของคำหลักมีผลต่อประสิทธิภาพในการค้นหาคำหลัก โดยสามารถค้นหาคำหลักที่มีความยาวมาก ๆ (Long Keyword) ได้ความแม่นยำมากว่าคำหลักที่มีความยาวสั้น ๆ (Short Keyword) จึงได้มีการนำบริบทมาใช้เพื่อเพิ่มค่าความน่าจะเป็นในค้นหาคำหลักเข้าไปในโมเดลภาษา (Language Model) ซึ่งทำให้แก้ไขปัญหาเรื่องความแม่นยำในการค้นหาได้ ได้ความถูกต้อง 86.2% แต่วิธีการฝึกฝน (Training) โมเดลคำหลัก (Keyword Model) ใช้เวลาถึง 90 ชั่วโมง

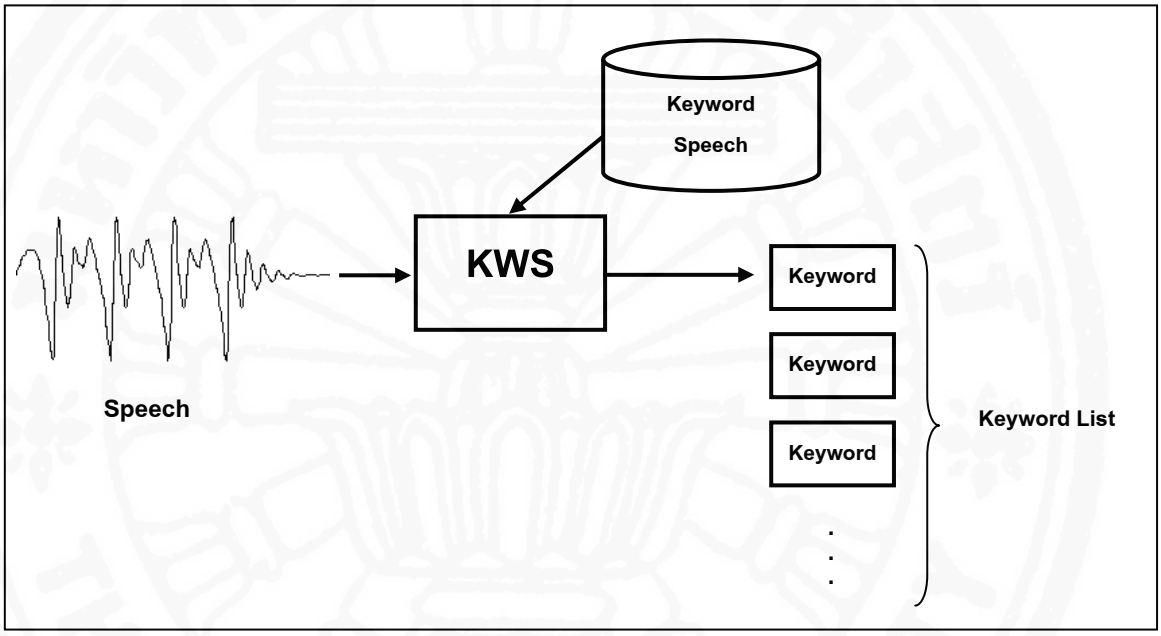
งานวิจัยในกลุ่มที่สองนี้สามารถใช้โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) เพื่อแยกคำที่ไม่ใช่คำหลักออกจากสัญญาณเสียงได้ ซึ่งสามารถแก้ปัญหาของกลุ่มงานวิจัยแรกได้ ข้อดีของวิธีการนี้ทำให้ไม่ต้องฝึกฝนระบบใหม่หากมีคำหลักใหม่เพิ่มเข้ามาในโดเมน แต่มีข้อเสียคือ

1. คำที่ค้นหาออกมาได้นั้น มีความเป็นไปได้ว่าจะไม่ใช่คำหลักที่ตรงกับโดเมนที่ต้องการ
2. คำที่ไม่ใช่คำหลักกลายเป็นคำหลักของโดเมน จะทำให้การค้นหาคำหลักนั้นมีข้อผิดพลาด
3. ต้องมีการฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) ถ้ามีคำที่ไม่ใช่คำหลักเพิ่มเข้ามาในระบบ

ดังนั้นในวิทยานิพนธ์นี้ จะแก้ปัญหาของงานวิจัยกลุ่มแรกที่ต้องการมีการฝึกฝนโมเดลคำหลัก (Keyword Model) ด้วยวิธีการค้นหาคำหลักด้วยการสร้างคำหลักแบบอัตโนมัติ ทำให้ไม่ต้องทำการฝึกฝนโมเดลเสียงใหม่ทุกครั้ง เมื่อมีคำหลักเพิ่มเข้ามาในระบบ เนื่องจากระบบค้นหาคำหลัก (Keyword Spotting System : KWS) โดยทั่วไปจะสร้างโมเดลคำหลักด้วยฐานข้อมูลข้อมูลเสียง และให้ผลลัพธ์เป็นรายการคำหลัก (Keyword list) ซึ่งทำให้มีข้อจำกัดในการเพิ่ม

คำหลักใหม่ เนื่องจากจะต้องทำการจัดเตรียมแฟ้มเสียงของคำหลัก แล้วนำไปฝึกฝน (re-training) ระบบใหม่ ดังภาพที่ 2.20

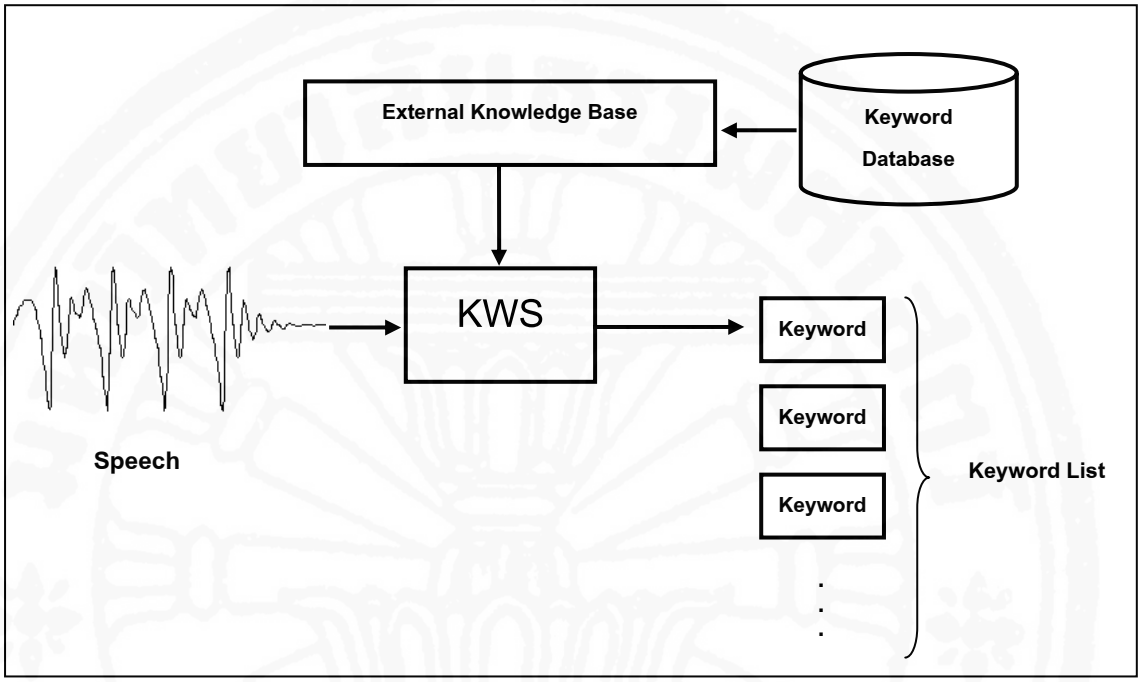
ภาพที่ 2.20
การสร้างโมเดลคำหลักจากฐานข้อมูลเสียง



แต่ในวิทยานิพนธ์นี้จะใช้แหล่งความรู้ภายนอก (External knowledge) โดยใช้ฐานข้อมูลคำหลัก (Keyword database) ที่ประกอบด้วยคำหลัก และลำดับหน่วยเสียงของคำหลักแต่ละคำ ในรูปของแฟ้มข้อความ (Text file) ดังนั้นถ้ามีคำหลักเพิ่มขึ้นใหม่ ก็เพียงแต่เพิ่มคำหลักและลำดับของหน่วยเสียงของคำหลักใหม่คำนั้นลงในฐานข้อมูล โดยไม่ต้องฝึกฝนระบบใหม่ ดังภาพที่ 2.21

ภาพที่ 2.21

การสร้างโมเดลคำหลักจากฐานข้อมูลข้อความ



ตารางที่ 2.6

สรุปกลุ่มงานวิจัยที่ใช้ Keyword Model ในการค้นหาคำหลัก

ปี	งานวิจัยของ	วิธีการค้นหาคำหลัก	ข้อดี	ข้อเสีย
1996	J. Junkawitsch, L. Neubauer, H. Höge, and G. Ruske	แก้ไขอัลกอริธึ่มสำหรับการค้นหาคำหลักด้วยการเก็บค่าความน่าจะเป็นของคำหลักทุกเส้นทางที่สามารถค้นหาได้	สามารถค้นหาคำหลัก ในเส้นทางค้นหาที่อาจจะมีคำหลักซ่อนอยู่ได้ ผลการทดลองได้ความถูกต้อง ในการค้นหา 94.4%	สิ้นเปลืองหน่วยความจำมาก เนื่องจากจะต้องมีการจัดเก็บค่าความน่าจะเป็นของเส้นทางทุกเส้นที่คำหลักสามารถผ่านได้
1999	Marius-Calin Silaghi and Herve Bourlard	แก้ไขอัลกอริธึ่ม โดยเปรียบเทียบสัญญาณเสียงที่ถอดรหัสได้กับโมเดลคำหลักว่ามีค่าความน่าจะเป็นใกล้เคียงกับโมเดลคำหลักใดมากที่สุด	สามารถค้นหาคำหลักได้รวดเร็ว โดยความถูกต้องที่ได้ 95%	ใช้หน่วยความจำมาก ถ้ามีโมเดลคำหลักที่ใช้เปรียบเทียบมาก
2003	Ya-Dong Wu and Bao-Long Liu	เปรียบเทียบสัญญาณเสียงที่ถอดรหัสได้กับเทมเพลต (Template) ของชุดคำหลัก	สามารถค้นหาคำหลัก ได้ตรงกับเทม-เพลตแน่นอน โดยความถูกต้องในการค้นหาเท่ากับ 100%	เสียงพูดที่นำมาทดสอบเป็นเสียงพูดแบบเจาะจงผู้พูด

ตารางที่ 2.6 (ต่อ)

สรุปกลุ่มงานวิจัยที่ใช้ Keyword Model ในการค้นหาคำหลัก

ปี	งานวิจัยของ	วิธีการค้นหาคำหลัก	ข้อดี	ข้อเสีย
2003	Ru Haibo, Chen Yining, Liu Jia and Liu Runsheng	เปรียบเทียบสัญญาณเสียงที่ถอดรหัสได้กับโมเดลคำหลักที่ดีที่สุดที่ได้ฝึกฝนไว้แล้ว	การค้นหาคำหลักสามารถทำได้รวดเร็ว เพราะเพียงเปรียบเทียบว่าคำที่ค้นหาได้กับโมเดลคำหลักที่เตรียมไว้ว่าเหมือนหรือไม่เหมือน โดยความถูกต้องที่ได้ 95%	ถ้าโมเดลคำหลักที่ดีที่สุดไม่ใช่คำหลักที่ต้องการค้นหา จะต้องฝึกฝนระบบให้รู้จำคำหลักใหม่
2004	Thomas Fang Zheng, Jing Li, Zhanjiang Song and Mingxing Xu	เปรียบเทียบหน่วยเสียงของแฟ้มเสียงไปตามลำดับของสัญลักษณ์ทางเสียง ซึ่งการค้นหาจะดูว่าเสียงเริ่มต้นตรงกับเทมเพลต (Template) ที่กำหนดไว้หรือไม่ ถ้าใช่ก็จะทำการค้นหาไปจนถึงเสียงสุดท้าย	ค้นหาคำหลักได้เร็ว เพราะการค้นหาจะเปรียบคำกับเทมเพลตก่อน ได้ความถูกต้องในการค้นหาคำหลัก 88.0%	ถ้าสัญญาณเสียงที่ใช้ทดสอบมีการออกเสียงคำที่เป็นคำหลักไม่ชัดเจน จะทำให้การค้นหาผิดพลาด

ตารางที่ 2.6 (ต่อ)

สรุปกลุ่มงานวิจัยที่ใช้ Keyword Model ในการค้นหาคำหลัก

ปี	งานวิจัยของ	วิธีการค้นหาคำหลัก	ข้อดี	ข้อเสีย
2004	Shan-Ruei You et al.	ใช้โมเดลเสียงภาษาจีน กับ โมเดลเสียงภาษาอังกฤษร่วมกัน เพื่อค้นหา keyword ในภาษาจีนที่มีภาษาอังกฤษผสมอยู่	ทำให้สามารถแยกเสียงภาษาจีน และภาษาอังกฤษออกจากกันได้ และได้ความถูกต้องในการค้นหา 97%	ใช้ได้เฉพาะการค้นหาคำหลักในสัญญาณเสียง ที่มีการเจาะจงตัวผู้พูด (Speaker Dependent) และจะต้องพูดเป็นคำโดด (Isolated Word)

ตารางที่ 2.7

สรุปกลุ่มงานวิจัยที่ใช้ Non-Keyword Model ในการค้นหาคำหลัก

ปี	งานวิจัยของ	วิธีการค้นหาคำหลัก	ข้อดี	ข้อเสีย
1985	Alan L. Higgins and Robert E. Wohlford	เปรียบเทียบเฟรมของสัญญาณเสียงที่เข้ามากับเทมเพลต (Template) ของคำที่ไม่ใช่คำหลักที่ได้มีการกำหนดไว้ก่อน	ได้ความถูกต้องในการค้นหาคำหลัก 79%	สิ้นเปลื้องหน่วยความจำมากในการระหว่างกระบวนเปรียบเทียบสัญญาณเสียงที่เข้ามากับเทม-เพลตที่กำหนดไว้
1990	Richard C. Rose and Douglas B. Paul	ค้นหาคำหลักด้วยการเปรียบเทียบสัญญาณเสียงที่เข้ากับโมเดลคำหลัก (Keyword Model) และโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model)	จากการทดลองการรู้จำคำหลักสำหรับการพูดสนทนาแบบมาตรฐานได้ความถูกต้องถึง 82%	ต้องทำการแยกแฟ้มเสียงที่จะนำมาฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword) ไม่ให้มีคำหลักที่กำหนดไว้อยู่ในแฟ้มเสียงเหล่านั้นเสียงก่อน
1995	K. M. Knill and S. J. Young	ใช้โมเดลภาษา (Language Model) ที่สร้างจากแฟ้มข้อความ นำมาช่วยแยกคำที่ไม่ใช่คำหลักออกจากเสียง	ไม่ต้องจัดเตรียมแฟ้มเสียงจำนวนมากเพื่อฝึกฝนโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword model) ผลการทดลองได้ความถูกต้อง 91.5%	สำหรับภาษาที่ไม่มีการแบ่งคำหรือมีการแบ่งคำไม่ชัดเจน จะทำให้สิ้นเปลืองเวลาในการจัดเตรียมแฟ้มข้อความ

ตารางที่ 2.7(ต่อ)

สรุปกลุ่มงานวิจัยที่ใช้ Non-Keyword Model ในการค้นหาคำหลัก

ปี	งานวิจัยของ	วิธีการค้นหาคำหลัก	ข้อดี	ข้อเสีย
2001	Chen Yining, Liu Jing, Zhong Lin, Lin Jia and Liu Runsheng	ได้ใช้วิธีการแยกคำที่ไม่ใช่คำหลัก (Non-keyword) ด้วยการใช้ไวยากรณ์ (Grammar) ร่วมกับโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model)	สามารถเพิ่มรายการของคำที่ไม่ใช่คำหลักได้ด้วยการกำหนดไวยากรณ์เพิ่มเติมจากผลการทดลองเฉพาะชื่อบุคคลความถูกต้อง 88.15% ส่วนการทดลองค้นหาคำหลักจากประโยค ได้ความถูกต้อง77.57%	แต่วิธีการใช้ไวยากรณ์ไม่สามารถแยกคำที่ไม่ใช่คำหลักได้ดีเท่ากับโมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model)
2006	Hamed Ketabdar, Jithendra Vepa, Samy Bengio and Herve Boulard	ใช้วิธีการเปรียบเทียบลำดับหน่วยเสียงของสัญญาณเสียงที่ใช้ทดสอบ กับโมเดลคำหลัก (Keyword Model) และ โมเดลคำที่ไม่ใช่คำหลัก (Non-keyword Model) ว่าตรงกับโมเดลใดมากที่สุด	ได้ความถูกต้อง 98.3%	สิ้นเปลืองหน่วยความจำมากเนื่องจากระบบทำการจัดเก็บคะแนนของคำแต่ละคำที่เปรียบเทียบกับโมเดลทั้งสองไว้ก่อน จากนั้น จึงนำคะแนนที่มาได้พิจารณาหาคะแนนสูงสุดของคำนั้นว่าเป็นคำใกล้เคียงกับคำหลัก หรือคำที่ไม่ใช่คำหลักมากที่สุด

ตารางที่ 2.7(ต่อ)

สรุปกลุ่มงานวิจัยที่ใช้ Non-Keyword Model ในการค้นหาคำหลัก

ปี	งานวิจัยของ	วิธีการค้นหาคำหลัก	ข้อดี	ข้อเสีย
2006	JianEn Liang, Meng Meng, XiaoRui Wang, Peng Ding and Bo Xu	ใช้บริบทและความยาวของคำมาใช้เพื่อเพิ่มค่าความน่าจะเป็นในค้นหาคำหลักเข้าไปในโมเดลภาษา (Language Model)	นำเสนอสามารถค้นหา long keyword ได้ดีกว่า โดยได้ความถูกต้อง 86.2%	การฝึกฝน (Training) โมเดลคำหลัก (Keyword Model) ใช้เวลาถึง 90 ชั่วโมง