บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทนี้จะกล่าวถึงเนื้อหา 2 ส่วน คือ ทฤษฎีที่เกี่ยวข้องและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องกับระบบการสืบค้นข้อมูลโดยเสียงภาษาไทยมีรายละเอียดดังนี้

1. การรู้จำเสียงพูดต่อเนื่องภาษาไทย (Continuous Speech Recognition)

เทคนิคหรือวิธีการรู้จำเสียงต่อเนื่องภาษาไทยที่ได้นำมาใช้กับงานวิจัยนี้มีหลายวิธีด้วยกัน ประกอบด้วย ระบบเสียงภาษาไทย การสกัดลักษณะสำคัญของเสียง(Feature Extraction) แบบจำลองหน่วยเสียง (Acoustic Model) แบบจำลองภาษา (Language Model) การถอดรหัส (Decoder) ถูกรวบรวมไว้ในส่วนนี้

1.1 ระบบเสียงภาษาไทย (Thai Phonology)

ระบบเสียงในภาษาไทยมีการออกเสียงในระดับพยางค์อยู่ในรูปแบบ $/C_i V (Cf)^T /$ โดยที่ C_i คือ พยัญชนะต้น ซึ่งอาจเป็นพยัญชนะเดี่ยวหรือควบกล้ำ V คือสระ ซึ่งอาจจะเป็นได้ทั้งสระเดี่ยวหรือสระผสม Cf คือเสียงตัวสะกด อาจจะไม่มีก็ได้ T คือเสียงวรรณยุกต์ ซึ่งคำอ่านของคำในภาษาไทยมีจำนวนมากที่เขียนไม่ตรงตามรูปแบบ โดยเฉพาะคำที่มาจากภาษาต่างประเทศ สัญลักษณ์แทนเสียงในภาษาไทยแสดงในตารางที่ 2.1

ชนิดของเสียง		สัญลักษณ์เสียง
พยัญชนะต้น	เดี่ยว	k,kh,ng,c,ch,s,j,d,t,th,n,b,p,ph,f,m,r,l,w,h,z
	ควบกล้ำ	pr,pl,tr,kr,kl,kw,phr,phl,thr,khr,khl,khw,br,bl,fr,fl,dr
สระ	เดี่ยว	a,i,v,u,e,x,o,@,q aa,ii,vv,uu,ee,xx,oo,@@,qq
	ผสม	ia,iia,va,vva,ua,uua
ตัวสะกด		k^,ng^,j^,t^,n^,p^,m^,w^,ch^,f^,l^,s^
หน่วยเสียงเงียบ		sil

ตารางที่ 2.1 ตารางสัญลักษณ์แทนเสียงในภาษาไทย

1.2 การสกัดลักษณะสำคัญของเสียง (Feature Extraction)

การสกัดลักษณะสำคัญของเสียงจะใช้วิธีของ Mel frequency cepstrum coefficient (MFCC) ซึ่งเป็นวิธีที่นิยมมากที่สุดในปัจจุบัน เพราะจะให้ผลที่ดีกว่าการสกัดลักษณะสำคัญของเสียงวิธีอื่นๆ

การแปลงฟูริเยร์ช่วงเวลาสั้น (Short-time discrete Fourier transform: short time DFT) จะแสดงบนแต่ละหน้าต่างสัญญาณ (windowed signal) ที่ได้รับจากการประมวลผลก่อนหน้า การแปลงฟูริเยร์ช่วงเวลาสั้นของแต่ละหน้าต่างสัญญาณถูกประมวลผลโดย แถบของตัวกรองเมล (mel filter banks) แถบของตัวกรองเป็นแถบความถี่ผ่าน ความถี่ที่ต่ำกว่า 1 kHz. เกิดจากกระบวนการของแถบตัวกรองความถี่ที่มากกว่า 1 kHz. สำหรับ ตัววัดเมล (mel scale) จะใช้ในระบบรู้จำเสียงที่สัมพันธ์กับตัววัดการรับรู้ของมนุษย์ แถบของตัวกรองเป็นรูปแบบตัววัดสำหรับความถี่ที่ต่ำกว่า 1 kHz. และ ลอการิทึมสำหรับความถี่ที่มากกว่า 1 kHz. ดังที่แสดงในสมการที่ (1)

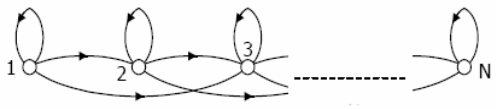
$$mel(f) = 2595 * \log 10(1 + \frac{f}{700}) \qquad (1)$$

1.3 แบบจำลองหน่วยเสียง (Acoustic Model)

ในงานวิจัยนี้จะอาศัยแบบจำลองหน่วยเสียง HMM (Hidden Markov model) โดยรูปแบบของ HMM ที่ใช้ในระบบ จะใช้แบบจำลองซ้ายไปขวา (left-right model) กับ N สถานะ ดังแสดงในภาพที่ 2.1

ภาพที่ 2.1

แบบจำลอง HMM



HMM ประกอบด้วย ค่าความน่าจะเป็นในการเปลี่ยนแปลงสถานะ (state transition probability) A โดย a_{ij} เป็นสัมประสิทธิ์ในการเปลี่ยนแปลงสถานะ (state transition coefficient), ค่าความน่าจะเป็นของสถานะการสังเกตที่เกิดขึ้น (Observation Symbol Probability Distribution) B โดย $b_j(k)$ เป็นสัมประสิทธิ์ของการสังเกต

สำหรับแต่ละสถานะ j จะมีลักษณะดังนี้

1) ทิศทางของค่าความน่าจะเป็นในการเปลี่ยนแปลงสถานะ A โดย a_{ji} เป็นความน่าจะเป็นของการเปลี่ยนแปลงสถานะ i มีค่าดังนี้

   $a_{ji} = 0$ เมื่อ $i < j$ สำหรับ $i > j + 2$

2) สถานะการสังเกตที่เกิดขึ้น B โดย $b_j(O)$ จะมีรูปแบบดังที่แสดงในสมการที่ (2)

$$b_j(O) = \sum_{k=1}^{M} c_{jk} N(O, m_{jk}, U_{jk}) \qquad (2)$$

ซึ่ง M เป็นจำนวนของ Mixture, O เป็นทิศทางของการสังเกตในมิติ D, c_{jk} เป็นสัมประสิทธิ์ mixture สำหรับ mixture ที่ k ในสถานะ j และ $N(O, m_{jk}, U_{jk})$ แสดงถึงความหนาแน่นของฟังก์ชันกับค่าเฉลี่ยของทิศทาง (mean vector) m_{jk} และค่าเบี่ยงเบนเมทริกซ์ (covariance matrix) U_{jk}

ในกระบวนการเรียนรู้จะใช้วิธี Forward-Backward Algorithm และ Baum-Welch ปรับค่าความน่าจะเป็นต่างๆให้เข้ากับลำดับของการสังเกต

ความแตกต่างของจำนวนสถานะเป็นพารามิเตอร์ที่ส่งผลกระทบถึงประสิทธิภาพของอัตราการเรียนรู้ และจำนวนขององค์ประกอบแบบเกาส์ (Gaussian mixture) เป็นพารามิเตอร์ที่สำคัญสำหรับ HMM ด้วย

1.4 แบบจำลองภาษา (Language Model)

งานวิจัยนี้จะใช้แบบจำลองภาษาเอ็นแกรม (N-gram language model) ในการสร้างแบบจำลองภาษา เพราะแบบจำลองภาษาเอ็นแกรมจะใช้สำหรับระบบรู้จำเสียงพูดต่อเนื่อง ที่ครอบคลุมคำศัพท์จำนวนมาก (Large vocabulary continuous speech recognition: LVCSR) ที่ใช้คำศัพท์มากกว่า 1,000 คำศัพท์ และไม่มีไวยากรณ์ตายตัว เช่น การถอดความจากข่าว การพิมพ์ตามคำพูด

ถ้าให้การเรียงลำดับของคำ (Word sequence) $W = \{w_1, \ldots, w_M\}$ โดย $P(W)$ คือความน่าจะเป็นของการเรียงลำดับคำ จะแสดงในสมการที่ (3)

$$P(W) = \prod_{n=1}^{M} P(w_n \mid w_{n-1,\ldots,} w_1) \tag{3}$$

ถ้า w_i ขึ้นอยู่กับหนึ่งคำก่อนหน้า w_{i-1} จะเรียก 2-gram (Bigram) Model จะแสดงในสมการที่ (4)

$$P(W) = \prod_{n=1}^{M} P(w_n \mid w_{n-1}) \tag{4}$$

ถ้า w_i ขึ้นอยู่กับสองคำก่อนหน้า w_{i-1} w_{i-2} จะเรียก 3-gram (Trigram) Model จะแสดงในสมการที่ (5)

$$P(W) = P(w_1)P(w_2 \mid w_1)\sum_{i=3}^{M} P(w_i \mid w_{i-1}, w_{i-2}) \tag{5}$$

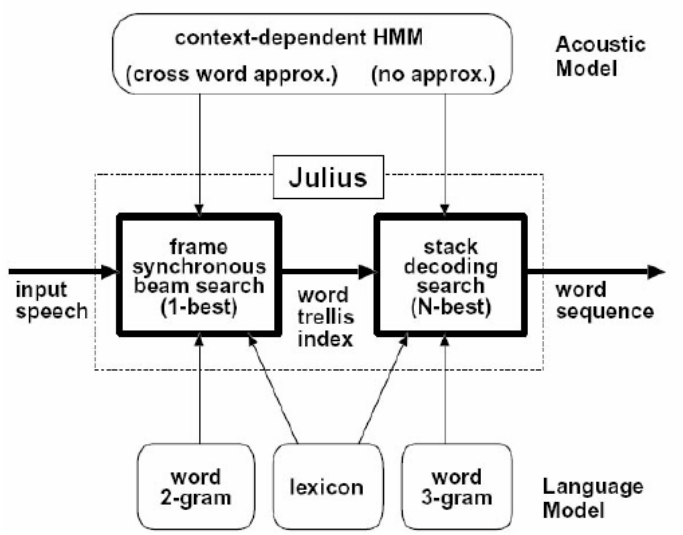
โดยที่ $P(w_i \mid w_{i-1}, w_{i-2})$, $P(w_i \mid w_{i-1})$ และ $P(w_i)$ คำนวณจาก การเรียนรู้ข้อความ (Training Text)

1.5 การถอดรหัส (Decoder)

งานวิจัยนี้ การถอดรหัสของเสียงจะใช้วิธีการค้นหาหลายขั้น (Multi-pass search) คือมีการผ่านกระบวนการค้นหามากกว่าหนึ่งครั้ง ก่อนที่จะได้คำตอบ โดยจะใช้ Julius ซึ่งเป็นตัวถอดรหัสแบบ Two-pass เป็นตัวถอดรหัส หลักการทำงานของ Julius แสดงในภาพที่ 2.2

ภาพที่ 2.2

หลักการทำงานของ Julius



Viterbi เป็น อัลกอริทึม ที่มีประสิทธิภาพในการหาคำตอบที่ดีที่สุด สามารถสรุปได้ดังนี้

1) Initialization Step

$$\delta_1(t) = \pi_i b_i(O_1), \quad 1 \le i \le N$$
$$\psi_1(i) = 0 \tag{6}$$

2) Recursion Step

$$\delta_t(t) = \max_{1 \le i \le N} \lfloor \delta_{t-1}(i) a_{ij} \rfloor b_j(O_t)$$
$$\psi_1(i) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] \tag{7}$$

โดยที่ $1 \le j \le N$, $2 \le t \le T$

3) Termination Step

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$
$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)] \quad (8)$$

4) State Sequence Backtracking Step

$$q_T^* = \psi_{t+1}(q_{t+1}^*) \quad t = T-1,\ T-2, \ldots, 1 \quad (9)$$

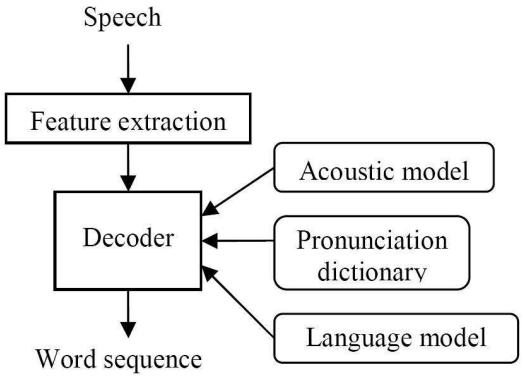
2. ระบบรู้จำเสียงพูดต่อเนื่องที่ครอบคลุมคำศัพท์จำนวนมาก (LVCSR)

2.1 โครงสร้างของระบบ LVCSR

โครงสร้างมาตรฐานของ LVCSR ดังแสดงในภาพที่ 2.3 ประกอบด้วย

1) ส่วนสกัดค่าลักษณะสำคัญ (Feature extraction) จะแบ่งสัญญาณเสียงออกเป็นส่วนย่อยๆ เรียกว่า เฟรม (Frame) และสกัดเวกเตอร์ค่าลักษณะสำคัญ (Feature vector) ออกมาจากสัญญาณเสียงแต่ละเฟรม
2) แบบจำลองหน่วยเสียง (Acoustic model) อาศัยแบบจำลอง HMM (Hidden Markov model) 1 แบบจำลองต่อหน่วยเสียง 1 หน่วย ซึ่ง 1 HMM จะแทนหน่วยเสียงที่พิจารณาหน่วยเสียงรอบข้างด้วย
3) พจนานุกรมเสียงอ่าน (Pronunciation dictionary) รายการของคำศัพท์ที่ระบบสามารถรู้จำได้ แต่ละคำจะมีเสียงอ่านกำกับในรูปลำดับของหน่วยเสียง (Phoneme sequence)
4) แบบจำลองภาษา (Language model) อาศัย N-gram model ของคำ หรือกลุ่มคำ (word class) ในการคำนวณความน่าจะเป็นในการเกิดประโยค (Language probability)
5) ส่วนถอดรหัส (Decoder) เป็นส่วนที่รับลำดับของ Feature vector จากส่วนสกัดค่าลักษณะสำคัญเข้ามา ภายใต้คำศัพท์ที่มีในพจนานุกรม

ภาพที่ 2.3
โครงสร้างระบบ LVCSR



2.2 การสร้างระบบ LVCSR

ในการสร้างระบบ LVCSR สำหรับภาษาใดๆ มีสิ่งที่ต้องเตรียม ได้แก่

1. คลังข้อมูลเสียงพูด (Speech corpus) ซึ่งประกอบด้วยเสียงพูดประโยคที่ครอบคลุมหน่วยเสียงที่เกิดขึ้นในภาษานั้นๆ คลังข้อมูลนี้จะใช้ฝึกฝนแบบจำลองหน่วยเสียง
2. พจนานุกรมเสียงอ่าน ประกอบด้วยคำศัพท์ที่ต้องการให้ระบบรู้จำได้
3. คลังข้อความ (Text corpus) ขนาดใหญ่ประกอบด้วยประโยคจำนวนมากที่ครอบคลุมคำศัพท์ที่ปรากฏในพจนานุกรมเสียงอ่าน คลังข้อความนี้จะใช้ในการฝึกฝนแบบจำลองภาษา

3. เทคนิคการค้นคืนข้อมูลทั่วไป (Information Retrieval Techniques)

โดยทั่วไประบบค้นคืนข้อมูลจะใช้เทคนิคการค้นคืนโดยใช้คำสำคัญ (Keyword-Based Retrieval) หลักการทำงานคือผู้ใช้ต้องระบุคำต่างๆ ที่ต้องการค้นหา จากนั้นระบบจะค้นคืนผลลัพธ์เป็นรายการที่มีเอกสารที่มีคำเหล่านั้นปรากฏอยู่

ระบบค้นคืนข้อมูลทั่วไปมีส่วนประกอบหลัก 6 ส่วนที่สำคัญ แสดงไว้ดังภาพที่ 2.4 ได้แก่

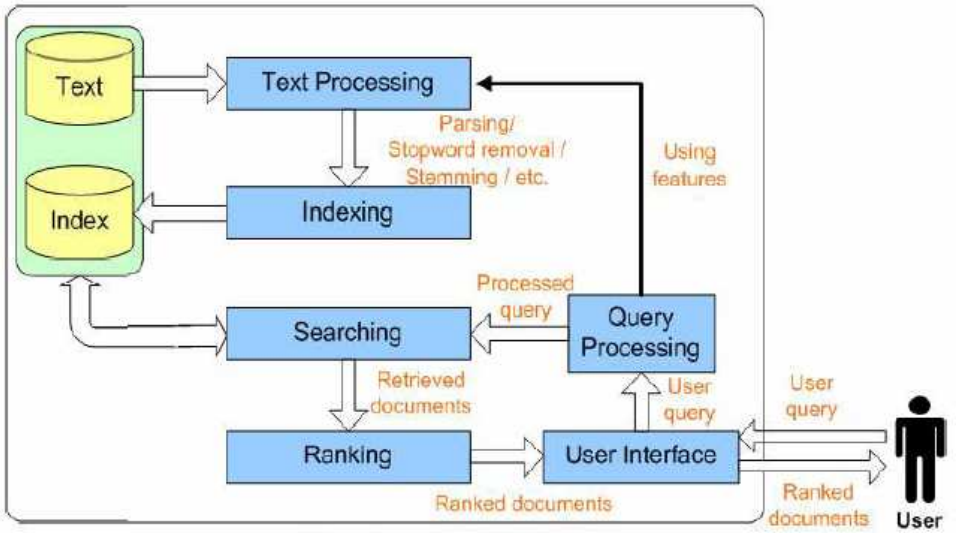
1. หน่วยประมวลข้อความ (Text Processing Module)

มีหน้าที่ในการวิเคราะห์และประมวลข้อความสำหรับการสร้างดัชนีคำ (Index Terms)

2. หน่วยสร้างดัชนี (Indexing Module)
   มีหน้าที่ในการสร้างชุดดัชนีของคำที่ได้จากการวิเคราะห์และประมวลข้อความ
3. หน่วยค้นคืนเอกสาร (Searching Module)
   มีหน้าที่ในการค้นคืนรายการเอกสาร โดยการตรวจสอบคำที่ผู้ใช้ระบุกับชุดดัชนี คำที่ระบบสร้างเก็บไว้ล่วงหน้า
4. หน่วยจัดลำดับผลลัพธ์จากการค้นคืน (Ranking Module)
   มีหน้าที่ในการจัดเรียงลำดับของเอกสารที่ได้จากการค้นคืน
5. หน่วยประมวลคิวรี่จากผู้ใช้ (Query Processing Module)
   ผู้ใช้ส่งคิวรี่ให้แก่ระบบจากนั้นระบบต้องทำการประมวลคิวรี่ เพื่อให้รู้ความต้องการค้นคืนของผู้ใช้
6. หน่วยเชื่อมต่อระหว่างผู้ใช้และระบบ (User Interface Module)
   มีหน้าที่ในการรับคิวรี่จากผู้ใช้และส่งผ่านให้ระบบประมวลผลและมีหน้าที่ในการแสดงผลลัพธ์จากการค้นคืนให้กับผู้ใช้

ภาพที่ 2.4

ส่วนประกอบของระบบค้นคืนข้อมูลทั่วไป



งานวิจัยที่เกี่ยวข้อง

Hiromitsu N., Seiichi N. (2000) ทำระบบสำหรับการค้นหาข้อมูลเสียงข่าวโดยใช้คำสำคัญและความคล้ายคลึงระหว่างคำ โดยใช้วิธี novel method ซึ่งตัดคำที่ไม่สำคัญออกโดยจับกลุ่มความคล้ายคลึงของคำและเรียนรู้คะแนนของคำสำคัญ จะทำให้มีเปอร์เซ็นต์ความแม่นยำมากขึ้น ได้ใช้คลังข้อมูลข่าวของ Japanese NHK broadcast news ระหว่างวันที่ 1 มิถุนายน ถึง 14 กรกฎาคม 1996 จำนวนข้อความ 7,099 มาทำการทดลอง โดยคลังข้อมูลข่าวนี้ครึ่งหนึ่งจะมีเสียงเพลงเป็นเสียงรบกวนประโยค

Amarin D., Asanee K., (2004) เสนอการรู้จำเสียงพูดตัวเลขภาษาไทยที่เชื่อมต่อกัน โดยใช้ผู้พูด 100 คน (ชาย 50 คน หญิง 50 คน) อายุระหว่าง 20-28 ปี เป็นชุดฝึกฝน และ ผู้พูด50 คน (ชาย 25 คน หญิง 25 คน) เป็นชุดทดสอบ โดยใช้ HMM จะได้อัตราการรู้จำ 75.25% สำหรับการรู้ความยาวคำ และ 70.33% สำหรับการไม่รู้ความยาวของคำ

C. Haruechaiyasak et al., (2006) เสนอ Sansarn Look ซึ่งเป็นแพลตฟอร์มสำหรับการสืบค้นข้อมูลในภาษาไทย โดยการเลือกดัชนีสำหรับภาษาไทยนั้น มี 2 ขั้นตอน คือ 1. inverted index เป็นการตัดคำโดยแยกเป็นคำแต่ละคำ ตัวเลข และอักขระพิเศษ 2. suffix array เป็นการดูโครงสร้างของคำในประโยค

Yung H., Jeong S., Kyung M., (2007) ทำระบบค้นหาคำสำคัญในข่าวออนไลน์ และตัดบทความที่มีคำสำคัญนั้นออกมา ได้ใช้คลังข้อมูลข่าวของ Korean broadcast news ใช้เทคนิค LVCSR, Phonme recognizer, whole-word model และ DTW-based confidence measure โดยที่สามารถตรวจจับคำที่ไม่ถูกต้องได้ 50%

Michal Fapso (2007) ทำการค้นหาข้อมูลเสียง ระบบนี้จะเสนอ LVCSR, Phoneme recognizer และทั้งสองแบบรวมกัน โดยใช้ค่าถ่วงน้ำหนักเป็นตัววัดความถูกต้อง ซึ่งผลลัพธ์ที่ได้ปรากฏว่า LVCSR รวมกับ Phoneme recognizer จะให้ผลความถูกต้องมากที่สุด เพราะ LVCSR ไม่สามารถค้นหาคำที่ไม่มีในพจนานุกรม (OOV) แต่ระบบ Phoneme recognizer สามารถค้นหาคำที่ไม่มีในพจนานุกรมได้

Boxing C., Marcello F., Mauro C., (2007) ทำการแปลภาษาจีนเป็นภาษาอังกฤษ โดยข้อมูลที่ใช้ทดสอบคือ NIST Chinese-to-English สำหรับวิธี word-based n-gram Language Model จะช่วยเพิ่มประสิทธิภาพของ N-best ในการถอดรหัสให้มากขึ้น

Jongtaveesataporn M., Wutiwiwatchai C., Iwano K., Furui S., (2008) ทำฐานข้อมูลเสียงข่าวภาษาไทยให้ใหญ่ขึ้นซึ่งประกอบด้วยข้อมูลเสียงข่าว 17 ชั่วโมงและข้อมูลถอดความเสียง

ข่าว 35 ชั่วโมง โดยใช้โปรแกรม transcriber และทดสอบด้วย DARPA HUB-4 จากการทดสอบปรากฏว่าขนาดของคลังข้อมูลคำที่ได้จากการถอดความยังมีขนาดเล็ก แบบจำลองทางภาษาที่ได้จึงยัง ทำนาย 2-grams และ 3-grams ได้ไม่ดีนัก

Niran A., Choochart H., Sanparith M., (2008) ทำ Thai Q-Cor (Thai Query Correction) สามารถยืนยันความถูกต้องและความคลาดเคลื่อนของคิวรี่ Thai Q-Cor เป็นการรวมวิธี Word Approximation และ Soundex สำหรับการแก้ไขคิวรี่ภาษาไทย โดย Word Approximation จะแก้ปัญหาความผิดพลาดในการพิมพ์จะใช้เทคนิค beam search, Soundex จะแก้ไขความผิดพลาดในการเรียนรู้ จะใช้เทคนิค phoneme search คะแนนของการแก้ไขจะได้ 88.8% สำหรับ Word Approximation และ 80.0% สำหรับ Soundex

Nawaporn L., Worapog K., (2008) นำเสนอเทคนิคการสืบค้นข้อมูลในภาษาไทย ซึ่งโดยทั่วไปสำหรับการสืบค้นข้อมูลในภาษาไทยนั้นจะใช้วิธี STC (Suffix tree clustering) แต่ผลลัพธ์ที่ได้จากการสืบค้นนั้นยังได้ประโยคที่ไม่สมบูรณ์นัก ในงานวิจัยจะนำเสนอ 3 ขั้นตอน คือ 1. การแยกคำในประโยคโดยใช้ maximal matching algorithm และตัดคำหยุด (stop word) 2. ใช้ STC และเรียงลำดับของกลุ่มคำใหม่ และ 3. สังเคราะห์กลุ่มคำใหม่ ผลที่ได้ในการสืบค้นข้อมูลจะได้ประโยคที่สมบูรณ์มากขึ้น

และในปัจจุบันนี้การสืบค้นข้อมูลด้วยเสียงได้พัฒนาไปถึงการให้บริการสืบค้นข้อมูลบนโทรศัพท์มือถือ ซึ่งจะมี Tellme ของ Microsoft ได้ให้บริการค้นหาข้อมูลที่เป็นลักษณะเฉพาะ เช่น ข้อมูลภาพยนตร์ แผนที่ หรือ บริการบอกเส้นทาง, OneSearch ของ Yahoo โดยสืบค้นข่าวสาร รูปภาพ พยากรณ์อากาศ และข้อมูลการเงิน แบบทันเหตุการณ์, VoiceSearch ของ Google ซึ่งเป็นแอพพลิเคชั่นบนไอโฟน โดยลักษณะการค้นหานั้นจะทำการแปลงเสียงผู้ใช้เป็นไฟล์ดิจิตอลแล้วจึงส่งไปยังเซิร์ฟเวอร์ ของ google เพื่อค้นหา

อย่างไรก็ตาม งานวิจัยนี้ไม่สามารถเปรียบเทียบกับงานวิจัยเก่าๆที่เคยทำมาได้เนื่องด้วยจุดประสงค์การใช้งานที่แตกต่างกัน โดยข้อมูลที่นำมาใช้ทำแบบจำลองหน่วยเสียงและแบบจำลองภาษามีความแตกต่างกัน