การรู้จำเสียงพูดแบบทนทานต่อเสียงรบกวนสำหรับภาษาไทย โดยใช้อัลกอริทึม N-best LIMABEAM

Noise Robust Speech Recognition for Thai Language Using N-best LIMABEAM Algorithm

รุ่งโรจน์ กอปัญญาพิพัฒน์ และรัชฎา คงคะจันทร์*

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

Rungroj Kopanyapipat and Rachada Kongkachandra*

Department of Computer Science, Faculty of Science and Technology, Thammasat University,

Rangsit Centre, Khlong Nueng, Khlong Luang, Pathum Thani 12120

บทคัดย่อ

บทความนี้เสนอวิธีรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนในสภาพแวดล้อมจริง ซึ่งประสิทธิภาพในการรู้จำสามารถเพิ่มขึ้นได้ โดยการใช้ไมโครโฟนอาเรย์ (microphone array) แล้วใช้อัลกอริทึม N-best LIMABEAM ในการรู้จำเสียงพูด ด้วยอัลกอริทึมนี้สามารถที่จะทำให้ได้ค่าลักษณะสำคัญของเสียงที่มีเสียงรบกวนที่ดีขึ้น ทำให้ได้ค่าสมมุติฐานจากการถอดรหัสเสียงที่ดีขึ้นในขั้นตอนการรู้จำในขั้นแรก แล้วหาค่าการถอดรหัสเสียงที่ดีที่สุดในแต่ละค่าสมมุติฐานจากการถอดรหัสเสียงด้วยวิธีการจากอัลกอริทึม LIMABEAM แล้วสุดท้ายได้ผลลัพธ์ที่ดีที่สุดจากชุดค่าสมมุติฐานของ N-best จากผลการทดลองอัลกอริทึม N-best LIMABEAM ได้ค่าความถูกต้องอยู่ที่ 27.22 % ซึ่งดีกว่าอัลกอริทึม LIMABEAM 20.12 % และเสียงที่มีเสียงรบกวน 9.47 % ในสภาพแวดล้อมที่ถูกกำหนดเสียงรบกวนและเสียงกังวาน

คำสำคัญ : การรู้จำเสียงพูดแบบทนทาน; เสียงพูดภาษาไทย; การประมวลผลเสียงพูด; การประมวลผลแบบหลายช่องรับสัญญาณ

Abstract

In this paper, we propose robust speech recognition for Thai language in real noisy environments. Performance of speech recognition can be increased by using a microphone array and N-best extension of the LIMABEAM algorithm is used for recognition. We show that this algorithm can be used to optimize the noisy acoustic features using the N-best hypothesized

***ผู้รับผิดชอบบทความ :** rdkpansy@gmail.com

DOI 10.14456/tstj.2016.83

transcriptions generated at a first recognition step and then apply LIMABEAM algorithm in each N-best hypothesized transcriptions to get the recognition result. The resulting N-best hypotheses list is automatically re-ranked. Results shows improvements over LIMABEAM algorithm with considerable amount of noise and limited reverberation.

Keywords: robust speech recognition; Thai speech; speech processing; microphone array processing

1. บทนำ

ระบบรู้จำเสียงสำหรับภาษาไทย (Thai speech recognition) ได้มีการศึกษาและทำวิจัยมามากกว่า 40 ปี แล้ว ซึ่งปัจจุบันเทคโนโลยีการรู้จำ speech recognition นี้มีประโยชน์อย่างมาก สามารถนำมาใช้ในชีวิตประจำวันได้หลากหลายอย่างดังต่อไปนี้

1.1 ช่วยในการสั่งงาน [1,2] รถเข็นคนพิการที่สามารถสั่งการได้ด้วยเสียงพูด จะช่วยให้คนพิการเคลื่อนที่ไปในทิศทางต่าง ๆ ด้วยคำสั่งเสียงพูด ทำให้เกิดความสะดวกสบายในการเดินทาง คำสั่งเสียงพูดสำหรับควบคุมอุปกรณ์ไฟฟ้าภายในบ้าน เป็นต้น

1.2 ช่วยในการค้นหาข้อมูล [3] สามารถใช้เสียงพูดในการค้นหาข้อมูลที่ต้องการได้ ดังตัวอย่าง application ในโทรศัพท์ smartphone ได้แก่ Google Voice, S Voice (Samsung), Siri (Apple) เป็นต้น

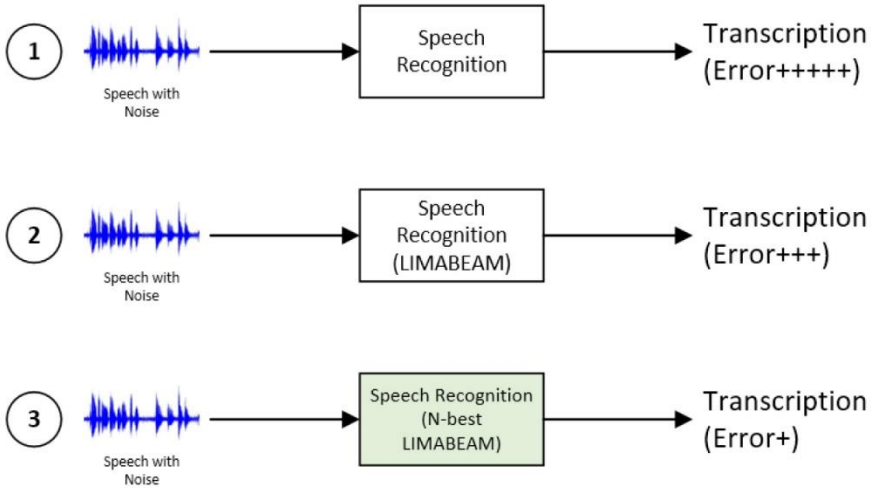
1.3 ช่วยให้ทำงานได้หลายอย่างพร้อมกัน เช่น ในขณะขับรถ สามารถสั่งให้โทรศัพท์โทรออกได้โดยที่ไม่ต้องใช้มือในการกดโทร

1.4 ลดอันตรายในการทำงาน [4] เช่น ในโรงงานอุตสาหกรรม หรือในขณะขับรถ ช่วยให้ผู้ที่ปฏิบัติหน้าที่ในสถานการณ์ที่ไม่สามารถใช้วิธีการกดหรือพิมพ์เพื่อสั่งงานได้ หรือใช้ได้ไม่สะดวก โดยการใช้เสียงพูดในการสั่งการแทน

อย่างไรก็ตาม ระบบรู้จำเสียงภาษาไทยที่มีอยู่ในปัจจุบันก็ยังทำงานได้ไม่ถูกต้อง 100 % ซึ่งก็เกิดจากความผันแปรในด้านต่าง ๆ ที่ขึ้นอยู่กับผู้พูดที่มีความแตกต่างในการพูดของแต่ละคนที่ไม่เหมือนกัน สำเนียงการพูดที่แตกต่างกันออกไปในแต่ละคน ผู้พูดที่เป็นชายหรือหญิงก็จะให้เสียงที่แตกต่างกันออกไป ความผันแปรในด้านโดเมนของเรื่องที่จะพูด การพูดในกลุ่มของเรื่องที่ระบบเคยเรียนรู้ไปแล้ว ก็จะได้ผลที่แตกต่างกับการพูดในกลุ่มของเรื่องใหม่ที่ระบบยังไม่ได้เรียนรู้ ความผันแปรในด้านสภาพแวดล้อม ซึ่งในสภาพแวดล้อมที่แตกต่างกันก็จะให้ผลที่แตกต่างกันออกไป เช่น การใช้งานระบบรู้จำในสภาพแวดล้อมที่มีเสียงรบกวน เช่น เสียงแอร์ในห้อง เสียงเครื่องจักรทำงาน เสียงคนสนทนารอบข้าง เสียงรถยนต์ข้างถนน เสียงกังวานในห้อง ระยะห่างของผู้พูดกับเครื่องบันทึกเสียงที่อยู่ห่างเกินไป ซึ่งก็จะได้ความถูกต้องแตกต่างกันไป

เป็นที่รู้กันว่าระบบรู้จำนั้นยังทำงานได้ไม่ดีพอในสภาพแวดล้อมจริง ซึ่งเป็นสภาพแวดล้อมแบบเปิดที่จะมีเสียงรบกวนรอบข้างตลอดเวลา ทั้งที่เป็นเสียงเครื่องจักรหรือเสียงของผู้พูดที่เราไม่สนใจ ที่ผ่านมามีงานวิจัยทางด้านภาษาไทย เช่น งานของ มนตรี และเฉลิมภัณฑ์ [5] งานของ Hansakunbuntheung และคณะ [6] แต่ส่วนใหญ่จะเป็นแบบไมโครโฟนเดียว (single microphone) และสามารถจัดการได้เฉพาะเสียงรบกวนที่เป็นเครื่องจักร ดังนั้นงานวิจัยนี้จึงนำวิธีการรู้จำเสียงที่ทนทานต่อเสียงรบกวนใด ๆ หรือเสียงที่ไม่ได้อยู่ในความสนใจ โดยการใช้อัลกอริทึม N-best LIMABEAM (likelihood maximizing beamforming) ที่มีช่องทางในการรับเสียงมากกว่า 1 ช่องทาง แทนที่จะเป็นช่องทางเดียว และเลือกผลลัพธ์ที่ดีที่สุดจาก N-best hypotheses list แทนที่จะใช้ผลลัพธ์ที่ดี

สุดเพียงอันเดียว ซึ่งจะทำให้ได้ผลลัพธ์ที่ดีขึ้น จากการทดลองแสดงให้เห็นว่าอัลกอริทึม LIMABEAM สามารถพัฒนาให้ดีขึ้นได้ โดยสามารถสรุปเปรียบเทียบแนวทางการทำวิจัยได้ดังรูปที่ 1



รูปที่ 1 การเปรียบเทียบแนวทางการทำวิจัย

จากรูปที่ 1 วิธีที่ 1 การรู้จำเสียงพูดแบบวิธีปกติที่ไม่มีอัลกอริทึมที่ช่วยในการรู้จำที่ทนทานต่อเสียงรบกวน ก็จะได้ผลลัพธ์ transcription ที่มีค่าความผิดพลาดมากที่สุด ส่วนวิธีที่ 2 นั้นก็จะมีอัลกอริทึมที่ช่วยในการรู้จำแบบทนทานต่อเสียงรบกวน โดยใช้อัลกอริทึม LIMABEAM ซึ่งก็จะช่วยให้ได้ผลลัพธ์ transcription ที่มีค่าความผิดพลาดน้อยกว่าวิธีแรก และ ส่วนวิธีที่ 3 นั้นก็จะใช้อัลกอริทึมที่ทนทานต่อเสียงรบกวนเช่นกัน โดยใช้อัลกอริทึม N-best LIMABEAM ซึ่งเป็นอัลกอริทึมที่พัฒนาต่อยอดจากอัลกอริทึม LIMABEAM ก็จะช่วยให้ได้ผลลัพธ์ transcription ที่มีค่าความผิดพลาดที่น้อยที่สุดใน 3 วิธี

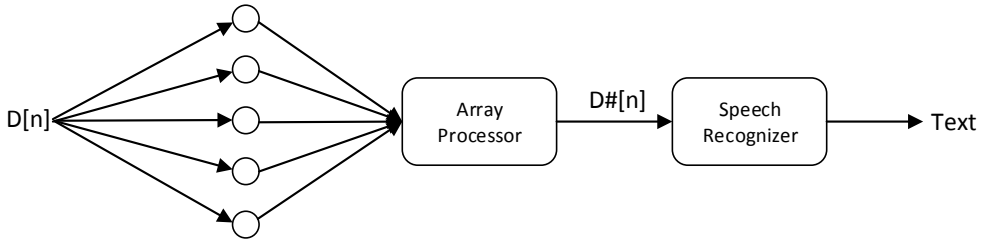
2. Robust Speech Recognition

การรู้จำแบบทนทานต่อเสียงรบกวนนั้น มีหลายวิธีที่ใช้ในการทำให้การรู้จำนั้นทำงานได้ดีในสภาพแวดล้อมที่มีเสียงรบกวน เช่น งานวิจัยของมนตรี และเฉลิมภัณฑ์ [5] นี้เป็นการวิเคราะห์เชิงความถี่ของเสียง โดยวิธีการลบเชิงสเปกตรัม อีกวิธีหนึ่งที่ทำให้การรู้จำทนทานต่อเสียงรบกวน คือ microphone array processing

Microphone array processing เป็นการแก้ปัญหาเรื่องการรับสัญญาณเสียงที่ไม่ได้คุณภาพเนื่องจากมีเสียงรบกวน และด้วยระยะห่างของไมโครโฟนกับผู้พูดที่มีผลต่อคุณภาพสัญญาณเสียง วิธีการนี้จะช่วยลดการบิดเบือนของสัญญาณเสียง และช่วยเพิ่มคุณภาพของสัญญาณ โดยการใช้ไมโครโฟนหลายตัวในการรับสัญญาณ แทนที่จะเป็นไมโครโฟนเพียงตัวเดียว สัญญาณที่ได้ก็จะไปผ่านกระบวนการ array processsing ซึ่งเป็นการรวมสัญญาณที่ได้รับจากแต่ละไมโครโฟนนำมารวมกันเพื่อให้ได้สัญญาณเสียงที่ดี

โดยปกติแล้วการทำงานของระบบรู้จำที่ใช้ไมโครโฟนอาเรย์ จะมีการแยกการทำงานเป็นอิสระต่อกันเป็น 2 ส่วน คือ ส่วนของการทำ array processing และส่วนของการรู้จำ ตามรูปที่ 2 ส่วนของ array processing เป็นส่วนที่จะพิจารณาในขั้นตอน pre-

processing เพื่อที่จะทำให้คุณภาพของสัญญาณเสียงนั้นดีขึ้น ก่อนที่จะส่งไปจำแนกคุณลักษณะสำคัญของเสียงและรู้จำต่อไป



รูปที่ 2 กระบวนการโดยทั่วไปของระบบรู้จำที่ใช้ไมโครโฟนอาเรย์โดยวัตถุประสงค์ของ array processing เพื่อที่จะให้ได้สัญญาณเสียงที่มีคุณภาพดีที่สุด

หลักการนำคลื่นเสียงหลักเพื่อมาใช้ใน microphone array processing นี้เรียกว่า beamforming ซึ่งโดยวิธีทั่วไปของ beamforming นี้เรียกว่า delay-and-sum (Johnson & Dudgeon, 1993) ด้วยวิธีนี้สัญญาณที่ได้รับมาจากไมโครโฟนอาเรย์จะมีการจัดเรียงสัญญาณเสียงของแต่ละไมโครโฟน เพื่อที่จะปรับในเรื่องของความยาวของสัญญาณเสียงที่แตกต่างกันระหว่างแหล่งที่มาของเสียงของไมโครโฟนแต่ละตัว ซึ่งก็มีการต่อยอดจากวิธี delay-and-sum beamforming ไปเป็นวิธี filter-and-sum beamforming ในแต่ละไมโครโฟนก็จะมีการเชื่อมต่อตัวกรองสัญญาณเสียง หลังจากผ่านตัวกรองสัญญาณแล้วก็จะนำมารวมกันเป็นสัญญาณเสียงเดียว

3. LIMABEAM Algorithm

ปัญหาของ microphone array processing ในการรู้จำนั้น เป็นเรื่องของการแยกประเภทของรูปแบบของสัญญาณเสียงที่ถูกแปลงไปเป็นลำดับของชุดค่าคุณลักษณะ (feature vector) ที่จะต้องทำให้ได้กลุ่มของสัญญาณเสียงที่ถูกต้องมากที่สุด อัลกอริทึม LIMABEAM นั้นทำขึ้นมาเพื่อเพิ่มประสิทธิภาพของไมโครโฟนอาเรย์ โดยวิธีการหาชุดของ array parameter ที่มีโอกาสสูงสุดที่จะได้เป็นกลุ่มของสัญญาณเสียงที่ถูกต้อง แล้วใช้ข้อมูลจากการรู้จำ นำกลับมาใช้ในการหา array parameters อีก ซึ่งเป็นการทำให้ประสิทธิภาพของการรู้จำนั้นดีขึ้น อัลกอริทึม LIMABEAM ได้มีการนำเอากระบวนการ filter-and-sum array processing มาใช้ในการช่วยชดเชยเสียงที่ผิดปกติจากเสียงรบกวน และเสียงก้องกังวานได้เป็นอย่างดี โดยสามารถแสดงสูตรทางคณิตศาสตร์ได้ดังต่อไปนี้

$$x[k] = \sum_{m=1}^{M} h_m[k] * s_m[k] \quad (1)$$

โดย $s_m[k]$ คือ สัญญาณเสียงที่แยกตามช่วงเวลาของไมโครโฟนตัวที่ m; $h_m[k]$ คือ ตัวกรองสัญญาณแบบ FIR ของไมโครโฟนตัวที่ m; $x[k]$ คือ ผลลัพธ์ของ beamformer;* คือ convolution; k คือ index ของเวลา

ค่าสัมประสิทธิของชุดตัวกรองสัญญาณเสียง finite impulse response (FIR) ทั้งหมดของไมโครโฟนทุกตัวสามารถแสดงได้เป็นแบบ super-vector h ในแต่ละ frame สามารถรู้จำคุณลักษณะของเสียงได้จากฟังก์ชั่น h ดังนี้

$$y_L(h) = \log_{10}(W \mid FFT(x(h))|^2) \quad (2)$$

โดย $x(h)$ คือobservationvector;$| FFT(x(h))|^2$

คือ vector ของ power spectrum component; W คือ Mel filter matrix; $y_L(h)$ คือ vector ของ log filter bank energy (LFBE)

ค่าสัมประสิทธิ cepstral ได้มาจากการผ่าน DCT rotation ดังนี้

$$y_C(h) = DCT(y_L(h)) \quad (3)$$

ค่าที่ได้จากอัลกอริทึม LIMABEAM นั้นได้มาจากชุดของตัวกรองสัญญาณเสียง FIR ของแต่ละไมโครโฟน ซึ่งความเป็นไปได้มากที่สุดของ $y_L(h)$ ซึ่งจะให้ผลของ state sequence ออกมาเป็นค่าสมมุติฐานของ transcription ที่ถูกประเมินแล้ว ซึ่งแสดงได้ดังนี้

$$\hat{h} = \arg\max_h P(y_L(h)|w) \quad (4)$$

โดย w คือ ค่าสมมุติฐานของ transcription; $P(y_L(h)|w)$ คือ ความน่าจะเป็นของ observation feature จะได้เป็น transcription ที่ถูกพิจารณาแล้ว; $\hat{h}$ คือ FIR parameter ของ super-vector

การทำ optimization จะถูกทำผ่าน non-linear conjugate gradient ส่วน state sequence สามารถประเมินได้จากผลลัพธ์ของ array beamformer (unsupervised LIMABEAM) ซึ่งจะทำงานได้ดีในสภาพแวดล้อมที่มีเสียงรบกวน (noisy) ถึงแม้ว่าจะมีเพียง 1 ช่องทางในการรับสัญญาณเสียงก็ตาม

4. N-best LIMABEAM Algorithm

LIMABEAM อัลกอริทึมนั้นเป็นการเพิ่มความน่าจะเป็นของการทำให้ได้ค่าสมมุติฐานของ transcription จากการทำขั้นตอนการรู้จำเพียงแค่ครั้งเดียว ส่วนในงานวิจัยนี้จะใช้วิธีการทำแบบ N-best เข้ามาประยุกต์ใช้ร่วมกับการรู้จำด้วยอัลกอริทึม LIMABEAM ซึ่งเพิ่มประสิทธิภาพของการรู้จำของอัลกอริทึม LIMABEAM ให้ดียิ่งขึ้น โดยในขั้นตอนแรกจะรู้จำให้ได้ค่าสมมุติฐานของ transcription ตามค่า N-best ที่กำหนด แล้วนำค่าสมมุติฐานของ transcription ที่ได้ไปรู้จำต่อด้วยอัลกอริทึม LIMABEAM จนได้เป็นผลลัพธ์ของแต่ละชุด N-best แล้วเลือกผลลัพธ์ที่ดีที่สุดจากชุด N-best ทุกตัว ก็จะได้เป็นผลลัพธ์สุดท้ายเป็นค่าสมมุติฐานของ transcription ของอัลกอริทึมนี้

ขั้นตอนของการทำ N-best LIMABEAM ในแต่ละชุด ข้อมูลชุดสัญญาณเสียงจะถูกจัดเรียงสัญญาณแล้วไปผ่านขั้นตอนการ optimization แล้วนำชุดข้อมูลสัญญาณเสียงที่ได้ไปผ่านตัวกรองสัญญาณ นำค่าที่ได้ไปหาค่าคุณลักษณะสำคัญอันใหม่ แล้วค่อยส่งไปรู้จำต่อก็จะได้ค่าสมมุติฐานของ transcription ที่ถูกเลือก ซึ่งสามารถแสดงได้ดังสูตรนี้

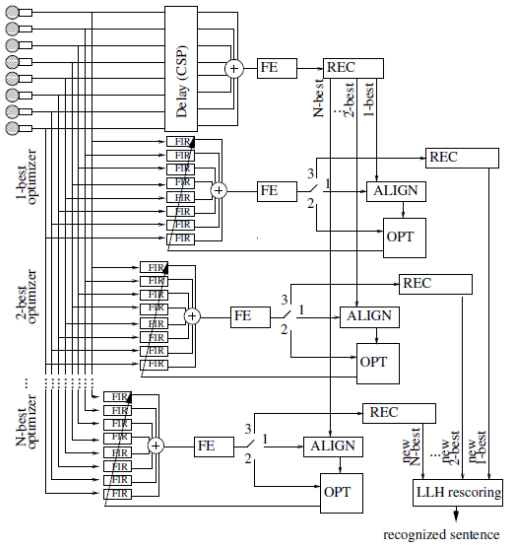
$$\hat{n} = \arg\max_n P(y_C(\hat{h}_n)|\hat{w}_n) \quad (5)$$

โดย $\hat{w}_n$ คือ transcription ที่ได้มาจากกระบวนการรู้จำในครั้งที่ 2; $\hat{n}$ คือ เป็น index ของ transcription ที่มีความน่าจะเป็นสูงสุด ซึ่ง คือ $\hat{w}_n$

การ optimization นั้นถูกทำเสร็จเรียบร้อยในขั้นตอนการทำ log filter bank energies (LFBE) domain ส่วนกระบวนการรู้จำนั้น ถูกทำในขั้นตอน cepstral domain ค่าของความน่าจะเป็นได้ถูกปรับในขั้นตอน cepstral domain เช่นเดียวกัน ภาพรวมของระบบที่จะนำ N-best มาใช้สามารถแสดงดังรูปที่ 3

จากรูปที่ 3 สัญญาณเสียงที่ได้มาจากกระบวนการของไมโครโฟนอาเรย์ ผ่านวิธีปกติแบบ delay and sum (D&S) หลังจากนั้นก็มาทำสกัดหาค่าคุณลักษณะสำคัญ feature extraction (FE) แล้วตามด้วยขั้นตอนการรู้จำ recognition (REC) ในครั้งแรกกระบวนการรู้จำโดยใช้ HMM ได้สร้างค่าสมมุติฐานของ transcription เป็นแบบ N-best ขึ้นมา แล้วใช้อัลกอริทึม LIMABEAM ในขั้นตอนแรกจะเลือก state sequence โดยใช้ Viterbi อัลกอริทึมหรือเรียกว่า Viterbi alignment (switch to 1: ALIGN) แล้วแก้ไข หลังจากนั้นค่าสัมประสิทธิของตัวกรอง FIR จะถูก

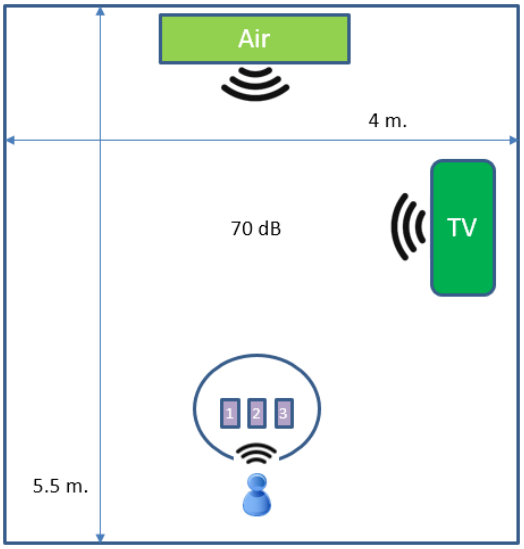
optimization โดยใช้ conjugate gradient (switch to 2: OPT) หลังจากที่ได้ค่าที่ convergence กันแล้ว ก็จะเข้าสู่การรู้จำ N-best feature ที่ได้มา (switch to 3: REC) และหลังจากนั้น ชุดของ transcription อื่น ๆ ก็ถูกดำเนินการตามไป สุดท้ายก็จะเปรียบเทียบค่าสมมุติฐานของ transcription ที่ได้จากแต่ละชุด N-best log-likelihood (LLH-rescoring) โดยการเลือกค่าที่สูงที่สุดก็จะได้เป็นประโยคที่ถูกเลือกออกมาเป็นผลลัพธ์ของอัลกอริทึมนี้



รูปที่ 3 Block diagram ของ N-best LIMABEAM

5. สภาพแวดล้อมที่ใช้ในการทดลอง

ห้องที่ใช้บันทึกเสียงสำหรับการทดลองมีขนาด 5.5x4 เมตร ในห้องมีเสียงรบกวนเป็นเสียงเครื่องปรับอากาศและเสียงโทรทัศน์ โดยวัดค่าความดังได้ 70 dB บันทึกเสียงโดยใช้โทรศัพท์ smartphone จำนวน 3 เครื่อง ประกอบด้วย Samsung Galaxy Note II, Samsung Galaxy Note III และ Samsung Galaxy Tab S ด้วยค่า sample rate 16 kHz และค่า bit-depth 16 bit วางเรียงห่างกัน 15 เซนติเมตร ระยะห่างระหว่างผู้พูดกับโทรศัพท์ 30 เซนติเมตร เนื่องด้วยระยะห่างระหว่างผู้พูดกับโทรศัพท์เป็นระยะ ห่างของการพูดผ่านไมโครโฟนที่ตั้งอยู่บนโต๊ะแบบปกติทั่วไป จึงทำให้ต้องจำกัดจำนวนไมโครโฟนให้ได้ตามขนาดของพื้นที่ตั้งบนโต๊ะ อีกทั้งด้วยขนาดของโทรศัพท์ที่มีขนาดใหญ่ ตามรูปที่ 4



รูปที่ 4 ห้องบันทึกเสียงการทดลอง

6. การทดลองและผลการทดลอง

การทดลองนี้ได้ใช้ HTK HMM-based recognizer เป็นเครื่องมือในการรู้จำเสียง และใช้โปรแกรม MATLAB ในการประมวลผลเสียง ซึ่งภาพรวมของการทดลองมีขั้นตอนในการทดลองตามรูปที่ 5 โดยมีรายละเอียดดังนี้

6.1 ขั้นการฝึกฝน (Train)

ข้อมูลเสียงเพื่อจัดทำเป็นฐานข้อมูลรูปแบบเสียง (acoustic model) ในแบบ HMM-based โดยบันทึกเสียงด้วย microphone ตัวเดียว และอยู่ในสภาพแวดล้อมที่ไม่มีเสียงรบกวน เพื่อให้ได้เสียงที่สะอาด (clean signal) การบันทึกเสียงนี้จะใช้บทสนทนาการจองโรงแรมเป็นรูปแบบประโยคในการฝึกฝน ซึ่งมีจำนวนประโยคทั้งหมด 455 ประโยค ใน

การบันทึกเสียงนี้ใช้เสียงพูดของผู้ชายจำนวน 231 ประโยค เสียงพูดของผู้หญิงจำนวน 214 ประโยค มีจำนวนคำทั้งหมด 4,069 คำ หลังจากบันทึกเสียงเสร็จแล้ว ก็นำไฟล์เสียงที่ได้ไปสกัดหาค่าคุณลักษณะสำคัญ (feature extraction) เมื่อได้ค่าคุณลักษณะสำคัญแล้วก็นำไป training โดยใช้ข้อมูลบทสทนา (transcription) ร่วมด้วย สุดท้ายก็จะได้เป็นฐานข้อมูลรูปแบบเสียงเพื่อเตรียมเอาไว้ใช้ในขั้นตอนการทดสอบต่อไป

6.2 ขั้นการทดสอบ (Test)

โดยใช้อัลกอริทึม N-best LIMABEAM ขั้นตอนนี้จะบันทึกเสียงในสภาพแวดล้อมที่มีเสียงรบกวน เพื่อนำสัญญาณเสียงที่ได้ไปผ่านกระบวนการรู้จำเสียงโดยใช้อัลกอริทึม N-best LIMABEAM เริ่มจากการบันทึกเสียงโดยใช้ ไมโครโฟนอาเรย์ และอยู่ในสภาพแวดล้อมที่มีเสียงรบกวนรอบข้าง แล้วนำไฟล์เสียงไปเข้ากระบวนการรู้จำเสียงโดยใช้อัลกอริทึม N-best LIMABEAM ซึ่งจะมีการสกัดหาค่าคุณลักษณะสำคัญ (feature extraction) และการถอดค่าคุณลักษณะสำคัญ (decoding) โดยใช้ฐานข้อมูลรูปแบบเสียงที่ได้จัดทำไว้แล้วในขั้นตอนการฝึกฝน เป็นข้อมูลตั้งต้นในการรู้จำเสียง จนสุดท้ายก็จะได้ผลลัพธ์เป็นข้อมูลบทสทนา (transcription) ออกมาพร้อมค่าความถูกต้องแม่นยำ (% correct and % accuracy)

6.3 ขั้นการทดสอบ (Test)

โดยใช้อัลกอริทึม LIMABEAM ในขั้นตอนนี้จะนำไฟล์เสียงที่ได้จากการบันทึกเสียงในขั้นตอนการทดสอบโดยใช้อัลกอริทึม N-best LIMABEAM นำไปผ่านกระบวนการรู้จำเสียงโดยใช้อัลกอริทึม LIMABEAM โดยเริ่มจากการสกัดหาค่าคุณลักษณะสำคัญ (feature extraction) และการถอดค่าคุณลักษณะสำคัญ (decoding) โดยใช้ฐานข้อมูลรูปแบบเสียงที่ได้จัดทำไว้แล้วในขั้นตอนการฝึกฝน เป็นข้อมูลตั้งต้นในการรู้จำเสียง จนสุดท้ายก็จะได้ผลลัพธ์เป็นข้อมูลบทสทนา (Transcription) ออกมาพร้อมค่าความถูกต้องแม่นยำ (%Correct and %Accuracy)

6.4 ขั้นการทดสอบ (Test)

โดยใช้วิธีการรู้จำแบบปกติทั่วไป (recognition) ในขั้นตอนนี้จะนำไฟล์เสียงที่ได้จากการบันทึกเสียงในขั้นตอนการทดสอบโดยใช้อัลกอริทึม N-best LIMABEAM แต่นำเพียงไฟล์เสียงจากไมโครโฟนเพียงตัวเดียว นำไปผ่านกระบวนการรู้จำเสียง โดยเริ่มจากการสกัดหาค่าคุณลักษณะสำคัญ (feature extraction) และการถอดค่าคุณลักษณะสำคัญ (decoding) โดยใช้ฐานข้อมูลรูปแบบเสียงที่ได้จัดทำไว้แล้วในขั้นตอนการฝึกฝน เป็นข้อมูลตั้งต้นในการรู้จำเสียง จนสุดท้ายก็จะได้ผลลัพธ์เป็นข้อมูลบทสทนา (transcription) ออกมาพร้อมค่าความถูกต้องแม่นยำ (% correct and % accuracy)

6.5 ขั้นการเปรียบเทียบค่าความแม่นยำ (Accuracy Comparison)

ในขั้นนี้ก็จะนำผลลัพธ์ที่ได้จากการทดสอบด้วยอัลกอริทึม N-best LIMABEAM, LIMABEAM และ recognition มาเปรียบเทียบค่าความแม่นยำของผลลัพธ์ที่ได้ ว่าวิธีไหนจะมีความแม่นยำมากกว่ากัน ดังผลในตารางที่ 1

ตารางที่ 1 เปรียบเทียบค่าความถูกต้อง

วิธีการ	ความถูกต้อง	การพัฒนาเชิงสัมพันธ์
เสียงที่มีเสียงรบกวน	9.47 %	-
LIMABEAM	20.12 %	11.76 %
N-best LIMABEAM	27.22 %	19.61 %

ผลการทดลองการรู้จำเสียงที่มีเสียงรบกวนด้วยอัลกอริทึม N-best LIMABEAM ให้ค่าความถูกต้องดีที่สุดที่ 27.22 % ซึ่งมีค่าการพัฒนาเชิงสัมพันธ์อยู่ที่

19.61 % เทียบกับค่าของเสียงที่มีเสียงรบกวน โดยใช้ค่า N-best อยู่ที่ 10 ส่วนอัลกอริทึม LIMABEAM ให้ค่าความถูกต้องที่ 20.12 % ซึ่งมีค่าการพัฒนาเชิงสัมพันธ์อยู่ที่ 11.76 % ส่วนเสียงที่มีเสียงรบกวน ให้ค่าความถูกต้องที่ 9.47 %

7. อภิปรายผลการทดลอง

งานวิจัยนี้เป็นการทดลองกับเสียงพูดภาษาไทยที่เป็นประโยคบทสนทนา ซึ่งที่ผ่านมายังไม่มีการทำวิจัยในมุมมองนี้มาก่อน จากผลการทดลองอัลกอริทึม N-best LIMABEAM ให้ค่าความถูกต้องดีที่สุดที่ 27.22 % ซึ่งพัฒนาได้ดีขึ้นจากอัลกอริทึม LIMABEAM แต่ค่าความถูกต้องที่ได้ยังค่อนข้างต่ำอยู่ น่าจะเป็นผลมาจากฐานข้อมูลที่ใช้ในการฝึกฝนเพื่อรู้จำเสียงมีจำนวนที่น้อยเลยทำให้ได้โมเดลเสียงที่ไม่ละเอียดพอที่จะจำแนก observation sequence ที่ส่งเข้าไปจำแนกเสียงว่าเป็นเสียงของ phone ที่ถูกต้องที่สุด อย่างไรก็ตาม จากการทดลองในงานวิจัยนี้ ก็ยังให้ค่าความถูกต้องที่ดีกว่า อันเนื่องจากการใช้อัลกอริทึม N-best LIMABEAM ที่เป็นการเพิ่มโอกาสที่จะได้ค่าสมมุติฐานของ transcription ที่ดีที่สุดจากจำนวนชุดค่าสมมุติฐานที่ได้จากการทำ N-best ซึ่งวิธีนี้ก็จะช่วยเพิ่มผลลัพธ์ที่ให้ได้ค่าความถูกต้องมากยิ่งขึ้น และจากการทดลองยังพบว่าหากยิ่งเพิ่มระดับความดังของเสียงรบกวน ก็จะยิ่งทำให้ค่าความถูกต้องลดลง

การทดลองนี้ยังมีข้อจำกัดอยู่ โดยมีปัญหาจากการบันทึกเสียงในขั้นตอนการฝึกฝน และขั้นตอนการรู้จำ หากมีการออกเสียงที่ไม่ชัดเจน ก็จะทำให้ได้ผลลัพธ์ที่ไม่ดี การใช้ smartphone เป็นไมโครโฟนในการบันทึกเสียง พบว่า smartphone แต่ละเครื่องให้คุณภาพเสียงที่แตกต่างกัน อาจทำให้ได้คุณภาพเสียงหลังจากการรวมเสียงที่แตกต่างกัน การใช้เสียงในการทดสอบการรู้จำคนละเสียงกับที่ใช้ในการฝึกฝนก็มีผลทำให้ได้ค่าความถูกต้องที่ไม่ดี ซึ่งน่าจะเป็นผลมาจากฐานข้อมูลที่ใช้ในการฝึกฝนเพื่อรู้จำเสียงมีจำนวนที่น้อยและมีลักษณะของเสียงไม่หลากหลาย

งานวิจัยในอนาคต สามารถที่จะเพิ่มจำนวนไมโครโฟนให้มีมากยิ่งขึ้น เปลี่ยนเครื่องบันทึกเสียงจาก smartphone เป็นชุดไมโครโฟนชนิดเดียวกัน พัฒนากระบวนการในการรู้จำให้ใช้เวลาเร็วขึ้น ทดสอบกับสภาพแวดล้อมที่หลากหลายมากขึ้น และเพิ่มขนาดฐานข้อมูลเสียงที่ใช้ฝึกฝนให้มากขึ้น รวมถึงให้สามารถรู้จำเสียงพูดได้หลากหลายบุคคล

8. รายการอ้างอิง

[1] ณรงค์รัตน์ เลี้ยวรุ่งโรจน์, อนุพงษ์ ธรรมรักษาสิทธิ์ และโกสินทร์ จำนงไทย, 2546, รถเข็นคนพิการควบคุมด้วยระบบรู้จำเสียงพูด, ภาควิชาวิศวกรรมอิเล็กทรอนิกส์และโทรคมนาคม คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, กรุงเทพฯ.

[2] สกล ลิโมทัย, 2550, ระบบรู้จำเสียงพูดสำหรับควบคุมอุปกรณ์ไฟฟ้าภายในบ้าน, วิทยานิพนธ์ปริญญาโท, บัณฑิตวิทยาลัยมหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ

[3] Jiulong, S., Genqing, W., Zhihong, H., Xiliu, T., Jansche, M. and Moreno, P.J., 2010, Search by voice in Mandarin Chinese, Google China Engineering Center, No. 1 Zhongguancun East Road, Beijing 100084, PRC.

[4] Hataoka, N., Kokzibo, H., Obuchi, Y. and Amano, A., 1998, Development of robust speech recognition middleware on microprocessor, IEEE Int. Conf. ASSP'98, pp. 837-840.

[5] มนตรี โพธิโสโนทัย และเฉลิมภัณฑ์ ฟองสมุทร, 2553, วิธีการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนภายนอก, วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์, มหาวิทยาลัยบูรพา, ชลบุรี.

[6] Hansakunbuntheung, C., Rugchatjaroen, A. and Wutiwiwatchai, C., 2005, Space reduction of speech corpus based on quality perception for unit selection speech synthesis, Proc. SNLP, pp. 127-132.

[7] Seltzer M., 2003, Microphone Array Processing for Robust Speech Recognition, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

[8] Brayda, L., Wellekens, C. and Omologo, M., 2006, Improving robustness of a likelihood-based beamformer in a real environment for automatic speech recognition, Institut Eurecom, 2229 Route des Cretes, 06904 Sophia Antipolis, France. ITC-irst, Via Sommarive 18, 38050 Povo (TN), Italy.

[9] หน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, บทความ Thai speech recognition tutorial, แหล่งที่มา : http://tvis.nectec.or.th/speech/index.php?Itemid=78.