การลดสัญญาณรบกวนของการจดจำเสียงพูดด้วยการแปลงเวฟเล็ต โดยการประมาณค่าจุดเปลี่ยน

Denoise Speech Recognition Based on Wavelet Transform Using Threshold Estimation

ณัฐนันท์ ทัดพิทักษ์กุล[1,2] และบุญธีร์ เครือตราชู[2]

[1]ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

112 อุทยานวิทยาศาสตร์ประเทศไทย ถนนพหลโยธิน ต.คลองหนึ่ง อ.คลองหลวง จ.ปทุมธานี 12120

โทร 0-25646900 ต่อ 2257 โทรสาร 0-25646873 E-mail: Nattanun_t@notes.nectec.or.th

[2]ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง แขวงลาดกระบัง เขตลาดกระบัง กรุงเทพฯ 10520

โทร 02-7392400 ต่อ 125 โทรสาร 02-7392400 E-mail: Boontee@diamond.ce.kmitl.ac.th

บทคัดย่อ

บทความนี้นำเสนอการลดสัญญาณรบกวนด้วยวิธีเวฟเล็ตสำหรับการดึงลักษณะสำคัญ ในงานการประยุกต์การจดจำเสียงพูดแบบอัตโนมัติ การลดสัญญาณรบกวนด้วยวิธีเวฟเล็ตให้การประมาณค่าสัญญาณที่เหมาะสมใกล้เคียงการทำให้สัญญาณราบเรียบ เมื่อมีสัญญาณรบกวนเข้ามา ในบทความนี้มีการนำเสนอขั้นตอนก่อนกระบวนการของการดึงลักษณะสำคัญ โดยขั้นตอนก่อนกระบวนการเป็นการใช้เวฟเล็ตในการลดสัญญาณรบกวนเพื่อทำให้สัญญาณเสียงดีขึ้น โดยใช้วิธีการ soft-thresholding ในการลดสัญญาณรบกวนในสัญญาณเสียงพูด และวัดประสิทธิภาพการจดจำเสียงพูดด้วยวิธีเปอร์เซ็นต์ความถูกต้อง และเปอร์เซ็นต์ความแม่นยำ จากการทดสอบพบว่าสัมประสิทธิเซปสตรอลบนความถี่เมล ที่ผ่านการลดสัญญาณรบกวนด้วยวิธีเวฟเล็ตให้ประสิทธิภาพการจดจำเสียงพูดแบบอัตโนมัติได้ดีขึ้น

คำสำคัญ: การลดสัญญาณรบกวนของเสียงพูด, การประมาณค่าจุดเปลี่ยนของเวฟเล็ต

Abstract

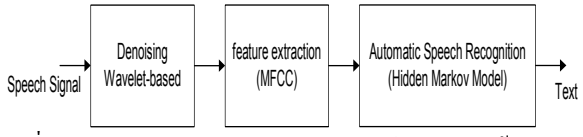
This paper presents the using of wavelet-based denoising for feature extraction in automatic speech recognition applications. Wavelet-based denoising has been found to give a nearly optimal estimation of the piecewise smooth signal that has been corrupted by noise. In this paper, we propose a pre-processing stage before the feature extraction. This pre-processing stage uses wavelet-based denoising to clean the noisy signal. Soft-thresholding scheme is implemented for denoising of the speech signal. To evaluate speech recognition efficiency, the percent correct and percent accuracy are used. As a result, Mel frequency cepstral coefficient features extracted after pre-processing stage based on wavelet denoising gives improve the speech recognition performance of the ASR system.

Keywords: speech denoising, wavelet thresholding

1.คำนำ

ในปัจจุบันเทคโนโลยีที่ทำให้มนุษย์และเครื่องคอมพิวเตอร์สามารถติดต่อสื่อสารกันได้กำลังเป็นที่สนใจ ซึ่งวิธีที่ทำการติดต่อสื่อสารกันนั้นมีอยู่ด้วยกันหลายวิธี แต่ในบทความนี้จะกล่าวถึงเฉพาะวิธีการจดจำเสียงพูด เนื่องจากวิธีนี้กำลังเป็นที่สนใจ โดยเทคนิคการจดจำเสียงพูดในบทความนี้เลือกใช้คือ แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model) และการดึงลักษณะสำคัญของเสียงพูดในบทความนี้ใช้คือ สัมประสิทธิเซปสตรอลบนความถี่เมล (Mel Frequency Cepstral Coefficients) แต่ปัญหาหนึ่งที่เกิดขึ้นในระบบการจดจำเสียงพูดคือสัญญาณรบกวนที่เข้ามาในขณะที่ทำการจดจำเสียงพูด ซึ่งจะส่งผลทำให้ประสิทธิภาพการจดจำเสียงพูดลดลง ดังนั้นในบทความนี้จึงนำเสนอการประยุกต์ใช้ การลดสัญญาณรบกวนด้วยการประมาณค่าจุดเปลี่ยนของเวฟเล็ต (wavelet thresholding) ซึ่งเป็นวิธีการลดสัญญาณรบกวนที่ขึ้นกับค่าจุดเริ่มเปลี่ยน (threshold) วิธีการนี้เป็นวิธีการที่ง่ายและมีประสิทธิภาพ Donoho [1] ได้แสดงให้เห็นว่าการลดสัญญาณรบกวนวิธีนี้มีคุณสมบัติที่เหมาะสมในการลดสัญญาณรบกวน โดยค่าสัมประสิทธิ์ที่ไม่สำคัญและขึ้นกับจุดเริ่มเปลี่ยนจะเปรียบได้กับสัญญาณรบกวน ขณะที่ค่าสัมประสิทธิ์ที่สำคัญจะเปรียบได้กับโครงสร้างหลักของสัญญาณ เห็นได้ว่าการลดสัญญาณรบกวนด้วยวิธีนี้ ได้ดีมากน้อยแค่ไหนจะขึ้นอยู่กับค่าจุดเริ่มเปลี่ยน เพราะถ้ามีค่ามากเกินไปโครงสร้างหลักของสัญญาณก็จะหายไปมาก แต่ถ้าน้อยเกินไปสัญญาณรบกวนก็ยังคงเหลืออยู่ในสัญญาณมากเกินไป ดังนั้นการเลือกค่าจุดเริ่มเปลี่ยน เป็นสิ่งสำคัญในการลดสัญญาณรบกวนด้วยวิธีนี้

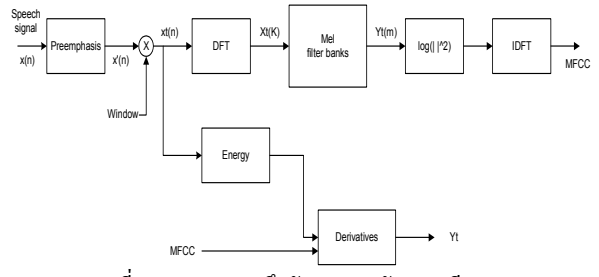
บทความนี้ได้นำเสนอการประยุกต์ใช้การประมาณค่าจุดเปลี่ยนของเวฟเล็ต ในการลดสัญญาณรบกวนของเสียงพูดก่อนเข้าระบบการจดจำด้วยเสียงพูด เป็นดังภาพที่ 1 และทำการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าจุดเริ่มเปลี่ยน 3 วิธี คือ วิธี Universal [2], วิธี BayShrink [3] และวิธี Generalized Cross Validation (GCV) [4] ในการลดสัญญาณรบกวนของเสียงพูด โดยสัญญาณรบกวนที่ใช้ทดสอบเป็นสัญญาณรบกวนจากสิ่งแวดล้อม



ภาพที่ 1 แผนภาพการลดสัญญาณรบกวนของเสียงพูดก่อนเข้าระบบการจดจำด้วยเสียงพูด ด้วยการประมาณค่าจุดเปลี่ยนของเวฟเล็ต

2. การดึงลักษณะสำคัญของเสียงพูด

การดึงลักษณะสำคัญของเสียงพูดในบทความนี้ใช้สัมประสิทธิเซปสตรอลบนความถี่เมล [8] ซึ่งเป็นที่นิยมเป็นลักษณะสำคัญ ที่ใช้กันในการจดจำเสียงพูด เป็นดังภาพที่ 2



ภาพที่ 2 แผนภาพการดึงลักษณะสำคัญของเสียงพูด

3. การลดสัญญาณรบกวนด้วยวิธีเวฟเล็ต

การแปลงเวฟเล็ตแบบดีสครีทแบบ 1 มิติ โดยใช้หลักการของฟิลเตอร์แบงค์ (filter bank) [5,7] มีหลักการวิเคราะห์สัญญาณด้วยฟิลเตอร์แบงค์ กำหนดให้ x(n) คือข้อมูลอินพุต $h_0(n)$ คือตัวฟิลเตอร์ที่กรองความถี่ต่ำ และ $h_1(n)$ คือตัวฟิลเตอร์ที่กรองความถี่สูง สามารถเขียนเป็น สมการการวิเคราะห์ได้ดังสมการที่ 1 และ 2

$$yA(k) = \sum_m h_0(2k-m)x(m) \quad (1)$$

$$yD(k) = \sum_m h_1(2k-m)x(m) \quad (2)$$

เมื่อ yA คือ สัมประสิทธิ์เวฟเล็ตขององค์ประกอบความถี่ต่ำ

yD คือ สัมประสิทธิ์เวฟเล็ตขององค์ประกอบความถี่สูง

มีหลักการสังเคราะห์สัญญาณด้วยฟิลเตอร์แบงค์กำหนดให้ $g_0(n)$ คือ ตัวฟิลเตอร์ที่กรองความถี่ต่ำ และ $g_1(n)$ คือ ตัวฟิลเตอร์ที่กรองความถี่สูง สามารถเขียนเป็นสมการการสังเคราะห์ได้ดังสมการที่ 3

$$x(n) = \sum_m yA(m)g_0(n-2m) + \sum_m yD(m)g_1(n-2m) \quad (3)$$

การประมาณค่าจุดเปลี่ยนของเวฟเล็ตแบบ soft-thresholding [1] เป็นวิธีการลดขนาดของค่าสัมประสิทธิ์ของเวฟเล็ต โดยทุกค่าที่ทำการลดขนาดจะขึ้นอยู่กับค่าจุดเริ่มเปลี่ยน ซึ่งมีวิธีการหาเป็นดังสมการที่ 4 และ 5

$$C = \begin{cases} \text{sgn}(c)(|c| - th) & |c| \geq th \\ 0 & |c| < th \end{cases} \quad (4)$$

$$C = \text{Soft}(c, th) \quad (5)$$

เมื่อ c คือสัมประสิทธิ์เวฟเล็ต

C คือสัมประสิทธิ์เวฟเล็ตที่ผ่านการทำ soft-thresholding

th คือค่าจุดเริ่มเปลี่ยน

การหาค่าจุดเริ่มเปลี่ยนมีความสำคัญต่อการลดสัญญาณรบกวน ด้วยวิธีการประมาณค่าจุดเปลี่ยนของเวฟเล็ตแบบ soft-thresholding เป็นอย่างมาก และในการหาค่าจุดเริ่มเปลี่ยนมีการประมาณค่าอยู่ด้วยกันหลายวิธี ซึ่งแต่ละวิธีก็ให้ผลต่างกันไป ดังนั้นในการนำมาใช้ต้องทำการทดสอบเสียก่อนว่าการประมาณค่าจุดเริ่มเปลี่ยนแบบใด ให้ประสิทธิภาพกับข้อมูลที่เราต้องการลดสัญญาณรบกวนได้ดีที่สุด ในบทความนี้มีการทดสอบวิธีการประมาณค่าจุดเริ่มเปลี่ยนที่นิยมใช้กัน 3 วิธีคือ วิธี Universal , วิธี BayShrink และวิธี Generalized Cross Validation (GCV) ซึ่งแต่ละวิธีเป็นดังสมการที่ 6, 7 และ 8 ตามลำดับ

- วิธี Universal

$$th_i = \frac{\text{median}(|yD_i|)}{0.6745}\sqrt{2 * \log_{10}(N_i)} \quad (6)$$

- วิธี BayeShrink

$$th_i = \frac{\sqrt{\log_{10}\left(\frac{N_i}{J}\right)} * \left[\frac{\text{median}(|yD_i|)}{0.6745}\right]^2}{\text{std}(yD_i)} \quad (7)$$

- วิธี GCV

$$GCV(i, th) = \frac{\frac{\sum(\text{Soft}(yD_i, th) - yD_i)^2}{N_i}}{\left[\frac{N_{0i}}{N_i}\right]^2} \quad (8)$$

$$th_i = \min[GCV(i, th)]$$

เมื่อ th_i คือจุดเริ่มเปลี่ยนที่ระดับแบนด์ย่อยที่ i

yD_i คือสัมประสิทธิ์เวฟเล็ตที่แบนด์ย่อยที่ i

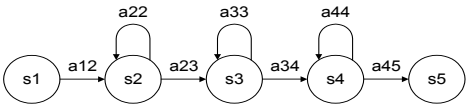
J คือจำนวนระดับการแปลงแบนด์ย่อยทั้งหมด

N_i คือจำนวนสัมประสิทธิ์เวฟเล็ตที่แบนด์ย่อยที่ i

N_{0i} คือจำนวนข้อมูลที่มีค่าเท่ากับศูนย์ของสัมประสิทธิ์เวฟเล็ตที่แบนด์ย่อยที่ i

4. การจดจำเสียงพูด

แบบจำลองที่ใช้ในการจดจำเสียงพูดในบทความนี้คือ แบบจำลองฮิดเดนมาร์คอฟแบบต่อเนื่อง ซึ่งงานวิจัยส่วนมากนิยมใช้เป็นแบบจำลองในการจดจำเสียงพูด ในการจดจำเสียงพูดเป็นการสร้างแบบจำลองฮิดเดนมาร์คอฟเป็นแบบหน่วยพื้นฐานของเสียง (phoneme) และแต่ละแบบจำลองจะใช้แบบจำลองฮิดเดนมาร์คอฟแบบ 5 สถานะ (state) มีการเปลี่ยนสถานะแบบซ้ายไปขวาและแต่ละสถานะเป็นแบบ 1 เกาส์เซียน เป็นดังภาพที่ 3 โดยสามารถดูวิธีการสอน และวิธีการทดสอบได้จาก HTK [9]



ภาพที่ 3 แผนภาพแสดง HMM แบบ 5 สถานะ ที่มีการเปลี่ยนแปลงสถานะแบบซ้ายไปขวา

5. คลังข้อมูลเสียงพูดไทย (Thai speech corpus)

คลังข้อมูลเสียงที่ใช้ในบทความนี้คือ คลังข้อมูลเสียง NECTEC-ATR [10] โดยคลังข้อมูลนี้ประกอบด้วยประโยค 398 ประโยค (เฉพาะชุดที่ใช้ในการทดสอบในบทความนี้) และใช้เจ้าของภาษาเป็นผู้พูดทั้งหมด 42 คน แบ่งออกเป็นผู้ชาย 21 คน และผู้หญิง 21 คน โดยในการบันทึกเสียงจะเก็บเป็นไฟล์แบบ 16 บิต แบบโมโน (mono) และมีอัตราการสุ่มสัญญาณเท่ากับ 16 kHz

6. การวัดประสิทธิภาพการจดจำเสียงพูด

การวัดประสิทธิภาพการจดจำเสียงพูดจะวัดด้วยค่าเปอร์เซ็นต์ความถูกต้อง (Percent correct, PC) และเปอร์เซ็นต์ความแม่นยำ (Percent accuracy, PA) และมีวิธีการหาดังสมการที่ 9 และ 10 ตามลำดับ

$$Percent\quad Correct = \frac{N-D-S}{N}\times 100\% \tag{9}$$

$$Percent\quad Accuracy = \frac{N-D-S-I}{N}\times 100\% \tag{10}$$

เมื่อ N คือจำนวนหน่วยเสียงข้อมูลที่ถูกต้องตามการอัดเสียง

D คือจำนวนหน่วยเสียงข้อมูลที่หายไป

S คือจำนวนหน่วยเสียงข้อมูลที่ผิด

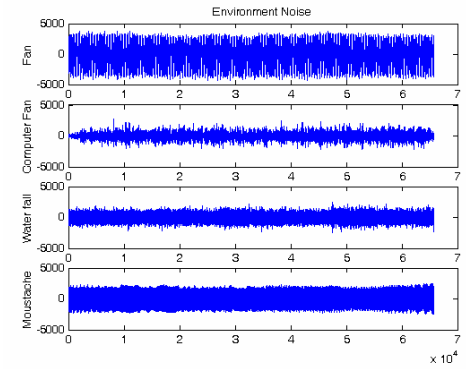
I คือจำนวนหน่วยเสียงข้อมูลที่เกินมา

7. การวิเคราะห์ผล

ในการทดสอบใช้ชุด DB2 ในคลังข้อมูล NECTEC-ATR และมีรายละเอียดของข้อมูลของชุดทดสอบเป็นดังตารางที่ 1 ซึ่งเป็นข้อมูลเสียงแบบการพูดต่อเนื่อง ในชุดการสอนระบบการจดจำเสียงใช้ผู้พูดทั้งหมด 34 คน (ผู้ชาย 17 คน และผู้หญิง 17 คน) และในชุดการทดสอบระบบใช้ผู้พูดทั้งหมด 8 คน (ผู้ชาย 4 คน และผู้หญิง 4 คน) ในการดึงลักษณะของเสียงพูดใช้สัมประสิทธิเซปสตรอลบนความถี่เมล 39 พารามิเตอร์ [9] และในส่วนการลดสัญญาณรบกวนด้วยเวฟเล็ตทำการแปลงเวฟเล็ตแบบ 4 แบนด์ย่อยทุกการทดสอบ และใช้ Daubechies4 (db4) [6] เป็นตระกูลเวฟเล็ตแม่ของฟิลเตอร์ในการแปลงเวฟเล็ต และสัญญาณรบกวนที่ใช้ทดสอบเป็นสัญญาณรบกวนจากสิ่งแวดล้อมคือ เสียงพัดลม (fan), เสียงน้ำตก (water fall), เสียงพัดลมจากคอมพิวเตอร์ (computer fan) และเสียงเครื่องโกนหนวด (moustache) เป็นดังภาพที่ 4 โดยแต่ละสัญญาณที่นำมาทดสอบจะมีอัตราส่วนสัญญาณต่อสัญญาณรบกวน (signal to noise ratio, SNR) เท่ากับ 10 dB

ตารางที่ 1 รายละเอียดชุดข้อมูล DB2

คุณสมบัติ	จำนวนข้อมูล
No. of sentences	398
No. of words	3,377
No. of syllables	5,501
No. of phones	14,472



ภาพที่ 4 สัญญาณรบกวนจากสิ่งแวดล้อม

จากตารางที่ 2 แสดงให้เห็นว่าการลดสัญญาณรบกวนด้วยวิธีการประมาณค่าจุดเปลี่ยนของเวฟเล็ตแบบ soft-thresholding โดยใช้การประมาณค่าจุดเริ่มเปลี่ยนด้วยวิธี GCV สามารถเพิ่มประสิทธิภาพการจดจำเสียงพูดเมื่อมีสัญญาณรบกวนเกิดขึ้น แต่วิธีการประมาณค่าจุดเริ่มเปลี่ยนด้วยวิธีอื่นกลับทำให้ประสิทธิภาพการจดจำเสียงพูดเมื่อมีสัญญาณรบกวนเกิดขึ้นลดลงมากกว่าตอนไม่มีส่วนการลดสัญญาณรบกวนจากเสียงพูด แต่ในกรณีการเพิ่มส่วนการลดสัญญาณรบกวนจากเสียงพูดกลับส่งผลทำให้ประสิทธิภาพการจดจำเสียงพูดที่ไม่มีสัญญาณรบกวนเกิดขึ้นลดลง ซึ่งวิธีประมาณค่าจุดเปลี่ยนด้วยวิธี GCV และ Bayshrink ทำให้ประสิทธิภาพการจดจำเสียงพูดลดลงเล็กน้อย แต่วิธีประมาณค่าจุดเปลี่ยนด้วยวิธี Universal กลับทำให้ประสิทธิภาพการจดจำเสียงพูดลดลงเป็นอย่างมาก

ตารางที่ 2 เปรียบเทียบประสิทธิภาพการจดจำเสียงพูด ในการลดสัญญาณรบกวนด้วยการประมาณค่าจุดเปลี่ยนของเวฟเล็ต

เสียงสัญญาณรบกวน	ไม่มีการลดสัญญาณรบกวน		GCV		Bayeshrink		Universal	
	PC	PA	PC	PC	PC	PA	PC	PA
ไม่มีสัญญาณรบกวน	60.55	53.52	60.35	52.64	60.25	52.64	45.90	34.67
เสียงน้ำตก	33.30	24.61	37.60	28.13	35.25	26.46	33.89	23.05
เสียงพัดลม	43.55	21.09	41.99	20.21	43.75	19.73	38.48	14.65
เสียงพัดลมคอมพิวเตอร์	47.85	38.28	48.73	38.67	48.34	37.30	40.53	23.14
เสียงเครื่องโกนหนวด	46.58	37.11	48.05	38.77	38.28	29.00	28.65	14.71

8. บทสรุป

บทความนี้เป็นการนำเสนอแนวิธีในการลดสัญญาณรบกวนจากสิ่งแวดล้อมของเสียงพูด เพื่อเพิ่มประสิทธิภาพของการจดจำเสียงพูดด้วยวิธีการประมาณค่าจุดเปลี่ยนของเวฟเล็ตแบบ soft-thresholding จากผลการทดสอบดังที่ได้นำเสนอพบว่าการลดสัญญาณรบกวนด้วยวิธีการประมาณค่าจุดเปลี่ยนของเวฟเล็ตแบบ soft-thresholding และใช้การประมาณค่าจุดเริ่มเปลี่ยนด้วยวิธี GCV ให้ประสิทธิภาพการจดจำเสียงพูดได้ดีกว่าการหาค่าจุดเริ่มเปลี่ยนด้วยวิธีอื่นที่มาผ่านกระบวนการจดจำเสียงพูดด้วยวิธีเดียวกัน

9. เอกสารอ้างอิง

[1] D. Donoho, “De-noise by soft-thresholding”, IEEE trans. Inform. Theory, vol 41,pp 613-627, Mar. 1995

[2] I. Johnstone and B. Silverman, “Wavelet threshold estimators for data with correlated noise”, J. Roy. Statist. Soc. B, vol.59, pp.319-351,1997

[3] S. G. Chang, B. Yu and M. Vetterli, “Adaptive Wavelet Thresholding for Image Denoising and Compression”, IEEE trans. On Image Processing. Vol.9, pp1532-1546, September 2000.

[4] M. Jansen and A. Bultheel, “Multiple Wavelet Threshold Estimation by Generalized Cross Validation for Images with Correlated Noise”, IEEE trans. On Image Processing. Vol 8, no 7, pp 947-953, July 1999.

[5] C. S. Burrus, R. A. Gopinath and H. Guo. “Introduction to Wavelets and Wavelet Transforms.” New York: Prentice-Hall International, Inc. 1998.

[6] S. G. Mallat. “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation.” IEEE Transaction on Patten and Machine Intelligence. Vol 11, July 1989. pp. 674-693.

[7] Olivier, R., and Martin, V. (1991). Wavelets and signal processing. IEEE SP Magazine. 14-38.

[8] C. Becchetti and L. P. Ricotti. “Speech Recognition Theory and C++ Implementation” New York: JOHN WILEY & SONS. 1999.

[9] S. Young, et al. “The HTK book (for HTK version 3.1)”, July,2000.

[10] S. kasuriya, et al. “Thai Speech Database for Speech Recognition (NECTEC-ATR Thai Speech Database)”, Proceedings of Oriental COCOSDA 2003. International Coordinating Committee on Speech Databases and Speech I/O System Assessment. October 2003. pp 105-111.

ณัฐนันท์ ทัดพิทักษ์กุล สำเร็จการศึกษาระดับปริญญาโท ด้านวิศวกรรมไฟฟ้า จากมหาวิทยาลัยเทคโนโลยีสุรนารี ในปี 2545 ปัจจุบันดำรงตำแหน่งเป็นผู้ช่วยนักวิจัย ประจำศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ มีความสนใจทางด้าน Speech Recognition, Wavelet Transform และ Digital Signal Processing

รศ. ดร. บุญธีร์ เครือตราชู สำเร็จการศึกษาระดับปริญญาเอก ด้านวิศวกรรมคอมพิวเตอร์ จาก Oregon State University ในปี 2530 ปัจจุบันดำรงตำแหน่งเป็นอาจารย์ประจำสาขาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง มีความสนใจทางด้าน data mining, pattern recognition, machine learning และ evolutionary algorithm