การรู้จำหน่วยเสียงสระเสียงเดี่ยวสำหรับภาษาไทยโดยการใช้ ทรานส์เฟอร์ฟังก์ชันของอวัยวะกำทอนเสียงบนสเกลบาร์ก

Unmixed Vowel Utterances Recognition in Thai spoken language using Vocal Tract Transfer functions on the Bark scale

นริศ บุญศักดิ์เฉลิม วรา คงคาวิทูร ไกรสิน ส่งวัฒนา

ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

3-2 ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพ 10520

โทร: 662 3269967 ; Email: kskraisi@kmitl.ac.th

ABSTRACT - In this paper we address the problem of unmixed vowel utterances recognition in Thai spoken language. The vowel utterances to be recognized are "i", "e", "ε", "ω", "γ", "a", "u", "o", and "⊃". Each utterance is represented by a 18-dimensional feature vector of Critical Band Intensities of the vocal tract transfer function which is determined from Linear Predictive Coding Coefficients. Using the K-Nearest Neighbor rule on feature vectors of sample utterance show the maximum accuracy of 96%.

KEYWORDS - Vowel Recognition, Bark scale, K-Nearest Neighbor

บทคัดย่อ – บทความนี้เสนอวิธีการในการรู้จำหน่วยเสียงสระเสียงเดี่ยวในภาษาไทย และวิธีการปรับปรุงระบบเพื่อให้ได้ประสิทธิภาพสูง โดยหน่วยเสียงสระที่จะทำการรู้จำ คือ เสียง อิ, เอะ, แอะ, อึ, เออะ, อะ, อุ, โอะ และเอาะ แต่ละหน่วยเสียงนั้นจะถูกแทนโดยเวคเตอร์ 18 มิติ โดยแต่ละมิติ คือ Critical Band Intensity ของ Transfer Function ของอวัยวะกำทอนเสียงในการกำเนิดเสียงนั้นๆ ซึ่งคำนวณจากสัมประสิทธิ์ที่ได้จากการทำ Linear Predictive Coding (LPC) การทดลองจะทำการเก็บตัวอย่างเสียงมาทำการคำนวณ Transfer Function ของอวัยวะกำทอนเสียงจาก LPC อันดับต่างๆ แล้วจึงใช้เทคนิค K-Nearest Neighbor ในการแยกแยะเวคเตอร์ ผลการทดลองพบว่าได้ความถูกต้องเฉลี่ยในการรู้จำสูงสุดเท่ากับ 96%

คำสำคัญ - การรู้จำเสียงสระ, สเกลบาร์ก, K-Nearest Neighbor

1. บทนำ

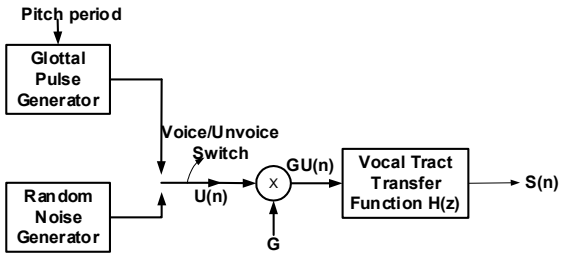
เสียงสระในภาษาไทยประกอบด้วยเสียงสระแท้ทั้งหมด 24 เสียง โดยประกอบไปด้วยเสียงสระเสียงเดี่ยวเสียงสั้นจำนวน 9 เสียง คือ อิ, เอะ, แอะ, อึ, เออะ, อะ, อุ, โอะ และเอาะ เสียงสระเสียงเดี่ยวเสียงยาว 9 เสียง คือ อี, เอ, แอ, อือ, เออ, อา, อู, โอ และออ และเสียงสระผสมอีก 6 เสียง คือ เอียะ, เอีย, เอือะ, เอือ, อัวะ และอัว เสียงสระเสียงเดี่ยวเสียงยาวนั้นเกิดจากการเปล่งหน่วยเสียงสระเสียงเดี่ยวเสียงสั้นในช่วงเวลาที่ยาวขึ้นและมีการหยุดของเสียงที่ช้ากว่าเสียงสระเสียงเดี่ยวเสียงสั้น ส่วนสระผสมนั้นเกิดจากการรวมกันของหน่วยเสียงสระเสียงเดี่ยว 2 ชนิด ดังนั้นในการที่จะรู้จำเสียงสระในภาษาไทยนั้นจะต้องเริ่มต้นจากการรู้จำหน่วยเสียงสระเสียงเดี่ยวให้ได้เสียก่อน

ในบทความนี้นำเสนอวิธีการรู้จำหน่วยเสียงสระเสียงเดี่ยวสำหรับภาษาไทยโดยอาศัย Transfer Function ของอวัยวะกำทอนเสียงบนสเกล Bark ซึ่งสามารถคำนวณหาได้โดยตรงจากสัมประสิทธิ์ที่ได้จากการทำ Linear Predictive Coding สำหรับธรรมชาติของการได้ยินของมนุษย์นั้นเราสามารถที่จะคาดคะเนระดับของการตอบสนองต่อเสียงที่ความถี่ใดๆ ของ

ระบบการได้ยินของมนุษย์ได้จาก Critical Band Intensity ดังนั้นในบทความนี้จึงได้ทำการคำนวณ Transfer Function ของอวัยวะกำทอนเสียง บน Critical Band Rate Scale หรือเรียกว่า สเกลบาร์กและใช้ Critical Band Intensity เป็น Feature Vector ในการแยกแยะและจัดกลุ่มเสียงสระแต่ละเสียงด้วยเทคนิค K-Nearest Neighbor

2. Feature Extraction

เนื่องจากความแตกต่างของเสียงสระแต่ละเสียงนั้นเกิดจากความแตกต่างของ Transfer Function ของอวัยวะกำทอนเสียงในการกำเนิดเสียงนั้น ดังในรูปที่ 1 ดังนั้นเราจึงสามารถรู้จำเสียงสระโดยการรู้จำ Transfer Function ของอวัยวะกำทอนเสียงได้



รูปที่ 1. แสดงแผนภาพระบบกำเนิดสัญญาณเสียงพูดของมนุษย์

2.1 การคำนวณ Transfer Function ของอวัยวะกำทอนเสียงจากสัมประสิทธิ์ที่ได้จากการทำ LPC

ในกระบวนการทำ LPC นั้นเริ่มต้นจากสัญญาณเสียงที่เวลา n ซึ่งมีค่าเป็น $s(n)$ นั้นสามารถถูกคาดคะเนได้จากการคำนวณเชิงเส้นของสัญญาณที่ผ่านมาจำนวน p ค่า [4]

$$s(n) \approx \sum_{k=1}^{p} a_k s(n-k) \qquad (1)$$

โดยที่ a_k เป็นสัมประสิทธิ์ของการทำ LPC ที่ใช้สัญญาณที่ผ่านมา k ค่าในการคำนวณ เราสามารถที่จะแปลงการประมาณ (1) ให้เป็นสมการโดยการรวมเอาสัญญาณการกระตุ้น $Gu(n)$ ตามรูปที่ 1 เข้าไปด้วยจะได้

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (2)$$

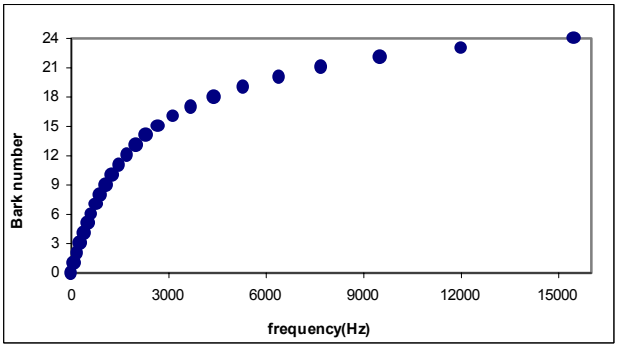
โดยที่ $u(n)$ เป็นสัญญาณการกระตุ้นที่ถูก normalize แล้ว และ G คืออัตราการขยาย และเมื่อแปลงสมการ(2) ให้มาอยู่ใน Z-domain จะได้เป็น

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + GU(z) \qquad (3)$$

ซึ่งจัดรูปให้สามารถคำนวณ Transfer Function ของอวัยวะกำทอนเสียงได้ดังนี้

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (4)$$

2.2 Critical Band Rate Scale หรือสเกลบาร์ก



รูปที่ 2. แสดงความสัมพันธ์ระหว่างลำดับของ Bark scale กับความถี่ Hz

สเกลบาร์ก หรือ Critical Band Rate Scale [7] นั้นเป็นสเกลที่ใช้ในการวัดสเปกตรัมในเชิงของ Psychoacoustics ซึ่งจะมีความใกล้เคียงกับระบบการได้ยินของมนุษย์มาก โดยค่าความเข้มพลังงานที่แสดงอยู่ในสเปกตรัมบนสเกลบาร์กนั้นจะแสดงให้เห็นถึงระดับการตอบสนองต่อเสียงที่ความถี่นั้นๆ ของระบบการได้ยินของมนุษย์ โดยมีความสัมพันธ์ระหว่างสเกลบาร์กกับสเกลความถี่ปกติเป็นดังกราฟในรูปที่ 2 และสมการประมาณความสัมพันธ์ดังกล่าวในสมการ 5 ซึ่งใช้ได้ตลอดช่วงความถี่ที่มนุษย์สามารถได้ยิน

$$B = 13 \arctan(0.76 f / KHz) + 3.5 \arctan(f / 7.5KHz)^2 \qquad (5)$$

โดยที่ f เป็นความถี่ในหน่วย Hz
และ B คือ Bark Number บนสเกลบาร์ก

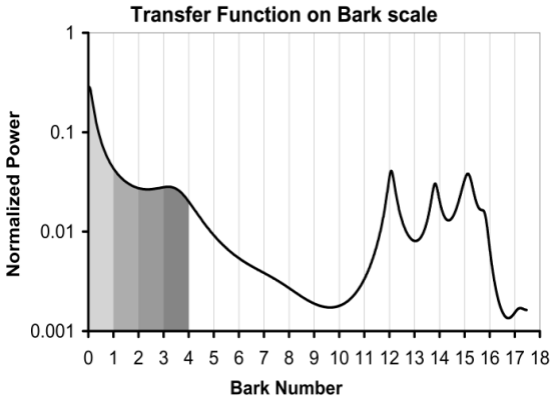
2.3 การใช้ Critical Band Intensity เป็น Feature Vector

ระดับการตอบสนองต่อเสียงที่ความถี่ใดๆ ของระบบการได้ยินเสียงของมนุษย์นั้น สามารถคาดคะเนได้จากการหาความเข้มของสัญญาณเสียงในแต่ละ Critical Band โดยสามารถคำนวณได้จากสเปกตรัมบนสเกลบาร์ก ดังนี้

$$I_B = \int_{B-0.5Bark}^{B+0.5Bark} \frac{dI}{dB} dB \qquad (6)$$

โดยที่ B คือ Bark Number

ซึ่งการหา Critical Band Intensity ในสมการที่ 6 ก็คือพื้นที่ใต้กราฟของแต่ละแถบความถี่ภายใน Transfer function บนสเกลบาร์ก ดังรูปที่ 3 ในบทความนี้เสียงจะถูกสุ่มที่ความถี่ 11.025 KHz ดังนั้นความถี่สูงสุดของเสียงที่สุ่มมาได้ คือ 5.50125 KHz ซึ่งอยู่ในบาร์กที่ 18 ดังนั้น Critical Band Intensity ทั้ง 18 ตัว (บาร์กที่ 1 ถึง 18) จะถูกนำไปใช้เป็นองค์ประกอบของ Feature Vector 18 มิติ ที่จะใช้เป็นตัวแทนของเสียงสระใดๆ



รูปที่ 3. แสดง การคำนวณ Critical band intensity

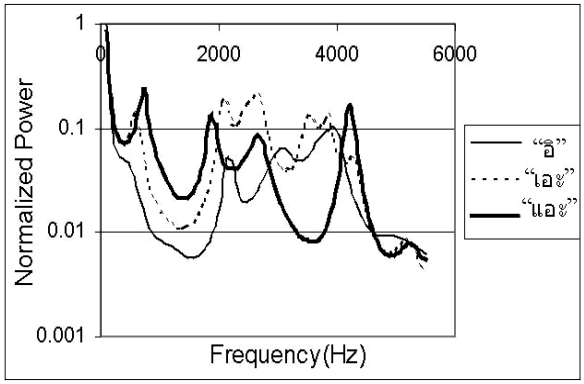
3. การแยกแยะโดยใช้เทคนิค K-Nearest Neighbor

KNN [5] เป็นเทคนิคที่นิยมใช้ในการทำ Pattern Classification โดยในบทความนี้มีเซตของ Feature Vector 18 มิติของ Critical Band Intensity ทั้ง 18 band จากเสียงตัวอย่างจำนวน N เสียง ซึ่งมีเสียงสระ 9 รูปแบบ โดยที่ แต่รูปแบบของเสียงสระนั้นจะสามารถอธิบายได้โดยเซตของเวกเตอร์ 18 มิติ

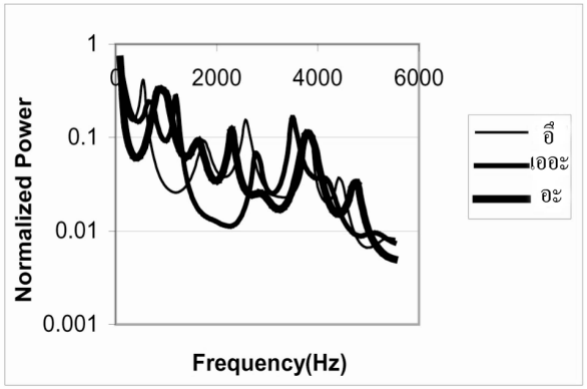
โดยที่ i อยู่ในช่วง 1 ถึง N และเมื่อมีเวกเตอร์เสียงที่ไม่ทราบกลุ่ม $I_{Unknown}=[i_i,i_2,...,i_{18}]$ และต้องการที่จะจัดเข้ากลุ่มใดกลุ่มหนึ่งใน 9 กลุ่ม ซึ่งสามารถที่จะทำโดยการคำนวณระยะห่าง $d(I_{Unknown},I_i)$ ซึ่งเป็นระยะห่างแบบ Euclidean ระหว่างเวกเตอร์ที่ไม่ทราบกลุ่มกับเวกเตอร์ทุกตัวในเซต แล้วจึงทำการคำนวณหาเวกเตอร์ที่ใกล้เคียงที่สุด 9 ตัวกับเวกเตอร์ที่ต้องการจัดกลุ่มและสามารถทำการจัดกลุ่มเวกเตอร์นั้นได้ โดยตรวจสอบดูว่าในเวกเตอร์ทั้ง 9 ตัวที่ใกล้เคียงที่สุดนั้นตรงกับกลุ่มใดมากที่สุดก็จะจัดให้เวกเตอร์ที่ต้องการทราบกลุ่มถูกจัดเข้ากับกลุ่มนั้น

4. การทดลอง

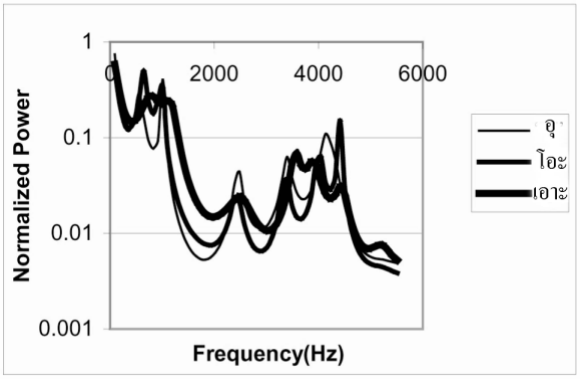
การทดลองในบทความนี้ได้ทำการเก็บตัวอย่างเสียง จากผู้ชาย 5 คน และผู้หญิง 5 คนในห้องทำงานปกติ โดยให้แต่ละคนออกเสียงเป็นคำๆ ที่มีเสียงพยัญชนะและวรรณยุกต์รวมอยู่ด้วย เสียงสระที่ประกอบไปด้วยหน่วยเสียงทั้งหมด 9 แบบ ทั้งเสียงสั้นและเสียงยาวรวมเป็น 18 เสียง สระผสมกับพยัญชนะทั้งหมด 24 เสียง และเสียงวรรณยุกต์ 5 เสียง รวมเป็น 9x2x24x5x10 = 21,600 เสียง เทคนิคในการแยกได้กล่าวไว้ในขั้นต้น Feature Extraction นั้นจะทำการคำนวณ Transfer Function ของอวัยวะกำทอนเสียงของแต่ละเฟรมของเสียงตามการคำนวณสัมประสิทธิ์ LPC คือใช้เฟรมเสียงขนาด 300 samples และขยับไปครั้งละ 100 samples ในเบื้องต้นจะใช้การคำนวณ LPC อันดับที่ 19 ตาม [2] ซึ่งตัวอย่างของ Transfer Function ของเสียงสระทั้ง 9 เสียง ถูกแสดงเป็นกลุ่มของ สระหน้า (อิ, เอะ, แอะ) สระกลาง (อึ, เออะ, อะ) และสระหลัง (อุ, โอะ, เอาะ)[3] ในรูปที่ 2(a), 2(b) และ 2(c) ตามลำดับ



(a) สระหน้า



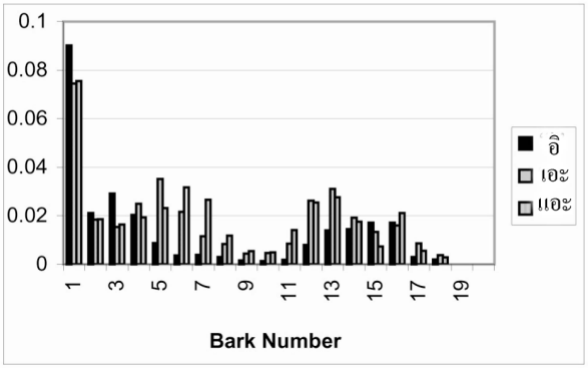
(b) สระกลาง



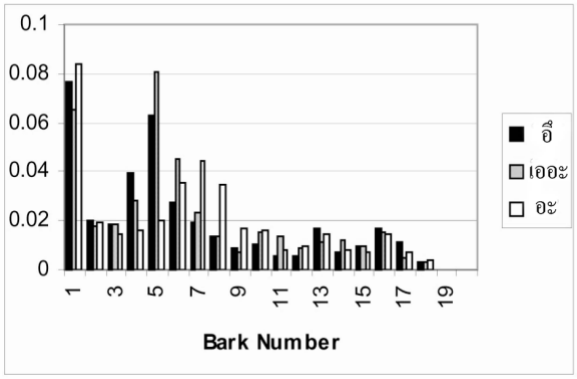
(c) สระหลัง

รูปที่ 4.แสดงตัวอย่าง Transfer Function ของอวัยวะกำทอนเสียง ของเสียงสระเสียงเดี่ยวทั้ง 9 เสียง

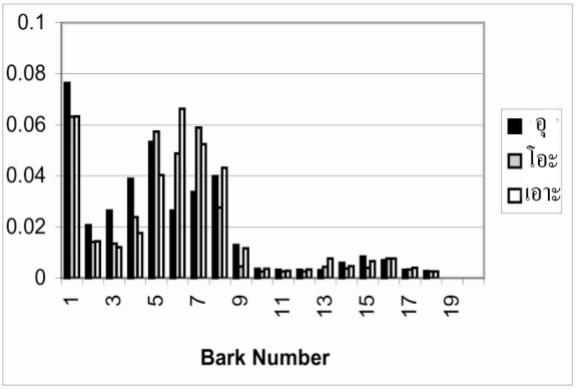
การแบ่งแยกสระหน้า สระกลาง และสระหลังนั้นเป็นการแบ่งตามลักษณะตำแหน่งของลิ้น ในขณะที่กำลังออกเสียงนั้นๆ ในรูปที่ 4 จะเห็นถึงความแตกต่างของ Transfer Function ที่มากพอที่จะแยกแยะ Transfer Function ของแต่ละเสียงออกจากกันได้ หลังจากนั้นเราจะทำการแปลง Transfer Function ดังกล่าวให้มาอยู่ในสเกลบาร์กแล้วจึงทำการคำนวณ Critical Band Intensity ของแต่ละ Band โดยใช้ สมการที่ (6) ผลที่ได้คือ Critical Band Intensity 18 ค่า ซึ่งจะถูกใช้เป็น Feature Vector 18 มิติ โดยตัวอย่างของ Critical Band Intensity ทั้ง 18 Band ของเสียงสระหน้า สระกลาง และสระหลังแสดงในรูปที่ 5(a), 5(b) และ 5(c)



(a) สระหน้า



(b) สระกลาง



(c) สระหลัง

รูปที่ 5.แสดงตัวอย่าง Critical Band Intensity ทั้ง 18 Band ของเสียงสระเสียงเดี่ยวทั้ง 9 เสียง

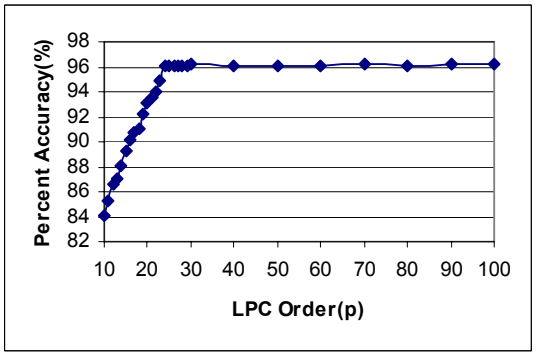
Feature Vector ที่ได้จากตัวอย่างเสียง 21,600 เสียงนั้นจะถูกใช้เป็น Training Set สำหรับสร้างแบบอ้างอิง และในการทดสอบแบบอ้างอิงนั้นใช้

วิธี KNN โดยที่จะไม่ใช้ Nearest Neighbor เพื่อจะไม่ให้เกิดการ self-matching ซึ่งจะทำให้ vector ที่นำเข้าไปทดสอบไม่เป็นส่วนหนึ่งของ training set ซึ่งได้ผลการทดลองดังตารางที่ 1

ตารางที่ 1. แสดงความถูกต้องในการรู้จำเสียงสระเสียงเดี่ยว

แบบอ้างอิง		สระหน้า			สระกลาง			สระหลัง			ความถูกต้อง
แบบทดสอบ		"อิ"	"เอะ"	"แอะ"	"อึ"	"เออะ"	"อะ"	"อุ"	"โอะ"	"เอาะ"	%
สระหน้า	"อิ"	2337	5	28	1	10	0	15	4	0	97.38
	"เอะ"	120	2220	40	0	20	0	0	0	0	92.50
	"แอะ"	60	137	2153	10	40	0	0	0	0	89.71
สระกลาง	"อึ"	10	10	0	2046	220	44	0	10	60	85.25
	"เออะ"	0	0	17	66	2284	33	0	0	0	95.17
	"อะ"	0	0	10	40	80	2217	43	0	10	92.38
สระหลัง	"อุ"	0	0	3	0	0	10	2300	57	30	95.83
	"โอะ"	0	0	2	0	80	10	120	2128	60	88.67
	"เอาะ"	0	10	10	20	0	40	22	75	2223	92.63
										เฉลี่ย	92.17

ในตารางที่ 1 นั้นคอลัมน์ทางซ้ายเป็นเสียงสระที่นำเข้ามาทดสอบ ส่วนแถวบนเป็นเสียงสระที่เป็นแบบอ้างอิง ตัวเลขในตาราง คือจำนวนของเสียงใน Test Set ที่ถูกจัดเข้ากับแบบอ้างอิงใดๆ โดยใช้วิธี KNN ส่วนคอลัมน์ขวาสุดจะเป็นเปอร์เซ็นต์ความถูกต้องของแต่ละเสียงและเปอร์เซ็นต์ความถูกต้องเฉลี่ยที่ได้ คือ 92.17% เราสามารถสังเกตเห็นได้ชัดว่าความผิดพลาดที่เกิดขึ้นในการรู้จำส่วนใหญ่เกิดขึ้นภายในกลุ่มสระเดียวกัน คือ ภายในกลุ่มสระหน้า สระกลาง และสระหลัง ซึ่งมีความคล้ายคลึงกันมากกว่าความผิดพลาดที่เกิดขึ้นข้ามกลุ่มของสระมาก เหตุผลคือลักษณะของช่องปากในการออกเสียงในกลุ่มสระเดี่ยวกันจะมีความคล้ายกันมากซึ่งทำให้ transfer function มีความคล้ายคลึงกันมาก สำหรับการทดลองเพื่อปรับปรุงประสิทธิภาพของระบบโดยการเปลี่ยนอันดับในการคำนวณ LPC เพื่อหาอันดับที่ทำให้ได้ประสิทธิภาพสูงสุด จะทำการสร้างแบบอ้างอิงที่ใช้ LPC อันดับที่ 10 ถึง 100 ได้ความถูกต้องในการรู้จำเฉลี่ยดังรูปที่ 6



รูปที่ 6. แสดงแปรียบเทียบเปอร์เซ็นต์ความถูกต้องในการรู้จำเฉลี่ยกับอันดับที่ใช้ในการคำนวณ LPC

จากรูปจะเห็นได้ว่า การคำนวณ LPC อันดับที่ต่ำเกินไป จะทำให้ ประสิทธิภาพในการรู้จำของระบบต่ำแต่ถ้าอันดับสูงเกินไปก็จะไม่ทำให้ประสิทธิภาพของระบบดีขึ้นเท่าใดนัก ดังนั้นจากรูปจะเห็นได้ว่า อันดับที่ทำให้ระบบเกิดประสิทธิภาพสูงสุดคืออันดับที่ 24 ให้ความถูกต้องประมาณ 96% ซึ่งมากกว่าการใช้อันดับที่ 19 จากผลการทดลองในตารางที่ 1 ประมาณ 4%

5. สรุป

การแปลง Transfer Function ของอวัยวะกำทอนเสียงไปบนสเกลบาร์กและการใช้เทคนิค KNN ในการจัดและแยกแยะกลุ่มของ Feature Vector นั้นสามารถใช้ในการสร้างแบบอ้างอิงและรู้จำหน่วยเสียงสระในภาษาไทยได้ความถูกต้องถึง 96% และการทดลองเพื่อปรับปรุงประสิทธิภาพของระบบแสดงให้เห็นว่าการใช้อันดับในการคำนวณ LPC ที่ต่ำเกินไปทำให้ประสิทธิภาพของระบบต่ำลงมากแต่การใช้อันดับที่สูงมากๆ ก็ไม่ได้ช่วยให้ความถูกต้องในการรู้จำมากขึ้นเท่าใดนัก แต่กลับทำให้ความซับซ้อนในการคำนวณมากขึ้น ส่วนจุดที่ทำให้ระบบมีประสิทธิภาพสูงสุดในการทดลองนี้อยู่ที่อันดับที่ 24 ผลของการรู้จำเสียงสระเดียวนั้นสามารถนำไปใช้ร่วมกับวิธีการอื่นๆ เช่น Hidden Makov Model เพื่อสร้างวิธีการรู้จำเสียงสระสั้นยาวและผสมต่อไป

เอกสารอ้างอิง

[1] ไกรสิน ส่งวัฒนา, อิทธิชัย อรุณศรีแสงไชย, จิตรลดา จารุมิศร์, “แบบจำลองเสียงวรรณยุกต์สำหรับภาษาไทยโดยใช้เทคนิคการคอนไตซ์พิทช์และ Hidden Makov Modeling”, งานประชุมวิชาการทางวิทยาการคอมพิวเตอร์แห่งชาติ 2541, คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

[2] สุนทร อรอินทร์, อัฐ เครือฟัก, “การประมวลผลเสียงพูดโดยการประมาณเชิงเส้น”, วิทยานิพนธ์ ภาควิชาวิศวกรรมไฟฟ้า สาขาวิชาวิศวกรรมโทรคมนาคม คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง 1995

[3] อมร ทวีศักดิ์, “สัทศาสตร์”, สถาบันวิจัยภาษาและวัฒนธรรมเพื่อพัฒนาชนบท มหาวิทยาลัยมหิดล 1999

[4] Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition”, Prentice Hall, 1993.

[5] Morton Nadler, Eric P. Smith, “Pattern Recognition Engineering”, John Wiley&Sons, 1993.

[6] David M. Haward, James Angus, “Acoustics and Psychoacoustics”, Prentice Hall, 1996

[7] E.Zwicker, H.Fastl, “Psychoacoustics : Facts and Models” Second Edition, Springer, 1999

นริศ บุญศักดิ์เฉลิม เกิดวันที่ 15 ตุลาคม 2517, สำเร็จการศึกษาครุศาสตร์อุตสาหกรรมบัณฑิต สาขาวิศวกรรมโทรคมนาคม คณะครุศาสตร์อุตสาหกรรม ปีการศึกษา 2541 ปัจจุบันกำลังศึกษาปริญญาโทคณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง และทำงานในตำแหน่งวิศวกร บริษัท โยโกกาวา (ประเทศไทย) จำกัด

วรา คงคาวิทูร เกิดเมื่อวันที่ 5 กุมภาพันธ์ 2520, สำเร็จการศึกษาวิศวกรรมบัณฑิต ปีการศึกษา 2541 และวิศวกรรมมหาบัณฑิต ปีการศึกษา 2543 สาขาวิศวกรรมโทรคมนาคม คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ปัจจุบันทำงานในตำแหน่งวิศวกร การสื่อสารแห่งประเทศไทย

ไกรสิน ส่งวัฒนา เกิดเมื่อวันที่ 16 มิถุนายน 2507, ได้รับปริญญาตรี, ปริญญาโท และปริญญาเอก จาก University of Wisconsin-Madison ประเทศอเมริกา ในปี พ.ศ. 2530, 2532 และ 2536 ตามลำดับ ขณะนี้เป็นผู้ช่วยศาสตราจารย์ประจำภาควิชาโทรคมนาคม สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง