การแยกเสียงร้องออกจากเสียงเพลงที่เก็บในช่องสัญญาณเดียว โดยการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ

นางสาวอังคนา จันทร์รุ่งอุทัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2551

SINGING VOICE SEPARATION FOR MONO-CHANNEL MUSIC USING NON-NEGATIVE MATRIX FACTORIZATION

Miss Angkana Chanrungutai

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2008

หัวข้อวิทยานิพนธ์	การแยกเสียงร้องออกจากเสียงเพลงที่เก็บในช่องสัญญาณเดียวโดยการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ
โดย	นางสาวอังคนา จันทร์รุ่งอุทัย
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์
(ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ นครทิพย์ พร้อมพูล)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.กัลยา นฤดมกุล)

อังคนา จันทร์รุ่งอุทัย : การแยกเสียงร้องออกจากเสียงเพลงที่เก็บในช่องสัญญาณเดียวโดยการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ. (SINGING VOICE SEPARATION FOR MONO-CHANNEL MUSIC USING NON-NEGATIVE MATRIX FACTORIZATION) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ, 90 หน้า.

การแยกเสียงร้องออกจากเสียงเพลง คือ การสกัดเอาเสียงร้องออกมาให้ได้โดยไม่มีเสียงดนตรี หรือให้เหลือน้อยที่สุดเท่าที่จะเป็นไปได้ งานวิจัยหลายด้านเกียวกับเสียงดนตรี เช่น การค้นคืนเพลงในรูปแบบไฟล์ทั่วไป การจับคู่เสียงร้องและเนื้อเพลง การรู้จำเนื้อเพลง และการระบุตัวผู้ร้อง ล้วนมีอุปสรรคที่สำคัญคือเสียงดนตรีที่อยู่ในเพลงนั้น ดังนั้นการแยกเสียงร้องออกจากเสียงเพลงจึงน่าจะมีส่วนช่วยงานวิจัยต่าง ๆ ดังกล่าว โดยเฉพาะอย่างยิงการแยกเสียงร้องสำหรับเสียงเพลงในแบบช่องสัญญาณเดียว ซึงจะมีผลคือสามารถรองรับได้กับเสียงเพลงในทุกรูปแบบ ไม่ว่าจะเป็นแบบช่องสัญญาณคู่ หรือไฟล์เพลงชนิดอืน ๆ และจะเพิมความเข้าใจเกียวกับองค์ประกอบของเสียงเพลงได้มากยิงขึ้นอีกด้วย ความพยายามในการแยกเสียงร้องที่ผ่านมา วิธีที่มีการศึกษาวิจัยในเร็ว ๆ นี้และให้ผลการแยกเสียงที่ดี คือวิธีการวิเคราะห์โสตตามภาวการณ์เชิงคำนวณ แต่ก็ยังคงจำกัดอยู่ที่แนวเพลงบางประเภท งานวิจัยนี้จึงได้นำเสนอถึงวิธีการแยกเสียงร้องออกจากเสียงเพลงสำหรับเพลงในรูปแบบเพิมเติม และมีการศึกษาในเชิงวิเคราะห์มากขึ้น โดยใช้ชุดข้อมูลทดลองต่าง ๆ และมาตรวัดแบบอัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน รวมทั้งการวัดค่าความถูกต้องของคอนทัวร์ระดับเสียงที่หาได้ ซึงค่าที่ได้จากผลการทดลองต่าง ๆ ให้ผลลัพธ์เป็นที่น่าพอใจ

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2551

4970697921 : MAJOR COMPUTER SCIENCE
KEYWORDS : MONO-CHANNEL MUSIC / NON-NEGATIVE MATRIX FACTORIZATION / SINGING VOICE SEPARATION / SOUND SOURCE SEPARATION

ANGKANA CHANRUNGUTAI : SINGING VOICE SEPARATION FOR MONO-CHANNEL MUSIC USING NON-NEGATIVE MATRIX FACTORIZATION. ADVISOR : CHOTIRAT RATANAMAHATANA, Ph.D., 90 pp.

Singing voice separation is an extraction of singing voice from a song snippet by minimizing instrumental sounds. Many music related research areas, such as music information retrieval, singing voice and lyrics alignment, lyrics recognition, and singer identification, have been encountered the main obstacle which is the instrumental sound within the song. Removal of such instrumental sound or separation of the singing voice out of the song could be very useful for these research areas. Especially, the singing voice separation for mono-channel music can support any song formats, e.g., stereo music or other types of music file format. Moreover, studying about mono-channel music can provide much better understanding on music components and it characteristics. The recent effort tending to be good for solving this problem is Computational Auditory Scene Analysis (CASA). However, this method is still limited to only some genres of music. This research, therefore, proposes a novel singing voice separation method using Non-negative Matrix Factorization (NMF), a matrix decomposition, for additional types of music by studying some instrumental sounds in greater detail. We use various datasets and measures, peak signal-to-noise ratio (PSNR) and the accuracy of pitch contour extracted from the separated singing voice, to evaluate our proposed work. The satisfactory of our work is confirmed by the experimental results.

Department : Computer Engineering Student's Signature :.....
Field of Study :Computer Science Advisor's Signature :.....
Academic Year : 2008

กิตติกรรมประกาศ

การทีวิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยดี อันดับแรกและสำคัญที่สุด ผู้วิจัยขอขอบคุณผู้ที่คอยดูแลเอาใจใส่ให้คำปรึกษา ทั้งในฐานะอาจารย์ที่ปรึกษา และในฐานะเทรนเนอร์ ที่เคียวเข็ญ ให้ผู้วิจัยสามารถพัฒนาไปในแนวทางที่นักวิจัยที่ดีควรจะเป็น ท่านคือ ผศ. ดร. โชติรัตน์ รัตนามหัทธนะ รวมทั้งคณาจารย์ผู้เป็นกรรมการการสอบวิทยานิพนธ์ ที่มีความกระตือรือร้นต่องานวิจัย และได้เสนอคำแนะนำด้วยวิสัยทัศน์ที่กว้างขวางอันเป็นประโยชน์ต่อวิทยานิพนธ์ฉบับนี้ ได้แก่ ศ. ดร. บุญเสริม กิจศิริกุล ผศ. นครทิพย์ พร้อมพูล จากภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย และ ผศ. ดร. กัลยา นฤดมกุล จากคณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล

นอกจากนี้ผู้วิจัยขอขอบคุณ ศ. ดร. ชิดชนก เหลือสินทรัพย์ จากภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ผู้ที่ให้คำแนะนำอันมีคุณค่าอย่างยิ่งต่อการทำวิจัยทางด้านการแยกเสียง และอาจารย์อีกสองท่านจากภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ รศ. ดร. ลัญฉกร วุฒิสิทธิกุลกิจ และ อ. สุวิทย์ นาคพีระยุทธ รวมทั้ง ดร. ทรงพล องค์วัฒนกุล จากภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยมหิดล ที่กรุณาสละเวลาให้คำปรึกษาและถ่ายทอดความรู้ที่จำเป็นต่อการทำงานวิจัยทางด้านนี้ให้กับผู้วิจัยเป็นอย่างดี

ผู้วิจัยขอขอบคุณคณาจารย์ทุกท่านจากภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ตลอดจนคณาจารย์ทุกท่านตั้งแต่อดีตจนถึงปัจจุบัน ที่ได้ประสิทธิ์ประสาทวิชาความรู้หลากหลายแขนงให้ผู้วิจัย

ขอขอบคุณเพื่อน ๆ และน้อง ๆ ในห้องปฏิบัติการภาควิชาวิศวกรรมคอมพิวเตอร์ ชั้น 18 สำหรับความช่วยเหลือต่าง ๆ กำลังใจ คำปรึกษา รวมทั้งคำแนะนำที่มีประโยชน์ และเป็นแรงผลักดันอันแข็งแกร่งมาโดยตลอด

ขอขอบคุณเพื่อน ๆ สมาชิกในบอร์ดคาราโอเกะที่ให้ความอนุเคราะห์ในด้านตัวอย่างเสียงร้องเพลงอันมีค่ายิ่งต่อการทดลองในงานวิจัยนี้ รวมทั้งการสอบถามด้วยความห่วงใยอย่างสม่ำเสมอ

ท้ายนี้ ผู้วิจัยขอขอบคุณคุณพ่อ คุณแม่ สมาชิกในครอบครัวทุกคน เพื่อน ๆ พี่ ๆ และน้อง ๆ ภาควิชาวิศวกรรมคอมพิวเตอร์ ที่คอยถามไถ่ ให้ความช่วยเหลือต่าง ๆ รวมทั้งให้กำลังใจผู้วิจัยตลอดสามปีที่ผ่านมา จนกระทั่งวิทยานิพนธ์ชิ้นนี้สำเร็จลุล่วงมาได้ด้วยดี

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

เป็นที่ทราบกันดีว่า ระบบการฟังของมนุษย์เรา สามารถทำการแยกเสียงที่ได้ยินตามแหล่งกำเนิดเสียงต่าง ๆ ได้โดยไม่ต้องใช้ความพยายามมากนัก จึงทำให้เราสามารถนำสารที่ได้ยิน เช่น เสียงร้องเพลง หรือเสียงพูด ซึ่งปะปนมากับเสียงอื่น ๆ ไปทำความเข้าใจต่อได้อย่างรวดเร็ว แตกต่างจากการแยกเสียงด้วยคอมพิวเตอร์ ซึ่งใช้ขั้นตอนวิธีที่มีความซับซ้อน และยังต้องการงานวิจัยที่ศึกษาในประเด็นนี้อย่างจริงจัง งานวิจัยชิ้นนี้จึงสนใจศึกษาปัญหาการแยกเสียงร้องออกจากเสียงเพลง ซึ่งเป็นหนึ่งในปัญหาที่สำคัญด้านการแยกเสียง

เสียงร้องที่อยู่ในเพลงนั้นบรรจุข้อมูลสำคัญต่าง ๆ มากมาย เช่น ข้อมูลทางภาษา ความหมาย จังหวะ ท่วงทำนอง เอกลักษณ์ของเสียงผู้ร้อง เป็นต้น จึงทำให้งานวิจัยทางด้านการแยกเสียงร้องออกจากเสียงเพลงมีความจำเป็นต่องานวิจัยอื่น ๆ อีกหลายด้าน เช่น การค้นคืนเพลง (Music Information Retrieval) การระบุผู้ร้อง (Singer Identification) การจับคู่เนื้อเพลงให้ตรงกับเสียงร้อง (Lyrics and Singing Voice Alignment) การเพิ่มคุณภาพเสียงดนตรี (Music Enhancement)

ในงานวิจัยด้านการค้นคืนเพลง แนวคิดเรื่องการสร้างระบบค้นหาเพลงโดยการร้องทำนอง (Query by Humming) [1, 2] เป็นที่สนใจมากขึ้น เนื่องจากระบบดังกล่าวช่วยให้ผู้ใช้ที่จดจำได้เพียงทำนองสั้น ๆ ของเพลง สามารถค้นหาเพลงนั้น ๆ ได้ ระบบค้นหาเพลงโดยการร้องทำนอง [1, 2] ส่วนใหญ่ให้ความสำคัญกับการพัฒนาในส่วนของการค้นคืน จึงใช้ฐานข้อมูลเป็นไฟล์เพลงมิดิ (MIDI) เพียงชนิดเดียว ซึ่งไฟล์เพลงชนิดนี้มีการแยกช่องสัญญาณต่าง ๆ ของเครื่องดนตรีออกจากกัน และมีการจัดเก็บข้อมูลเกี่ยวกับโน้ตดนตรีของเพลงนั้น ๆ อย่างไรก็ตาม รูปแบบไฟล์เพลงอันเป็นที่นิยม เช่น .WAV .MP3 .WMA นั้นเก็บข้อมูลในรูปแบบของคลื่นเสียง ซึ่งไม่มีการแยกช่องสัญญาณในส่วนของเสียงร้องและเสียงเครื่องดนตรีต่าง ๆ ออกจากกัน จึงทำให้ไม่สามารถสกัดคุณลักษณะที่เหมาะสมสำหรับระบบดังกล่าวได้ งานวิจัยที่พัฒนาระบบค้นคืนเพลงซึ่งใช้ฐานข้อมูลเพลงชนิดที่เป็นคลื่นเสียง [3] จึงจำเป็นต้องทำการแยกเสียงร้องออกจากเสียงเพลงก่อน เพื่อนำไปสร้างฐานข้อมูลสำหรับการค้นหาเพลงที่มีประสิทธิภาพ ในงานวิจัยด้านการระบุผู้ร้อง [4-6] มีการกล่าวถึงปัญหาสำคัญ คือ ต้องการข้อมูลเสียงร้องที่ไม่มีเสียงดนตรีประกอบเพื่อนำไปใช้สกัดคุณลักษณะของเสียงผู้ร้อง ซึ่งหาได้ยากในทางปฏิบัติ งานวิจัยด้านนี้จึงต้องพยายาม

แยกเสียงร้องออก [4] หรือพยายามลดเสียงดนตรีประกอบลงก่อน [5] เช่นเดียวกับในงานวิจัยด้านการจับคู่เนื้อเพลงให้ตรงกับเสียงร้อง [6] ซึ่งมีประโยชน์อย่างยิ่งต่อการใส่เนื้อเพลงสำหรับการทำเพลงคาราโอเกะโดยอัตโนมัติ โดยผู้วิจัยจะต้องแยกเสียงร้องออกก่อนที่จะทำการจับคู่ นอกจากนี้ในงานวิจัยด้านการปรับปรุงคุณภาพเสียงดนตรี หรืองานปรับแต่งผสมเสียงดนตรี สำหรับการบันทึกเสียงร้องประกอบเพลงโดยทั่วไป ซึ่งไม่ได้เก็บสัญญาณเสียงร้องแยกไว้นั้น ผู้ใช้อาจต้องการนำเสียงร้องไปผสมกับดนตรีลักษณะอื่น ๆ เพื่อสร้างสรรค์ผลงานเพลงชิ้นใหม่ขึ้นอีกมากมาย

แม้ว่าการแยกเสียงร้องออกจากเสียงเพลงจะมีความจำเป็นอย่างยิ่งต่องานวิจัยหลากหลายด้านดังที่ได้กล่าวแล้วข้างต้น แต่เมื่อศึกษางานวิจัยทางด้านการแยกเสียงอย่างจริงจังกลับพบว่ามีงานวิจัยด้านการแยกเสียงพูด (Speech Separation) ออกจากเสียงรบกวนอย่างกว้างขวาง [7] ในขณะที่เมื่อเปรียบเทียบกันแล้ว งานวิจัยที่เน้นเฉพาะการแยกเสียงร้องออกจากเสียงเพลงยังมีไม่มากนัก ดังนั้นงานวิจัยด้านนี้จึงต้องการการศึกษาที่เพิ่มมากขึ้น และเหตุผลที่ต้องแบ่งงานวิจัยทั้งสองด้านนี้ออกจากกัน เนื่องจากงานวิจัยทั้งสองนี้มีเป้าหมายในการแยกเสียงที่ต่างกัน ออกจากเสียงประกอบที่ต่างกัน คือ การแยกเสียงพูดออกจากสัญญาณรบกวน และการแยกเสียงร้องออกจากเสียงเพลง ซึ่งความแตกต่างที่สำคัญที่สุดคือ ลักษณะของเสียงประกอบนั้นเอง งานวิจัยด้านการแยกเสียงพูด โดยส่วนใหญ่แล้วเสียงประกอบเป็นเสียงรบกวนซึ่งไม่สัมพันธ์กับเสียงพูด ในขณะที่เสียงประกอบของการแยกเสียงร้องนั้น คือเสียงดนตรีที่ได้รับการประพันธ์เพื่อให้สอดคล้องกลมกลืนกับเสียงร้องตลอดทั้งเพลง ส่งผลให้การแยกเสียงร้องออกจากเสียงเพลงมีความท้าทายมากกว่า และการที่งานวิจัยนี้เลือกศึกษาการแยกเสียงร้องออกจากเสียงเพลงสัญญาณเดียว (Mono Channel) แม้ว่าในปัจจุบันจะมีเพลงช่องสัญญาณคู่ (Stereo Channel) ออกมามากมายนั้น เนื่องจากวิธีการที่ศึกษาได้จะสามารถยังประโยชน์ได้กว้างขวางกว่าในแง่ของการนำไปใช้ได้กับเพลงที่มีจำนวนช่องสัญญาณสูงขึ้น ซึ่งก็คือความสามารถในการแก้ปัญหาเดียวกันสำหรับเพลงช่องสัญญาณคู่ด้วยนั้นเอง

ขั้นตอนวิธีในการแยกเสียงร้องออกจากเสียงเพลงของงานวิจัยต่าง ๆ สามารถแบ่งออกได้เป็น 3 กลุ่มหลัก ๆ ได้แก่ การใช้แบบจำลองทางสถิติ (Statistical Model) [8, 9] การวิเคราะห์โสตตามภาวการณ์เชิงคำนวณ (Computational Auditory Scene Analysis) [5, 10-12] และการแยกส่วนประกอบเมทริกซ์ (Matrix Decomposition) [13, 14] และนอกจากนี้ยังมีวิธีการอื่น ๆ ที่ไม่ได้จัดอยู่ในกลุ่มใดกลุ่มหนึ่งอีกด้วย [3, 4, 6] ซึ่งแต่ละขั้นตอนวิธีต่างมีข้อดีและข้อจำกัดดังจะได้กล่าวต่อไปในบทที่ 2 ในส่วนงานวิจัยด้านการแยกเสียงร้องออกจากเสียงเพลง

จากที่กล่าวมาข้างต้น ร่วมกับการได้ศึกษาวิธีแก้ปัญหาในกลุ่มของการแยกส่วนประกอบเมทริกซ์ พบว่าวิธีการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ (Non-Negative Matrix Factorization หรือ NMF) ซึ่งเป็นหนึ่งในวิธีการในกลุ่มของการแยกส่วนประกอบเมทริกซ์ มีแนวโน้มที่จะให้ผลการแยกเสียงได้ดีบนเงื่อนไขที่น้อยกว่าวิธีการอื่น ๆ งานวิจัยนี้จึงมีการออกแบบขั้นตอนวิธีสำหรับการแยกเสียงร้องออกจากเสียงเพลง โดยเลือกใช้วิธีการดังกล่าว ซึ่งจะได้กล่าวถึงรายละเอียดในบทต่อไป โดยจะศึกษากับเสียงเพลงรูปแบบ .WAV ซึ่งเป็นแฟ้มข้อมูลรูปแบบมาตรฐาน ที่สามารถแปลงได้จากไฟล์เพลงอันเป็นที่นิยมชนิดอื่น ๆ เช่น .MP3 และ .WMA

1.2 วัตถุประสงค์ของการวิจัย

ออกแบบขั้นตอนวิธีสำหรับการแยกเสียงร้องออกจากเสียงดนตรีที่เก็บในรูปแบบไฟล์เพลงนามสกุล .WAV ช่องสัญญาณเดียว

1.3 ขอบเขตของการวิจัย

1. พัฒนาวิธีการแยกเสียงร้องออกจากเสียงเพลง บนฐานข้อมูลเพลงที่มีเสียงร้องของคนหนึ่งคน มีเครื่องดนตรีประกอบไม่เกิน 2 ชิ้น ขนาดความยาวไม่เกิน 10 วินาที
2. ทำการเปรียบเทียบเสียงร้องที่แยกได้เชิงปริมาณในโดเมนความถี่ โดยใช้ค่าอัตราส่วนสูงสุดต่อสัญญาณรบกวน
3. ทำการเปรียบเทียบเสียงร้องที่แยกได้ในแง่ของการนำไปใช้ประโยชน์ในด้านอื่น ๆ โดยการวัดค่าระยะห่างของคอนทัวร์ระดับเสียงของเสียงร้องที่แยกได้ และเสียงร้องต้นฉบับ

1.4 ประโยชน์ที่ได้รับ

เสียงร้องที่แยกได้จากเสียงเพลงแบบช่องสัญญาณเดียวที่ได้ศึกษาในงานวิจัยนี้สามารถนำไปใช้กับงานด้านอื่น ๆ ต่อได้โดยตรง ได้แก่

1. นำไปสร้างฐานข้อมูลเพลงสำหรับระบบการค้นหาเพลงโดยการร้องทำนอง เพื่อให้สามารถค้นหาเพลงในรูปแบบอันเป็นที่นิยมทั่วไป เช่น .MP3 .WMA .WAV ได้
2. นำไปใช้สกัดคุณลักษณะของเสียงร้องสำหรับงานวิจัยด้านการระบุผู้ร้อง
3. นำไปใช้ในงานวิจัยที่ทำการจับคู่เนื้อเพลงให้ตรงกับเสียงร้อง สำหรับการทำเพลงคาราโอเกะ

4. นำไปใช้ผสมกับเสียงดนตรีลักษณะอื่น ๆ เพื่อสร้างสรรค์ผลงานเพลงชิ้นใหม่ขึ้นได้

1.5 วิธีดำเนินการวิจัย

1. ศึกษาข้อมูลเอกสารและงานวิจัยที่เกี่ยวข้องกับการแยกเสียง โดยเฉพาะการแยกเสียงร้องออกจากเสียงเพลง
2. ศึกษาธรรมชาติของเสียง และวิธีการต่าง ๆ ในการวิเคราะห์คลื่นเสียง
3. วิเคราะห์เปรียบเทียบข้อดีและข้อจำกัดของวิธีการแยกเสียงร้องออกจากเสียงเพลงในงานวิจัยต่าง ๆ
4. เลือกวิธีการและออกแบบขั้นตอนวิธีในการแยกเสียงร้องออกจากเสียงเพลง
5. เลือกวิธีการประเมินผลที่เหมาะสมสำหรับการแยกเสียงร้อง
6. พัฒนาขั้นตอนวิธีในการแยกเสียงร้อง
7. ทดสอบและประเมินผลการทำงาน รวมทั้งประสิทธิภาพของขั้นตอนวิธีในการแยกเสียงกับฐานข้อมูลเพลงที่เลือก
8. ปรับปรุงขั้นตอนวิธีการแยกเสียงให้ดียิ่งขึ้น
9. สรุปผลการวิจัยและตีพิมพ์ผลการวิจัย
10. เรียบเรียงและจัดทำวิทยานิพนธ์

1.6 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 5 บท ดังนี้คือ บทที่ 1 บทนำ บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการทำงานวิจัยชิ้นนี้ บทที่ 3 กล่าวถึงการดำเนินงานวิจัยโดยอธิบายเป็นขั้นตอนต่าง ๆ อย่างละเอียด ส่วนในบทที่ 4 เป็นการทดลองและผลที่ได้จากการทดลองตามชุดข้อมูลต่าง ๆ และบทที่ 5 เป็นการสรุปผลการทดลองและข้อเสนอแนะของงานวิจัยอันจะเป็นประโยชน์ต่องานวิจัยอื่น ๆ ในอนาคต

1.7 ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของงานวิทยานิพนธ์นี้ ได้รับการตีพิมพ์เป็นบทความทางวิชาการ ดังนี้

- “ Singing Voice Separation for Mono-Channel Music Using Non-negative Matrix Factorization” โดย อังคนา จันทร์รุ่งอุทัย และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการระดับนานาชาติ “The First International Conference on Advanced Technologies for

Communications (ATC 2008)" ซึ่งจัดขึ้น ณ เมืองฮานอย ประเทศเวียดนาม ระหว่างวันที่ 6-9 ตุลาคม 2551 ดังภาคผนวก ก หน้า 71-74

- "Singing Voice Separation in Mono-Channel Music" โดย อังคนา จันทร์-รุ่งอุทัย และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการระดับนานาชาติ "International Symposium on Communications and Information Technologies 2008 (ISCIT 2008)" ซึ่งจัดขึ้น ณ เมืองเวียงจันทน์ ประเทศลาว ระหว่างวันที่ 21-23 ตุลาคม 2551 ดังภาคผนวก ก หน้า 75-80

นอกจากนี้ ยังมีผลงานตีพิมพ์จากงานวิจัยอื่นในขณะที่กำลังศึกษา คือ

- "การแปลผลและการบรรยายรูปภาพตารางสำหรับผู้พิการทางสายตา" โดย วงศ์ยศ เกิดศรี อังคนา จันทร์รุ่งอุทัย และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการ "11th National Computer Science and Engineering Conference (NCSEC 2007)" ซึ่งจัดขึ้น ณ กรุงเทพมหานคร ประเทศไทย ระหว่างวันที่ 19-21 พฤศจิกายน 2550 ดังภาคผนวก ข หน้า 82-89

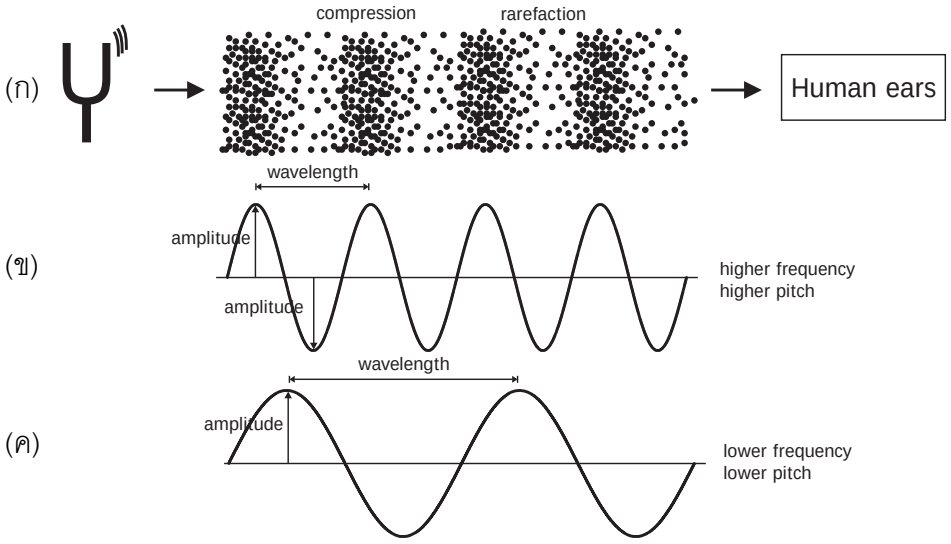
บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ ผู้วิจัยจะกล่าวถึง คลื่นเสียง เพื่อให้เข้าใจถึงข้อมูลเข้าของการแยกเสียง การแทนข้อมูลของคลื่นเสียง การแปลงแบบฟูเรียร์ซึ่งเป็นหัวใจของการศึกษาคลื่นเสียงในเชิงโดเมนความถี่ และวิธีการแยกเสียงร้องออกจากเสียงเพลง อันเป็นจุดมุ่งหมายของงานวิจัยนี้ รวมทั้งคำอธิบายศัพท์ที่เกี่ยวข้อง ตามลำดับ

2.1 คลื่นเสียง

เสียง หรือคลื่นเสียง เป็นคลื่นกลตามยาว ที่เกิดจากการสั่นของวัตถุ ซึ่งก่อให้เกิดการอัดและการขยายตัวของโมเลกุลของตัวกลาง เช่น อากาศ ให้กลายเป็นพลังงานและถ่ายทอดผ่านตัวกลางดังกล่าวไปเรื่อย ๆ เช่น เสียงที่เกิดจากการสั่นของส้อมเสียง การสั่นของสายกีตาร์ การสั่นของชุดสายเสียง (Glottis) ในลำคอของคน เป็นต้น และเมื่อคลื่นของช่วงอัดและช่วงขยายเคลื่อนที่เข้าสู่หูของคนจะเกิดการได้ยินเสียงขึ้น ดังรูปที่ 2.1 (ก)



รูปที่ 2.1 การเกิดขึ้นของเสียงและองค์ประกอบของเสียง [15]

การที่เราสามารถเปรียบเทียบได้ว่าเสียงแต่ละเสียงที่ได้ยินนั้นมีความแตกต่างกันทั้งในด้านเสียงสูง เสียงต่ำ เสียงดัง เสียงเบา หรือคุณภาพของเสียงมีลักษณะต่าง ๆ กัน ล้วนขึ้นอยู่กับองค์ประกอบพื้นฐานของเสียง [15] ซึ่งสามารถอธิบายได้ดังต่อไปนี้

2.1.1 **ความถี่ ความยาวคลื่น และระดับเสียง** (Frequency, Wavelength, and Pitch)

คลื่นเสียงมีคุณสมบัติสำคัญประการหนึ่ง คือ เสียงทุกชนิดเดินทางด้วยความเร็วเท่ากันในตัวกลางชนิดเดียวกัน และความเร็วของเสียงมีค่าเท่ากับผลคูณระหว่างความยาวคลื่นและความถี่ โดยที่ความยาวคลื่น (Wavelength) คือ ระยะห่างระหว่างยอดคลื่นที่ติดกัน และความถี่ (Frequency) คือ ความบ่อยของการเดินทางของคลื่น ซึ่งสามารถวัดได้ในหน่วยของจำนวนคลื่นในหนึ่งวินาที หรือเฮิรตซ์ (Hertz)

ด้วยความถี่และความยาวคลื่นนี้เอง ทำให้เราสามารถบอกความแตกต่างของเสียงที่ได้ยินว่า เสียงใดสูงหรือต่ำกว่ากัน เสียงที่มีความยาวคลื่นสั้นจะมีจำนวนคลื่นที่เดินทางมาสู่หูของเรามาก นั่นคือมีความถี่สูง และส่งผลให้เสียงที่ได้ยินนั้นมีระดับเสียง (Pitch) ที่สูงกว่าเสียงที่มีความยาวคลื่นสูงและความถี่ต่ำกว่า ดังรูปที่ 2.1 (ข) และรูปที่ 2.1 (ค) ตามลำดับ

ระดับเสียง (Pitch) คือ คำที่ใช้เรียกความถี่ของคลื่นในวงการดนตรี ซึ่งนอกจากนี้นักดนตรียังได้มีการตั้งชื่อระดับเสียงที่ใช้บ่อยเป็นชื่อโน้ตต่าง ๆ อีกด้วย

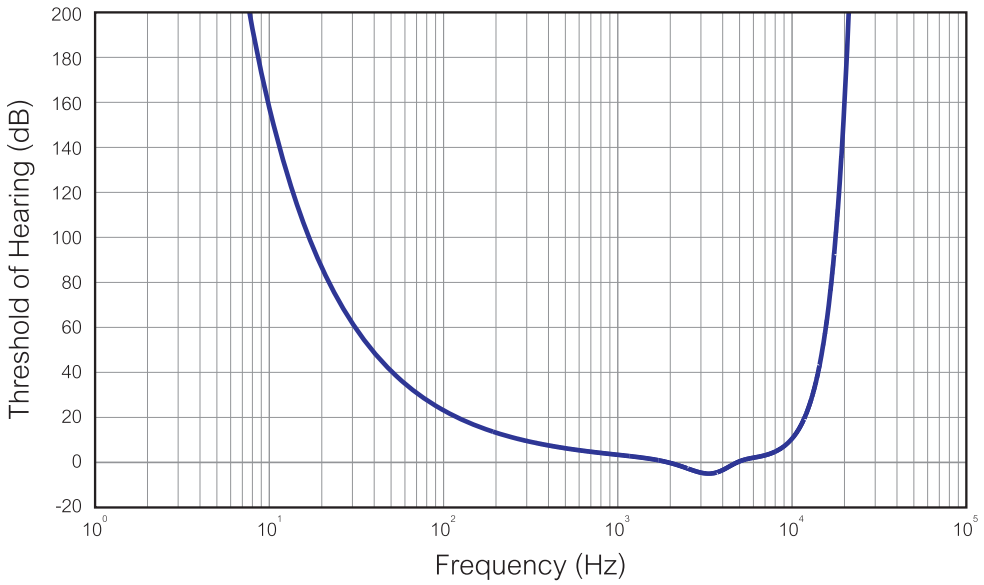
2.1.2 **แอมพลิจูด** (Amplitude)

นอกจากความแตกต่างในแง่ของระดับเสียงแล้ว เรายังสามารถใช้ความดังในการเปรียบเทียบเสียงต่าง ๆ ที่ได้ยิน ความดังของเสียงแปรผันตามการเปลี่ยนแปลงความดันของโมเลกุลของตัวกลาง ที่ถ่ายทอดมาถึงหูของเรา ซึ่งมีได้ทั้งค่าบวกและค่าลบ โดยการเปลี่ยนแปลงความดันนี้เรียกว่าแอมพลิจูด (Amplitude) จากรูปที่ 2.1 (ข) และ รูปที่ 2.1 (ค) แอมพลิจูด คือความสูงของคลื่นจากแนวสมดุล ซึ่งในการวัดความเปลี่ยนแปลงของความดันนี้ จะวัดในหน่วยของความต่างศักย์ (Voltage) หรือสามารถแปลงเป็นค่าความเข้มของเสียง ซึ่งวัดในหน่วยเดซิเบล (Decibel หรือ dB) อันเป็นค่าที่นิยมใช้ในการอธิบายความดังของเสียง เสียงดังคือเสียงที่มีแอมพลิจูดสูง และเสียงเบาคือเสียงที่มีแอมพลิจูดต่ำ

สำหรับเสียงที่มีแอมพลิจูดของเสียงต่ำกว่า 0 เดซิเบล จะถือว่าเป็นเสียงเงียบ อย่างไรก็ตาม ความดังของเสียงที่มนุษย์สามารถได้ยิน จะมีค่าไม่เท่ากันที่ความถี่ต่าง ๆ ดังสมการ (2.1) [16] ซึ่งสามารถแสดงได้ดังรูปที่ 2.2 ซึ่งระดับของเสียงที่มนุษย์ได้ยินที่ความถี่ต่าง ๆ มีค่าต่ำสุดที่ประมาณ -5 เดซิเบล ที่ความถี่ประมาณ 3300 เฮิรตซ์ ซึ่งเป็นความถี่ที่ระบบการฟังของมนุษย์ไวต่อการได้ยินมากที่สุด

$$T_{hearing}(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}\left(\frac{f}{1000}\right)^{4} \quad (2.1)$$

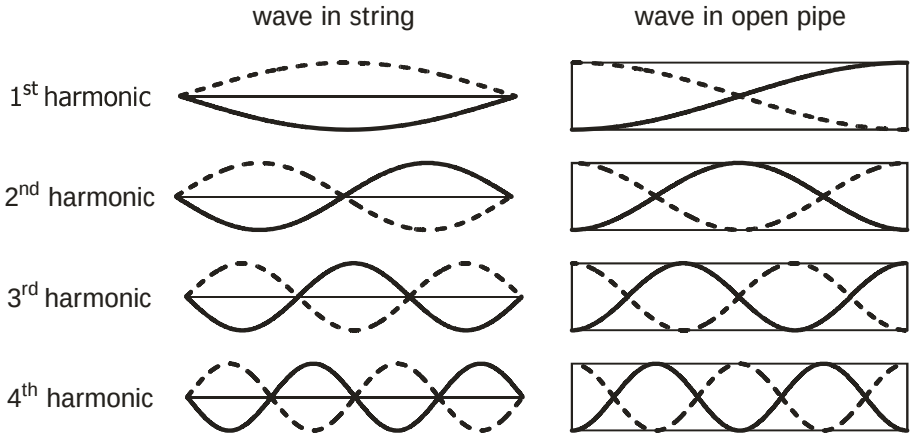
เมื่อ f คือ ค่าความถี่ในหน่วยเฮิรตซ์



รูปที่ 2.2 ค่าขีดแบ่งของการได้ยินเสียงของมนุษย์สำหรับเสียงความถี่ต่าง ๆ [16]

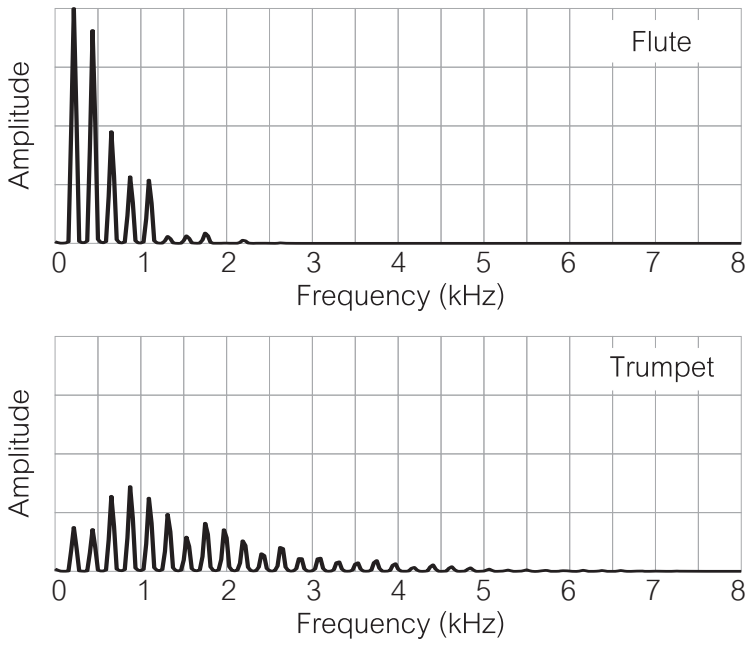
2.1.3 **ฮาร์มอนิกและแทมเบอร์** (Harmonics and Timbre)

ปัจจัยที่ทำเกิดความแตกต่างของคลื่นเสียง นอกจากในแง่ของระดับเสียง และความดังของเสียงแล้ว ยังมีปัจจัยที่สำคัญอีกประการหนึ่งซึ่งทำให้เราทราบได้ว่า เสียงที่ได้ยินเป็นเสียงของอะไร เช่น เสียงของฟลุทที่เล่นพร้อมกับทรัมเปต ด้วยระดับเสียงและความดังเท่ากัน



รูปที่ 2.3 ฮาร์มอนิกของคลื่นในเส้นเชือกและคลื่นในท่อปลายเปิด [15]

เสียงที่ได้ยินโดยทั่วไปนั้นไม่ได้มีเพียงความถี่เดียว แต่เป็นเสียงที่มีหลายความถี่ในเวลาเดียวกัน เช่น เมื่อดีดสายกีตาร์หนึ่งเส้น จะเกิดคลื่นที่มีความยาวคลื่นเป็นสองเท่าของสายกีตาร์นั้น เป็นคลื่นความถี่แรก เรียกว่าความถี่มูลฐาน (Fundamental Frequency) ซึ่งเป็นตัวแทนระดับเสียงหลักของเสียงที่เล่นนั้น ในขณะเดียวกัน คลื่นที่มีความถี่เป็นสองเท่า สามเท่า สี่เท่า และอื่น ๆ ของความถี่มูลฐาน จะดังขึ้นพร้อม ๆ กันไปด้วย ดังรูปที่ 2.3 ชุดของคลื่นที่มีความถี่เป็นจำนวนเท่าของความถี่แรกนี้เรียกว่า ฮาร์มอนิก (Harmonic) โดยคลื่นความถี่มูลฐาน เรียกว่า ฮาร์มอนิกที่หนึ่ง และคลื่นที่มีความถี่เป็นสองเท่า สามเท่า และสี่เท่าของความถี่มูลฐานนี้ เรียกว่า ฮาร์มอนิกที่สอง ฮาร์มอนิกที่สาม และฮาร์มอนิกที่สี่ ตามลำดับ

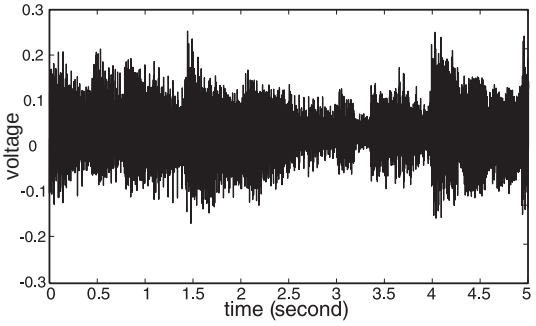


รูปที่ 2.4 สเปกตรัมของเสียงฟลุทและทรัมเปตที่มีความถี่มูลฐาน 220 เฮิรตซ์

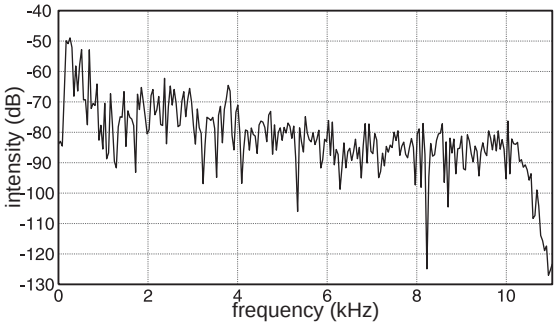
สำหรับปัจจัยที่ทำให้เสียงแต่ละชนิดมีคุณลักษณะหรือแทมเบอร์ (Timbre) ที่แตกต่างกัน คือ องค์ประกอบความถี่ หรือองค์ประกอบฮาร์มอนิก (Harmonic Content) ซึ่งก็คือองค์ประกอบความถี่ที่นอกเหนือจากความถี่มูลฐานในเสียงแต่ละชนิดนั้นเอง ตัวอย่างเช่น รูปที่ 2.4 เป็นองค์ประกอบทางความถี่ของเครื่องดนตรี 2 ชนิด คือ ฟลุทและทรัมเปต ซึ่งมีความดังเท่ากัน และเล่นด้วยโน้ตตัวเดียวกัน ที่ความถี่มูลฐาน 220 เฮิรตซ์ จะเห็นว่าเสียงของเครื่องดนตรีทั้งสองมีค่าแอมพลิจูดสูงค่าแรกที่ความถี่เดียวกัน คือ ที่ความถี่ 220 เฮิรตซ์ และที่ฮาร์มอนิกต่าง ๆ คือ ที่ 440 660 880 1120 เฮิรตซ์ เป็นต้น ในขณะที่สัดส่วนของค่าแอมพลิจูดของแต่ละความถี่แตกต่างกัน นั่นคือมีองค์ประกอบทางความถี่ที่ต่างกัน ซึ่งส่งผลให้ได้ยินเสียงที่ต่างกัน

2.2 การแทนข้อมูลของคลื่นเสียง

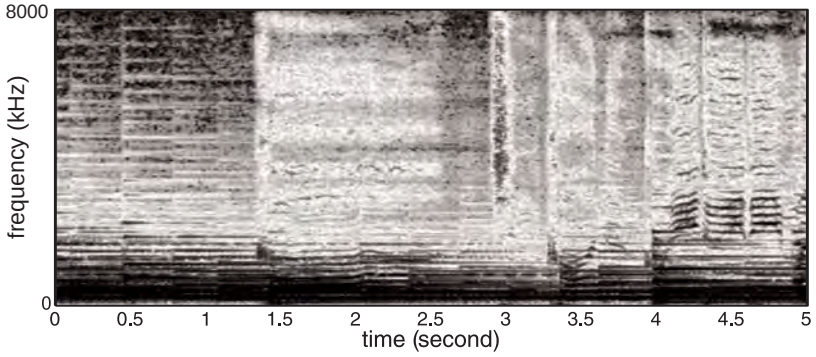
หลังจากที่ได้ทราบถึงองค์ประกอบของเสียงแล้ว ต่อมาเราสามารถมาดำเนินการกับเสียงที่ถูกเก็บเป็นรูปแบบดิจิทัลในคอมพิวเตอร์ได้ โดยรูปแบบการแทนข้อมูลของเสียงที่นิยมนั้นมี 2 ลักษณะ ได้แก่



(ก)



(ข)



(ค)

รูปที่ 2.5 การแทนข้อมูลเสียงในรูปแบบต่าง ๆ (ก) คลื่นเสียง (ข) สเปกตรัม (ค) สเปกโทรแกรม

2.2.1 การแทนข้อมูลในรูปแบบคลื่นเสียง (Wave) คือ การแทนข้อมูลของคลื่นเสียงในเชิงเวลา โดยแสดงได้เป็นความสัมพันธ์ระหว่างความดัง หรือแอมพลิจูด (Amplitude) ของเสียงกับเวลา ซึ่งมักใช้หน่วยเป็นค่าความต่างศักย์ (Voltage) ของการเปลี่ยนแปลงความดัน ดังรูปที่ 2.5 (ก)

2.2.2 การแทนข้อมูลในรูปแบบสเปกตรัม (Spectrum) คือ การแทนข้อมูลของคลื่นเสียงในเชิงความถี่ โดยแสดงได้เป็นความสัมพันธ์ระหว่างแอมพลิจูดกับความถี่ที่เกิดขึ้น ซึ่งหน่วยที่ใช้วัดความดังจะเป็นเดซิเบล (Decibel) อันแปลงได้จากค่าความต่างศักย์ (Voltage) ดังรูปที่ 2.5 (ข)

จากการแทนข้อมูลทั้งสองรูปแบบนี้ พบว่ามีข้อมูลสามชนิด คือ แอมพลิจูด ความถี่ และเวลา ซึ่งสามารถแสดงความสัมพันธ์ได้ด้วยสเปกโทรแกรม (Spectrogram) ดังรูปที่ 2.5 (ค) สเปกโทรแกรมเป็นการแสดงรูปร่างของคลื่นเสียงในเชิงความถี่และเวลา โดยใช้ค่าสีระดับต่าง ๆ แทนแอมพลิจูดของแต่ละความถี่ ณ ช่วงเวลาสั้น ๆ ซึ่งข้อมูลเชิงความถี่สามารถคำนวณได้จากข้อมูลเชิงเวลาด้วยการแปลงแบบฟูเรียร์ ดังจะได้กล่าวต่อไป

2.3 **การแปลงแบบฟูเรียร์** (Fourier Transformation)

ในปี ค.ศ. 1822 Joseph Fourier ได้เสนอการแปลงแบบฟูเรียร์ (Fourier Transformation) ซึ่งสามารถแยกองค์ประกอบของสัญญาณออกเป็นผลรวมของฟังก์ชันไซน์และฟังก์ชันโคไซน์ และแปลงกลับเป็นสัญญาณเดิมได้ ด้วยการแปลงแบบฟูเรียร์ผกผัน (Inverse Fourier Transformation)

โดยปกติ การแปลงฟูเรียร์จะใช้กับสัญญาณแอนะล็อกซึ่งมีความต่อเนื่อง แต่สำหรับในงานวิจัยนี้เป็นการประมวลผลสัญญาณดิจิทัล ซึ่งเป็นสัญญาณที่ไม่ต่อเนื่อง จึงได้ใช้การแปลงแบบฟูเรียร์ไม่ต่อเนื่อง หรือ DFT (Discrete Fourier Transformation) ดังสมการที่ (2.2)

$$S(f) = \sum_{t=0}^{N-1} s(t) e^{-i2\pi ft/N} \tag{2.2}$$

ซึ่งข้อมูลเชิงความถี่นี้ สามารถแปลงกลับเป็นสัญญาณเดิมได้ด้วยการแปลงแบบฟูเรียร์ไม่ต่อเนื่องผกผัน หรือ IDFT (Inverse Discrete Fourier Transformation) ดังสมการที่ (2.3)

$$s(t) = \frac{1}{N} \sum_{f=0}^{N-1} S(f) e^{i2\pi ft/N} \tag{2.3}$$

โดยที่ $S(f)$ และ $s(t)$ แทนสัญญาณในเชิงโดเมนความถี่และโดเมนเวลา ตามลำดับ และ N แทนจำนวนจุดข้อมูลคลื่นเสียงดิจิทัล

สมการการแปลงแบบฟูเรียร์เหล่านี้ ถือเป็นหัวใจของการวิเคราะห์คลื่นเสียง เมื่อพิจารณาสมการที่ (2.2) จะเห็นว่าผลลัพธ์ของฟังก์ชัน DFT เป็นจำนวนเชิงซ้อน ดังสมการที่ (2.4) ซึ่งสามารถมองในรูปแบบของสมการคลื่นที่ประกอบด้วยแอมพลิจูดและมุมเฟส ดังแสดงได้ด้วยสมการที่ (2.5)

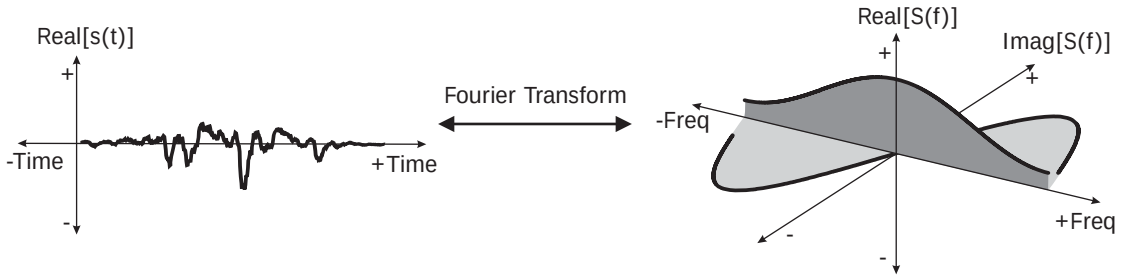
$$S(f) = a + bi \tag{2.4}$$

$$S(f) = A(f) e^{i\phi(f)} \tag{2.5}$$

โดยที่ $A(f)=|S(f)|=\sqrt{a^2+b^2}$ และ $\phi(f)=\tan^{-1}\frac{b}{a}$ คือ แอมพลิจูดและมุมเฟสที่ความถี่ f ใด ๆ ตามลำดับ ซึ่งการแสดงผลแบบสเปกตรัมและสเปกโทรแกรมนั้น เป็นการแสดงค่าแอมพลิจูด หรือ $A(f)$ นั้นเอง

คุณสมบัติที่น่าสนใจของการแปลงแบบฟูเรียร์ทั้งการแปลงแบบปกติและแบบผกผัน คือ การเป็นฟังก์ชันเฮอร์มิเชียน (Hermitian Function) ซึ่งเป็นฟังก์ชันเชิงซ้อนที่มีคุณลักษณะ 2 ประการ ได้แก่ ส่วนจริงเป็นฟังก์ชันคู่ หรือมีความสมมาตรในแกนตั้ง และส่วนจินตภาพเป็นฟังก์ชันคี่ หรือมีความสมมาตรที่จุดกำเนิด และมีคุณสมบัติในการแปลง ซึ่งแสดงไว้ในรูปที่ 2.6 ดังนี้

- การแปลงแบบฟูเรียร์ของฟังก์ชันจริง $s(t)$ ใด ๆ จะได้ผลลัพธ์เป็นฟังก์ชันเฮอร์มิเชียน
- การแปลงแบบฟูเรียร์ของฟังก์ชันเฮอร์มิเชียน $S(f)$ ใด ๆ จะได้ผลลัพธ์เป็นฟังก์ชันจริง



รูปที่ 2.6 คุณสมบัติของการแปลงแบบฟูเรียร์

การที่สามารถทำความเข้าใจถึงคุณสมบัติความเป็นฟังก์ชันเฮอร์มิเชียน จะช่วยในการแปลงคลื่นเสียงจากโดเมนความถี่มาเป็นโดเมนเวลาได้ซึ่งเป็นฟังก์ชันจำนวนจริง โดยไม่สูญเสียข้อมูลในส่วนจินตภาพที่เกิดขึ้นได้จากการแปลงแบบฟูเรียร์แบบผกผัน

2.4 การแยกเสียงร้องออกจากเสียงเพลง

ในหัวข้อนี้ ผู้วิจัยจะกล่าวถึง การเปรียบเทียบระหว่างเสียงร้องเพลงและเสียงพูด ข้อกำหนดต่าง ๆ ของงานวิจัยด้านการแยกเสียงที่ผ่านมา อันจะส่งผลต่อความซับซ้อนของงานวิจัยด้านการแยกเสียงร้อง วิธีการแยกเสียงร้องออกจากเสียงเพลงที่ใช้ในงานวิจัยก่อนหน้านี้ รวมทั้งข้อดีและข้อจำกัดของแต่ละวิธี และรายละเอียดของวิธีการที่งานวิจัยนี้เลือกใช้ ตามลำดับ

2.4.1 การเปรียบเทียบระหว่างเสียงร้องเพลงและเสียงพูด

งานวิจัยด้านการแยกเสียงร้องออกจากเสียงเพลงนั้น มีความเกี่ยวข้องกับงานวิจัยทางด้านการแยกเสียงพูด เนื่องจากเสียงร้องเพลงและเสียงพูดเกิดมาจากแหล่งกำเนิดเสียงเดียวกัน อย่างไรก็ตาม ด้วยแนวคิด สภาพแวดล้อม และปัจจัยของงานวิจัยด้านการแยกเสียงพูดรวมทั้งธรรมชาติของเสียงเอง ทำให้มีความแตกต่างในการแยกเสียงหลายประการ ซึ่ง Li และ Wang [12] ได้เปรียบเทียบเสียงพูดและเสียงร้องเพลงไว้ดังตารางที่ 2.1

ตารางที่ 2.1 การเปรียบเทียบระหว่างเสียงพูดและเสียงร้องเพลง

ข้อเปรียบเทียบ		เสียงพูด	เสียงร้องเพลง
เสียงพูด/ เสียงร้องเพลง	แหล่งกำเนิดเสียง	ช่องสายเสียงของคน	ช่องสายเสียงของคน
	ช่วงของระดับเสียง	แคบ (80–400 เฮิรตซ์)	กว้าง (80–1400 เฮิรตซ์)
	การเปลี่ยนแปลงของระดับเสียง	ช้า	รวดเร็ว
	สัดส่วนของเสียงก้องต่อเสียงไม่ก้อง	น้อยกว่า	มากกว่า
	เสียงเงียบ	มากกว่า	น้อยกว่า
	ระดับเสียงที่ต่อเนื่อง	น้อยกว่า	มากกว่า
เสียงประกอบ	ลักษณะของเสียง	เป็นอิสระต่อเสียงพูด	สัมพันธ์กับเสียงร้องเพลง
	ฮาร์มอนิก	น้อยกว่า	มากกว่า
	จังหวะ	น้อยกว่า	มากกว่า

จากตารางที่ 2.1 แม้ว่าเสียงพูดและเสียงร้องจะมีแหล่งกำเนิดมาจากช่องสายเสียง (Glottis) ของคนเช่นเดียวกัน แต่ธรรมชาติของเสียงและเสียงประกอบนั้น มีความแตกต่างอย่างมีนัยสำคัญ อันจะส่งผลต่อความซับซ้อนของงานวิจัยด้านการแยกเสียง โดยความแตกต่างเหล่านี้ ได้แก่ การที่เสียงพูดมีช่วงของระดับเสียงที่แคบกว่าเสียงร้องเพลง ทำให้ความเป็นไปได้ทั้งหมดในการเลือกชุดความถี่ที่ต่าง ๆ กันของเสียงพูด มาสังเคราะห์สัญญาณเสียงใหม่นั้น มีน้อยกว่าของเสียงร้อง

ในส่วนของเสียงประกอบ สำหรับงานการแยกเสียงพูด เสียงประกอบจะมีความเป็นอิสระจากเสียงพูด เนื่องจากพื้นฐานของปัญหาการแยกเสียงพูด คือปัญหางานเลี้ยงค็อกเทล

(Cocktail Party Problem) ซึ่งมีเสียงรบกวนที่ไม่เกียวข้องกับเสียงพูดปะปนอยู่ ในขณะที่เสียงประกอบของเสียงร้องเพลงนั้นมักเป็นเสียงดนตรี ซึ่งได้รับการประพันธ์มาให้สอดคล้องกลมกลืนกันตลอดท้งเพลง รวมท้งยังมีเสียงเครื่องดนตรีหลายชนิด ทีอยู่ในย่านความถีตรงกับเสียงร้องเพลงอีกด้วย

จากความแตกต่างทีสำคัญเหล่านี้ งานวิจัยทางด้านการแยกเสียงร้องออกจากเสียงเพลง จึงไม่สามารถนำข้นตอนวิธีทีใช้สำหรับการแยกเสียงพูดมาใช้ได้โดยตรง และด้วยความแตกต่างเหล่านี้เอง งานวิจัยทางด้านการแยกเสียงร้องจึงมีความท้าทายอย่างยิง และทำให้งานวิจัยเหล่านี้ต้องมีการใช้ข้อกำหนดต่าง ๆ เข้ามาช่วยในการหาคำตอบ

2.4.2 ข้อกำหนดต่าง ๆ ของงานวิจัยที่ผ่านมา

ข้อกำหนดทีได้มีการนำมาใช้ในงานวิจัยด้านการแยกเสียงต่าง ๆ อันจะมีนัยต่อข้อจำกัดและความซับซ้อนของงาน ดังต่อไปนี้

1) ความเป็นอิสระต่อกันทางสถิติ[1] (Statistical Independence) [17] ของแหล่งกำเนิดเสียง หมายถึง เสียงจากแหล่งกำเนิดเสียงหนึง ไม่มีผลต่อเสียงจากแหล่งกำเนิดอืน หรือการได้ยินเสียงของแหล่งกำเนิดเสียงหนึง จะไม่ส่งผลต่อความสามารถในการคาดเดาลักษณะของแหล่งกำเนิดเสียงอีกแหล่งหนึงได้ ข้อกำหนดนีมักใช้ในงานแยกเสียงพูด และเป็นไปได้ยากในการแยกเสียงเพลง ทีเสียงจากแต่ละแหล่งนันขึนต่อกัน ตัวอย่างเช่น เสียงเครื่องดนตรีทีเล่นในจังหวะเดียวกัน หรือในเมโลดีทีสัมพันธ์กัน (คีย์เดียวกัน)
2) เสียงทีต้องการแยกออกจากกันต้องไม่อยู่ในย่านความถีเดียวกัน ณ เวลาเดียวกัน สำหรับเสียงดนตรีนัน โอกาสทีข้อกำหนดนีเป็นจริงมีน้อย เนืองจากเสียงของเครื่องดนตรีและเสียงร้องในเพลงเดียวกัน มีโอกาสใช้ระดับเสียง หรือความถีร่วมกัน ณ เวลาเดียวกันสูง
3) จำนวนและชนิดของเสียงประกอบในเพลง เช่น การใช้เฉพาะเครื่องดนตรีกำกับจังหวะ การใช้เฉพาะกีตาร์ หรือเสียงดนตรีในเพลงทัวไป ซึงหากมีจำนวนชนิดของเสียงประกอบในเพลงมาก ความซับซ้อนในการแยกเสียงจะสูง และถ้าเสียง

[1] ความเป็นอิสระต่อกันทางสถิติ (Statistical Independence) ของสองเหตุการณ์ หมายความว่า การเกิดขึ้นของเหตุการณ์หนึ่งไม่ได้ส่งผลให้เกิดอีกเหตุการณ์หนึง [17]

ประกอบมีความถี่อยู่ในย่านเดียวกับเสียงที่ต้องการแยก จะยิ่งทำให้การแยกเสียงมีความซับซ้อนมากขึ้นไปอีก

4) จำนวนตัวรับสัญญาณ (Sensor) หรือช่องสัญญาณ (Channel) อันเป็นข้อมูลเข้าของการแยกเสียง ช่องสัญญาณสำหรับเสียงเพลงโดยทั่วไปจะมีสองแบบ คือ ช่องสัญญาณเดียว และช่องสัญญาณคู่ ซึ่งขั้นตอนวิธีการแยกเสียงบางอย่าง มีการใช้ประโยชน์จากการบันทึกเสียงแบบช่องสัญญาณคู่ หรือช่องสัญญาณสเตอริโอ (Stereo) อันจะช่วยลดความซับซ้อนของการแยกเสียงที่ต้องพิจารณาจากสัญญาณเสียงโดยตรงไปได้มาก
5) ผลลัพธ์ของการแยกเสียงที่คาดหวัง มีสองประเภทคือ การแยกเฉพาะเสียงที่ต้องการ หรือการแยกเสียงทั้งหมดออกจากกัน ซึ่งการแยกเสียงในกรณีหลังนั้นย่อมมีความซับซ้อนกว่า เนื่องจากผลลัพธ์มีความหลากหลายมากกว่า

2.4.3 วิธีการแยกเสียงร้องออกจากเสียงเพลง

จากการศึกษางานวิจัยการแยกเสียงร้องออกจากเสียงเพลง ดังตารางที่ 2.2 พบว่าขั้นตอนวิธีที่ใช้สามารถแบ่งออกได้เป็น 3 กลุ่มใหญ่ ๆ ได้แก่ การใช้แบบจำลองทางสถิติ การวิเคราะห์โสตตามภาวการณ์เชิงคำนวณ การแยกส่วนประกอบเมทริกซ์ และวิธีการอื่น ๆ ที่ไม่สามารถจัดให้เข้ากลุ่มใด ๆ ใน 3 กลุ่มนี้ได้ ดังรายละเอียดต่อไปนี้

2.4.3.1 การใช้แบบจำลองทางสถิติ (Statistical Modeling)

การแยกเสียงร้องด้วยวิธีการทางสถิติ จะต้องมีการเรียนรู้แบบจำลองของเสียงร้องและเสียงดนตรีที่ยังไม่ได้นำมาผสมกันก่อน แล้วจึงใช้การประมาณค่าเพื่อหาสเปกตรัมของเสียงร้องที่ต้องการ

Ozerov และคนอื่น ๆ [9] ใช้แบบจำลองเกาส์เซียนมิกซ์เจอร์ (Gaussian Mixture Models) ในการเรียนรู้สเปกตรัมของเสียงร้องบริสุทธิ์ (Pure Voice) และเสียงดนตรีประกอบ แล้วประมาณสเปกตรัมของเสียงร้องที่ต้องการด้วยวิธีวัดความควรจะเป็นสูงสุด (Maximum Likelihood) ระหว่างเสียงเพลงผสมอันเป็นข้อมูลเข้าและแบบจำลองสเปกตรัมที่เรียนรู้ไว้ ในขณะที่ Tsai และ Wang [8] ใช้ข้อกำหนดที่ว่าเสียงดนตรีประกอบและเสียงร้องมีความเป็นอิสระต่อกันทางสถิติ และสร้างแบบจำลองเฟ้นสุ่ม (Stochastic Models) ซึ่งเรียนรู้จากเสียงดนตรีประกอบและเสียงร้องของผู้ร้องหลายคน โดยในการแยกเสียงร้อง แบบจำลองเสียงร้องของผู้ร้องที่เหมาะสมจะถูกเลือกนำมาใช้คำนวณหาค่าสัญญาณเสียงร้องที่ต้องการ

จุดเด่นของวิธีการใช้แบบจำลองทางสถิติ คือ ความสามารถในการแยกเสียงได้กับเพลงหลายประเภท เนื่องจากวิธีการนี้สามารถเรียนรู้แบบจำลองเสียงร้องและเสียงดนตรีที่เข้ามาใหม่ได้เสมอ อย่างไรก็ตาม ความจำเป็นของการเรียนรู้จากเสียงร้องบริสุทธิ์ หรือเสียงร้องที่ไม่มีดนตรีประกอบนี้ ยังเป็นข้อจำกัดที่สำคัญ เนื่องจากในความเป็นจริง โอกาสที่จะหาเสียงร้องบริสุทธิ์ของผู้ร้องโดยทั่วไปได้นั้นมีน้อยมาก

ตารางที่ 2.2 งานวิจัยทางด้านการแยกเสียงร้องออกจากเสียงเพลง

งานวิจัย	ขอบเขตของปัญหาการแยกเสียง		วิธีการ
	เสียงข้อมูลเข้า	เสียงข้อมูลออก	
Ozerov et al., 2005 [9]	Mono Audio	Vocal, Music	Statistical Modeling
Tsai et al., 2006 [8]	Mono Audio	Vocal	Statistical Modeling
Meron et al., 1998 [10]	Mono (Vocal+Piano)	Vocal, Piano	CASA (Using Musical Score)
Zhang et al., 2006 [11]	Mono (Vocal+Instr.)	Vocal, Instrument	CASA (Harmonic Structure)
Fujihara et al., 2005 [5]	Mono Audio	Vocal	CASA (Using Pitch Contour)
Li et al., 2007 [12]	Mono Audio	Vocal	CASA (Using Pitch Contour)
Feng et al., 2002 [13]	Stereo, Mono Audio	Vocal	Matrix Decomposition
Vembu et al., 2005 [14]	Mono Audio	Vocal	Matrix Decomposition
Mesaros et al., 2007 [4]	Mono Audio	Vocal	Transcription and Resynthesis
Wong et al., 2007 [6]	Stereo Audio	Vocal	Center Pan Extraction
Duda et al., 2007 [3]	Stereo Audio	Vocal	Center Pan Extraction

2.4.3.2 **การวิเคราะห์โสตตามภาวการณ์เชิงคำนวณ** (Computational Auditory Scene Analysis **หรือ** CASA)

การวิเคราะห์โสตตามภาวการณ์เชิงคำนวณ มีจุดมุ่งหมายเพื่อแยกเสียงผสมที่มีจำนวนช่องสัญญาณไม่เกินสองช่องด้วยวิธีการเชิงคำนวณ โดยพยายามเลียนแบบระบบการฟังของมนุษย์ ซึ่งมีการแยกสัญญาณเสียงที่ได้ยินเป็นความถี่ต่าง ๆ และใช้การประมวลผลเพื่อเลือกความถี่ต่าง ๆ ตามข้อมูลช่วยเหลือ อันเป็นสมมติฐานเริ่มต้นของผู้วิจัยในการแยกเสียงที่ได้ยิน

สำหรับงานวิจัยทางด้านการแยกเสียงเพลงที่ใช้วิธีนี้ ได้เริ่มต้นขึ้นในปี ค.ศ. 1998 โดย Meron และ Hirose [10] ได้เสนอวิธีการแยกเสียงร้องและเสียงเปียโนออกจากกัน เนื่องจากเสียงเปียโนนั้น มีโครงสร้างฮาร์มอนิกที่แน่นอนกว่าเสียงร้องของคน ผู้วิจัยจึงใช้โน้ตเพลง (Musical Score) ของเสียงเปียโนไปทำการปรับแนว (Alignment) ให้ตรงกับเสียงเปียโนในเสียงผสม และพยายามสังเคราะห์เสียงเปียโน แล้วนำไปลบออกจากเสียงผสม เพื่อให้ได้สัญญาณเสียงร้องออกมา ซึ่งข้อดีของการลบเสียงเปียโนออก คือ การไม่ต้องกำหนดแบบจำลองของเสียงร้องซึ่งมีความแปรผันตามผู้ร้อง อย่างไรก็ตาม เสียงเปียโนที่สังเคราะห์ขึ้นยังมีความแตกต่างจากเสียงเปียโนที่เล่นจริง ทำให้ยังได้เสียงร้องที่มีเสียงเปียโนปนอยู่

ต่อมา Zhang, Y.-G. และ Zhang, C.-S. [11] ได้ขยายแนวคิดนี้ไปใช้กับเครื่องดนตรีหลายชนิดขึ้น และไม่ใช้โน้ตเพลงมาช่วย โดยทำการสร้างแบบจำลองโครงสร้างฮาร์มอนิก (Harmonic Structure Models) สำหรับเครื่องดนตรีแต่ละชนิด รวมทั้งเสียงร้องไว้ เพื่อใช้ติดตามเสียงของเครื่องดนตรีชนิดนั้น ๆ ในเพลงและทำการแยกเสียงออกเป็นแหล่งกำเนิดเสียงต่าง ๆ ได้ ถึงกระนั้นก็ตาม การระบุเครื่องดนตรีทั้งหมดที่เล่นในเพลงทั่วไปไม่ใช่เรื่องง่าย อีกทั้งเสียงร้องเพลงของแต่ละคนมีโครงสร้างฮาร์มอนิกที่แตกต่างกัน และเครื่องดนตรีบางชนิดที่ไม่มีฮาร์มอนิก หรือบอกทำนองไม่ได้ เช่น กลอง มักปรากฏอยู่ในเพลง จึงเป็นอุปสรรคสำคัญอย่างยิ่งต่อการแยกเสียงเพลงด้วยวิธีนี้

ในทางกลับกัน แทนที่จะติดตามเสียงของเครื่องดนตรี ซึ่งในเพลงทั่วไปนั้นไม่อาจทราบชนิดและจำนวนที่แน่ชัดได้ง่ายนัก งานวิจัยของ Fujihara และคนอื่น ๆ [5] รวมทั้ง Li และ Wang [12] จึงใช้การติดตามเมโลดี (Melody) ของเสียงร้อง เพื่อสกัดหาโครงสร้างฮาร์มอนิกของเสียงร้องที่ต้องการออกมา แล้วจึงสังเคราะห์กลับเป็นสัญญาณเสียงร้องต่อไป ซึ่งเมโลดีของเสียงร้องนี้ ได้จากการตรวจหาระดับเสียงที่เด่นชัด (Predominant Pitch Detection) ในเสียงเพลงผสม

สำหรับวิธีการหลัง สามารถให้ผลการแยกเสียงได้เป็นที่น่าพอใจ โดยใช้เมโลดีของเสียงร้อง ในการสกัดโครงสร้างฮาร์มอนิกที่สัมพันธ์กัน กล่าวอีกนัยหนึ่งคือ วิธีการนี้ได้ใช้เมโลดีของเสียงร้องในการเลือกความถีฮาร์มอนิกที่สัมพันธ์กัน ซึ่งชนิดของเสียงที่มีฮาร์มอนิกคือเสียงก้อง (Voiced Sound) เท่านั้น ดังนั้นสำหรับเพลงที่มีเสียงไม่ก้อง (Unvoiced Sound) อยู่มาก อาจให้ผลการแยกเสียงที่ไม่น่าพอใจนัก อีกทั้งการที่ไม่สามารถบอกองค์ประกอบทางความถีของเสียงร้องได้แน่นอน ทำให้การเลือกความถีที่ไม่ครอบคลุมมีโอกาสเกิดขึ้นได้

2.4.3.3 **การแยกส่วนประกอบเมทริกซ์** (Matrix Decomposition)

การแยกส่วนประกอบเมทริกซ์ที่มีการนำมาใช้ในงานวิจัยด้านการแยกเสียงร้อง คือ การวิเคราะห์ส่วนประกอบอิสระ หรือ ICA (Independent Component Analysis) [18] และการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ หรือ NMF (Non-Negative Matrix Factorization) [19] ซึ่งแต่ละวิธีการจะเป็นการหาส่วนประกอบในรูปของผลคูณของ 2 เมทริกซ์ โดยเริ่มจากการสุ่มค่าเริ่มต้น แล้วทำการวนรอบการทำงานเพื่อปรับค่าสมาชิกของเมทริกซ์ไปจนกระทั่งครบตามจำนวนรอบที่กำหนด หรือผลคูณของเมทริกซ์ตัวประกอบและเมทริกซ์ตั้งต้นมีค่าเข้าใกล้กันด้วยเกณฑ์วัดบางอย่าง เช่น การวัดระยะทางแบบยุคลิด (Euclidean Distance) โดยมีรายละเอียดดังต่อไปนี้

1) การวิเคราะห์ส่วนประกอบอิสระ

การวิเคราะห์ส่วนประกอบอิสระ หรือ ICA (Independent Component Analysis) เป็นวิธีการสำหรับแยกสัญญาณผสม ให้อยู่ในรูปผลบวกของสัญญาณจากแหล่งกำเนิดต่าง ๆ โดยสมมติว่าสัญญาณจากแหล่งต่าง ๆ มีความเป็นอิสระต่อกันทางสถิติ (Statistical Independent) ให้ $x_i(t)$ เป็นแอมพลิจูดของสัญญาณที่บันทึก ณ เวลา t ใด ๆ ด้วยไมโครโฟน i โดยแหล่งกำเนิดสัญญาณแต่ละแหล่งมีการส่งสัญญาณด้วยปริมาณที่แตกต่างกันในเวลาต่าง ๆ ซึ่งสามารถแสดงได้ดังสมการที่ (2.6)

$$x_i(t) = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{in}s_n \tag{2.6}$$

โดยที่ a_{ik} เป็นระยะจากแหล่งกำเนิดเสียงที่ s_k และไมโครโฟน i เมื่อ $k = 1, 2, \ldots, n$ ซึ่งสามารถแสดงในรูปของเมทริกซ์ หรือแบบจำลอง ICA ได้ดังสมการที่ (2.7)

$$X \approx AS \tag{2.7}$$

เมื่อ X ประกอบด้วยเวกเตอร์หลัก (Column Vector) ของสัญญาณต่าง ๆ ที่บันทึกได้จากไมโครโฟนแต่ละตัว A เป็นค่าสัมประสิทธิ์ และ S ประกอบด้วยเวกเตอร์แถว (Row Vector) ของสัญญาณจากแหล่งกำเนิดเสียงต่าง ๆ

ปัญหาการแยกเสียงที่มีการใช้วิธี ICA ส่วนใหญ่มีพื้นฐานมาจากปัญหางานเลี้ยงค็อกเทล (Cocktail Party Problem) ซึ่งมีสัญญาณเสียงจากแหล่งต่าง ๆ ของผู้คนปะปนกัน และหากต้องการแยกสัญญาณจากแหล่งกำเนิดเสียง k แหล่ง จะต้องใช้ไมโครโฟนอย่างน้อย k ตัว ซึ่งในงานวิจัยการแยกเสียงร้องออกจากเสียงเพลงของ Feng และคนอื่น ๆ [13] ที่ใช้วิธี ICA นี้ จึงใช้ฐานข้อมูลเป็นเพลงช่องสัญญาณคู่ แต่ในกรณีของเพลงช่องสัญญาณเดียว วิธี ICA แบบดั้งเดิมดังที่ได้กล่าวข้างต้น จะไม่สามารถใช้แยกเสียงได้

อย่างไรก็ตาม Vembu และ Baumann [14] ได้พยายามใช้วิธี ICA ในงานวิจัยด้านการแยกเสียงร้องที่ใช้เพลงช่องสัญญาณเดียว โดยต้องมีการแปลงสัญญาณเสียงเป็นข้อมูลเชิงความถี่และเวลาหรือสเปกโทรแกรม โดยถือว่าแต่ละความถี่เป็นสัญญาณข้อมูลเข้าของวิธี ICA แต่ด้วยความที่มีจำนวนความถี่มากเกินไป ทำให้ไม่เหมาะกับการแยกเสียงด้วยวิธีนี้ จึงต้องพยายามลดมิติของสัญญาณลงด้วยวิธีการบางอย่าง ก่อนการคำนวณด้วยวิธี ICA ต่อไป นอกจากนี้ Vembu และ Baumann ยังมีการเสนอแนะว่าสามารถใช้วิธี NMF แทนการใช้วิธี ICA ได้ตั้งแต่ในขั้นตอนของการลดมิติของสัญญาณ

2) การหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ

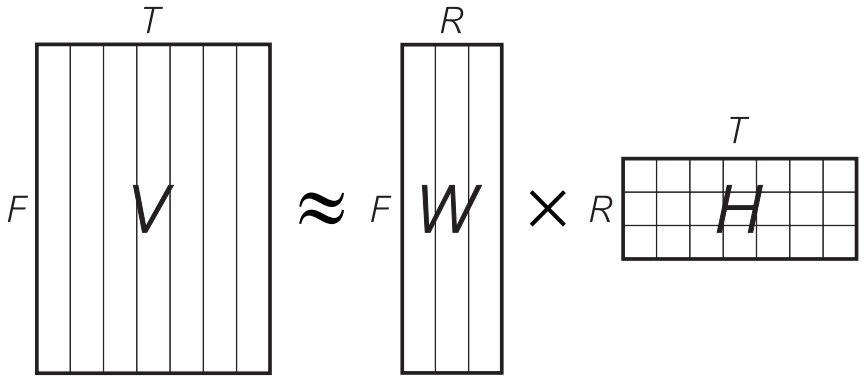
การหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ หรือ NMF (Non-Negative Matrix Factorization) จะทำการหาตัวประกอบของเมทริกซ์ V ให้อยู่ในรูปของผลคูณของเมทริกซ์ W และ H ดังสมการที่ (2.8) โดยที่สมาชิกในเมทริกซ์จะมีค่าไม่น้อยกว่าศูนย์

$$V_{F\times T} \approx W_{F\times R} H_{R\times T} \tag{2.8}$$

โดยที่ค่า R คือจำนวนเต็มบวกใด ๆ ซึ่งน้อยกว่า F หรือ T [20] ที่เลือกให้กับการหาตัวประกอบด้วยวิธีนี้ สำหรับปัญหาการแยกเสียงสมาชิกตำแหน่ง (f,t) ใด ๆ ของเมทริกซ์ V คือแอมพลิจูดของแต่ละความถี่ f ที่เวลา t ต่าง ๆ เมื่อ $1\le f\le F$ และ $1\le t\le T$ หรือค่าที่นำมาแสดงใน สเปกโทรแกรมนั่นเอง

สมการที่ (2.8) สามารถอธิบายได้โดยใช้พีชคณิตเชิงเส้น จากรูปที่ 2.7 เมื่อมองเมทริกซ์ V เป็นคอลัมน์เวกเตอร์ความยาว F จำนวน T เวกเตอร์ การหาตัวประกอบด้วยวิธีนี้จะ

เป็นการหาค่าเมทริกซ์ W ซึ่งประกอบด้วยเวกเตอร์ฐานหลัก (Basis Vector) ความยาว F จำนวน R เวกเตอร์ และเมทริกซ์ H ซึ่งประกอบด้วยค่าสัมประสิทธิ์ของเวกเตอร์ฐานหลักทั้งหมด สำหรับคำนวณได้เป็นแต่ละคอลัมน์เวกเตอร์ของ V



สมาชิกแต่ละตัวมีค่าไม่เป็นลบ

รูปที่ 2.7 การหาตัวประกอบด้วยวิธี NMF

หลังจากการประมาณค่าเมทริกซ์ W และ H ในการแยกเสียง จะเป็นการเลือกเฉพาะเวกเตอร์ฐานหลักของแหล่งกำเนิดเสียงที่ต้องการจากเมทริกซ์ทั้งสอง เช่น เลือกหลักที่ r และแถวที่ r ของเมทริกซ์ W และ H นำมาคูณกัน จะได้เมทริกซ์ใหม่ขนาด $F \times T$ เพื่อใช้ในการคำนวณค่าสเปกตรัมของเสียงที่ต้องการด้วยวิธีการต่าง ๆ ต่อไป

งานวิจัยที่มีการนำวิธี NMF มาทำการแยกเสียงนั้น เริ่มด้วยการนำมาใช้ในการแยกเสียงดนตรี เช่น Wang และ Plumbley [21] ทำการแยกเสียงเครื่องดนตรีชนิดต่าง ๆ ออกจากกันด้วยวิธี NMF โดยกำหนดค่าพารามิเตอร์ R เป็นจำนวนต่าง ๆ เพื่อให้วิธีการดังกล่าวคำนวณเมทริกซ์ตัวประกอบออกมา แล้วผู้วิจัยจะต้องเลือกเวกเตอร์ฐานหลักเอง เพื่อนำมาสร้างสเปกตรัมของเสียง และสังเคราะห์เสียงด้วยการแปลงแบบฟูเรียร์ผกผันต่อไป นอกจากนี้ งานวิจัยอื่น ๆ ที่มีการนำ NMF มาใช้นั้นจะทำเพื่อการวิเคราะห์สเปกตรัมของสัญญาณประเภทต่าง ๆ เช่น คลื่นแม่เหล็กไฟฟ้า รังสี และนอกจากนี้ยังมีการนำมาใช้ในงานวิจัยเกี่ยวกับรู้จำสื่อประสม (Multimedia) รูปแบบอื่น ๆ นอกจากเสียงอีกด้วย เช่น รูปภาพ วีดีโอ เป็นต้น

จากการวิเคราะห์วิธีการหาตัวประกอบเมทริกซ์นี้ พบว่าวิธีการดังกล่าว มีแนวโน้มที่จะให้ผลการแยกเสียงได้เป็นที่น่าพอใจ โดยเฉพาะวิธี NMF ซึ่งสามารถแยกเสียงได้แม้จะเป็นเสียงเพลงช่องสัญญาณเดียว โดยการประมาณค่าเมทริกซ์ตัวประกอบจากข้อมูลสเปกตรัมของเสียง ซึ่งจะได้ส่วนประกอบของเสียงมาครบถ้วน ทั้งยังสามารถคำนวณได้โดยไม่ต้องทำการลดมิติ

ของข้อมูลอีกด้วย งานวิจัยนี้จึงเลือกใช้วิธี NMF ในการแยกเสียง และจะกล่าวถึงรายละเอียดในหัวข้อ 2.4.4

2.4.3.4 วิธีการอื่น ๆ

สำหรับงานวิจัยด้านการแยกเสียงที่จัดให้อยู่ในกลุ่มวิธีการอื่น ๆ จะกล่าวถึงกระบวนการทำงานในมุมกว้างเท่านั้น เนื่องจากใช้วิธีการ หรือข้อมูลเข้าชนิดที่อยู่นอกเหนือความสนใจของงานวิจัยชิ้นนี้

Mesaros และคนอื่น ๆ [4] มีวิธีการคล้ายกับ CASA คือพยายามหาเมโลดีของเพลง แต่สิ่งที่ต่างกันคืองานวิจัยนี้ไม่ได้ทำการแยกเสียง กล่าวคือ ผู้วิจัยใช้เมโลดีของเพลงเพื่อประมาณค่าความถี่โน้ต แล้วทำการสังเคราะห์สัญญาณไซน์ที่มีฮาร์มอนิกต่าง ๆ ตามค่าความถี่นั้นขึ้นมาโดยตรง

งานวิจัยของ Wong และคนอื่น ๆ [6] และ Duda และคนอื่น ๆ [3] มีการใช้ประโยชน์จากกระบวนการบันทึกเสียงแบบสเตอริโอ (Stereo) ซึ่งบรรจุเสียงร้องลงในช่องสัญญาณทั้งสอง (Center Pan) เท่า ๆ กัน และบรรจุเสียงเครื่องดนตรีต่าง ๆ ในช่องสัญญาณซ้ายและขวาด้วยอัตราส่วนที่แตกต่างกัน โดยในขั้นตอนแรกจะทำการตัดเสียงส่วนกลาง (Center Pan Removal) เพื่อให้ได้เฉพาะเสียงดนตรีประกอบ โดยนำสัญญาณเสียงซ้ายลบกับสัญญาณเสียงขวา เช่นเดียวกับการตัดเสียงร้องในโปรแกรมคาราโอเกะทั่วไป ต่อจากนั้น ในงานวิจัยทั้งสองนี้มองว่าเสียงดนตรีประกอบเป็นเสมือนเสียงรบกวน จึงใช้ขั้นตอนวิธีสำหรับลดเสียงรบกวนต่าง ๆ ซึ่งหลังจากขั้นตอนนี้และการปรับแต่งอีกเล็กน้อย จะได้เสียงร้องตามที่ต้องการ

นอกจากนี้ วิธีการที่ไม่ได้ใช้สำหรับการแยกเสียงร้องโดยตรง แต่สามารถนำมาปรับใช้กับการแยกเสียงร้องออกจากเสียงเพลงได้ คือ การตัดสัญญาณรบกวน (Noise Removal) เมื่อพิจารณาเสียงดนตรีประกอบเป็นสัญญาณรบกวน โดยวิธีนี้มีขั้นตอนการทำงานประกอบด้วยการรับข้อมูลสัญญาณรบกวน (Noise Profile) [22] แล้วสร้างสเปกตรัมความถี่ (Frequency Spectrum) ของสัญญาณรบกวนนั้น เพื่อทำการกรองข้อมูลเสียงเพลง โดยวิธีการนี้เรียกว่าการทำประตูสเปกตรัมสัญญาณรบกวน (Spectrum Noise Gating) ตัวอย่างซอฟต์แวร์ที่มีฟังก์ชันนี้ใช้งาน ได้แก่ Audacity® [23] Adobe® Audition® [24] แม้ว่าวิธีนี้ค่อนข้างจะมีความแปรผัน โดยขึ้นกับสัญญาณรบกวนหรือเสียงดนตรีในกรณีนี้ ที่ผู้ใช้เป็นผู้เลือก ซึ่งผลการแยกเสียงร้องอาจให้ผลที่ไม่ดีเท่าที่ควร หากเลือกช่วงเสียงดนตรีที่มีลักษณะทางความถี่ หรือระดับเสียงที่เล่นต่างจากในเสียงผสม อย่างไรก็ตาม วิธีนี้ค่อนข้างมีความยืดหยุ่นต่อรูปแบบของเพลงประเภทต่าง ๆ

ตามแต่ข้อมูลเสียงดนตรีที่ป้อนให้ และผลการแยกเสียงยังอยู่ในระดับที่น่าพอใจ หากมีการกำหนดเสียงดนตรีให้เหมาะสมกับในเพลง ด้วยข้อได้เปรียบทางด้านความยืดหยุ่นต่อรูปแบบของเพลงดังกล่าว งานวิจัยนี้จึงเห็นว่าวิธีการตัดสัญญาณรบกวน มีความเหมาะสมที่จะนำมาเปรียบเทียบผลการแยกเสียงร้องออกจากเสียงเพลงกับงานวิจัยนี้ โดยจะมีการทดลองในบทที่ 4

2.4.4 วิธีการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ และการประยุกต์เข้ากับปัญหาการแยกเสียงร้องออกจากเสียงเพลง

วิธีการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ หรือวิธี NMF ได้มีการกล่าวถึงไปแล้วบางส่วนในหัวข้อ 2.4.3.3 ในหัวข้อนี้จะเป็นการอธิบายเพิ่มเติม รวมทั้งกล่าวถึงแนวทางการประยุกต์เข้ากับปัญหาที่งานวิจัยนี้สนใจ นั่นคือการแยกเสียงร้องออกจากเสียงเพลง

จากที่กล่าวไปแล้ว ว่าวิธี NMF จะเป็นการหาตัวประกอบของเมทริกซ์ V ในรูปผลคูณของสองเมทริกซ์คือ W และ H ดังสมการที่ (2.8) โดยที่สมาชิกในเมทริกซ์เหล่านี้จะต้องมีค่าไม่น้อยกว่าศูนย์ ในการประมาณค่า W และ H นั้น Lee และ Seung [19] ได้เสนอวิธีการคำนวณความคล้ายหรือฟังก์ชันค่าใช้จ่าย (Cost Function) ระหว่าง V และ WH ไว้สองแบบ คือ การวัดระยะห่างแบบยุคลิด (Euclidean Distance) และการวัดการลู่ออก (Divergence) ระหว่าง V และ WH ดังสมการที่ (2.9) และ (2.10) ตามลำดับ

$$\|V - WH\|^2 = \sum_{ft} \left(V_{ft} - (WH)_{ft}\right)^2 \tag{2.9}$$

$$D(V \,||\, WH) = \sum_{ft} \left(V_{ft} \log \frac{V_{ft}}{(WH)_{ft}} - V_{ft} + (WH)_{ft} \right) \tag{2.10}$$

โดยที่ $1 \le f \le F$ และ $1 \le t \le T$

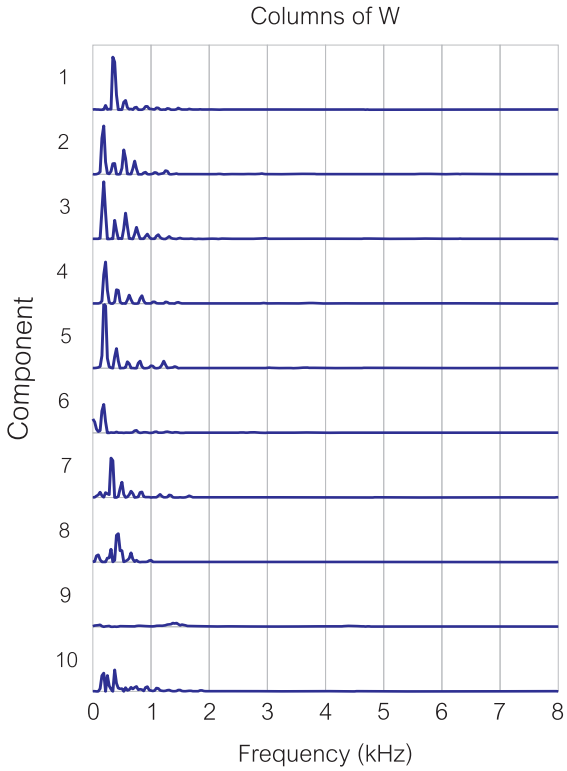
ค่าที่คำนวณได้จากสมการทั้งสองนี้มีค่าต่ำสุดคือศูนย์ และจะเป็นศูนย์ก็ต่อเมื่อ $V = WH$ และจากฟังก์ชันค่าใช้จ่ายทั้งสองนี้เอง นำมาซึ่งกฎการปรับค่า (Update Rules) สองแบบ ตามแต่ละฟังก์ชัน เพื่อลดค่าที่คำนวณได้จากฟังก์ชันค่าใช้จ่ายลงในแต่ละรอบของการปรับ โดยในงานวิจัยนี้ได้เลือกใช้การวัดระยะห่างแบบยุคลิดเป็นฟังก์ชันค่าใช้จ่าย เนื่องจากเป็นฟังก์ชันค่าใช้จ่ายพื้นฐานของวิธีการ NMF อย่างไรก็ตาม ฟังก์ชันค่าใช้จ่ายนี้สามารถเปลี่ยนแปลงได้ตามความเหมาะสม โดยฟังก์ชันดังกล่าวมีกฎการปรับค่าดังสมการที่ (2.11)

$$H_{rt} \leftarrow H_{rt} \frac{(W^T V)_{rt}}{(W^T WH)_{rt}} \qquad W_{fr} \leftarrow W_{fr} \frac{(VH^T)_{fr}}{(WHH^T)_{fr}} \tag{2.11}$$

โดยที่ $1 \le f \le F, 1 \le r \le R$ และ $1 \le t \le T$

บทพิสูจน์การลู่เข้าของกฎการปรับค่า สามารถอ่านเพิ่มเติมได้ในงานวิจัย [19]

ในการอธิบายแนวคิดการประยุกต์เอาวิธี NMF เข้ากับการแยกเสียงร้องออกจากเสียงเพลงนั้น สิ่งที่ควรทำความรู้จักก่อนคือ เงื่อนไขของวิธี NMF ที่สมาชิกของเมทริกซ์จะต้องมีค่าไม่เป็นลบ เงื่อนไขนี้ทำให้เวกเตอร์ฐานหลักในเมทริกซ์ W มีคุณสมบัติความมากเลขศูนย์ (Sparseness)

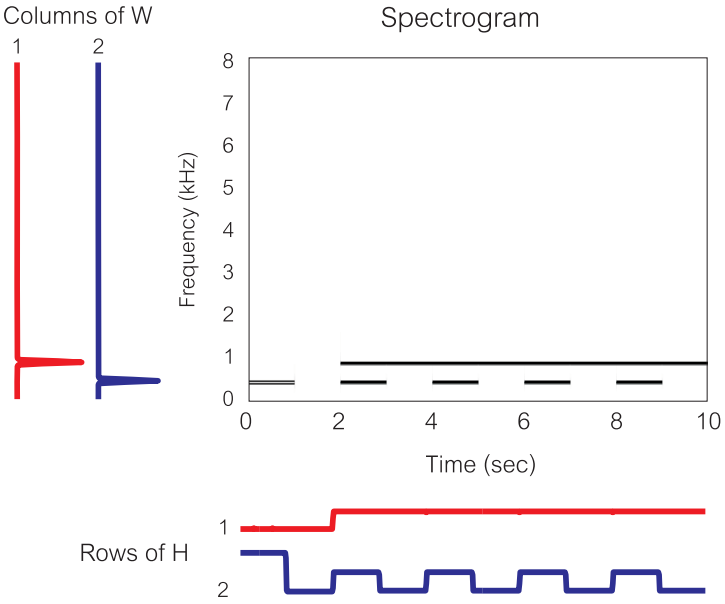


รูปที่ 2.8 คุณสมบัติความมากเลขศูนย์ (Sparseness) ของตัวอย่างเวกเตอร์ฐานหลักจากเมทริกซ์ W ที่ได้จากวิธี NMF เมื่อกำหนดจำนวนเวกเตอร์ฐานหลัก $R = 10$

คุณสมบัติความมากเลขศูนย์ที่กล่าวถึงนี้ คือ การที่แต่ละเวกเตอร์ฐานหลัก ซึ่งในงานวิจัยนี้จะเรียกว่าสเปกตรัมฐานหลัก มีองค์ประกอบทางความถี่ที่แสดงค่าออกมา (Active) [25] เป็นจำนวนไม่มากจากความถี่ทั้งหมด ดังรูปที่ 2.8 เนื่องจากเงื่อนไขที่สมาชิกของเมทริกซ์ต้องมีค่าไม่เป็นลบ และตัวดำเนินการที่ใช้ในสมการที่ (2.8) มีเพียงการบวกเท่านั้น ทำให้สเปกตรัมฐานหลักแสดงความถี่ออกมาเป็นจำนวนไม่มาก เพื่อให้ค่าที่ไม่เป็นลบเหล่านั้นสามารถกระจายไปยังสเปกตรัมฐานหลักได้ทั้งหมด โดยเฉพาะอย่างยิ่ง ความถี่ที่มักจะเกิดขึ้นในเวลาเดียวกันมักจะถูกนำมารวมเข้าด้วยกัน กล่าวคือ ตัวสเปกตรัมฐานหลักจะสะท้อนให้เห็นถึงฮาร์มอนิกหรือแทมเบอร์

ของโน้ตแต่ละตัวสำหรับเสียงเครื่องดนตรีได้ ซึ่งในกรณีนี้ จำนวนเต็ม R ในการหาตัวประกอบด้วยวิธี NMF สามารถกำหนดให้เป็นจำนวนโน้ตทั้งหมดของเครื่องดนตรีแต่ละชิ้นได้ ดังที่ได้กล่าวถึงในงานวิจัย [20, 21]

อย่างไรก็ตาม ปัญหาการแยกเสียงร้องนั้นมีความซับซ้อนยิ่งกว่าการแยกเสียงเครื่องดนตรี เนื่องจากความถี่ของเสียงร้องที่เปล่งออกมามีความแตกต่างกันในแต่ละคำและยังขึ้นกับบุคคล ทำให้ความถี่ที่เกิดขึ้นพร้อมกัน หรือความถี่ที่ปรากฏในสเปกตรัมฐานหลักมีความหลากหลาย ซึ่งต่างจากเสียงดนตรีที่ชุดของความถี่ค่อนข้างคงที่ ในงานวิจัยนี้จึงใช้ประโยชน์จากเสียงดนตรีที่มีความคงที่ทางความถี่มากกว่า ในการเลือกสเปกตรัมฐานหลักของเสียงดนตรีที่ไม่ต้องการออกไป



รูปที่ 2.9 ตัวอย่างการหาตัวประกอบเมทริกซ์ด้วยวิธี NMF

นอกจากนี้ ประโยชน์ของผลจากการใช้วิธี NMF คือ ความเข้าใจง่ายและไม่กำกวม เนื่องจากสเปกตรัมฐานหลักในเมทริกซ์ W และค่าสัมประสิทธิ์ของแต่ละสเปกตรัมฐานหลักในเมทริกซ์ H เป็นค่าบวกหรือศูนย์ทั้งหมด ทำให้สามารถบอกได้ว่า สเปกตรัมของเสียงผสมประกอบด้วยชุดความถี่ใด ๆ ที่เวลาใดบ้าง และยังบอกได้ถึงปริมาณของแต่ละความถี่ที่มากหรือน้อยเช่นเดียวกับสเปกตรัมปกติ เช่น จากรูปที่ 2.9 เมทริกซ์ W และ H อันเป็นผลจากการหาตัวประกอบด้วยวิธี NMF นั้น สามารถอธิบายได้ว่าองค์ประกอบความถี่ของเวกเตอร์ฐานหลักที่ 1 (หลักที่ 1 ของ W) จะเกิดขึ้นเมื่อเวลาผ่านไปตามสัดส่วนค่าสัมประสิทธิ์ในแถวที่ 1 ของเมทริกซ์ H และเป็นเช่นนี้เรื่อยไปสำหรับทุก ๆ เวกเตอร์ฐานหลัก

2.5 คำอธิบายศัพท์ที่เกี่ยวข้อง

- เพลง (Music)

เสียงที่มักเกิดจากเครื่องดนตรีต่าง ๆ ทั้งที่มีระดับเสียงและไม่มีระดับเสียง ซึ่งเสียงจากเครื่องดนตรีที่มีระดับเสียงนั้นจะประกอบด้วยตัวโน้ตเดียวที่มีระดับเสียงที่ค่อนข้างแน่นอนหลาย ๆ ตัวมาต่อกัน สำหรับเพลงที่งานวิจัยนี้สนใจ จะมีเสียงร้องเพลงเป็นส่วนประกอบ

- เสียงร้องบริสุทธิ์ (Pure Voice)

เสียงร้องเพลงที่ไม่มีเสียงดนตรีประกอบ

- เพลงช่องสัญญาณเดียว (Mono-Channel Music หรือ Monaural Music)

เพลงช่องสัญญาณเดียว เป็นเพลงที่บันทึกในช่องสัญญาณเดียว ได้มาจากการใช้ไมโครโฟนอันเดียวในการบันทึก ทำให้ได้สายสัญญาณเดียว หรือการบันทึกจากไมโครโฟนหลายอันแล้วนำเสียงมาผสมเข้าด้วยกันให้ได้สายสัญญาณเดียว

- แฟ้มข้อมูลแบบ .WAV (Wave File)

แฟ้มข้อมูลแบบ .WAV เป็นรูปแบบแฟ้มข้อมูลเสียงชนิดหนึ่ง ซึ่งเกิดจากการพัฒนาของ Microsoft และ IBM เพื่อใช้เป็นมาตรฐานในการเก็บข้อมูลเสียงในเครื่องคอมพิวเตอร์ .WAV เป็นรูปแบบแฟ้มข้อมูลหลักในการเก็บข้อมูลเสียงดิจิทัลบนระบบปฏิบัติการวินโดวส์ ซึ่งเป็นที่แพร่หลายมากขึ้นเนื่องจากความนิยมที่สูงขึ้นในระบบปฏิบัติการวินโดวส์และปริมาณโปรแกรมสำหรับระบบนี้ โปรแกรมสมัยใหม่เกือบทั้งหมดที่สามารถเปิดหรือบันทึกเสียงดิจิทัลจะรองรับรูปแบบแฟ้มข้อมูลชนิดนี้ และรูปแบบแฟ้มข้อมูลอันเป็นที่นิยมในปัจจุบัน เช่น .MP3 และ .WMA สามารถแปลงเป็นแฟ้มข้อมูลชนิดนี้ได้อีกด้วย ในงานวิจัยนี้จึงทำงานกับแฟ้มข้อมูลแบบ .WAV นี้ ซึ่งเป็นพื้นฐานของแฟ้มข้อมูลเสียงโดยทั่วไป

ในบทนี้ได้มีการกล่าวถึงทฤษฎีเกี่ยวกับเสียงที่มีความสำคัญต่อการทำวิจัยด้านการแยกเสียงร้องออกจากเสียงเพลง รวมทั้งงานวิจัยที่เกี่ยวข้อง เพื่อให้ทราบข้อมูลเบื้องต้น และแนวทางสำหรับการแก้ปัญหาดังกล่าว บทต่อไปจะกล่าวถึงรายละเอียดที่งานวิจัยนี้นำเสนอ โดยนำเอาข้อมูลที่กล่าวถึงในบทนี้ไปประยุกต์ใช้

บทที่ 3

วิธีดำเนินงานวิจัย

การแยกเสียงร้องออกจากเสียงเพลงเป็นปัญหาที่มีความท้าทายอย่างยิ่ง แม้จากที่ได้กล่าวในบทที่ 1 ถึงความจำเป็นที่ต้องมีการแยกเสียงร้องออกจากเสียงเพลง เพื่อใช้ประโยชน์ในงานวิจัยต่าง ๆ มากมาย แต่ด้วยความซับซ้อนของเสียงเพลง ที่มีความหลากหลายในแง่ของเสียงร้อง คือแม้จะเป็นเสียงร้องของนักร้องคนเดียวกัน แต่หากออกเสียงคำที่ต่างกัน รูปแบบองค์ประกอบทางความถี่ก็จะต่างกันไปด้วย ดังนั้น รูปแบบองค์ประกอบทางความถี่ของเสียงร้องจึงมีหลากหลาย รวมทั้งเสียงเครื่องดนตรีที่บรรเลงประกอบในเพลง ล้วนมีผลให้งานวิจัยด้านการแยกเสียงร้องออกจากเสียงเพลงมีความท้าทายสูง งานวิจัยที่ศึกษาทางด้านนี้อย่างจริงจังจึงยังคงมีจำกัด

อย่างไรก็ตาม จากการศึกษางานวิจัยที่เกี่ยวข้อง และวิธีการแก้ปัญหาการแยกเสียงร้องออกจากเสียงเพลงดังที่ได้กล่าวในบทที่ 2 พบว่าวิธีการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ หรือวิธี NMF มีแนวโน้มที่จะนำมาใช้ช่วยในการแก้ปัญหาได้ดีกว่าวิธีอื่น ๆ ที่ใช้ในงานวิจัยที่ผ่านมา โดยสำหรับในบทนี้ จะเป็นการอธิบายถึงขั้นตอนการดำเนินงานวิจัย รวมถึงขั้นตอนวิธีการแก้ปัญหาการแยกเสียงร้องออกจากเสียงเพลงที่นำเสนอโดยนำเอาวิธี NMF เข้ามาช่วย

ในบทนี้ ผู้วิจัยจะอธิบายขั้นตอนการดำเนินงานวิจัยใน 4 หัวข้อหลัก โดยเริ่มจาก 1) การวิเคราะห์และกำหนดแนวทางในการแก้ปัญหาการแยกเสียงร้องออกจากเสียงเพลง โดยอ้างอิงจากความรู้ที่ได้ศึกษาในบทที่ 2 เมื่อได้ถึงวิธีการหลักสำหรับการแก้ปัญหาแล้ว จึงสามารถนำเอาวิธีการนั้นมาใช้ใน 2) การออกแบบขั้นตอนวิธีโดยรวม อย่างไรก็ตาม แม้วิธีการดังกล่าวเคยมีการนำมาใช้แก้ปัญหาการแยกเสียงดนตรี แต่ยังไม่เคยมีการใช้กับปัญหาการแยกเสียงร้องออกจากเสียงเพลง ดังนั้นในการวิจัยจึงต้องมีขั้นตอน 3) การทดสอบวิธีการที่จะนำมาใช้ เพื่อทดสอบว่าวิธี NMF สามารถนำมาใช้ในการแยกเสียงได้ และสุดท้ายจึงเป็นการอธิบายถึง 4) ขั้นตอนวิธีการในการแยกเสียงร้องที่นำเสนอ ดังรายละเอียดต่อไปนี้

3.1 การวิเคราะห์ปัญหาและการกำหนดแนวทางการแยกเสียงร้องออกจากเสียงเพลง

ปัญหาที่งานวิจัยนี้สนใจคือ การแยกเสียงร้องออกจากเสียงเพลงแบบช่องสัญญาณเดียว โดยที่แหล่งกำเนิดเสียงมีทั้งเสียงร้องและเสียงดนตรีประกอบ ซึ่งปัญหานี้สามารถเทียบได้กับการแก้ระบบสมการทางคณิตศาสตร์ ที่มีจำนวนตัวแปรที่ไม่ทราบค่า หรือ

สัญญาณเสียงจากแหล่งกำเนิดต่าง ๆ มากกว่าจำนวนสมการ หรือจำนวนสัญญาณเสียงผสมที่เป็นข้อมูลเข้าของปัญหานี้ ซึ่งมีเพียงสัญญาณเดียว โดยระบบสมการที่มีจำนวนตัวแปรไม่ทราบค่ามากกว่าจำนวนสมการ (Under-Determined System) นี้ เป็นปัญหาที่มีความท้าทายอย่างยิ่ง

งานวิจัยต่าง ๆ ที่ผ่านมา ซึ่งแก้ปัญหาโดยรับมือกับความท้าทายของการแยกเสียงร้องออกจากเสียงเพลงแบบช่องสัญญาณเดียวนี้ ยังคงมีข้อจำกัดในหลายด้านที่แตกต่างกันออกไปซึ่งถือเป็นขอบเขตของงานวิจัยนั้น ๆ ดังที่ได้กล่าวไว้แล้วในหัวข้อ 2.4.3 วิธีการแยกเสียงร้องออกจากเสียงเพลง เช่น วิธีการใช้แบบจำลองทางสถิติ ยังคงต้องมีการเรียนรู้จากเสียงร้องของนักร้องคนเดียวกัน วิธีการวิเคราะห์โสตตามภาวการณ์เชิงคำนวณ (CASA) ประกอบด้วยขั้นตอนหลายขั้นตอนซึ่งล้วนส่งผลต่อความถูกต้องในการแยกเสียงอย่างมีนัยสำคัญ ในปัจจุบันจึงยังไม่มีวิธีการใดที่สามารถบอกได้ว่า จะถูกนำมาแก้ปัญหาการแยกเสียงร้องโดยเฉพาะ งานวิจัยนี้จึงมีความสนใจที่จะแก้ปัญหาดังกล่าวนี้ ด้วยขอบเขตของปัญหาการแยกเสียงร้องสำหรับ

1. เสียงเพลงผสมแบบช่องสัญญาณเดียว
2. การแยกเสียงร้องได้โดยมีการเรียนรู้เสียงร้องล่วงหน้าให้น้อยที่สุด
3. การแยกเสียงร้องออกมาโดยจะต้องไม่มีความถี่ส่วนหนึ่งส่วนใดขาดหายไป อย่างไรก็ตาม การหลงเหลืออยู่ของเสียงดนตรีเพียงเบา ๆ ยังถือว่าไม่ใช่ปัญหาใหญ่ เนื่องจากเสียงร้องนั้นมีความสำคัญมากกว่า

จากมุมมองดังกล่าว และจากการวิเคราะห์วิธีการต่าง ๆ พบว่าเราสามารถนำวิธีการในกลุ่มของการแยกส่วนประกอบเมทริกซ์ (Matrix Decomposition) มาใช้ในการแยกเสียงได้ ซึ่งวิธีการที่งานวิจัยนี้เลือกใช้ คือ การหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ หรือ NMF (Non-negative Matrix Factorization) ซึ่งจากการทดลองเบื้องต้นนั้น วิธีการนี้มีแนวโน้มที่จะสามารถแยกเสียงได้แม้จะเป็นเสียงเพลงช่องสัญญาณเดียว ด้วยการประมาณค่าตัวประกอบทั้งหมดของสเปกโทรแกรมในรูปของผลคูณระหว่าง 2 เมทริกซ์ได้ โดยไม่ทำให้ข้อมูลทางความถี่หายไป ต่างจากวิธี CASA ซึ่งมีการเลือกเฉพาะความถี่ที่สัมพันธ์กับเมโลดี จึงอาจทำให้ได้ความถี่มาไม่ครอบคลุม รวมทั้งยังต้องขึ้นกับขั้นตอนก่อนหน้าในการหาเมโลดีของเสียงร้องเพลง ซึ่งสามารถมีความผิดพลาดอันเกิดจากเสียงดนตรีที่บรรเลงในเพลงได้ และสำหรับวิธีการในกลุ่มของการแยกส่วนประกอบเมทริกซ์ด้วยกัน วิธีการวิเคราะห์ส่วนประกอบอิสระ หรือ ICA (Independent Component Analysis) นั้นเป็นวิธีที่น่าสนใจ แต่การที่ต้องมีการลดมิติของเมทริกซ์ จึงย่อมมีโอกาสที่ส่วนประกอบของเสียงร้องจะถูกลดทอนไปในขั้นตอนนี้ได้

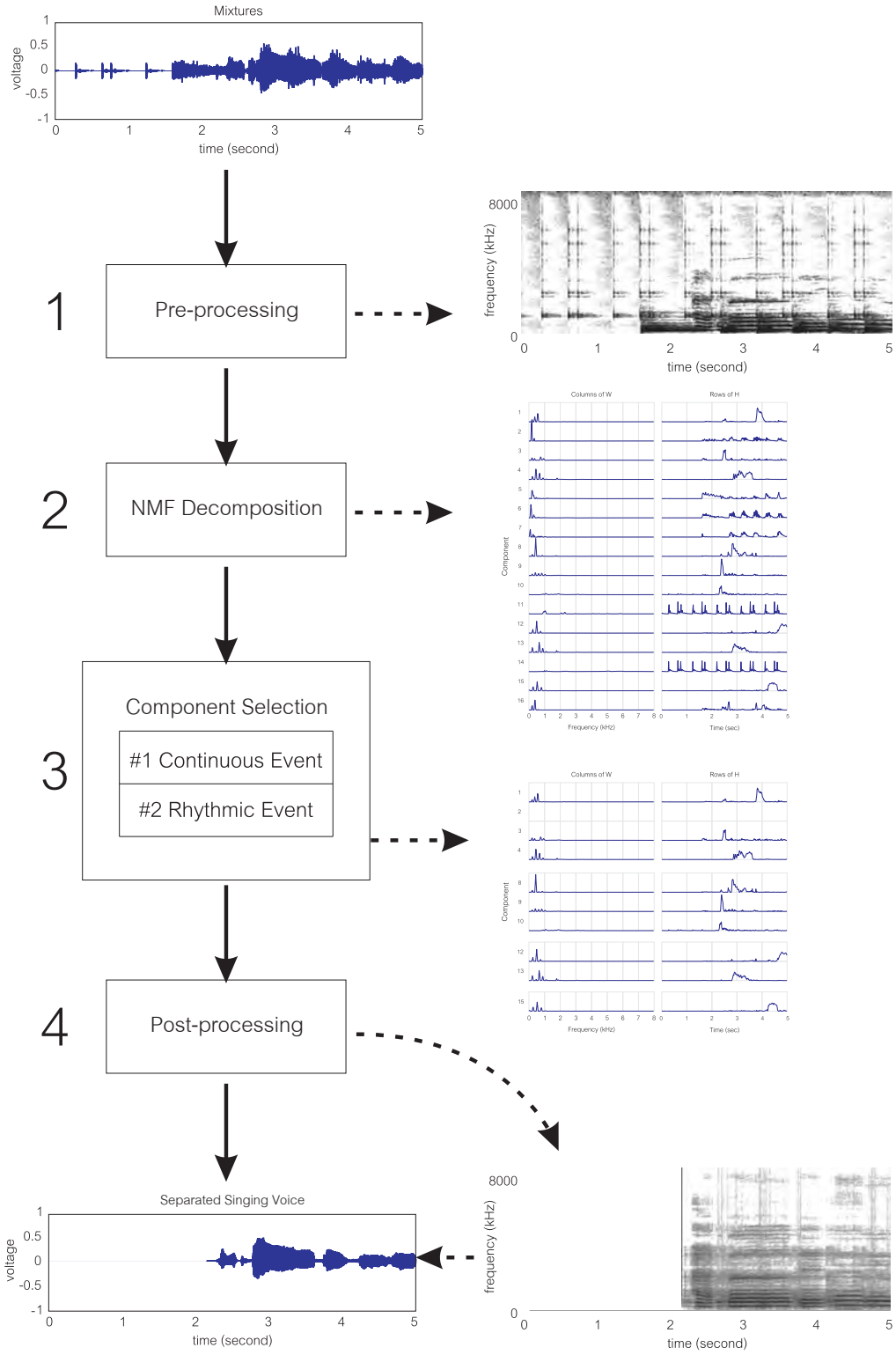
ข้อดีอีกประการหนึ่งของวิธี NMF คือการไม่ต้องเรียนรู้จากเสียงร้องบริสุทธิ์ของผู้ร้องคนเดียวกัน ซึ่งเสียงร้องบริสุทธิ์นี้มีความจำเป็นต่อวิธีการใช้แบบจำลองทางสถิติ และนอกจากนี้ วิธี NMF ยังให้ผลลัพธ์ของการแยกตัวประกอบต่าง ๆ ที่มีค่ามากกว่าหรือเท่ากับศูนย์ ซึ่งช่วยให้ง่ายต่อการตีความอีกด้วย

ดังนั้น จากที่กล่าวมาข้างต้น วิธี NMF จึงเหมาะสมต่อการนำมาใช้ในการแยกเสียงร้องออกจากเสียงเพลงเป็นอย่างยิ่ง ซึ่งขั้นตอนการแยกเสียงร้องออกจากเสียงเพลงโดยใช้วิธี NMF จะกล่าวถึงต่อไป

3.2 การออกแบบขั้นตอนวิธีโดยรวมสำหรับการแยกเสียงร้องออกจากเสียงเพลง

จากการทีงานวิจัยนี้ได้เลือกใช้วิธี NMF ซึ่งเป็นวิธีการหาตัวประกอบเมทริกซ์ มาช่วยในการแก้ปัญหาการแยกเสียงร้องออกจากเสียงเพลง งานวิจัยนี้จึงสามารถกำหนดขั้นตอนการทำงานหลัก ๆ ได้เป็น 4 ขั้นตอน คือ การประมวลผลก่อน (Pre-Processing) การหาตัวประกอบด้วยวิธี NMF (NMF Decomposition) การเลือกองค์ประกอบ (Component Selection) การประมวลผลหลัง (Post-Processing) โดยขั้นตอนการทำงานทั้งหมดสามารถสรุปเป็นแผนภาพได้ดังรูปที 3.1

จากรูปที 3.1 ขั้นตอนวิธีการแยกเสียงร้องออกจากเสียงเพลงในงานวิจัยนี้ จะรับข้อมูลเข้าเป็นสัญญาณเสียงผสม ซึ่งในที่นี้คือเสียงเพลงสัญญาณเดียวไปผ่านกระบวนการประมวลผลก่อน (Pre-Processing) เพื่อเตรียมสัญญาณเสียงให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปหาตัวประกอบด้วยวิธี NMF ในขั้นตอนการหาตัวประกอบด้วยวิธี NMF (NMF Decomposition) นั้น จะมีการคำนวณเพื่อหาค่าจำนวนตัวประกอบที่เหมาะสม แล้วจึงคำนวณหาตัวประกอบ หลังจากนั้น ในขั้นตอนการเลือกองค์ประกอบ (Component Selection) ระบบจะใช้เกณฑ์ต่าง ๆ เพื่อทำการเลือกตัวประกอบเมทริกซ์ที่มีแนวโน้มเป็นองค์ประกอบของเสียงร้อง และเมื่อตัวประกอบเมทริกซ์ที่เลือกได้มาผ่านกระบวนการประมวลผลหลัง (Post-Processing) และสร้างกลับเป็นสัญญาณเสียง ก็จะได้ผลลัพธ์เป็นเสียงร้องเพลงที่แยกแล้วตามต้องการ



รูปที่ 3.1 ขั้นตอนวิธีในการแยกเสียงร้องออกจากเสียงเพลงที่นำเสนอ

จากขั้นตอนหลัก ๆ ทั้งสี่ จะมีสองขั้นตอนที่มีความสำคัญและเป็นหัวใจของงานวิจัยด้านการแยกเสียงชิ้นนี้ นั่นคือ ขั้นตอนที่ 2 และ 3 การหาตัวประกอบด้วยวิธี NMF และการเลือกองค์ประกอบ ตามลำดับ ซึ่งมีความเกี่ยวข้องกับการนำเอาวิธี NMF มาใช้ในการ

แก้ปัญหา อย่างไรก็ตาม ก่อนที่จะนำวิธี NMF ไปใช้ในการแยกเสียง ในงานวิจัยนี้ต้องการเห็นผลการแยกเสียงที่ดีจาก NMF เพื่อที่จะมั่นใจถึงการนำเอาวิธีการดังกล่าวไปใช้ในการแยกเสียงต่อไปได้ ซึ่งการทดสอบวิธีการ NMF ได้กล่าวอธิบายในหัวข้อที่ 3.3 และหลังจากนั้นจึงจะเป็นการกล่าวถึงรายละเอียดของขั้นตอนวิธีที่นำเสนอ

3.3 การทดสอบวิธีการ NMF

การทดสอบวิธีการ NMF มีวัตถุประสงค์เพื่อแสดงว่าวิธี NMF ซึ่งเป็นวิธีที่งานวิจัยนี้ได้เลือกมาใช้ในการแก้ปัญหาการแยกเสียงร้องออกจากเสียงเพลงนั้น สามารถที่จะแยกเสียงได้จริง โดยจะทำการทดลองเบื้องต้นกับเพลงขนาดสั้น ที่มีคำร้องจำนวนไม่มาก และแยกตัวประกอบเป็น 16 ตัวประกอบ แล้วพิจารณาผลลัพธ์ที่ดีที่สุดจากรูปแบบการจับกลุ่มองค์ประกอบทั้งหมดที่เป็นไปได้ และเพื่อที่จะบอกได้ว่ารูปแบบการจับกลุ่มองค์ประกอบแบบใดเป็นคำตอบที่ดีที่สุด มาตรวัดที่ใช้ในการทดลองนี้ คือ อัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (Peak Signal-to-Noise Ratio หรือ PSNR) [26] (รายละเอียดเพิ่มเติมดูบทที่ 4)

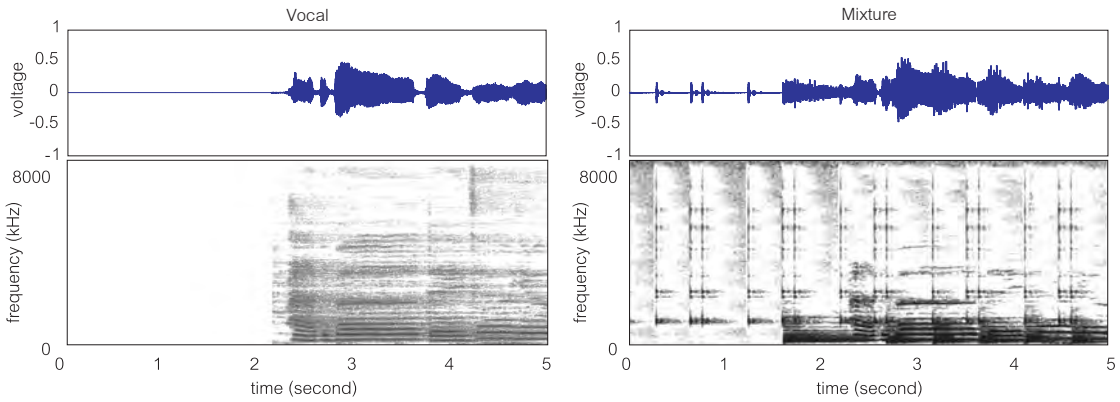
เมื่อ $Spectrogram_1$ และ $Spectrogram_2$ เป็นสเปกโทรแกรมขนาด $F \times T$ ของสัญญาณเสียงร้องที่แยกได้ และเสียงร้องต้นฉบับ ตามลำดับ PSNR ในหน่วยเดซิเบลของสเปกโทรแกรมทั้งสองสามารถคำนวณได้ดังสมการที่ (3.1) ซึ่งผลลัพธ์การเปรียบเทียบหากเสียงทั้งสองมีความใกล้เคียงกันมาก PSNR จะมีค่าสูง

$$PSNR = 20\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{3.1}$$

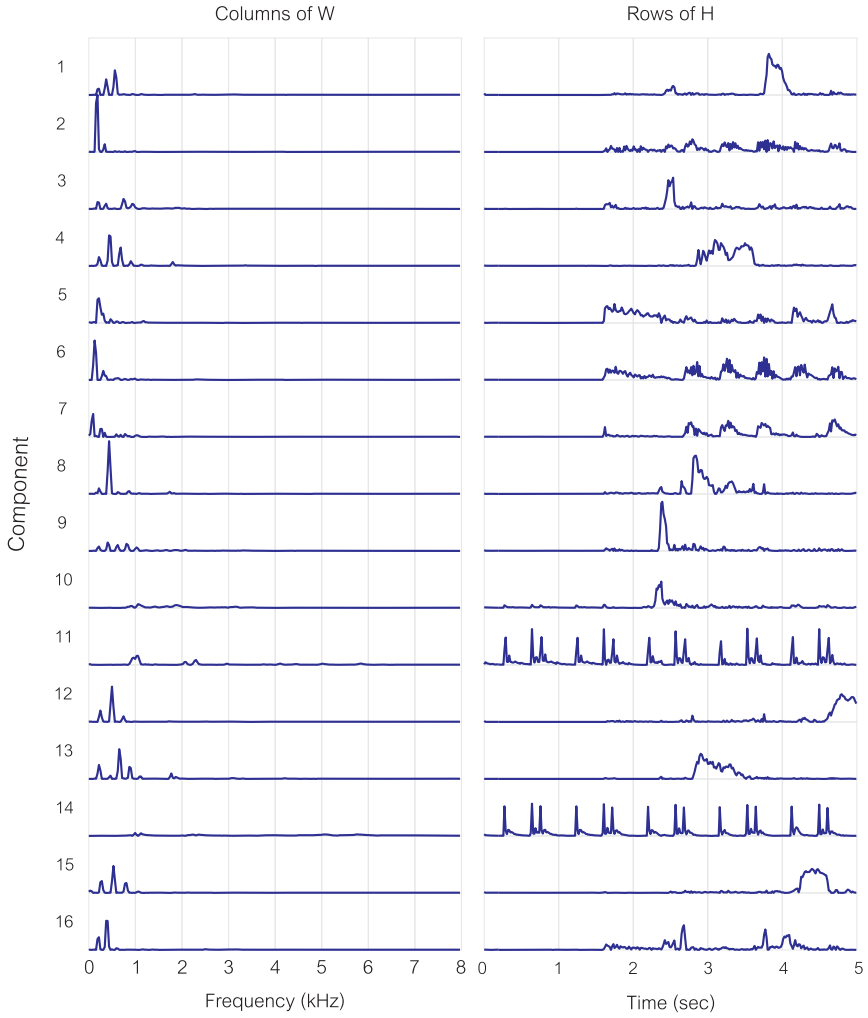
โดยที่ MAX_I คือค่าสูงสุดของสเปกโทรแกรมทั้งสอง และค่า MSE คือค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) ดังสมการที่ (3.2)

$$MSE = \frac{1}{FT}\sum_{i=1}^{F}\sum_{j=1}^{T}\left(Spectrogram_1(i,j) - Spectrogram_2(i,j)\right)^2 \tag{3.2}$$

ข้อมูลเสียงผสมที่ใช้ในการทดสอบวิธี NMF และข้อมูลเสียงร้องต้นฉบับ แสดงได้ดังรูปที่ 3.2 (ก) และ (ข) ตามลำดับ และผลการหาตัวประกอบแสดงดังรูปที่ 3.3



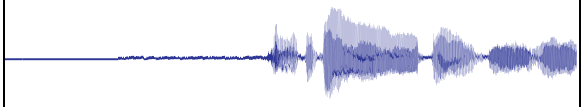
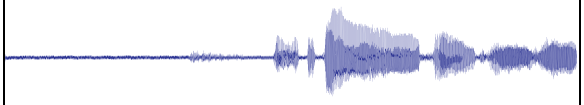
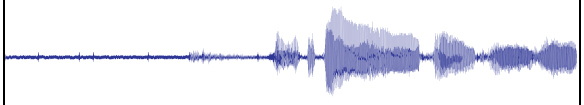
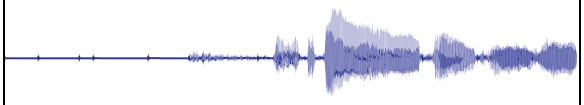
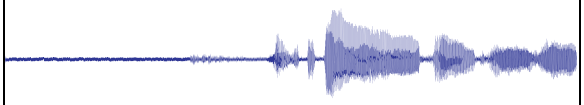
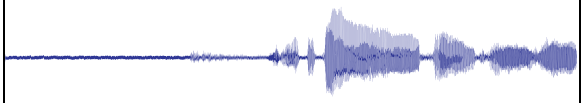
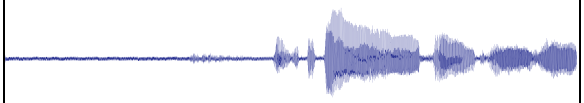
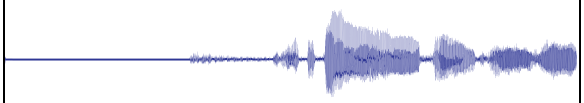
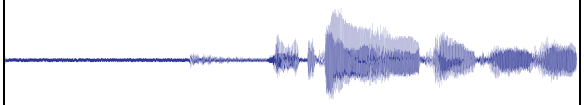
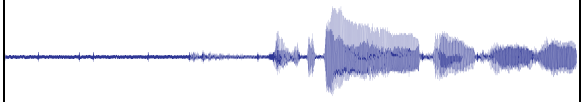
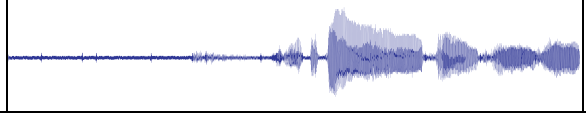
รูปที่ 3.2 คลื่นเสียงและสเปกโทรแกรมของ (ก) เสียงผสม และ (ข) เสียงร้อง

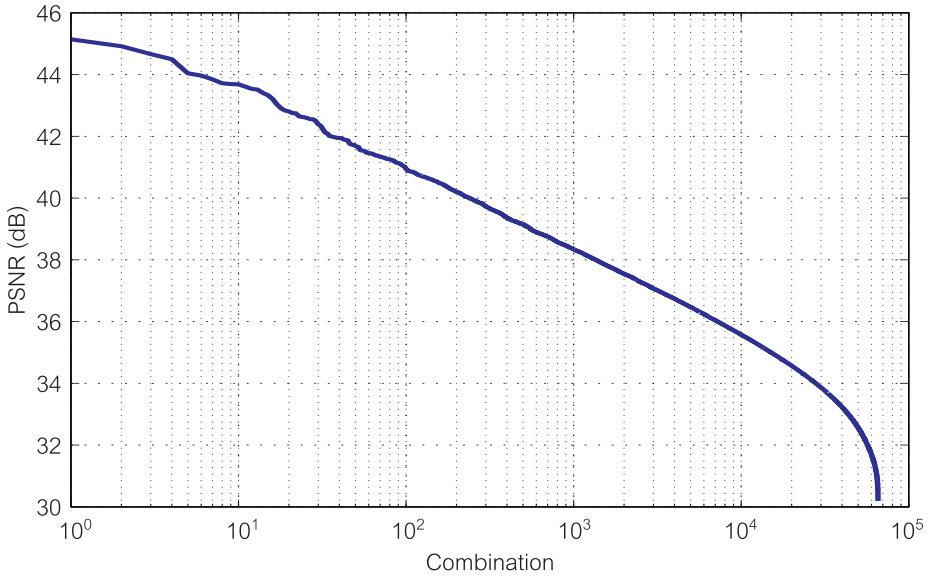


รูปที่ 3.3 ผลการหาตัวประกอบด้วยวิธี NMF ของสัญญาณเสียงผสม
เมื่อกำหนดจำนวนตัวประกอบเท่ากับ 16

ผลลัพธ์ที่ดีที่สุด 10 อันดับแรก จากการคำนวณค่า PSNR ของทุกรูปแบบที่เป็นไปได้ของ 16 ตัวประกอบ แสดงได้ดังตารางที่ 3.1 ซึ่งค่า PSNR ระหว่างเสียงร้องที่แยกได้และเสียงร้องต้นฉบับ คำนวณจากเสียงช่วงครึ่งหลังเท่านั้น และค่า PSNR ทั้งหมดเรียงตามลำดับจากมากไปน้อยแสดงได้ดังรูปที่ 3.4

ตารางที่ 3.1 ผลการจับกลุ่มตัวประกอบที่มีค่า PSNR ระหว่างข้อมูลเสียงร้องที่แยกได้และเสียงร้องต้นฉบับมากที่สุด 10 อันดับแรก

เสียงร้อง	คลื่นเสียง	ตัวประกอบ	PSNR (dB)
เสียงร้องต้นฉบับ		-	-
1		1,3,4,8,9,10, 12,13,15,16	45.1386
2		1,3,4,8,9, 12,13,15,16	44.9197
3		1,3,4,8,9,10, 12,13,14,15,16	44.6648
4		1,3,4,8,9, 12,13,14,15,16	44.4956
5		1,4,8,9,10, 12,13,15,16	44.0415
6		1,3,4,8,10, 12,13,15,16	43.9645
7		1,4,8,9, 12,13,15,16	43.8398
8		1,3,4,8, 12,13,15,16	43.7160
9		1,3,4,7,8,9,10, 12,13,15,16	43.6895
10		1,4,8,9,10, 12,13,14,15,16	43.6802



รูปที่ 3.4 ค่า PSNR ของรูปแบบการจัดกลุ่มทั้งหมดของ 16 ตัวประกอบ เรียงจากมากไปน้อย

จากตารางที่ 3.1 เมื่อพิจารณาจากคลื่นเสียงร้องต้นฉบับและคลื่นเสียงร้องผลลัพธ์อันดับที่หนึ่ง จะเห็นว่าวิธีการ NMF สามารถแยกตัวประกอบของเสียงร้องออกมาได้ โดยให้รูปแบบของคลื่นเสียงที่ใกล้เคียงกับเสียงร้องต้นฉบับมาก นอกจากนี้ เสียงร้องที่ให้ค่า PSNR ที่สูงในอันดับอื่น ๆ ก็มีลักษณะของคลื่นเสียงที่คล้ายกับเสียงร้องต้นฉบับมากเช่นเดียวกัน จึงเห็นได้ว่ารูปแบบของตัวประกอบที่ดีสามารถเป็นไปได้หลากหลาย โดยนอกเหนือจากตัวอย่างที่มีค่า PSNR สูงสุด 10 อันดับแรก รูปที่ 3.4 ยังแสดงให้เห็นอีกว่า รูปแบบการจัดกลุ่มองค์ประกอบอื่น ๆ ยังสามารถให้ค่า PSNR สูงใกล้เคียงกันในอันดับต้น ๆ อีกด้วย

จากการทดลองนี้ทำให้สามารถสรุปได้ว่า วิธี NMF สามารถแยกตัวประกอบของเสียงร้องออกมาได้จริง และผลการจัดกลุ่มในทุกรูปแบบที่มีค่า PSNR สูงใกล้เคียงกันมีหลายรูปแบบ และการที่รูปแบบของตัวประกอบที่ดีมีหลายรูปแบบนี้เอง ทำให้การเลือกตัวประกอบของเสียงร้องมีความยืดหยุ่น ซึ่งถือเป็นข้อดีของวิธีการนี้

จากผลลัพธ์ที่ดีในการทดลองนี้ ในหัวข้อถัดไปจึงเป็นการนำเอาวิธีการนี้ไปใช้ โดยจะอธิบายถึงขั้นตอนการแยกเสียงร้องออกจากเสียงเพลง ตั้งแต่รับข้อมูลเสียงผสมจนกระทั่งได้เสียงร้องที่แยกแล้วเป็นผลลัพธ์ในหัวข้อถัดไป

3.4 ขั้นตอนการแยกเสียงร้องออกจากเสียงเพลงในงานวิจัยนี้

หลังจากที่ได้วิเคราะห์ปัญหา กำหนดแนวทางการแก้ปัญหา เลือกวิธีการ และทดสอบวิธีการที่เลือกเรียบร้อยแล้ว ในหัวข้อนี้จะเป็นการอธิบายขั้นตอนการแยกเสียงร้องออกจาก

เสียงเพลงที่นำเสนอในงานวิจัยนี้ ซึ่งประกอบด้วย 4 ขั้นตอนหลัก ๆ ได้แก่ การประมวลผลก่อน การหาตัวประกอบด้วยวิธี NMF การเลือกตัวประกอบของเสียงร้อง และการประมวลผลหลัง โดยจะนำเสนอตัวอย่างของเสียงผสมที่ผ่านขั้นตอนต่าง ๆ จนกระทั่งแยกได้เป็นเสียงร้องออกมา ดังรายละเอียดต่อไปนี้

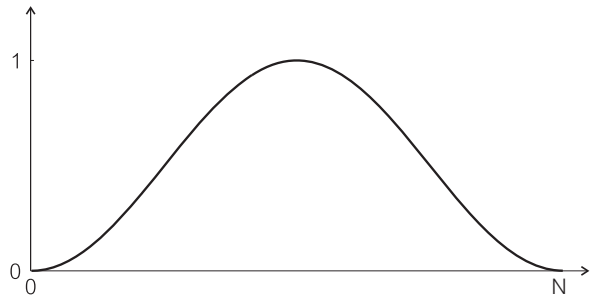
3.4.1 **การประมวลผลก่อน** (Pre-Processing)

ในขั้นตอนนี้มีเป้าหมายคือ การเตรียมข้อมูลเสียงนำเข้าให้อยู่ในรูปแบบที่เหมาะสมต่อการหาตัวประกอบด้วยวิธี NMF โดยข้อมูลเสียงผสมอันประกอบด้วยเสียงร้องเพลงและเสียงดนตรีจะถูกนำมาแทนข้อมูลในรูปแบบสเปกโทรแกรม หรือเมทริกซ์ค่าไม่เป็นลบ V ด้วย 3 ขั้นตอนย่อย ดังรายละเอียดต่อไปนี้

3.4.1.1 **การทำหน้าต่างเลื่อน**

การทำหน้าต่างเลื่อน เป็นขั้นตอนปกติของการประมวลผลสัญญาณดิจิทัล โดยแบ่งข้อมูลเสียงเป็นเฟรม (Frame) ตามหน้าต่างเลื่อน (Sliding Window) ให้มีการซ้อนเหลื่อมกัน (Overlap) โดยขนาดหน้าต่างเลื่อนและขนาดของการซ้อนเหลื่อมที่ใช้คือ $windowSize = 32$ มิลลิวินาทีและ $overlapSize = 16$ มิลลิวินาที ตามลำดับ ซึ่งเป็นค่าที่ใช้ในการประมวลผลทางเสียงโดยทั่วไป

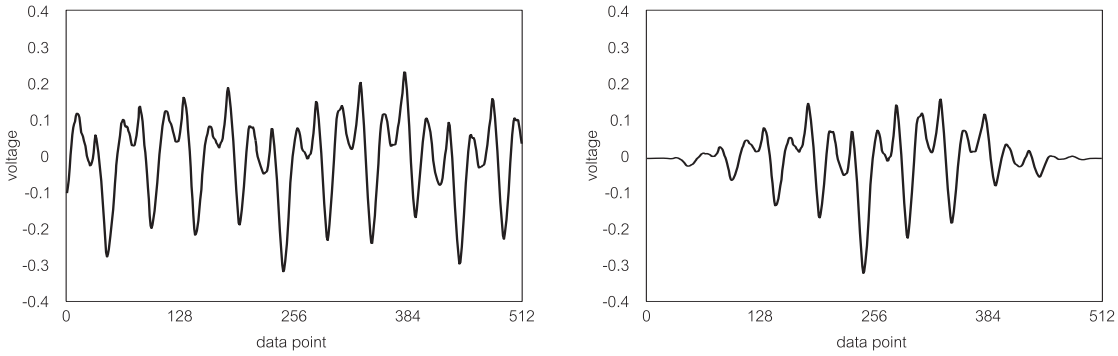
หลังจากนั้น นำข้อมูลเสียงแต่ละเฟรมมาผ่านฟังก์ชันหน้าต่าง (Window Function) ซึ่งในงานวิจัยนี้ได้เลือกใช้หน้าต่างฮานน์ (Hann Window) เพื่อลดความไม่ต่อเนื่องของสัญญาณเสียงระหว่างเฟรม ดังสมการที่ (3.3) ซึ่งแสดงเป็นกราฟได้ดังรูปที่ 3.5 และสัญญาณเสียงแต่ละเฟรมที่ผ่านขั้นตอนนี้ แสดงได้ดังรูปที่ 3.6



รูปที่ 3.5 หน้าต่างฮานน์ (Hann Window) ขนาด N จุดข้อมูล

$$w(n) = 0.5\left(1-\cos\left(2\pi\frac{n}{N}\right)\right), 0 \le n \le N \tag{3.3}$$

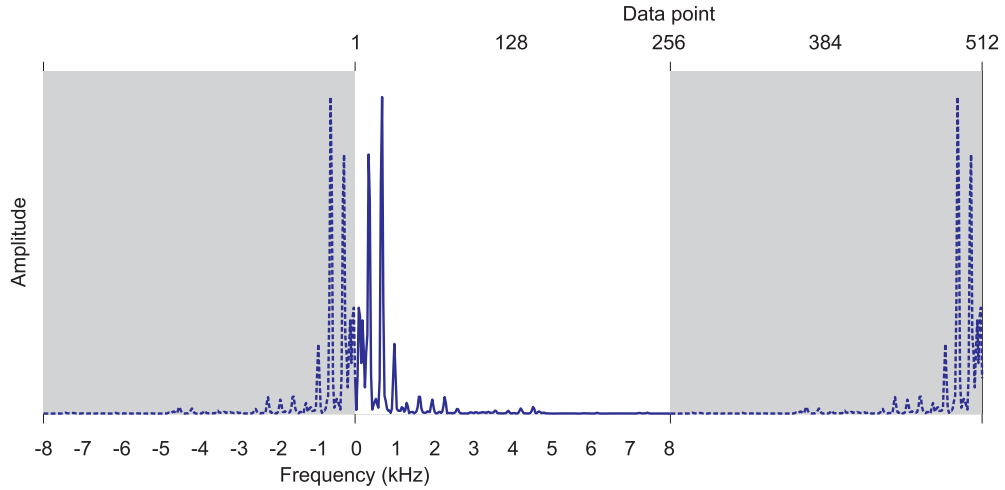
เมื่อ $N = frameSize - 1$ โดยที่ $frameSize = sr \times windowSize$ และ sr คือ อัตราการชักตัวอย่าง (Sampling Rate) ของเสียง เช่น เสียงที่มีอัตราการชักตัวอย่าง 16000 เฮิรตซ์ จะมี $frameSize = 16000 \times 32 \times 10^{-3} = 512$ จุดข้อมูล



รูปที่ 3.6 ตัวอย่างสัญญาณเสียงข้อมูลเข้าในเฟรมขนาด 512 จุดข้อมูล
และสัญญาณเสียงเดียวกันที่ผ่านฟังก์ชันหน้าต่างฮานน์

3.4.1.2 การแปลงข้อมูลให้อยู่ในรูปแบบสเปกตรัมด้วยการแปลงแบบฟูเรียร์ไม่ต่อเนื่อง

เมื่อได้ข้อมูลเสียงที่แบ่งเป็นเฟรมเรียบร้อยแล้ว จะเป็นการแปลงข้อมูลแต่ละเฟรมด้วยการแปลงแบบฟูเรียร์ไม่ต่อเนื่อง (Discrete Fourier Transformation) ผลลัพธ์ที่ได้สำหรับแต่ละเฟรม คือ ลำดับของจำนวนเชิงซ้อนซึ่งแสดงด้วยค่าขนาด (Magnitude) ได้ดังรูปที่ 3.7 (กลาง) และ (ขวา) สำหรับในกรณีนี้จะมีทั้งหมด 512 จุดข้อมูล (แกน y ด้านบน) นอกจากค่าที่ได้จากการแปลงแบบฟูเรียร์ไม่ต่อเนื่องแล้ว รูปที่ 3.7 ยังแสดงข้อมูลทางความถี่ที่อยู่ด้านความถี่ลบ (ซ้าย) ซึ่งเป็นข้อมูลเดียวกับข้อมูลในรูป (ขวา) อันเป็นรูปแบบข้อมูลที่แท้จริง ที่มีความสมมาตรกันที่ความถี่ 0 Hz (แกน y ด้านล่าง) ดังที่ได้กล่าวถึงฟังก์ชันเฮอร์มิเชียนในหัวข้อที่ 2.3 เรื่องการแปลงแบบฟูเรียร์



รูปที่ 3.7 ค่าขนาด (Magnitude) ของจำนวนเชิงซ้อนที่ได้จากการแปลงแบบฟูเรียร์ไม่ต่อเนื่องของเสียงในหนึ่งเฟรม ซึ่งประกอบด้วยสเปกตรัมของเสียงเฟรมนั้น (กลาง) และส่วนสมมาตรของสเปกตรัม (ขวา) ซึ่งมีค่าเท่ากับสเปกตรัมของความถี่ลบ (ซ้าย)

ในขั้นตอนนี้จะเก็บข้อมูลเพียงครึ่งเดียว คือ ข้อมูลในรูปที่ 3.7 (กลาง) นั่นคือจุดข้อมูลที่ 1 ถึง 256 และจุดข้อมูลที่ 257 อีกจุดหนึ่ง ซึ่งประกอบด้วยค่าพลังงาน (Energy) ของเสียงในเฟรมนั้น ๆ (จุดที่ 1) และองค์ประกอบความถี่ในด้านบวก (จุดที่ 2 ถึง 257) โดยที่จะเก็บข้อมูลในรูปของจำนวนเชิงซ้อนเพื่อใช้ในขั้นตอนการสร้างคลื่นเสียงร้องกลับมา

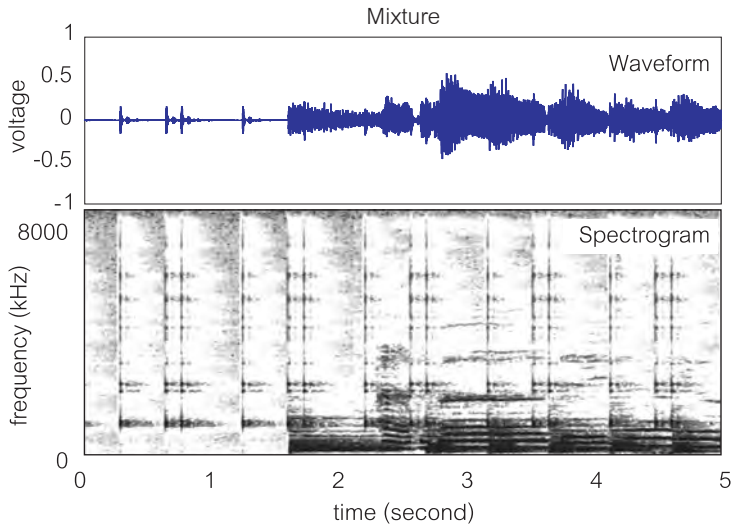
3.4.1.3 การสร้างเมทริกซ์ที่สมาชิกมีค่าไม่เป็นลบ

จากข้อมูลสเปกตรัมของทุกเฟรมเมื่อนำมาต่อกันจะได้เมทริกซ์หรือสเปกโทรแกรมของจำนวนเชิงซ้อน ในขั้นนี้จะเป็นการสร้างเมทริกซ์ที่สมาชิกไม่เป็นลบ หรือเมทริกซ์ V เพื่อเป็นข้อมูลเข้าของขั้นตอนการหาตัวประกอบด้วยวิธี NMF ต่อไป ซึ่งค่าที่ตำแหน่ง (f,t) ใด ๆ จะเป็นการหาค่าขนาดของจำนวนเชิงซ้อนนั้น ๆ ดังสมการที่ (3.4)

$$V(f,t) = \sqrt{Spectrum(f,t) \times \overline{Spectrum(f,t)}} \tag{3.4}$$

เมื่อ $\overline{a+bi} = a-bi$ เป็นค่าสังยุค (Conjugate) ของจำนวนเชิงซ้อน

ตัวอย่างคลื่นเสียงข้อมูลเข้าและเมทริกซ์ V หรือสเปกโทรแกรมที่ได้จากในขั้นตอนนี้ แสดงได้ดังรูปที่ 3.8 (ก) และ (ข) ตามลำดับ และเมื่อได้เมทริกซ์ค่าไม่เป็นลบแล้ว ในขั้นตอนต่อไปจะเป็นการหาตัวประกอบด้วยวิธี NMF



รูปที่ 3.8 (ก) คลื่นเสียงข้อมูลเข้า (ข) เมทริกซ์ V หรือสเปกโทรแกรมจากคลื่นเสียงนี้

3.4.2 การหาตัวประกอบด้วยวิธี NMF (NMF Decomposition)

การหาตัวประกอบด้วยวิธี NMF เป็นการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบจากสเปกตรัมของเสียงเพลง จากการศึกษาและวิเคราะห์วิธีการหาตัวประกอบเมทริกซ์ค่าไม่เป็นลบ หรือวิธี NMF แม้ว่าวิธีนี้จะเป็นทางเลือกที่ดีในการแยกเสียงร้องออกจากเสียงเพลง อย่างไรก็ตาม วิธีนี้ก็ยังคงมีข้อจำกัดซึ่งได้มีการกล่าวถึงบ้างแล้ว คือ

- วิธีนี้ต้องการค่าจำนวนตัวประกอบที่เหมาะสมสำหรับข้อมูลเข้า
- วิธีนี้ต้องการวิธีการเลือกตัวประกอบที่เหมาะสมสำหรับการแยกเสียงร้อง

ในงานวิจัยนี้ได้มีการศึกษาถึงธรรมชาติของข้อจำกัดแต่ละข้อ คือจำนวนตัวประกอบที่เหมาะสม (R) และวิธีการเลือกตัวประกอบที่เหมาะสมสำหรับงานวิจัยนี้จะกล่าวถึงต่อไป

3.4.2.1 การหาค่า R

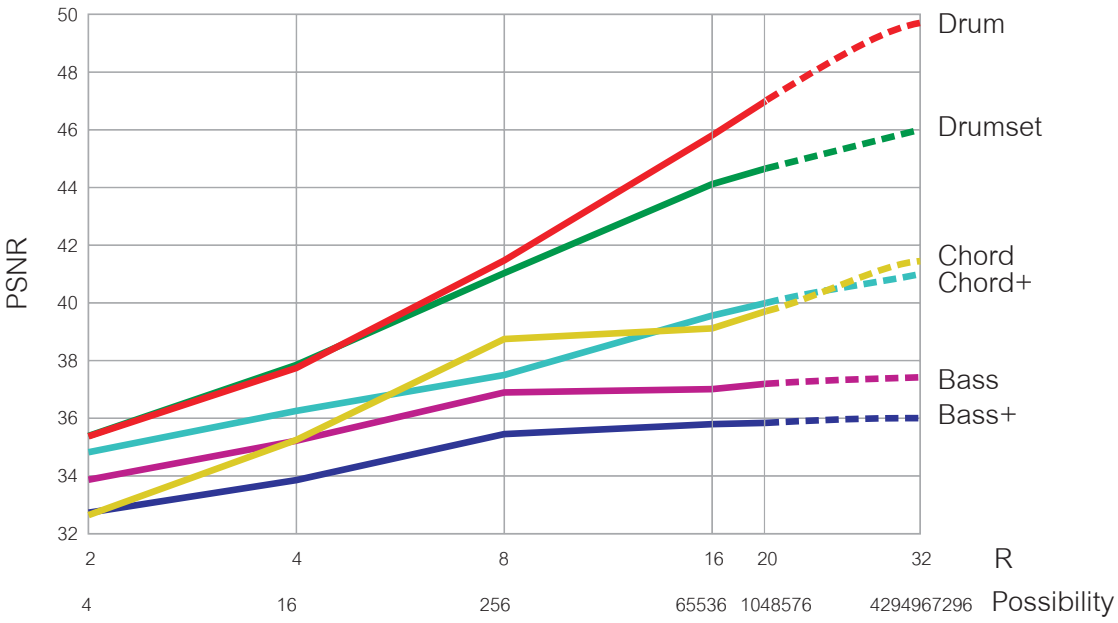
จำนวนตัวประกอบหรือค่า R ที่ใช้จะขึ้นอยู่กับความซับซ้อนของเสียงผสมนั้น ๆ กล่าวคือ จำนวนของแหล่งกำเนิดเสียง แต่จะไม่ส่งผลอย่างมีนัยสำคัญต่อความคลาดเคลื่อนในการหาตัวประกอบของเสียง ตราบใดที่มีค่าไม่น้อยเกินไป

ในการหาค่าจำนวนตัวประกอบที่เหมาะสมหรือค่า R งานวิจัยนี้ได้ทำการทดลองเบื้องต้น เพื่อประเมินว่า

- เมื่อค่า R มากขึ้น ผลการหาตัวประกอบจะดีขึ้นหรือไม่
- ถ้าผลการหาตัวประกอบดีขึ้น จะเป็นข้อมูลเสียงชนิดใด
- ถ้าผลการหาตัวประกอบไม่ดีขึ้น จะเป็นข้อมูลเสียงชนิดใด

ในขั้นตอนนี้ จะออกแบบการทดลองเบื้องต้น โดยทำการหาตัวประกอบ NMF ด้วยค่าจำนวนตัวประกอบ R เป็น 2 4 8 16 และ 20 และคำนวณค่าอัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (PSNR) ระหว่างรูปแบบของการจับกลุ่มตัวประกอบทั้งหมดและเสียงร้องต้นฉบับ เพื่อหาผลการหาตัวประกอบที่ดีที่สุดสำหรับ R นั้น ๆ

ชุดข้อมูลทดลองนี้ประกอบด้วย ข้อมูลเสียงผสมระหว่างเสียงร้อง (ผู้หญิง หรือผู้ชาย ที่ร้องเพลงท่อนเดียวกัน) และเสียงดนตรี (เสียงกลอง เสียงกลองชุด เสียงดนตรีให้ทำนองแบบทีละโน้ต เสียงดนตรีให้ทำนองแบบทีละโน้ตที่มีระดับเสียงสูงขึ้น เสียงดนตรีให้ทำนองแบบคอร์ด หรือเสียงดนตรีให้ทำนองแบบคอร์ดที่มีจำนวนโน้ตที่เล่นมากขึ้น) จำนวน 12 ข้อมูล และผลการทดลองแสดงได้ดังรูปที่ 3.9



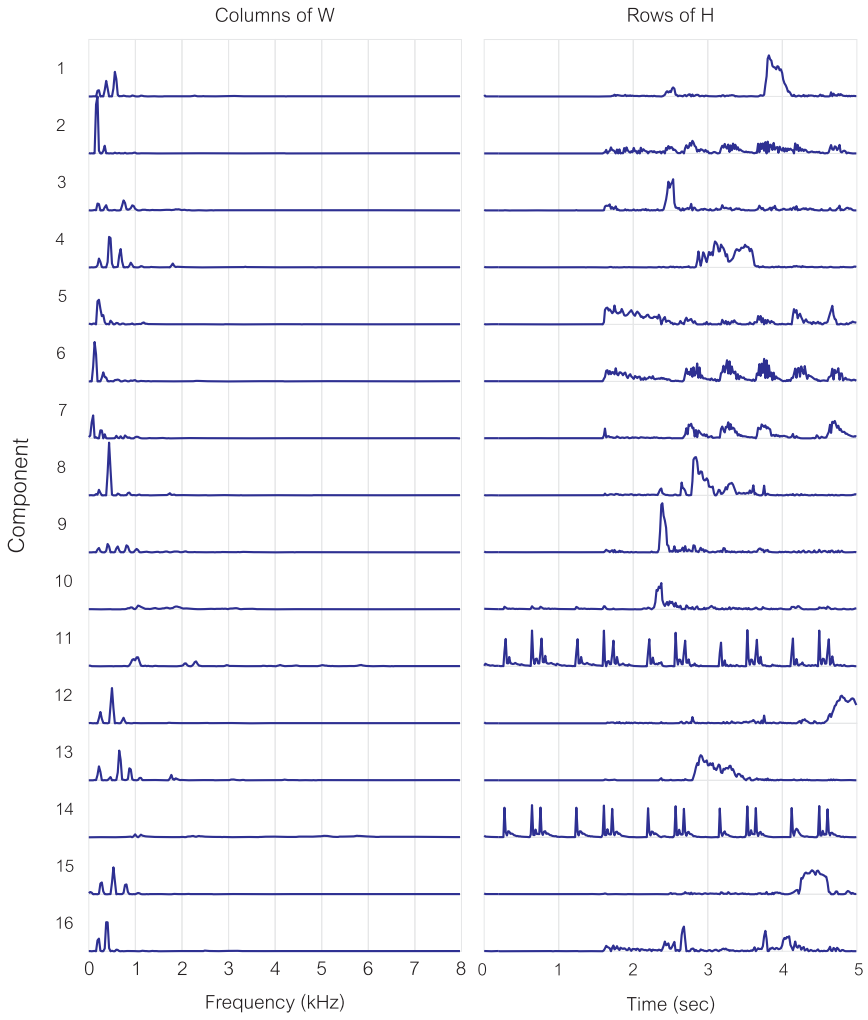
รูปที่ 3.9 ค่า PSNR เฉลี่ยของผลลัพธ์ที่ดีที่สุดจากเสียงผสมเมื่อมีการเปลี่ยนแปลงค่า R

จากรูปที่ 3.9 จะเห็นว่าค่า PSNR ของผลลัพธ์ที่ดีที่สุดของเสียงผสมใด ๆ จะมีค่าสูงขึ้นเมื่อ R เพิ่มมากขึ้น อย่างไรก็ตาม ในบางกรณี คือสำหรับข้อมูลเสียงร้องผสมกับเสียงเครื่องดนตรีให้ทำนองแบบทีละโน้ต และเครื่องดนตรีให้ทำนองแบบคอร์ด ค่า R ที่เพิ่มขึ้นกลับไม่ได้ช่วยให้ผลของการหาตัวประกอบดีขึ้นมากนัก ในขณะที่ค่า R ที่เพิ่มขึ้นจะยิ่งทำให้มีรูปแบบของการจัด

กลุ่มตัวประกอบที่เป็นไปได้มีค่าเพิ่มขึ้นแบบเอกซ์โพเนนเชียล ซึ่งอาจส่งผลต่อความถูกต้องในการเลือกตัวประกอบได้ ด้วยเหตุนี้ งานวิจัยนี้จึงเลือกใช้ค่า R เท่ากับ 16

3.4.2.2 การหาตัวประกอบ

ในขั้นตอนย่อยนี้จะเป็นการหาตัวประกอบด้วยวิธี NMF จากเมทริกซ์ข้อมูลเข้า V เมื่อกำหนดจำนวนตัวประกอบเท่ากับ R ที่ได้จากขั้นตอนที่ผ่านมา ซึ่งได้ผลลัพธ์จากการประมาณค่าเป็นเมทริกซ์ W หรือสเปกตรัมฐานหลัก (Basis Spectrum) จำนวน R สเปกตรัม และเมทริกซ์ H หรือสัมประสิทธิ์แสดงการกระจายตัวของสเปกตรัมฐานหลักแต่ละตัว ดังรูปที่ 3.10 เป็นตัวอย่างของการหาตัวประกอบของเสียงผสมจริงเมื่อ R เท่ากับ 16



รูปที่ 3.10 ผลการหาตัวประกอบด้วยวิธี NMF ของสัญญาณเสียงผสมตัวอย่าง

หลังจากที่ได้สเปกตรัมฐานหลัก และสัมประสิทธิ์แสดงการกระจายตัว ในเมทริกซ์ W และ H ตามลำดับ ในขั้นตอนต่อไปจะเป็นการเลือกตัวประกอบเมทริกซ์ของเสียงร้องจากผลลัพธ์ของขั้นตอนนี้

3.4.3 **การเลือกตัวประกอบเมทริกซ์** (Component Selection)

การเลือกตัวประกอบเมทริกซ์นั้น ถือเป็นขั้นตอนสำคัญของงานวิจัยนี้ และมีความท้าทายเป็นอย่างยิ่ง ในการที่จะคิดหาวิธีการเลือกตัวประกอบที่เป็นส่วนของเสียงร้อง นั่นคือ การเลือกเกณฑ์ที่จะใช้บอกว่า ตัวประกอบใด ๆ เป็นตัวประกอบของเสียงร้อง และการเปลี่ยนเกณฑ์นั้นจากนามธรรมมาให้อยู่ในรูปของการโปรแกรม

จากงานวิจัยที่ผ่านมา [21] ที่ใช้วิธี NMF ในการแยกเสียงดนตรี โดยมีการกำหนดจำนวนตัวประกอบเป็น 32 และ 64 ซึ่งทำให้มีวิธีการเลือกจัดกลุ่มเป็นจำนวนมาก โดยในขั้นตอนการเลือกตัวประกอบนี้ ยังไม่เป็นที่เปิดเผย อย่างไรก็ตาม งานวิจัยนี้พยายามออกแบบให้ขั้นตอนต่าง ๆ ในการแยกเสียงเป็นไปโดยอัตโนมัติ ซึ่งในขั้นตอนนี้ผู้วิจัยเห็นว่าสามารถใช้ความรู้ทางด้านเสียงและดนตรีเข้ามาช่วยในการเลือกตัวประกอบที่น่าจะเป็นส่วนของเสียงร้องได้

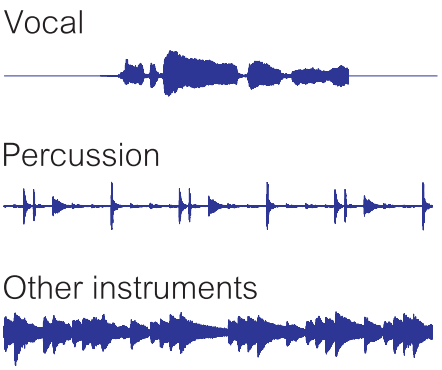
ผลลัพธ์จากการหาตัวประกอบในขั้นตอนที่ผ่านมา ได้แก่ เมทริกซ์ W ซึ่งแสดงรายละเอียดทางความถี่ และเมทริกซ์ H ซึ่งแสดงรายละเอียดทางเวลา โดยรายละเอียดทางความถี่สามารถใช้สำหรับการพิจารณาถึงรูปแบบของความถี่ที่เกิดขึ้นว่ามีลักษณะเป็นของเสียงร้องหรือเสียงดนตรี อย่างไรก็ตาม เครื่องดนตรีหลายชนิด สามารถผลิตเสียงที่มีความถี่ในย่านเดียวกับเสียงร้อง และอาจมีรูปแบบทางความถี่คล้ายกับเสียงร้องได้ ดังนั้น งานวิจัยนี้จึงเน้นการใช้รายละเอียดทางเวลาเป็นเกณฑ์ในการพิจารณาว่าตัวประกอบใด ๆ เป็นตัวประกอบของเสียงร้องหรือเสียงดนตรี

ลำดับของการทำงานในขั้นตอนนี้ประกอบด้วย การใช้เกณฑ์การเลือกตัวประกอบต่าง ๆ เพื่อกำหนดค่าคะแนนให้กับสเปกตรัมฐานหลักแต่ละตัว การรวมคะแนนจากเกณฑ์การเลือกตัวประกอบแต่ละแบบ และการเลือกตัวประกอบของเสียงร้องตามคะแนนที่ออกมา ดังรายละเอียดต่อไปนี้

3.4.3.1 **เกณฑ์การเลือกตัวประกอบ**

เมื่อพิจารณาถึงเสียงผสมต่าง ๆ เสียงเพลงส่วนใหญ่ มักประกอบด้วย เสียง 3 กลุ่ม ได้แก่ เสียงร้อง เสียงเครื่องดนตรีให้จังหวะ และเสียงเครื่องดนตรีให้ทำนอง ซึ่งมีตัวอย่าง

รูปคลื่นดังรูปที่ 3.11 เสียงเครื่องดนตรีให้จังหวะ จะมีการเล่นอย่างเป็นจังหวะ และเสียงเครื่องดนตรีให้ทำนอง มักจะมีการเล่นอย่างต่อเนื่อง งานวิจัยนี้จึงพยายามใส่เงื่อนไขให้กับความเป็นเสียงดนตรีทั้งสองประเภท คือ ความเป็นจังหวะ และความต่อเนื่องของเสียง หรือปริมาณที่ปรากฏของเสียงซึ่งจะมีมากกว่าของทั้งเสียงร้องและเสียงดนตรีประเภทให้จังหวะ



รูปที่ 3.11 ตัวอย่างคลื่นเสียงของเสียงร้อง เสียงเครื่องดนตรีประเภทให้จังหวะ และเสียงเครื่องดนตรีประเภทให้ทำนอง

จากเงื่อนไขทั้งสอง ขั้นตอนนี้จะเป็นการคิดค่าคะแนนให้กับแต่ละแถวของเมทริกซ์ H ซึ่งเป็นข้อมูลรายละเอียดทางเวลา เพื่อการนำไปใช้ประมวลผลในขั้นตอนต่อไปดังรายละเอียดต่อไปนี้

- เกณฑ์ความเป็นจังหวะของเสียงดนตรี

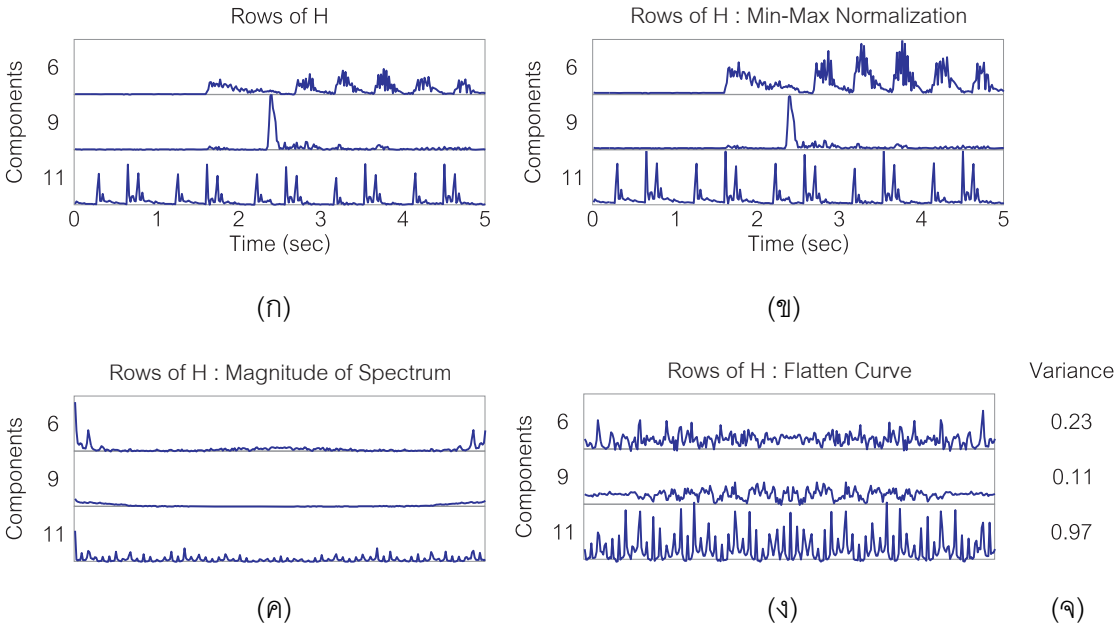
เครื่องดนตรีหลักในเพลงทั่ว ๆ ไป คือ เครื่องให้จังหวะ (Percussion) เช่น กลอง กลองชุด ซึ่งโดยปกติจะมีการเล่นอย่างเป็นจังหวะ โดยอาจยกเว้นในเพลงที่ร้องประกอบเครื่องดนตรีชนิดใดชนิดหนึ่งโดยเฉพาะ จึงจะไม่มีเครื่องดนตรีชนิดนี้ ซึ่งวิธีคิดค่าคะแนนในเกณฑ์นี้แสดงได้ดังรหัสเทียมในตารางที่ 3.2

ตารางที่ 3.2 รหัสเทียมของการคิดค่าคะแนนด้วยเกณฑ์ความเป็นจังหวะของเสียงดนตรี

```
Pseudocode Criterion#1 Rhythmic Event
01 Input : H (matrix of T by R)
02 Output : score1 (matrix of 1 by R)
03
04 For i = 1 to R,
05     Temp1 = H(i,:);              % Each row i of H
06     Temp1 = Temp1-min(Temp1);    % Set min to zero
07     Temp1 = Temp1./max(Temp1);   % Min_Max normalization
08     Temp1 = fft(Temp1);          % Apply Fourier transformation
09     Temp1 = sqrt(Temp1.*conj(Temp1)); % Calculate magnitude values
10     Temp2 = smooth(Temp1);       % Smooth curve of Temp1
```

```
11      Temp1 = Temp1./Temp2;        % Flatten Temp1
12      score1(i) = var(Temp1);      % Return variance of Temp1
13 End For
```

จากตารางที่ 3.2 การพิจารณาความเป็นจังหวะของตัวประกอบ `i` ใด ๆ จะพิจารณาในแถวที่ `i` ของเมทริกซ์ H โดยเริ่มจากการปรับค่าของแต่ละแถวให้อยู่ในระดับ 0 ถึง 1 (บรรทัดที่ 06-07) แล้วจึงใช้การแปลงแบบฟูเรียร์เข้ามาช่วย (บรรทัดที่ 08) ซึ่งจะได้สเปกตรัมของข้อมูลแต่ละแถวออกมา แล้วจึงคำนวณค่าขนาดของจำนวนเชิงซ้อน (Magnitude) (บรรทัดที่ 09) โดยใช้สมการที่ (3.4) ซึ่งหากตัวประกอบนั้นมีความเป็นจังหวะ สเปกตรัมที่ได้จะมีลักษณะเป็นคาบ อย่างไรก็ตาม ที่ความถี่ต่ำ ๆ ของสเปกตรัมจะมีแอมพลิจูดสูงกว่าที่ความถี่สูง ๆ ในขั้นต่อมา (บรรทัดที่ 10-11) จึงเป็นการปรับค่าสเปกตรัมนี้ด้วยการหารด้วยกราฟเดียวกันที่มีการปรับความเรียบแล้ว และผลคะแนนสำหรับเกณฑ์นี้คือ ค่าความแปรปรวน (Variance) ของสเปกตรัม ซึ่งหากมีค่าสูงจะแสดงว่าตัวประกอบนั้นมีความเป็นจังหวะสูงและมีโอกาสเป็นเสียงดนตรีสูง และเพื่อให้เห็นภาพชัดเจนยิ่งขึ้น รูปที่ 3.12 (ก) เป็นตัวอย่างแถวที่ 6 9 และ 11 ของเมทริกซ์ H ซึ่งพิจารณาแล้วว่าเป็นตัวประกอบของเสียงดนตรีแบบให้ทำนอง เสียงร้อง และเสียงดนตรีแบบให้จังหวะตามลำดับ ตัวอย่างการคิดค่าคะแนนด้วยเกณฑ์ความเป็นจังหวะของเสียงดนตรี ของตัวประกอบทั้งสามนี้ แสดงได้ดังรูปที่ 3.12 (ข) ถึง (จ) โดยจะเห็นว่าแถวที่ 11 ของเมทริกซ์ H ซึ่งมีความเป็นจังหวะมากที่สุดจะมีค่าความแปรปรวนสูงที่สุด



รูปที่ 3.12 ตัวอย่างการคิดค่าคะแนนด้วยเกณฑ์ความเป็นจังหวะของเสียงดนตรี
(ก) แถวที่ 6 9 11 ของเมทริกซ์ H (ข) การปรับค่าให้อยู่ในระดับ 0 ถึง 1
(ค) ค่าขนาดของสเปกตรัม (ง) สเปกตรัมที่ปรับเรียบแล้ว (จ) ค่าความแปรปรวนที่ได้

- เกณฑ์ความต่อเนื่องของเสียงดนตรี

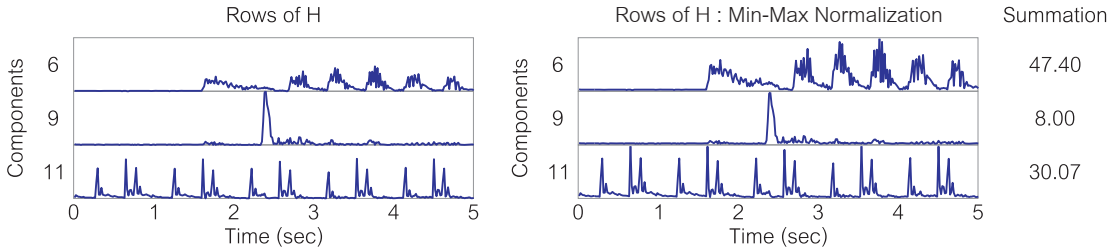
สำหรับความต่อเนื่องของเสียงดนตรี เครื่องดนตรีโดยทั่วไปนอกเหนือจากเครื่องให้จังหวะ มักได้รับการประพันธ์มาเพื่อให้บรรเลงไปอย่างต่อเนื่องในเพลง และเครื่องดนตรีบางชนิดยังมีการบรรเลงไปในเวลาเดียวกันอีกด้วย จึงส่งผลให้ตัวประกอบที่หาได้จากวิธี NMF ไม่สามารถแยกโน้ตแต่ละตัวของเครื่องดนตรีแต่ละชิ้นออกมาได้อย่างถูกต้อง อย่างไรก็ดี เสียงของเครื่องดนตรีเหล่านี้มักมีการรวมอยู่ในตัวประกอบที่แยกได้ และมีพฤติกรรมที่เกิดขึ้นอย่างต่อเนื่องในช่วงที่มีเสียงดนตรีชนิดเดียวกันบรรเลงอยู่ ดังนั้น งานวิจัยนี้จึงใช้เกณฑ์นี้สำหรับการเลือกตัวประกอบของเสียง

อย่างไรก็ตาม ความหมายของความต่อเนื่องของเสียงดนตรีตามที่ปรากฏในแถวของเมทริกซ์ H ต่างจากความต่อเนื่องในความหมายทางคณิตศาสตร์ โดยในที่นี้จะหมายถึง การปรากฏอยู่ตลอดทั้งท่อนของเพลงที่ทำการแยกเสียง ซึ่งวิธีการคิดค่าคะแนนแสดงได้ดังตารางที่ 3.3

ตารางที่ 3.3 รหัสเทียมของการคิดค่าคะแนนด้วยเกณฑ์ความต่อเนื่องของเสียงดนตรี

```
Pseudocode Criterion#2 Continuous Event
01 Input : H (matrix of T by R)
02 Output : score2 (matrix of 1 by R)
03
04 For i = 1 to R,
05     Temp1 = H(i,:);             % Each row i of H
06     Temp1 = Temp1-min(Temp1);   % Set min to zero
07     Temp1 = Temp1./max(Temp1);  % Min_Max normalization
08     score2(i) = sum(Temp1);     % Return summation of Temp1
09 End For
```

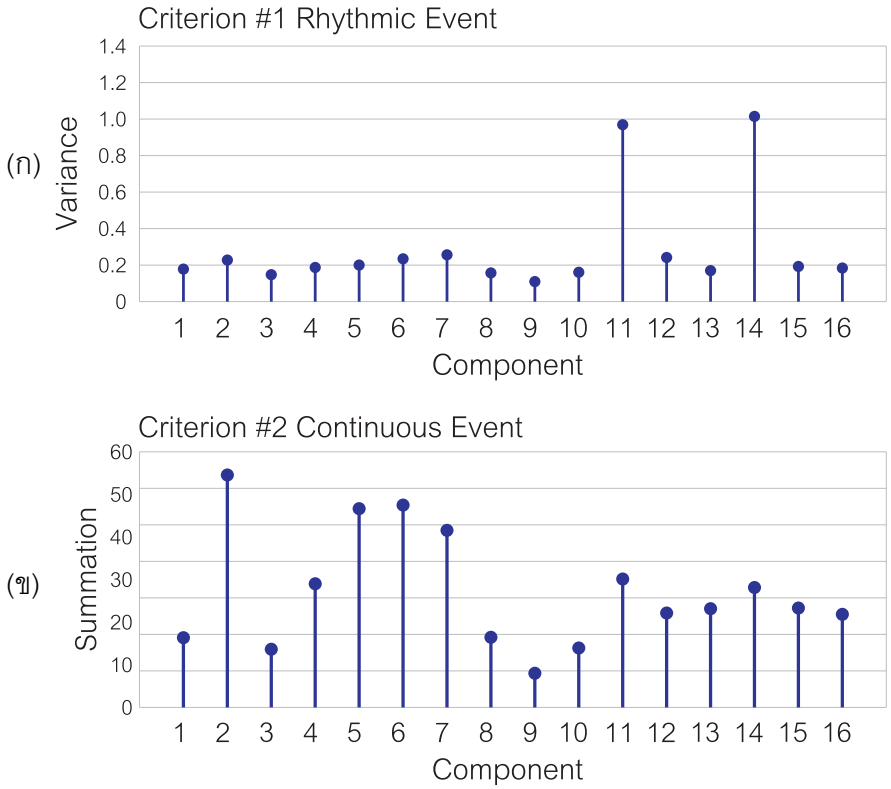
จากตารางที่ 3.3 การพิจารณาความต่อเนื่องของตัวประกอบ `i` ใด ๆ จะพิจารณาในแถวที่ `i` ของเมทริกซ์ H เฉพาะในส่วนของเสียงผสมที่ไม่มีเสียงร้องตอนต้น โดยในขั้นแรกจะทำการปรับค่าให้ตัวเลขอยู่ในช่วง 0 ถึง 1 (บรรทัดที่ 06-07) แล้วจึงคำนวณหาพื้นที่ใต้กราฟโดยการหาผลรวมของตัวเลขในแถวนั้น (บรรทัดที่ 08) ซึ่งหากค่าผลรวมนี้มีค่าสูงแสดงว่าตัวประกอบนั้นมีความต่อเนื่องสูงและมีโอกาสเป็นเสียงดนตรีสูง ตัวอย่างของการคิดค่าคะแนนด้วยเกณฑ์ความต่อเนื่องของเสียงดนตรี แสดงได้ดังรูปที่ 3.13



รูปที่ 3.13 ตัวอย่างการคิดค่าคะแนนด้วยเกณฑ์ความต่อเนื่องของเสียงดนตรี

(ก) แถวที่ 6 9 11 ของเมทริกซ์ H (ข) การปรับค่าให้อยู่ในช่วง 0 ถึง 1 (ค) ผลรวมของกราฟ

คะแนนของตัวประกอบทั้งหมด จากเกณฑ์การเลือกตัวประกอบทั้งสอง ซึ่งก็คือเกณฑ์ความเป็นจังหวะ และเกณฑ์ความต่อเนื่องของเสียงดนตรี แสดงได้ดังรูปที่ 3.14 โดยสำหรับการรวมคะแนนจะกล่าวถึงต่อไป



รูปที่ 3.14 คะแนนของตัวประกอบทั้งหมดที่คำนวณได้จากเกณฑ์ต่าง ๆ

(ก) เกณฑ์ความเป็นจังหวะของเสียงดนตรี (ข) เกณฑ์ความต่อเนื่องของเสียงดนตรี

3.4.3.2 การรวมคะแนนจากเกณฑ์การเลือกตัวประกอบ

ในขั้นตอนนี้ค่าคะแนนจากเกณฑ์การเลือกตัวประกอบทั้งสองในขั้นตอนที่ผ่านมาจะถูกนำมารวมกัน ซึ่งค่าคะแนนรวมของตัวประกอบใด ๆ หมายความถึงตัวประกอบนั้น ๆ มีความ

เป็นเสียงร้องหรือเสียงดนตรีมากน้อยเพียงใด ซึ่งจากขั้นตอนที่แล้ว คะแนนที่มีค่ามากหรือน้อยกว่าจะแสดงถึงโอกาสที่ตัวประกอบนั้นจะมีโอกาสเป็นเสียงดนตรีหรือเสียงร้องได้มากกว่า ตามลำดับ โดยวิธีการรวมค่าคะแนนแสดงได้ดังรหัสเทียมในตารางที่ 3.4

ตารางที่ 3.4 รหัสเทียมของการรวมคะแนนจากเกณฑ์การเลือกตัวประกอบ

```
Pseudocode Scoring
01 Input : score1 (matrix of 1 by R), score2 (matrix of 1 by R)
02 Output : scoretotal (matrix of 1 by R)
03
04 score1 = score1-min(score1);    % Set min to zero
05 score1 = score1./max(score1);   % Min_Max normalization
06 score2 = score2-min(score2);    % Set min to zero
07 score2 = score2./max(score2);   % Min_Max normalization
08 scoretotal = score1+score2;     % Sum scores
09 Return scoretotal;
```

จากตารางที่ 3.4 สำหรับเกณฑ์แต่ละตัวจะทำการหารด้วยค่าสูงสุดเพื่อปรับค่าให้อยู่ในระดับที่เท่ากันคืออยู่ในช่วง 0 และ 1 (บรรทัดที่ 04-07) และค่าคะแนนรวมของตัวประกอบแต่ละตัวคือผลรวมจากค่าเงื่อนไขทั้งสอง (บรรทัดที่ 08)

3.4.3.3 การเลือกตัวประกอบของเสียงร้อง

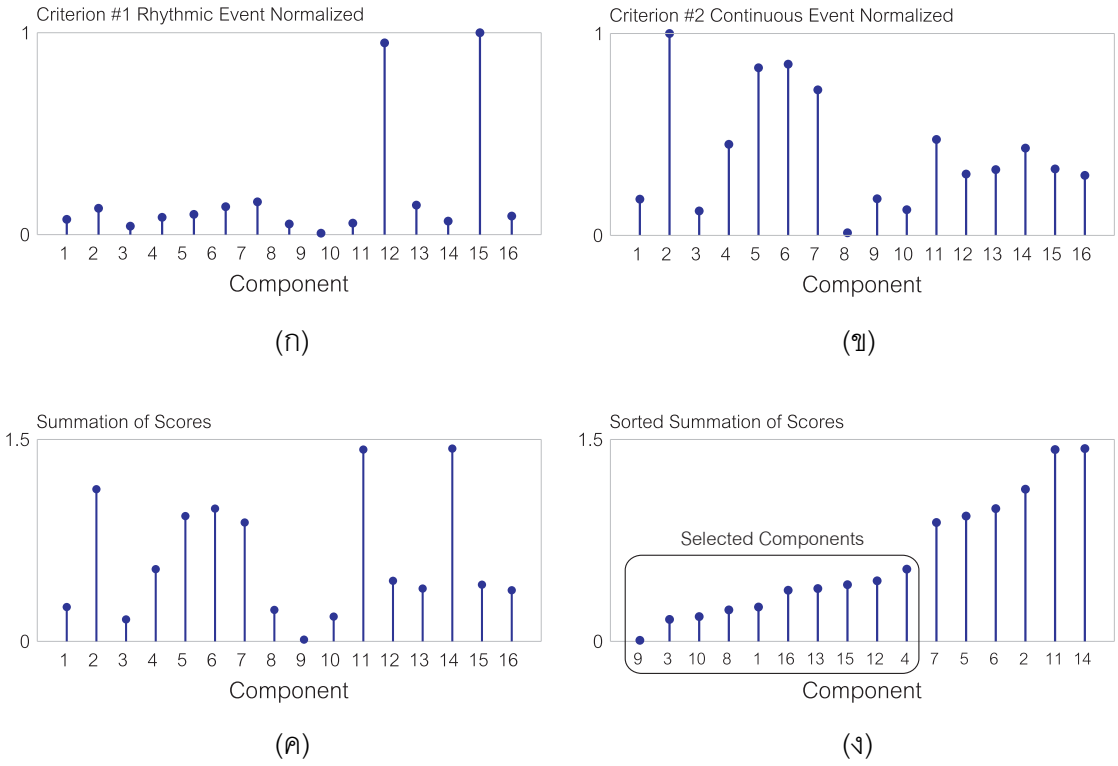
เมื่อได้ค่าคะแนนสำหรับตัวประกอบทั้งหมดแล้ว ในขั้นนี้จะเป็นการเรียงลำดับและแบ่งกลุ่มของตัวประกอบที่มีคะแนนต่ำและสูง โดยกลุ่มที่มีคะแนนต่ำจะถูกเลือกมาในฐานะที่เป็นตัวประกอบของเสียงร้องสำหรับประมวลผลในขั้นตอนต่อไป วิธีการเลือกตัวประกอบของเสียงร้องแสดงได้ดังรหัสเทียมในตารางที่ 3.5

ตารางที่ 3.5 รหัสเทียมของการเลือกตัวประกอบของเสียงร้อง

```
Pseudocode Component Selection
01 Input : scoretotal (matrix of 1 by R)
02 Output : components (matrix of 1 by number_of_selected_components)
03
04 For i = 1 to R,
05     scoretotal(2,i) = i;
06 End For
07 scoretotal = sort_by1st_row(scoretotal);
08
09 For i = 1 to R,
10     var1(i) = var(scoretotal(1,1:i));
11     var2(i) = var(scoretotal(1,i+1:R));
12 End For
13 var2(R) = 0;
14 vartotal = var1+var2;
15
16 [minv,ind] = min(vartotal);
```

```
17 components = [];
18 For i = 1 to indexMinVar,
19     components = [components,scoretotal(2,i)];
20 End For
21 Return components;
```

จากตารางที่ 3.5 จากคะแนนรวมที่ได้จากขั้นตอนก่อนหน้านี้ จะต้องมีการกำหนดค่าดัชนี (Index Number) ให้ก่อน (บรรทัดที่ 05) แล้วจึงเรียงลำดับคะแนนจากน้อยไปมาก (บรรทัดที่ 06) ต่อจากนั้นแบ่งกลุ่มตัวประกอบเป็นสองกลุ่มแล้วจึงคำนวณผลรวมของค่าความแปรปรวน (Variance) จากทั้งสองกลุ่ม (บรรทัดที่ 09-14) สองกลุ่มใดที่ให้ผลรวมค่าความแปรปรวนน้อยที่สุด จึงเลือกกลุ่มที่มีคะแนนต่ำกว่ามาเป็นตัวประกอบของเสียงร้อง (บรรทัดที่ 16-20) โดยสามารถแสดงได้ดังรูปที่ 3.15



รูปที่ 3.15 การเลือกตัวประกอบของเสียงร้อง

(ก) คะแนนจากเกณฑ์ความเป็นจังหวะของเสียงดนตรีที่ปรับให้อยู่ในช่วง 0 ถึง 1

(ข) คะแนนจากเกณฑ์ความต่อเนื่องของเสียงดนตรีที่ปรับให้อยู่ในช่วง 0 ถึง 1

(ค) ผลรวมคะแนนจากเกณฑ์ทั้งสอง (ง) คะแนนที่เรียงลำดับและตัวประกอบที่เลือกได้

หลังจากขั้นตอนนี้ตัวประกอบที่เลือกมาได้แก่ {1, 3, 4, 8, 9, 10, 12, 13, 15, 16} โดยผลลัพธ์จากขั้นตอนนี้กำหนดให้ W' และ H' แทนเมทริกซ์ของสเปกตรัมฐานหลัก และ

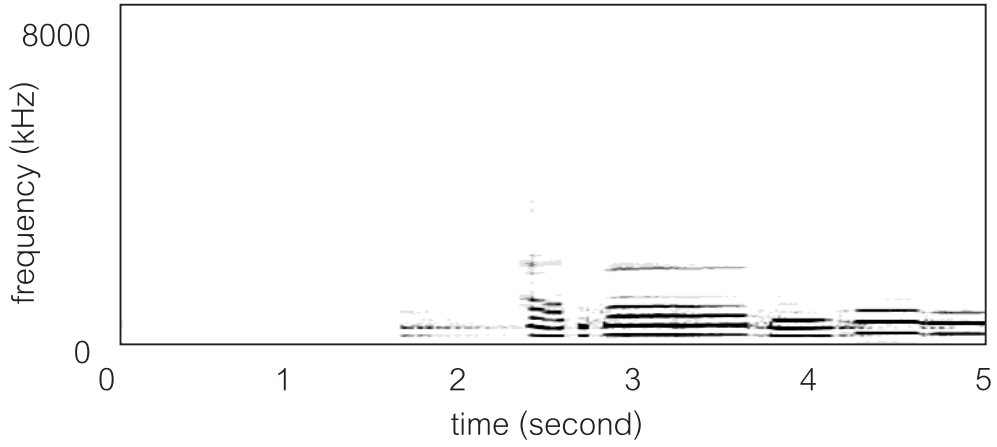
เมทริกซ์ของค่าสัมประสิทธิ์ของสเปกตรัมฐานหลักแต่ละตัวที่เลือกมา ตามลำดับ ในขั้นตอนต่อไปจะเป็นการนำตัวประกอบเหล่านี้ไปประมวลผลและสร้างสัญญาณเสียงร้องอันเป็นผลลัพธ์ออกมา

3.4.4 การประมวลผลหลัง

การประมวลผลหลังเป็นการสร้างสัญญาณเสียงร้องผลลัพธ์และ มีประมวลผลต่อเพื่อให้ได้เสียงที่ดีขึ้น โดยประกอบด้วยการสร้างสเปกโทรแกรมจากตัวประกอบที่เลือกในขั้นตอนที่ผ่านมา แล้วจึงสร้างกลับเป็นสัญญาณเสียง และหลังจากนั้นเป็นการลดสัญญาณเสียงเงียบ ดังรายละเอียดต่อไปนี้

3.4.4.1 การสร้างสเปกโทรแกรม

หลังจากที่ได้ตัวประกอบของเสียงร้องตามแล้ว ในส่วนนี้จะเป็นการคำนวณผลคูณของ W' และ H' แสดงได้ดังรูปที่ 3.16



รูปที่ 3.16 ผลคูณ $W'H'$ ของตัวประกอบของเสียงร้องที่เลือก

แม้ว่าสเปกโทรแกรมที่ได้จะสามารถนำมาสร้างเสียงกลับได้ทันทีโดยการแปลงแบบฟูเรียร์ไม่ต่อเนื่องผกผันหรือ IDFT (Inverse Discrete Fourier Transformation) อย่างไรก็ตาม เสียงที่ได้จะไม่เป็นธรรมชาติ เนื่องจากข้อมูลเชิงความถี่จะต้องประกอบด้วยส่วนสำคัญ 2 ส่วนคือ แอมพลิจูดและมุมเฟส ในขณะที่ผลคูณนั้นเป็นเพียงข้อมูลของแอมพลิจูดเท่านั้น ดังนั้น ในงานวิจัยนี้จึงใช้ผลคูณนั้นเป็นตัวกรองโดยกำหนดให้ $Filter = W'H'$ และนำตัวกรองหรือ $Filter$ นี้มาคูณกับสเปกโทรแกรมต้นต้นของเสียงผสมที่ได้จากขั้นตอนที่ 3.4.1.2 ซึ่งเป็นจำนวนเชิงซ้อน โดย ณ ตำแหน่ง (f,t) ใด ๆ ของสเปกโทรแกรมใหม่ คำนวณได้ดังสมการ (3.5)

$$NewSpectrum(f,t) = \frac{Filter(f,t) * Spectrum(f,t)}{\left|Spectrum(f,t)\right|} \tag{3.5}$$

เมื่อ $1 \le f \le F$ และ $1 \le t \le T$

3.4.4.2 การแปลงแบบฟูเรียร์ผกผัน และการสร้างสัญญาณเสียงกลับ

ในการแปลงสเปกโทรแกรมที่ผ่านการกรองในขั้นตอนก่อนหน้านี้ เป็นคลื่นเสียงซึ่งเป็นฟังก์ชันจำนวนจริง ด้วยการแปลงแบบฟูเรียร์ จะต้องทำให้รายละเอียดทางความถี่เหล่านี้เป็นฟังก์ชันเฮอร์มิเชียนก่อน สำหรับข้อมูลสเปกตรัมแต่ละเฟรม ซึ่งมีความยาว F จะสามารถสร้างสัญญาณเสียงขนาด $2(F-1)$ จุดข้อมูล โดยเริ่มจากการเติมอีกครึ่งหนึ่งของสเปกตรัมนั้น ๆ ด้วยค่าสังยุค (Conjugate) ของจำนวนเชิงซ้อนดังสมการที่ (3.6) สำหรับ

$$S(j) = \begin{cases} \overline{Spectrum(2F-j)} & ; j > F \\ Spectrum(j) & ; j \le F \end{cases} \tag{3.6}$$

เมื่อ $1 \le j \le 2(F-1)$ และ $\overline{a+bi} = a-bi$ เป็นค่าสังยุค (Conjugate) ของจำนวนเชิงซ้อน

แล้วจึงแปลงสเปกตรัม S ใด ๆ ด้วยการแปลงแบบฟูเรียร์ผกผัน ซึ่งจะได้คลื่นเสียงของแต่ละเฟรม และนำมาต่อกันตามความซ้อนเหลื่อมที่ได้กำหนดไว้ในเบื้องต้นนั้นคือ 16 มิลลิวินาที ซึ่งผลลัพธ์จะได้เป็นสัญญาณเสียงออกมา อย่างไรก็ตาม ในบางส่วนที่สัญญาณมีค่าแอมพลิจูดที่ต่ำมากอาจส่งผลต่อคุณภาพเสียงได้ในลักษณะของสัญญาณรบกวน ดังนั้นจึงต้องมีการปรับเสียงที่ไม่ได้ยินให้เป็นเสียงเงียบ

3.4.4.3 การปรับเสียงที่ไม่ได้ยินให้เป็นเสียงเงียบ

จากที่กล่าวถึงในบทที่ 2 เสียงเงียบคือเสียงที่มีแอมพลิจูดของเสียงต่ำกว่า 0 เดซิเบล สำหรับความถี่อย่างไรก็ตาม จากรูปที่ 2.2 ระดับของเสียงที่มนุษย์จะได้ยินที่ความถี่ต่าง ๆ ซึ่งมีค่าต่ำสุดที่ประมาณ –5 เดซิเบล ในงานวิจัยนี้จึงปรับให้สัญญาณในเฟรมที่มีแอมพลิจูดของเสียงต่ำกว่า –20 เดซิเบล เป็นศูนย์ โดยแอมพลิจูดหรือระดับความดังของเสียง (Volume) ในหน่วยเดซิเบลของเสียงในเฟรมใด ๆ สามารถคำนวณได้ดังสมการ (3.7) [27]

$$Volume = 10 \times \log_{10}\left(\sum_{i=1}^{n}\left(s_i^{\,2}\right)\right) \tag{3.7}$$

โดยที่ s_i คือค่าสัญญาณเสียงในจุดข้อมูลที่ i และ n คือจำนวนจุดข้อมูลในเฟรมนั้น ๆ

บทนี้ได้กล่าวถึงขั้นตอนการดำเนินงานวิจัย จนกระทั่งได้ขั้นตอนวิธีซึ่งสามารถแยกเสียงร้องออกจากเสียงเพลงได้ สำหรับบทต่อไปจะเป็นการนำเอาวิธีการที่นำเสนอไปประเมินผล และทดสอบกับเสียงเพลงจริง และเปรียบเทียบผลกับวิธีการที่ผ่านมา

บทที่ 4

การทดลองและวิเคราะห์ผลการทดลอง

ในบทนี้จะเป็นการทดสอบขั้นตอนวิธีการแยกเสียงร้องออกจากเสียงเพลงที่นำเสนอ ซึ่งก่อนจะกล่าวถึงการทดลอง ผู้วิจัยจะเริ่มต้นด้วยการกล่าวถึงมาตรวัดที่ใช้ในงานวิจัยนี้ คู่แข่งขันที่งานวิจัยนี้จะใช้เปรียบเทียบ แล้วจึงเป็นการทดลองร่วมกับการวิเคราะห์ผลการทดลองและข้อสรุปจากการทดลองดังรายละเอียดต่อไปนี้

4.1 มาตรวัด

งานวิจัยทางด้านการแยกเสียงโดยส่วนใหญ่ มักนำเสนอผลการแยกเสียงในลักษณะรูปภาพของคลื่นเสียง [9, 28] หรือการให้ฟังตัวอย่าง [28] ซึ่งน้อยงานที่จะประเมินผลการวิจัยในเชิงปริมาณระหว่างเสียงต้นฉบับและเสียงที่แยกได้ [21] อย่างไรก็ตาม งานวิจัยนี้ต้องการที่จะให้มีการประเมินผลในเชิงปริมาณร่วมด้วย โดยในการวัดประเมินผลความถูกต้องของการแยกเสียงนั้น จำเป็นต้องใช้เพลงที่มีเสียงร้องบริสุทธิ์ เพื่อให้สามารถทำการเปรียบเทียบความคล้ายกันระหว่างเสียงร้องที่แยกได้และเสียงร้องต้นฉบับ โดยในงานวิจัยนี้จะเปรียบเทียบความคล้ายกัน โดยเลือกใช้มาตรวัดคือ อัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (Peak Signal-to-Noise Ratio หรือ PSNR) และนอกจากนี้เพื่อเป็นการยืนยันถึงความสามารถในการนำวิธีการที่นำเสนอไปใช้ในงานวิจัยด้านอื่นต่อได้ เช่น งานวิจัยด้านการค้นคืนเสียงเพลงด้วยการร้องทำนอง การเปรียบเทียบความถูกต้องของคอนทัวร์ระดับเสียง (Pitch Contour) ที่คำนวณได้จากเสียงร้องที่แยกได้ จึงถูกเลือกมาใช้เป็นมาตรวัดอีกตัวหนึ่งด้วย ดังรายละเอียดต่อไปนี้

4.1.1 อัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (PSNR)

อัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (Peak Signal-to-Noise Ratio หรือ PSNR) [26] เป็นมาตรวัดแบบหนึ่งซึ่งมีพื้นฐานมาจากการคิดค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error หรือ MSE) เช่นเดียวกับอัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (Signal-to-Noise Ratio หรือ SNR) [29] ซึ่งเป็นมาตรวัดสำหรับการเปรียบเทียบสัญญาณเสียงแบบดั้งเดิม โดย SNR และ PSNR จะมีค่าในหน่วยเดซิเบล ซึ่งคำนวณได้จากสมการ (4.1) และ (4.2) ตามลำดับ

$$SNR = 20\log_{10}\left(\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} s_O(i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_O(i)-s_I(i))^2}}\right) \quad (4.1)$$

$$PSNR = 20\log_{10}\left(\frac{\sqrt{\max_{i=1}^{n}(s_O(i)^2)}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_O(i)-s_I(i))^2}}\right) \quad (4.2)$$

โดยที่ s_O และ s_I คือ สัญญาณเสียงที่ได้ และสัญญาณเสียงต้นฉบับ ตามลำดับ

จากสมการทั้งสอง จะเห็นว่าจุดที่แตกต่างระหว่าง SNR และ PSNR คือ อัตราส่วนของสัญญาณซึ่ง SNR จะได้มาจากค่าเฉลี่ยยกกำลังสองของสัญญาณที่ได้ และ PSNR ได้มาจากค่าสูงสุดของสัญญาณ ดังนั้นแนวโน้มของการเปรียบเทียบด้วย SNR และ PSNR จึงเป็นไปในทิศทางเดียวกัน แต่มีค่าที่แตกต่างกัน

อย่างไรก็ตาม แม้ SNR จะเป็นมาตรวัดดั้งเดิมที่ใช้ในการเปรียบเทียบสัญญาณเสียง แต่งานวิจัยบางส่วนพยายามหลีกเลี่ยงการเปรียบเทียบ SNR ระหว่างสัญญาณเสียงทั้งสอง [28] เนื่องจากการใช้ SNR แบบดั้งเดิมนั้นเป็นการเปรียบเทียบสัญญาณเสียงโดยตรง ในขณะที่สัญญาณเสียงส่วนใหญ่มีแอมพลิจูดไม่สม่ำเสมอ จึงทำให้มีข้อจำกัดต่อการนำไปใช้เพื่อให้ได้ผลการคำนวณตรงตามความต้องการให้มากที่สุด การใช้ SNR โดยส่วนใหญ่จึงมักนำไปอธิบายถึงคุณภาพของเสียงที่ได้จากอุปกรณ์ต่าง ๆ เช่น ชุดเครื่องเสียง ลำโพง หรือเครื่องเล่นซีดี หรือคุณภาพของการส่งสัญญาณผ่านเครือข่ายสื่อสาร และเนื่องจากการคำนวณ SNR จะต้องใช้ค่าเฉลี่ยของสัญญาณ มาตรฐานในการคำนวณ SNR จึงใช้สัญญาณไซน์ซึ่งมีแอมพลิจูดที่คงที่ในการคำนวณ [29]

สำหรับค่า PSNR มักนำมาใช้ในการเปรียบเทียบสัญญาณระหว่างสัญญาณต้นแบบและสัญญาณอันเกิดจากการบีบอัดแบบสูญเสียข้อมูล (Lossy Compression) ซึ่งเป็นที่นิยมในงานวิจัยด้านการบีบอัดภาพ (Image Compression) และเนื่องจาก PSNR ใช้สัญญาณสูงสุดที่เป็นไปได้ จึงเหมาะกับการนำมาเปรียบเทียบคุณภาพของเสียงใด ๆ ที่สร้างขึ้นใหม่และอาจมีการสูญเสียข้อมูล เพื่อบอกระดับความถูกต้องหรือความคล้ายกันของผลลัพธ์ได้

จากที่กล่าวมาข้างต้น งานวิจัยนี้จึงเลือกใช้ PSNR ในการเปรียบเทียบระหว่างเสียงร้องต้นฉบับและเสียงร้องที่แยกได้ อย่างไรก็ตาม รูปแบบของข้อมูลที่ใช้ในการเปรียบเทียบที่

เป็นคลื่นเสียงในเชิงโดเมนเวลานั้นยังไม่เหมาะสมนักต่องานวิจัยที่มีลักษณะของการเข้ารหัส และถอดรหัสของสัญญาณ หรือมีการแปลงสัญญาณภายในขั้นตอนวิธีการประมวลผล เนื่องจากจะส่งผลให้สัญญาณมีการเปลี่ยนเฟสได้

จากที่ได้กล่าวไว้แล้วในบทที่สองถึงการแทนข้อมูลของเสียง ซึ่งมีสองชนิดคือ แบบคลื่นเสียง และแบบสเปกตรัม โดยการเปรียบเทียบโดยใช้รูปแบบของคลื่นเสียงโดยตรงแบบที่ใช้คำนวณ SNR แบบดั้งเดิมนั้น เสียงโดยทั่วไปอาจมีการเปลี่ยนแปลงข้อมูลทางเฟส ส่งผลให้การคิดผลต่างระหว่างข้อมูลแบบคลื่นเสียงโดยตรงมีค่าสูง และทำให้ค่า SNR หรือ PSNR มีค่าต่ำในขณะที่เสียงที่ได้ยินไม่แตกต่างกัน

การแทนข้อมูลของคลื่นเสียงอีกชนิดหนึ่งคือ สเปกตรัม ซึ่งเป็นข้อมูลที่สามารถแปลงไปกลับกับสัญญาณเสียงได้ด้วยการแปลงแบบฟูเรียร์ เพื่อหลีกเลี่ยงกับปัญหาของการแทนข้อมูลแบบคลื่นเสียงดังกล่าว นั่นคือ รูปแบบการแทนข้อมูลในโดเมนความถี่นี้จะไม่ไวต่อการเปลี่ยนเฟสหรือการเลื่อนทางเวลา อย่างไรก็ตาม จากการทดลองย่อย พบว่าการคำนวณ PSNR จากข้อมูลสเปกตรัมนี้ ค่อนข้างไวต่อคลื่นเสียงที่มีช่วงเสียงเงียบ นั่นคือ ถ้าหากมีการกรองเสียงในช่วงเวลาที่ไม่ต้องการออกไป คือปรับค่าในทางโดเมนเวลาให้เป็นศูนย์หรือเสียงเงียบ ค่า PSNR ที่คำนวณได้จะตำลงทันที ซึ่งไม่สอดคล้องกับผลลัพธ์ที่ควรจะเป็นนั่นคือค่า PSNR ที่ได้ควรจะสูงขึ้น

อย่างไรก็ตาม ข้อมูลทั้งสองลักษณะนี้ต่างมีข้อดี กล่าวคือ ข้อมูลแบบคลื่นเสียงในโดเมนเวลาจะให้ค่าที่ดี เมื่อไม่มีปัญหาการเปลี่ยนเฟสหรือการเลื่อนทางเวลา ในขณะที่ข้อมูลแบบสเปกตรัมในโดเมนความถี่จะช่วยกำจัดปัญหาที่เกิดขึ้นกับข้อมูลทางเวลาได้ ดังนั้นในงานวิจัยนี้จึงต้องการใช้ประโยชน์ของการแทนข้อมูลทั้งสองแบบในการนำมาใช้เพื่อเปรียบเทียบเสียง โดยใช้วิธีการแทนข้อมูลแบบเวลา-ความถี่ หรือสเปกโทรแกรมในการเปรียบเทียบเสียง ซึ่งเป็นการคำนวณความถี่ในช่วงเวลาสั้น ๆ และเป็นรูปแบบข้อมูลชนิดเดียวกับที่ใช้ในการประมวลผลในขั้นตอนวิธีการแยกเสียง

กล่าวโดยสรุป ในงานวิจัยนี้เลือกใช้การวัดผลแบบ PSNR กับข้อมูลเสียงที่มีการแทนข้อมูลในรูปแบบของสเปกโทรแกรม ซึ่งจากสูตรคำนวณ PSNR ในสมการที่ (4.2) สามารถนำมาเขียนใหม่ได้ เมื่อ $Spectrogram_1$ และ $Spectrogram_2$ เป็นสเปกโทรแกรมขนาด $F \times T$ ของสัญญาณเสียงร้องที่แยกได้ และเสียงร้องต้นฉบับ ตามลำดับ PSNR ในหน่วยเดซิเบลของสเปกโทรแกรมทั้งสองสามารถคำนวณได้ดังสมการที่ (4.3) ซึ่งผลลัพธ์การเปรียบเทียบหากเสียงทั้งสองมีความใกล้เคียงกันมาก PSNR จะมีค่าสูง

$$PSNR = 20\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{4.3}$$

โดยที่ MAX_I คือค่าสูงสุดของสเปกโทรแกรมทั้งสอง และค่า MSE คือค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) ดังสมการที่ (4.4)

$$MSE = \frac{1}{FT}\sum_{i=1}^{F}\sum_{j=1}^{T}\left(Spectrogram_1(i,j) - Spectrogram_2(i,j)\right)^2 \tag{4.4}$$

หมายเหตุ จากการทดลอง ค่า PSNR ที่ได้โดยทั่วไปจะอยู่ในช่วง 30 ถึง 60 เดซิเบล โดยขึ้นกับปัจจัยต่าง ๆ เช่น ความยาวของสัญญาณเสียงที่นำมาเปรียบเทียบ กล่าวคือ ถ้าสัญญาณเสียงมีความยาวมากกว่า ค่า MSE ที่คำนวณได้ก็จะมีโอกาสสูงขึ้น ทำให้ค่า PSNR มีค่าต่ำกว่าเสียงที่มีความยาวน้อยกว่า สำหรับเสียงที่มีความยาวประมาณ 10 วินาที ค่า PSNR ประมาณ 38-39 เดซิเบล ขึ้นไปจะให้เสียงที่มีคุณภาพในระดับที่น่าพอใจ

4.1.2 การเปรียบเทียบความถูกต้องของคอนทัวร์ระดับเสียง

นอกจากค่า PSNR ซึ่งเป็นการเปรียบเทียบความคล้ายระหว่างเสียงร้องที่แยกได้และเสียงร้องต้นฉบับแล้ว ในงานวิจัยนี้ต้องการมาตรวัดอื่น เพื่อสื่อถึงความเป็นไปได้ของการนำวิธีการแยกเสียงที่นำเสนอไปใช้ประโยชน์กับงานวิจัยด้านอื่น ๆ ต่อไปได้จริง โดยเน้นไปที่งานวิจัยทางด้านการค้นคืนเพลงด้วยการร้องทำนอง ซึ่งต้องการคอนทัวร์ระดับเสียง (Pitch Contour) ของเสียงร้องในเพลงไปใช้ในการค้นคืน ในงานวิจัยนี้ จึงมีการสกัดคอนทัวร์ระดับเสียงจากเสียงร้องเพื่อนำมาเปรียบเทียบกัน

ในขั้นตอนการสกัดคอนทัวร์ระดับเสียงนั้น งานวิจัยนี้ได้เลือกใช้โปรแกรม Praat [30] ซึ่งเป็นโปรแกรมที่นิยมในการประมวลผลทางเสียง และสามารถเขียนคำสั่งการตรวจหาระดับเสียงของสัญญาณได้ง่าย และเป็นอัตโนมัติ ซึ่งผลการสกัดคอนทัวร์ระดับเสียงจะได้เป็นความถี่มูลฐานของเสียงร้องในหน่วยของเฮิรทซ์

หลังจากนั้น ในขั้นตอนการเปรียบเทียบคอนทัวร์ระดับเสียง เพื่อให้สามารถวัดความถูกต้องเป็นเปอร์เซ็นต์ได้ จึงมีการแจงหน่วย (Quantization) ของความถี่ให้อยู่ในรูปแบบของโน้ตที่เป็นเลขจำนวนเต็มด้วยสมการ (4.5) [31] แล้วจึงนำมาเปรียบเทียบความถูกต้องของโน้ตที่ได้ด้วยสมการ (4.6) โดยตัวอย่างของการเปรียบเทียบ แสดงได้ดังรูปที่ 4.1

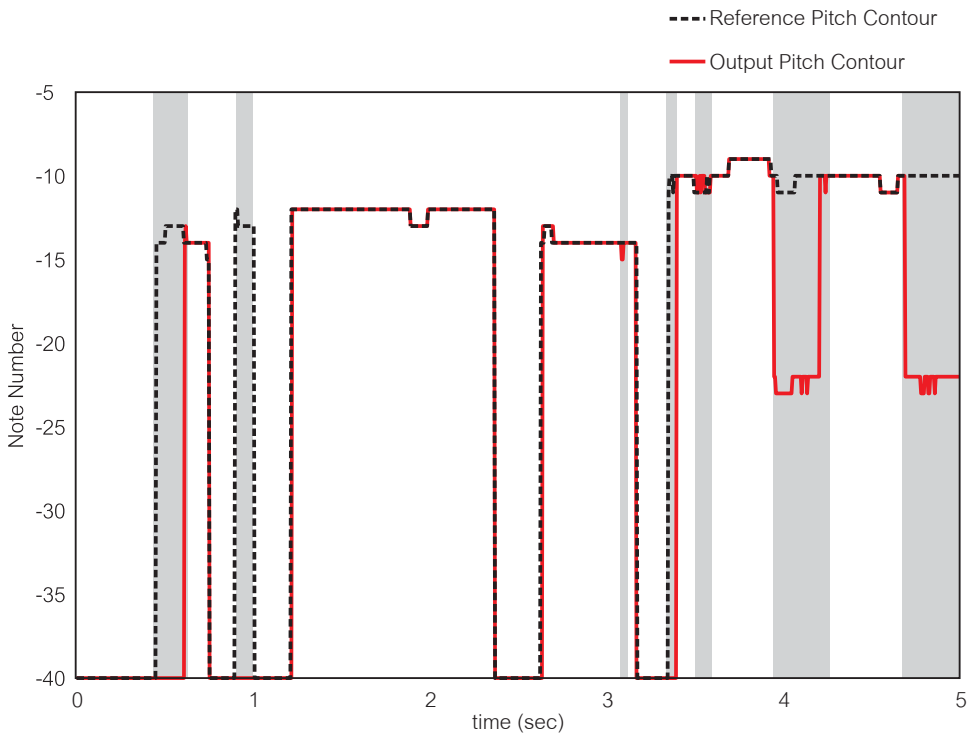
$$n_i = round\left(12\log_2\frac{pitch_i}{440}\right) \tag{4.5}$$

เมื่อ $pitch_i$ คือ ค่าระดับเสียงในหน่วยเฮิรทซ์ ทีเฟรม i ของเสียงที่คำนวณได้จากโปรแกรม Praat

$$Accuracy = \frac{\sum_{i=1}^{N} similarity(i)}{N} \times 100 \tag{4.6}$$

$$similarity(i) = \begin{cases} 1 & ; n_i = m_i \\ 0 & ; n_i \neq m_i \end{cases} \tag{4.7}$$

เมื่อ n_i และ m_i คือโน้ตของเสียงร้องที่แยกได้และโน้ตของเสียงร้องต้นฉบับ ตามลำดับ และ $1 \leq i \leq N$



รูปที่ 4.1 การคำนวณความถูกต้องของคอนทัวร์ระดับเสียงของเสียงร้องที่แยกได้ ส่วนที่แรเงาคือส่วนที่คอนทัวร์ระดับเสียงมีความแตกต่างกัน

4.2 คู่แข่งขัน

ในการเปรียบเทียบความถูกต้องของการแยกเสียงร้องออกจากเสียงเพลง งานวิจัยนี้เลือกเปรียบเทียบกับวิธีการในกลุ่มของ CASA ซึ่งเป็นวิธีการที่ดีในการแยกเสียงร้องออกจากเสียงเพลงวิธีหนึ่ง โดยเลือกใช้วิธีการลดสัญญาณรบกวน (Noise Removal) ของโปรแกรม Audacity® [23]

จากที่ได้กล่าวไว้ในบทที่ 2 ถึงวิธีการลดสัญญาณรบกวนแล้ว ว่าวิธีนี้คือการพิจารณาเสียงเพลงในรูปของโดเมนความถี่ และใช้ตัวช่วยในการเลือกความถี่ โดยตัวช่วยในที่นี้คือการเก็บข้อมูลสัญญาณรบกวน ในรูปแบบของสเปกตรัมความถี่ต่าง ๆ แล้วจึงนำไปทำการกรองที่เรียกว่าการทำประตูสเปกตรัมสัญญาณรบกวน (Noise Gating) หรือการกรองสัญญาณแบบรอยบาก (Notch Filtering) คือการกำหนดความถี่ที่ไม่สามารถผ่านไปได้ ซึ่งจากการศึกษาวิธีการนี้พบว่าเป็นวิธีที่มีความยืดหยุ่นกว่าวิธีอื่น ๆ ในกลุ่มเดียวกัน กล่าวคือ สามารถทำงานได้กับรูปแบบของเพลงทั่วไปได้ ซึ่งต่างจากวิธีการในกลุ่มของ CASA วิธีอื่น ซึ่งค่อนข้างจำกัดสำหรับใช้กับแนวเพลงบางประเภทเท่านั้น [28] อีกทั้งยังเป็นโปรแกรมประเภทฟรีแวร์ (Freeware) และมีการเปิดเผยอัลกอริทึมซึ่งทำให้ผู้วิจัยสามารถทำความเข้าใจถึงกลไกการทำงานได้

อย่างไรก็ตาม การที่วิธีนี้ยังไม่ได้ถูกนำมาใช้ในงานวิจัยต่าง ๆ ต่อนั้น เนื่องมาจากวิธีการนี้ยังคงมีความแปรผันสูง โดยขึ้นกับสัญญาณรบกวน ซึ่งในที่นี้จะใช้เสียงดนตรี ที่ผู้ใช้ต้องเป็นผู้เลือก ให้ตรงกับช่วงความถี่ที่อยู่ในสัญญาณเสียงผสมให้มากที่สุด โดยสำหรับงานวิจัยนี้ผู้วิจัยจะพยายามเลือกกรณีตัวแทนของเสียงดนตรีที่เหมาะสมกับวิธีการของโปรแกรม Audacity® ในการเปรียบเทียบด้วย

สำหรับเนื้อหาในส่วนต่อเป็นจะเป็นการกล่าวถึงการทดลองในงานวิจัยนี้ โดยการนำเอาวิธีการของคู่แข่งขันไปเปรียบเทียบกับวิธีการแยกเสียงร้องที่นำเสนอ และใช้ตัววัดคือ PSNR และการเปรียบเทียบคอนทัวร์ระดับเสียง

4.3 การทดลองและวิเคราะห์ผลการทดลอง

จากการที่ได้ศึกษาประเภทของเครื่องดนตรีตามรูปแบบของความถี่ พบว่าเครื่องดนตรีสามารถแบ่งได้ตามจำนวนความเป็นไปได้ในการสร้างความถี่เสียงได้เป็น 3 ประเภท คือ

- ประเภทที่ไม่มีการเปลี่ยนแปลงระดับเสียง เช่น กลอง กลองชุด
- ประเภทที่มีการเปลี่ยนแปลงระดับเสียงแบบเล่นทีละโน้ต เช่น ฟลุท ทรัมเปต
- ประเภทที่มีการเปลี่ยนแปลงระดับเสียงแบบเล่นเป็นคอร์ด เช่น เปียโน กีตาร์

งานวิจัยนี้จึงทำการทดลองเพื่อประเมินผลการแยกเสียง โดยใช้ข้อมูลเสียงผสมเป็นเสียงร้องท่อนเดียวกันของผู้หญิงและผู้ชาย ผสมกับเสียงดนตรีหลากหลายชนิดตามแต่ละประเภทข้างต้น ซึ่งนอกจากจะนำมาพิจารณาโดยรวมของเครื่องดนตรีแต่ละประเภทแล้ว ยังต้องการนำมาพิจารณาในแง่มุมอื่น ๆ อีก โดยที่

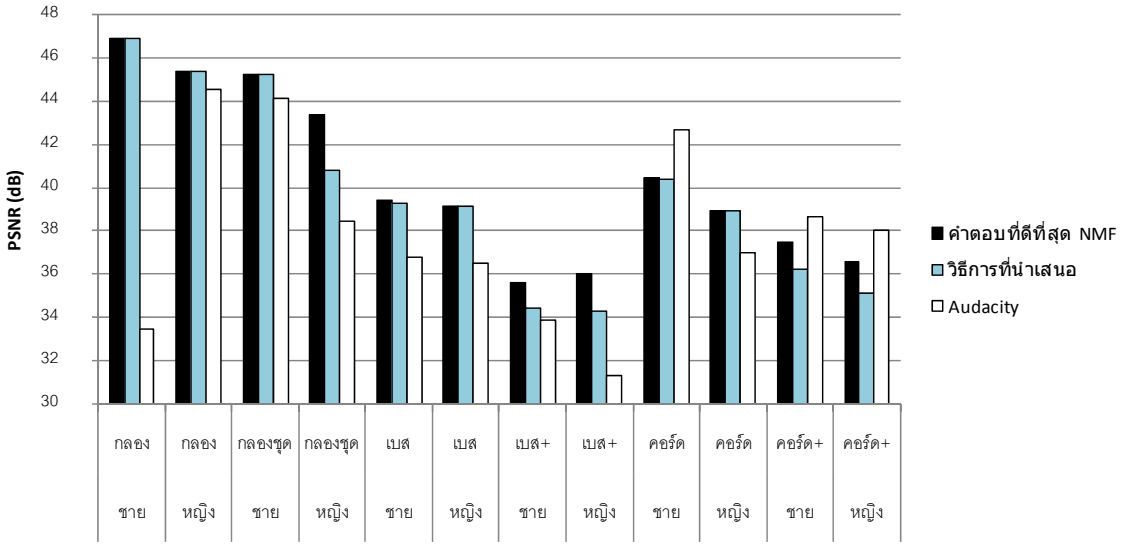
- ประเภทแรก ประกอบด้วย เครื่องเคาะจังหวะ (ใช้สัญลักษณ์ในการทดลองว่า กลอง) และกลองชุด เพื่อพิจารณาถึงผลของจำนวนรูปแบบของเสียงที่มีต่อความถูกต้องของการแยกเสียง
- ประเภทที่สอง ประกอบด้วย เบสที่เล่นด้วยเสียงต่ำ (ใช้สัญลักษณ์ในการทดลองว่า เบส) และเครื่องดนตรีชนิดเดิมที่เพิ่มความถี่เสียงให้สูงขึ้นเป็นสองเท่า (Octave) (ใช้สัญลักษณ์ในการทดลองว่า เบส+) เพื่อพิจารณาถึงผลของระดับเสียงของเครื่องดนตรีที่มีต่อความถูกต้องของการแยกเสียง
- ประเภทที่สาม ประกอบด้วย เสียงเปียโนที่เล่นเป็นคอร์ดและมีความหลากหลายของโน้ตน้อย (ใช้สัญลักษณ์ในการทดลองว่า คอร์ด) และเสียงเปียโนที่เล่นเป็นคอร์ดและมีความหลากหลายของโน้ตสูง (ใช้สัญลักษณ์ในการทดลองว่า คอร์ด+) เพื่อพิจารณาถึงผลของจำนวนโน้ตที่มีต่อความถูกต้องของการแยกเสียง

สำหรับการทดลองแรก สืบเนื่องจากการทดลองเบื้องต้นในบทที่ 3 หัวข้อ 3.4.2.1 งานวิจัยนี้จะเริ่มต้นที่การผสมเสียงร้องและเสียงเครื่องดนตรีแต่ละชนิด โดยประเมินวิธีการเลือกตัวประกอบที่นำเสนอในงานวิจัยนี้ เปรียบเทียบกับรูปแบบที่ดีที่สุดของการหาตัวประกอบด้วยวิธี NMF และวิธีการลดสัญญาณรบกวนของโปรแกรม Audacity® ซึ่งจะต้องมีการเรียนรู้เสียงดนตรีเพื่อให้วิธีที่ต้องการเปรียบเทียบให้ผลในระดับที่ดี งานวิจัยนี้จึงให้ Audacity® ได้ทำการเรียนรู้เสียงดนตรีจริงจากเสียงผสม โดยผลการทดลองแสดงได้ดังตารางที่ 4.1 ซึ่งสามารถแสดงเป็นกราฟได้ดังรูปที่ 4.2

ตารางที่ 4.1 ค่า PSNR จากการแยกเสียงโดยการหาผลลัพธ์ที่ดีที่สุดของวิธี NMF วิธีการที่นำเสนอ และวิธีการของโปรแกรม Audacity® ของเครื่องดนตรี 1 ชิ้น

ชุดข้อมูล	เสียงร้อง	เสียงดนตรี	PSNR (เดซิเบล)		
			ผลลัพธ์ที่ดีที่สุดของ NMF	วิธีการที่นำเสนอ	Audacity
1	ชาย	กลอง	46.91	46.91	33.44
2	หญิง	กลอง	45.32	45.32	44.52
3	ชาย	กลองชุด	45.20	45.20	44.07
4	หญิง	กลองชุด	43.36	40.78	38.43

5	ชาย	เบส	39.36	39.24	36.77
6	หญิง	เบส	39.13	39.13	36.48
7	ชาย	เบส+	35.60	34.43	33.84
8	หญิง	เบส+	35.99	34.27	31.28
9	ชาย	คอร์ด	40.45	40.34	42.64
10	หญิง	คอร์ด	38.92	38.92	36.99
11	ชาย	คอร์ด+	37.48	36.18	38.63
12	หญิง	คอร์ด+	36.54	35.07	37.99



รูปที่ 4.2 ค่า PSNR ของผลลัพธ์ที่ดีที่สุดของวิธี NMF วิธีการที่นำเสนอ และวิธีการลดสัญญาณรบกวนของ Audacity®

จากรูปที่ 4.2 เมื่อพิจารณาโดยรวมจะเห็นว่าในข้อมูลสัญญาณเสียงผสมส่วนใหญ่ PSNR ของผลลัพธ์ที่ดีที่สุดของวิธี NMF นั้นจะให้ค่าสูงกว่าวิธีอื่น ๆ ที่สามารถแยกเสียงร้องออกจากเสียงเพลงได้เช่นกัน นั้นแสดงว่าวิธี NMF สามารถนำมาแยกเสียงร้องออกจากเสียงเพลงได้ และให้ผลลัพธ์ที่ดีอีกด้วย นอกจากนี้เมื่อใช้วิธีการเลือกตัวประกอบที่นำเสนอในงานวิจัยนี้ ยังสามารถเลือกตัวประกอบที่ดีได้เท่ากับคำตอบที่ดีที่สุด จากจำนวนรูปแบบที่เป็นไปได้จำนวนทั้งหมด $2^{16} = 65536$ รูปแบบ ในขณะที่วิธีการลดสัญญาณรบกวนของโปรแกรม Audacity® จะให้ผลลัพธ์การแยกเสียงที่ดีกว่าวิธีการที่นำเสนอในบางกรณี ซึ่งให้ผลต่างไม่มากนัก และค่อนข้างมีความแปรผันสูง เช่น ในกรณีของเสียงกลองผสมกับเสียงผู้ชาย และเสียงกลองผสมกับเสียงผู้หญิงกลับให้ผลลัพธ์ที่มากน้อยต่างกัน

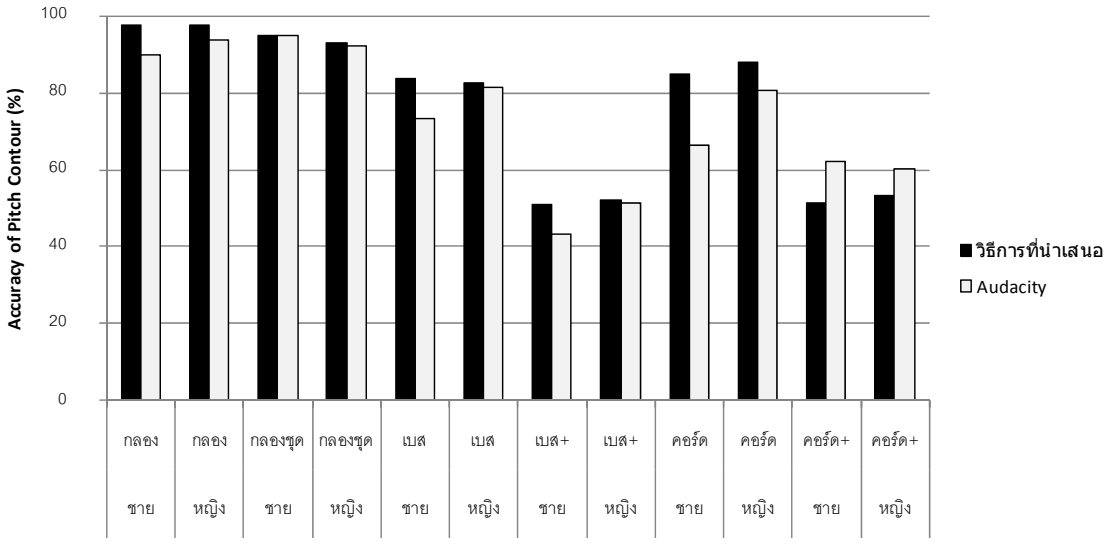
นอกจากนี้จากรูปที่ 4.2 เมื่อพิจารณาเฉพาะวิธีการแยกเสียงร้องออกจากเสียงเพลงที่นำเสนอ ยังสามารถวิเคราะห์ได้เป็นประเด็นต่าง ๆ ดังต่อไปนี้

1. ในกรณีของเสียงเครื่องดนตรีที่ไม่มีการเปลี่ยนแปลงระดับเสียง ได้แก่ เสียงกลอง และเสียงกลองชุด จะเห็นว่า การเพิ่มรายละเอียดของเสียงจะทำให้ผลลัพธ์ของการหาตัวประกอบมีค่าต่ำลง อย่างไรก็ตาม เมื่อเปรียบเทียบผลลัพธ์ของวิธีการที่นำเสนอจะมีค่า PSNR เท่ากับคำตอบที่ดีที่สุดของการหาตัวประกอบ NMF ในข้อมูลชุดที่ 1 2 และ 3 และมีค่าต่ำกว่าเล็กน้อยในข้อมูลชุดที่ 4 นั่นแสดงให้เห็นว่าการหาตัวประกอบของเสียงผสมที่เกิดจากเครื่องดนตรีประเภทนี้ให้ผลที่ดี รวมทั้งวิธีการเลือกตัวประกอบสำหรับเสียงเครื่องดนตรีประเภทนี้มีความถูกต้องเหมาะสม
2. กรณีของเสียงเครื่องดนตรีที่มีการเปลี่ยนแปลงระดับเสียงแบบทีละโน้ต ในงานวิจัยนี้ใช้เสียงเบส และเบส+ นั่นคือมีการเพิ่มระดับเสียงเป็นสองเท่า จากการทดลองนี้ จะเห็นว่า ระดับเสียงของเครื่องดนตรีที่ต่ำจะให้ผลดีกว่าเครื่องดนตรีที่มีเสียงสูงกว่า ซึ่งในความเป็นจริงนั้น เป็นผลมาจาก เครื่องดนตรีเสียงสูงนี้มีเสียงอยู่ในย่านความถี่เดียวกับเสียงร้อง จึงทำให้การหาตัวประกอบ NMF มีความคลาดเคลื่อน หรือไม่ตรงตามความต้องการนัก
3. กรณีของเสียงเครื่องดนตรีที่มีการเล่นเป็นคอร์ด จำนวนของตัวโน้ตที่เพิ่มมากขึ้นจะส่งผลให้ค่า PSNR ของการหาตัวประกอบต่ำลง
4. ผลลัพธ์ของเสียงผสมที่เกิดจากเสียงเครื่องดนตรีที่ให้ทำนอง คือเสียงเครื่องดนตรีอื่น ๆ นอกเหนือจากเครื่องให้จังหวะ จะมีค่าไม่แตกต่างกันมากนัก นั่นแสดงให้เห็นว่า แท้จริงแล้วเสียงที่มีผลให้การแยกเสียงร้องออกจากเสียงเพลงด้วยวิธีนี้มีค่าไม่สูงนัก คือ เสียงเครื่องดนตรีประเภทที่ให้ค่าทำนอง ซึ่งค่า PSNR มีความแตกต่างอย่างมากจากเสียงผสมที่เกิดจากเสียงเครื่องดนตรีที่ให้จังหวะ
5. จากผลลัพธ์ทั้งหมดโดยส่วนใหญ่ แม้จะเห็นว่าในเสียงร้องของผู้ชายมักให้ผลลัพธ์การแยกเสียงที่ดีกว่าเสียงร้องของผู้หญิง อย่างไรก็ตาม จากการวิเคราะห์แล้วพบว่า แม้จะเป็นการร้องของผู้หญิงหรือผู้ชาย แต่หากมีการร้องในเพลงเดียวกัน ก็จะใช้เสียงคีย์เดียวกันในการร้อง ยกเว้นแต่กรณีที่ร้องต่างกันระดับคู่แปด (Octave) ซึ่งการแยกเสียงควรจะให้ผลที่ใกล้เคียงกัน เนื่องจากความหลากหลายในเสียงดนตรียังคงเป็นเช่นเดิม และมีความถี่ของเสียงดนตรีในย่านเดิม

6. จากการทีเสียงร้องของผู้ชายให้ผลการแยกเสียงทีดีกว่าเสียงผู้หญิง เมือลองพิจารณาทีไฟล์เสียงร้องพบว่า ยังมีปัจจัยอืนทีส่งผลต่อการแยกเสียงอีก นันคือ ระดับความดังของเสียงร้อง กล่าวคือ เสียงผู้ชายจะดังกว่าเสียงของผู้หญิงเล็กน้อย แต่เมือทำการฟัง จะให้ความรู้สึกถึงความดังทีไม่แตกต่างกัน

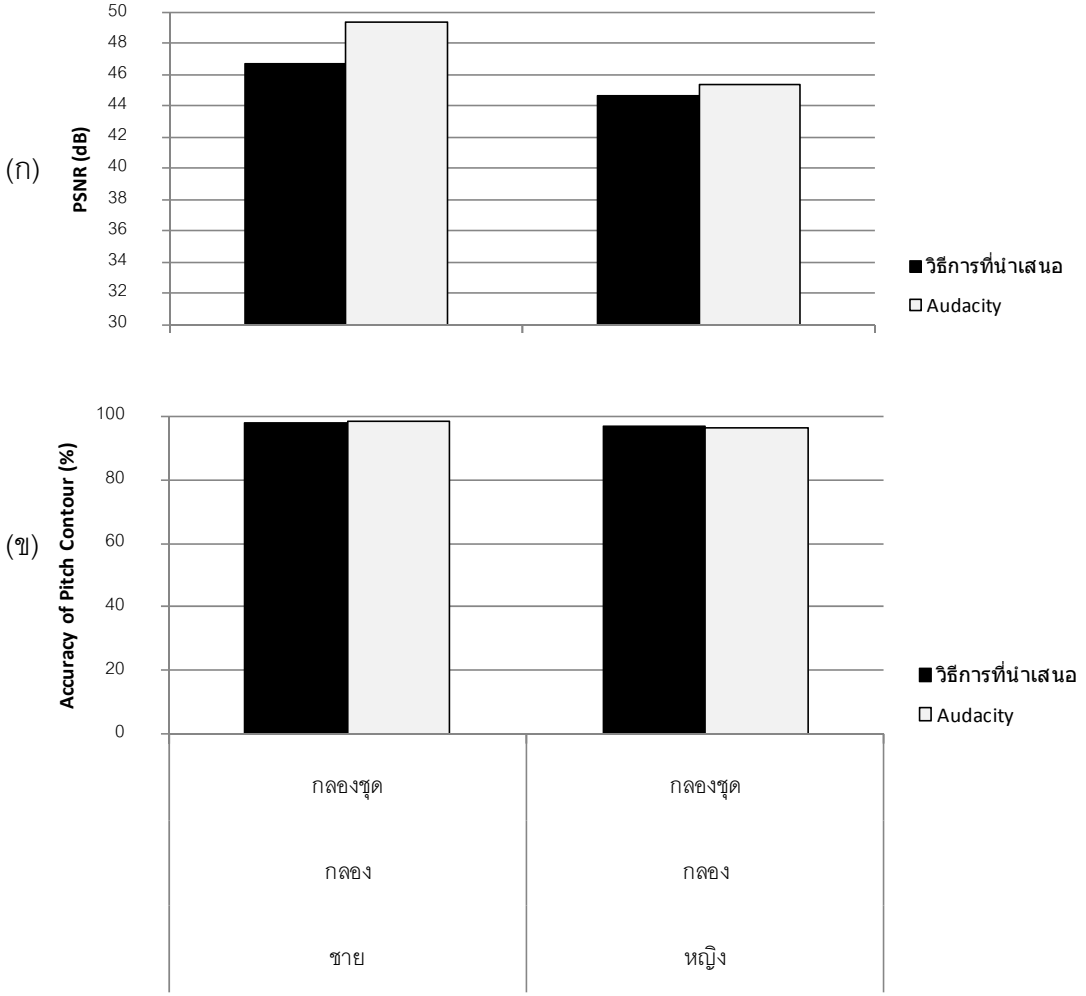
จากประเด็นต่าง ๆ ทีกล่าวมาข้างต้น สามารถสรุปได้ว่า ปัจจัยทีทำให้การแยกเสียงร้องออกจากเสียงเพลงได้ผลทีดีนันคือ รูปแบบทางความถีของเครืองดนตรี ซึงหากยิงมีความหลากหลายน้อย จะยิงทำให้การแยกเสียงได้ผลทีดีขึน และย่านความถีของเสียงดนตรีไม่ควรทีจะตรงกับย่านความถีของเสียงร้องนัก และนอกจากนียิงเสียงร้องมีระดับความดังมาก ผลการแยกเสียงก็จะดีขึน

นอกจากนีเมือพิจารณาความถูกต้องของคอนทัวร์ระดับเสียง ดังรูปที 4.3 นันก็เป็นไปในทิศทางเดียวกับค่า PSNR ทีได้ นันคือสำหรับวิธีการทีนำเสนอจะให้คอนทัวร์ระดับเสียงทีมีความถูกต้องสูงกว่า 90% เมือเป็นเสียงผสมของเครืองดนตรีชนิดให้จังหวะ ในข้อมูลชุดที 1 2 3 และ 4 และมีความถูกต้องสูงกว่าวิธีการของ Audacity® ในผลการทดลองส่วนใหญ่ ดังนันจากผลการทดลองนี แสดงให้เห็นว่าวิธีการทีนำเสนอสามารถนำไปใช้กับการแยกเสียงได้ดีกว่าวิธีการทีนำมาเปรียบเทียบ โดยเน้นทีเครืองดนตรีทีมีการเปลียนแปลงทางความถีไม่มาก เช่น เสียงของเครืองให้จังหวะ



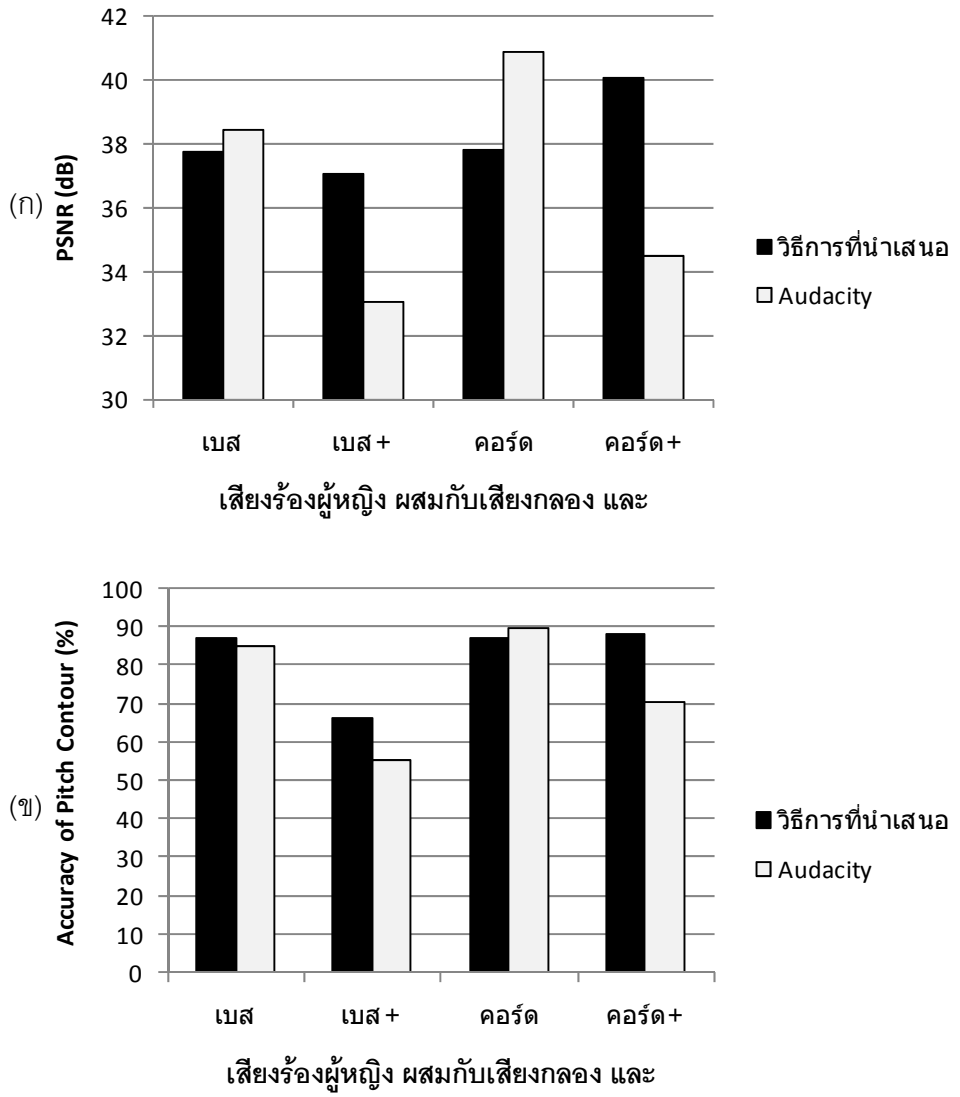
รูปที 4.3 ความถูกต้องของคอนทัวร์ระดับเสียงของวิธีการทีนำเสนอ
และวิธีการลดสัญญาณรบกวนของ Audacity®

เช่นเดียวกับการทดลองที่ผ่านมา การทดลองที่สองนี้จะทำการคำนวณการจัดกลุ่มของตัวประกอบทุกรูปแบบที่เป็นไปได้ เพื่อพิจารณาถึงความสามารถในการแยกเสียงร้องเมื่อเพลงมีรายละเอียดสูงขึ้น โดยรูปที่ 4.4 เป็นผลการทดลองของเสียงผสมกับเครื่องดนตรีให้จังหวะ ซึ่งเป็นไปตามที่คาดนั้นคือ วิธีการที่นำเสนอสามารถแยกเสียงร้องที่มีค่า PSNR และความถูกต้องของคอนทัวร์ระดับเสียงสูง และแม้ว่าค่า PSNR จะได้ต่ำกว่าวิธีการของ Audacity® เล็กน้อย อย่างไรก็ตาม การแยกเสียงยังให้ผลลัพธ์ที่ดี โดยพิจารณาจากความถูกต้องของคอนทัวร์ระดับเสียงที่สูง



รูปที่ 4.4 (ก) ค่า PSNR และ (ข) ค่าความถูกต้องของคอนทัวร์ระดับเสียง
ของวิธีการที่นำเสนอ และวิธีการลดสัญญาณรบกวนของ Audacity®
สำหรับเพลงที่ใช้เครื่องให้จังหวะ 2 ชนิด

นอกจากนี้จากรูปที่ 4.5 เมื่อนำเสียงกลอง (Drum) ไปผสมกับเสียงเครื่องดนตรีชนิดอื่น ๆ แล้ว ทำการแยกเสียงร้องของผู้หญิง ก็ยังคงให้ค่า PSNR ที่สูงในระดับดี แต่อาจมีเสียงเครื่องดนตรีปะปนมาบ้าง และสะท้อนให้เห็นที่ค่าความถูกต้องของคอนทัวร์ระดับเสียงที่ได้



รูปที่ 4.5 (ก) ค่า PSNR และ (ข) ค่าความถูกต้องของคอนทัวร์ระดับเสียง
ของวิธีการที่นำเสนอ และวิธีการลดสัญญาณรบกวนของ Audacity®
สำหรับเพลงที่ใช้เครื่องให้จังหวะผสมกับเสียงดนตรีชนิดอื่น

4.4 สรุปผลการทดลอง

จากการทดลองในบทนี้ ซึ่งประกอบด้วย การทดลองแยกเสียงร้องออกจากเสียงเพลงที่ผสมกับเครื่องดนตรีชนิดเดียว และผสมกับเครื่องดนตรีที่มากขึ้น โดยเปรียบเทียบกับ

วิธีการลดสัญญาณรบกวนของโปรแกรม Audacity® พบว่า ปัจจัยที่มีผลต่อการแยกเสียงด้วยวิธีการที่นำเสนอ คือ จำนวนรูปแบบทางความถี่ที่ปรากฏในเครื่องดนตรีแต่ละชนิด ความดังของเสียงร้องที่นำมาผสม และเมื่อรายละเอียดในเสียงเพลงผสมมีความสูงขึ้น เช่น มีจำนวนโน้ตมากขึ้น หรือมีจำนวนเครื่องดนตรีที่นำมาผสมมากขึ้น ก็เป็นเช่นดังที่คาดไว้คือ ผลการแยกเสียงมีความถูกต้องทั้งค่า PSNR และค่าความถูกต้องของคอนทัวร์ระดับเสียงที่ลดลง

แม้งานวิจัยนี้จะไม่ได้มีการใช้ข้อกำหนดต่าง ๆ ดังหัวข้อที่ 2.4.2 อย่างไรก็ตาม ผลการทดลองต่าง ๆ ที่ออกมาค่อนข้างจะสอดคล้องกับข้อกำหนดในหัวข้อดังกล่าว เช่น ในด้านของความถี่ของเสียงดนตรีที่มีความหลากหลายน้อย หรือการที่ความถี่ของเสียงดนตรีอยู่ในย่านความถี่เดียวกับเสียงร้องจะทำให้การแยกเสียงทำได้ยากขึ้น

อย่างไรก็ตาม เสียงผสมสำหรับเครื่องดนตรีชนิดที่ให้จังหวะซึ่งใช้ได้ผลดีในงานวิจัยที่นำเสนอนี้ จะยังคงสามารถใช้งานได้ เนื่องจากเครื่องดนตรีชนิดที่ให้จังหวะ เป็นเครื่องดนตรีที่มีความสำคัญและใช้กันมากในเพลงหลากหลายประเภท ซึ่งการแยกเสียงด้วยวิธีที่นำเสนอนี้จะสามารถนำไปประยุกต์ใช้ได้

บทที่ 5

สรุปผลการวิจัย และข้อเสนอแนะ

งานวิจัยนี้ได้เสนอวิธีการแยกเสียงร้องออกจากเสียงเพลงที่เก็บในช่องสัญญาณเดียว ซึ่งจากการวิเคราะห์ถึงแนวทางและวิธีการต่าง ๆ ของงานวิจัยที่ผ่านมา งานวิจัยนี้จึงได้เลือกใช้วิธีการหาตัวประกอบของเมทริกซ์ค่าไม่เป็นลบ และเสนอวิธีการเลือกองค์ประกอบหลักจากวิธีการดังกล่าวเพื่อนำไปสร้างเสียงร้องออกมา โดยได้มีการทดลองและวิเคราะห์ผลอย่างละเอียดไว้แล้วในบทที่ 4 ซึ่งจะเห็นได้ว่าวิธีที่นำเสนอสามารถให้ผลการแยกเสียงร้องเป็นที่น่าพอใจ ซึ่งผลการวิจัยสามารถสรุปได้ดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้เสนอวิธีการแยกเสียงร้องออกจากเสียงเพลงที่เก็บในช่องสัญญาณเดียว ที่สามารถทำการแยกเสียงได้จริง จากผลการทดลองในบทที่ 4 ที่มีการประเมินผลด้วยอัตราส่วนสัญญาณสูงสุดต่อสัญญาณรบกวน (PSNR) และค่าความถูกต้องของคอนทัวร์ระดับเสียง (Pitch Contour Accuracy) และจากการทดลอง ได้มีการศึกษาถึงปัจจัยที่มีผลต่อการแยกเสียง จากชุดข้อมูลเสียงผสมต่าง ๆ ที่ทำการทดลอง ซึ่งปัจจัยที่มีผลต่อการแยกเสียงด้วยวิธีการที่นำเสนอนั้น ได้แก่

- รูปแบบทางความถี่ที่ปรากฏของเสียงดนตรี ซึ่งหากมีรูปแบบทางความถี่น้อย การแยกเสียงด้วยวิธีที่นำเสนอจะมีความถูกต้องมาก ซึ่งกรณีนี้เห็นได้ชัดเจนจากผลการทดลองเสียงที่ผสมกับเครื่องดนตรีประเภทให้จังหวะ
- ระดับความดังของเสียง คือ หากเสียงร้องมีระดับความดังมากขึ้น ผลการแยกเสียงก็จะดีขึ้นตามไปด้วย
- ระดับความถี่ของเสียงของร้องและระดับเสียงดนตรี ซึ่งหากระดับเสียงของเสียงทั้งสองมีความแตกต่างกันมาก จะสามารถแยกเสียงได้ผลดีมากยิ่งขึ้น

สำหรับการทดลองโดยใช้เครื่องดนตรีที่มากขึ้น ซึ่งมีองค์ประกอบของเสียงมากยิ่งขึ้น จะส่งผลต่อการแยกเสียงโดยตรง อย่างไรก็ตาม ผลของการแยกเสียงในบางกรณียังคงอยู่ในระดับที่ยอมรับได้ โดยพิจารณาจากผลการเปรียบเทียบคอนทัวร์ระดับเสียงของเสียงร้องที่แยกได้และเสียงร้องต้นฉบับที่ยังคงมีความถูกต้องสูง

นอกจากนี้ จากการศึกษางานวิจัยด้านนี้ รวมทั้งได้ทำการค้นคว้าวิธีแก้ปัญหาการแยกเสียงร้องออกจากเสียงเพลงช่องสัญญาณเดียว พบว่างานวิจัยด้านนี้ยังคงเปิดกว้าง นั่นคือมีช่องทางให้พัฒนาได้อีกมาก โดยจะกล่าวถึงต่อไปในข้อเสนอแนะ

5.2 ข้อเสนอแนะ

จากวิธีการแยกเสียงร้องออกจากเสียงเพลงที่นำเสนอซึ่งสามารถทำการแยกเสียงได้กับเพลงที่เก็บในช่องสัญญาณเดียว อย่างไรก็ตาม ผู้วิจัยเห็นว่ายังมีข้อเสนอแนะบางประการที่จะสามารถช่วยให้งานวิจัยด้านการแยกเสียงร้องนี้มีประสิทธิภาพมากยิ่งขึ้นโดย

1. การหาตัวประกอบเมทริกซ์ที่ใช้ในงานวิจัยนี้ อาจแทนที่ด้วยวิธีการหาตัวประกอบวิธีอื่น ๆ หรือวิธีการหาตัวประกอบเมทริกซ์ค่าไม่เป็นลบ (Non-Negative Matrix Factorization) ที่มีการพัฒนาขึ้นมาใช้แทนวิธีการดั้งเดิม ซึ่งจะช่วยทั้งในเรื่องของการลดความซับซ้อนทางด้านเวลาลง และอาจให้ผลการหาตัวประกอบที่ดีขึ้นได้
2. เกณฑ์ที่ใช้ในการเลือกตัวประกอบของเสียงร้อง อาจมีการศึกษาเพิ่มเติมเพื่อให้มีเกณฑ์ที่มีความเหมาะสม หรือเฉพาะเจาะจงกับเสียงเพลงแต่ละประเภทมากขึ้น
3. เพื่อให้วิธีการแยกเสียงนี้ใช้ได้ดีมากยิ่งขึ้น ในส่วนปลีกย่อยอื่น ๆ เช่น ในส่วนของการเติมเสียงร้องตอนต้น อาจมีการศึกษา ถึงลักษณะเสียงร้องตอนต้นที่เหมาะสมกับการเติมสำหรับการแยกเสียงร้องในเพลงแต่ละท่อน เช่น ลักษณะทางความถี่ หรือรูปแบบของเสียงที่คล้ายคลึงกับเสียงร้องในเพลง
4. ในส่วนของการประมวลผลภายหลังหลังจากการเลือกตัวประกอบได้แล้วก็ยังสามารถหาวิธีอื่น ๆ มาใช้แทนการสร้างสเปกตรัมที่เป็นข้อมูลจำนวนเชิงซ้อน ของเสียงอาจช่วยให้ได้เสียงร้องที่มีคุณภาพดีขึ้นได้

รายการอ้างอิง

[1] Ghias, A., Logan, J., Chamberlin, D., and Smith, B.C. (1995). Query by humming: musical information retrieval in an audio database. Proceedings of 3rd ACM international conference on Multimedia. 231–236. San Francisco, California, United States: ACM Press.

[2] Dannenberg, R.B., Birmingham, W.P., Pardo, B., Hu, N., Meek, C., and Tzanetakis, G. (2007). A comparative evaluation of search techniques for query-by-humming using the MUSART testbed. Journal of the American Society for Information Science and Technology 58: 687-701.

[3] Duda, A., Nürnberger, A., and Stober, S. (2007). Towards query by singing/humming on audio databases. Proceedings of 8th International Conference on Music Information Retrieval, ISMIR. 331-334.

[4] Mesaros, A., Virtanen, T., and Klapuri, A. (2007). Singer identification in polyphonic music using vocal separation and pattern recognition methods. International Conference on Music Information Retrieval. Vienna, Austria.

[5] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H.G. (2005). Singer identification based on accompaniment sound reduction and reliable frame selection. Proceedings of 6th International Conference of Music Information Retrieval, ISMIR. 329-336.

[6] Wong, C.H., Szeto, W.M., and Wong, K.H. (2007). Automatic lyrics alignment for Cantonese popular music. Multimedia Systems, Springer.

[7] Pedersen, M.S., Larsen, J., Kjems, U., and Parra, L.C. (2007). A survey of convolutive blind source separation methods. Springer Press.

[8] Tsai, W.-H., and Wang, H.-M. (2006). Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. IEEE Transactions on Audio, Speech, and Language Processing 14: 330-341.

[9] Ozerov, A., Philippe, P., Gribonval, R., and Bimbot, F. (2005). One microphone singing voice separation using source-adapted models. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New York.

[10] Meron, Y., and Hirose, K. (1998). Separation of singing and piano sounds. Proceedings of International Conference on Spoken Language Processing III. 1059-1062.

[11] Zhang, Y.-G., and Zhang, C.-S. (2006). Separation of music signals by harmonic structure modeling. Advances in Neural Information Processing Systems. 1617-1624.

[12] Li, Y., and Wang, D.L. (2007). Separation of Singing Voice From Music Accompaniment for Monaural Recordings. IEEE Transactions on Audio, Speech, and Language Processing 15: 1475-1487.

[13] Feng, Y., Zhuang, Y., and Pan, Y. (2002). Popular song retrieval based on singing matching. Proceedings of 3rd IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing. 639-646. Springer-Verlag.

[14] Vembu, S., and Baumann, S. (2005). Separation of vocals from polyphonic audio recordings. 6th International Conference on Music Information Retrieval. 337-334.

[15] Schmidt-Jones, C. (2004). Sound, Physics and Music. Science and Technology.

[16] Huang, X., Acero, A., and Hon, H.-W. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR.

[17] Wikipedia. (2008). Statistical independence [Online]. Available from: http://en.wikipedia.org/wiki/Statistical_independence [28 July 2008]

[18] Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Networks 13: 411-430.

[19] Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference. 556-562. MIT Press.

[20] Bertin, N., Badeau, R., and Richard, G.e. (2007). Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007 (ICASSP 2007).

[21] Wang, B., and Plumbley, M.D. (2005). Musical audio stream separation by non-negative matrix factorization. Proceedings of the DMRN Summer Conference, Glasgow.

[22] Audacity-development-team. (2008). How Audacity Works [Online]. Available from: http://audacityteam.org/wiki [20 Mar 2009]

[23] Audacity-development-team. (2009). Audacity: Free audio editor and recorder (Version 1.3.7) [Computer program]. Dominic Mazzoni (Producer). Available from: http://audacity.sourceforge.net/

[24] Adobe-Systems-Incorporated. (2007). Adobe Audition (Version 3.0) [Computer program]. Adobe Systems Incorporated (Producer). Available from: http://www.adobe.com/products/audition

[25] Hoyer, P.O. (2004). Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5: 1457-1469.

[26] Wikipedia. (2008). Peak signal-to-noise ratio [Online]. Available from: http://en.wikipedia.org/wiki/PSNR [2 November 2008]

[27] Jang, R. (2009). Audio Signal Processing and Recognition [Online]. Available from: http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing [25 March 2009]

[28] Li, Y., and Wang, D.L. (2006). Singing voice separation from monaural recordings. Proceedings of ISMIR. 176-179.

[29] Wikipedia. (2008). Signal-to-noise ratio [Online]. Available from: http://en.wikipedia.org/wiki/Signal-to-noise_ratio [2 November 2008]

[30] Boersma, P., and Weenink, D. (2007). Praat: doing phonetics by computer (Version 4.5.12) [Computer program]. Institute of Phonetics Sciences of the University of Amsterdam (Producer). Available from: http://www.praat.org [2 November 2008]

[31] Nattiez, J.-J. (1990). Music and Discourse: Toward a Semiology of Music. Princeton University Press.

ภาคผนวก

ภาคผนวก ก

ผลงานตีพิมพ์จากงานวิจัย

บทความทางวิชาการเรื่อง “Singing Voice Separation for Mono-Channel Music Using Non-negative Matrix Factorization” โดยอังคนา จันทร์รุ่งอุทัย และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการระดับนานาชาติ “The First International Conference on Advanced Technologies for Communications (ATC 2008)” ณ เมืองฮานอย ประเทศเวียดนาม ระหว่างวันที่ 6-9 ตุลาคม 2551

บทความทางวิชาการเรื่อง “Singing Voice Separation in Mono-Channel Music” โดยอังคนา จันทร์รุ่งอุทัย และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการระดับนานาชาติ “International Symposium on Communications and Information Technologies 2008 (ISCIT 2008)” ซึ่งจัดขึ้น ณ เมืองเวียงจันทน์ ประเทศลาว ระหว่างวันที่ 21-23 ตุลาคม 2551

2008 International Conference on Advanced Technologies for Communications

Singing Voice Separation for Mono-Channel Music Using Non-negative Matrix Factorization

Angkana Chanrungutai and Chotirat Ann Ratanamahatana

Department of Computer Engineering, Chulalongkorn University
Phayathai Road, Pathumwan, Bangkok 10330 Thailand
E-mail: {g49ach, ann}@cp.eng.chula.ac.th

***Abstract*— As music has turned digital, much research has been shifted toward digital music processing. Singing voice separation is one of the active research areas since the singing voice itself contains abundant information within, including melody, singer's characteristic, lyrics, language, emotion, etc. These wide variety of resources are quite useful for Music Information Retrieval (MIR), singer identification, or even karaoke systems. However, this singing voice separation, especially in mono-channel environment, is a very challenging problem, whose existing methods are still impractical for real-world music. In this work, we propose an algorithm based on Non-negative Matrix Factorization (NMF) approach to decompose spectra of music, then provide criteria for automatic component selection. Our preliminary results have demonstrated its effectiveness in pitch extraction resulting from the separated singing voice.**

***Index Terms*—Mono-channel music, Non-negative Matrix Factorization, Singing voice separation, Sound source separation**

I. Introduction

With the widespread of digital music over the internet and other sources at present, music processing has increasingly become part of many people's daily routines, e.g., Music Information Retrieval (MIR) system, karaoke system using lyrics and singing voice alignment, singer identification, music enhancement, etc. Some MIR system such as a popular query-by-humming system is viable for only specific types of music, e.g., MIDI music format, since the melody of the song is separated in an individual channel and can readily be extracted for use. Unfortunately, most of the systems are still impractical for common digital music formats such as MP3, WMA, or WAV. Although it can be argued that these melodies can be extracted by a pitch detection algorithm, the instrumental sounds within the music always cause high errors during the detection. Therefore, singing voice separation must be performed in advance. Other than the usefulness for a query-by-humming system, this singing voice actually contains abundant information, e.g., lyrics, language, singer characteristics, emotion, etc. that could be useful for wide variety of research.

The singing voice separation is one of the challenging aspects of the sound source separation problem. The major complications lie in the fact that it is extremely difficult to determine the relationship between the number of observed signals and that of sources. In addition, the type of sources in the music mixture must also be discovered. Theoretically, singing voice is produced by vocal tract, which always varies by organ movements, making this singing voice separation problem more challenging comparing to instrumental sound that is much more stable. This perhaps explains the relatively few research works that are directed toward singing voice separation, regardless of the dire need in many digital music processing systems; existing approaches are still impractical for many genres of music.

One of the direct approaches for source separation, such as Independent Component Analysis (ICA) [2, 8], requires that the number of observed mixtures must be larger or equal to the number of sources. However, the number of sources cannot be easily determined or specified, and the number of sensors is also generally limited to one or two (mono or stereo) channels which is always less than the number of sources.

Other approaches are statistical modeling [5, 7], and computational auditory scene analysis (CASA) [4, 10]. For statistical modeling, models have to be learned from pure singing voice and instrumental sound in advance. Then the separation is done from the learned models. However, the acquisition of pure singing voice is very difficult in reality. Although CASA approach does not need pure singing voice for learning, it requires some cues to help the separation, such as harmonic structure of instrumental sounds [10] or estimated pitch of the singing voice itself [4]. However, it is infeasible for real world music to know the exact number of sources or to accurately estimate the pitch of singing voice while there are many other sounds producing pitch, including percussions.

In this work, we propose the singing voice separation method for mono-channel music which can also be applied to any types of digital music. We use Non-negative Matrix Factorization (NMF) [3] as a tool to decompose the music spectrum. The contribution of our research is to propose an NMF's component selection algorithm which is the crucial step in singing voice separation.

The rest of the paper is organized as follows. The NMF algorithm and our proposed singing voice separation method are explained in sections 2 and 3, respectively. The experimental results are given in section 4. Finally, we conclude and discuss the direction for future work in section 5.

978-1-4244-2681-2/08/$25.00 ©2008 IEEE

II. Non-Negative Matrix Factorization

A. Definition

Non-Negative Matrix Factorization (NMF) proposed by Lee and Seung [3] is an algorithm for multivariate data analysis where a matrix $V_{f\times t}$ is decomposed into the product of two matrices $W_{f\times r}$ and $H_{r\times t}$ as shown in eq. (1), where all elements in the matrices are non-negative, and m, n, and r are positive integer.

$$V \approx WH \tag{1}$$

To find W and H, they have designed the multiplicative update rules to minimize the cost function between V and WH. They also have proposed two cost functions, i.e., the square of the Euclidean distance shown in eq. (2) and the divergence of V and WH shown in eq. (3).

$$\|V - WH\|^2 = \sum_{ij} \left(V_{ij} - (WH)_{ij}\right)^2 \tag{2}$$

$$D(V \,||\, WH) = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right) \tag{3}$$

Due to space limitations, more information about NMF and the proofs of convergence can be found in [3].

B. NMF for singing voice separation

For music signal, we can use its spectrum as an input matrix of NMF, V (Figure 1), whose non-negative value at any position (f, t) is the amplitude of the signal at frequency-bin f and frame t.

In the aspect of linear algebra, considering the matrix V which contains t column vectors of length f, after a decomposition using NMF, we will get matrices W containing r basis vectors of length f and H containing the coefficients of each basis vector along the time axis.

The interesting part of NMF is the non-negative constraint that gives the sparseness to the result matrices. Consequently, we can think that each basis vector decomposed contains the frequencies that are usually produced simultaneously. For the instrumental sounds, the simultaneous frequencies can be considered as harmonics or timbres of each note. So if we know the total number of notes of each instruments, we will know the value of r as mentioned in [9]. However, it is much more complicated in the singing voice separation. Produced frequencies always vary across different utterance and person which is lacking consistency, unlike the sounds from instruments. Therefore, there are many combinations of the synchronous frequencies. For singing voice separation, we select the basis vectors that we assure to be of the instrumental sounds, instead of that of the singing voice.

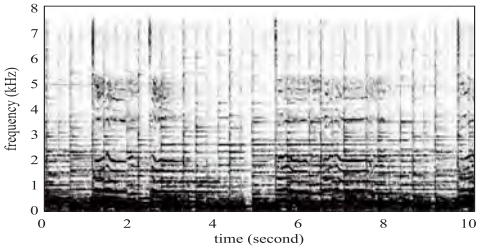


Figure 1 Spectrum of a music clips as an input matrix V of NMF, where dark color represents high amplitude and vice versa

III. Singing Voice Separation

Our proposed singing voice separation method using NMF are fivefold. Firstly, input music is preprocessed to the form that can be decomposed by NMF, and some parameters are prepared. Next, NMF is operated before the component selection. In the case that the selected components contain some frequencies of the instrumental sounds, we remove them with filter in the refinement stage. Finally, the singing voice is reconstructed. The details of our method are as follows.

A. Preprocessing

In this first stage, the mono-channel music is sampled at 16,000 Hz, and the spectrum of the music signal is constructed using Discrete Fourier Transform (DFT) with the Hamming window of 32 ms frame size and 16 ms overlap. Then the magnitude or amplitude of the spectrum is calculated to initialize the NMF input.

In addition, we initialize the number of basis components, r, as mentioned in section IIA. Our preliminary results reveal that they are not affected much by r values, as long as it is a positive integer that is not too small. According to our preliminary testing, r value of 64 deems to be generally good number.

B. Decomposition using NMF

The input matrix V and the number of basis components, r, from previous stage is now passed through the NMF in order to decompose into the result matrices W and H that are considered as the basis vectors and their coefficients, respectively. Examples of basis vectors of W and the coefficients in H along the time axis are shown in Figure 2.

C. Component Selection

In this stage, we remove the components that are not of the singing voice or are expected to be instrumental sounds. To achieve this, we employ two general ideas, i.e. rhythmic and continuous events. Firstly, the main instrument in typical music is percussion that is always played rhythmically (except in solo music or singing voice accompanied by some instruments). Other instruments are composed to be played concurrently with the music. Some instruments are played simultaneously, so the components decomposed by NMF might not separate each note of each instrumental sound correctly. However, we can conclude that most instruments except for percussion tend to be played continuously in time.

According to these ideas, we consider the coefficients in matrix H, which tells us how each component varies along the time axis. For instance, all the graphs in the right-hand column of Figure 2 are the coefficients of basis components of the left column. We can see that h_2 and h_5 are rhythmic; in other

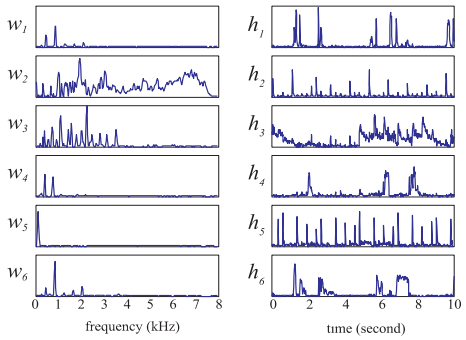


Figure 2 An example of components decomposed by NMF into matrices W (left) and H (right)

words, the sound of this component is played rhythmically, so we remove these two components out of the singing voice basis list.

Human singing voice usually does not last for long period in general music (unless it is a tiding note). Consequently, in the continuous event case, we can assume for general music clips that if the basis spectrum is lasting too long, it is most definitely not the spectrum of the singing voice. Based on our experiments, the duration threshold is set to four seconds. However, this parameter can be adjusted to suit various types of music. From Figure 2, h_3 is considered a continuous event, as relatively large values of coefficient of its component are produced during the 5-second until 10-second time slices. After this stage, the three components in Figure 2 are removed; the 2nd and 5th are removed by the rhythmic criterion, and the 3rd is removed by the continuous event criterion.

Although the components are manually selected in our preliminary tests, these criteria can be simply developed to be an automatic selection process.

D. Component Filtering

From the result of the previous stage, there are some frequencies of the basis vectors that are to be removed. Accordingly, we refine each component using the fact that human voice frequencies are never below 40 Hz and never over 2,000 Hz[4]. However, there exist some types of sound that produce frequencies over this threshold, e.g., fricative sound like 's' or 'f' whose frequencies can be above 2,000 Hz. Therefore, we only use the lower bound of 40 Hz to filter the selected components of *W*.

E. Signal Reconstruction

Once all the components or frequency sets of the singing voice are selected, we multiply *W* and *H* whose spectrogram of the components selected is shown in Figure 3. However, the elements of this matrix are still non-negative and lack of phase information. To fix this, we simply calculate the new spectrum of which the complex value in any position is the product of each individual element of this matrix and the original spectrum in the same position.

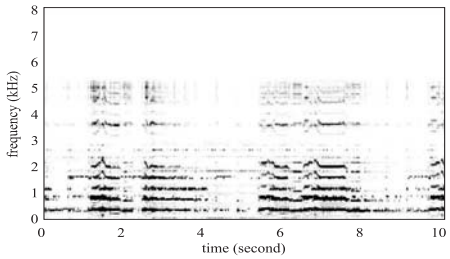


Figure 3 the spectrogram of WH after component selection, where dark color represents high amplitude and vice versa

IV. Experiments and Results

A. Experimental Setup

To avoid the copyright violation and to make quantitatively evaluation of our proposed method, we use our own mixtures as music test set. The music test set contains ten songs sung by various singers, and we crop each music into three clips. The input signal is sampled at 16,000 Hz and 16-bit depth. In addition, two songs are sung on MIDI backing tracks, while the rest are sung on real backing track.

The advantage of the MIDI backing track is that we can separate it into individual instrumental sound in order to analyze which type of instrumental sounds affect the performance of the separation method.

B. Metric

We have evaluated our singing voice separation method by the signal-to-noise ratio (SNR) [6] shown in eq. (4), which considers the separated singing voice, *s*, as a signal, and least square error between the separated singing voice and the original one, *x*, as noise.

$$SNR = 10\log_{10}\frac{\sum_{i=1}^{n} s(i)^2}{\sum_{i=1}^{n}\left(s(i)-x(i)\right)^2} \tag{4}$$

where *n* is the length of both singing voice signals.

C. Results

Real Backing Track

An example of separated singing voice signal is shown in Figure 4 and the average SNR results of separation on eight real backing tracks are shown in Table 1. We can readily see from the figure that the separated singing voice extracted by our proposed method is highly similar to the pure original singing voice. The accompanied instruments are extremely diverse, giving quite a wide range of results. This brings us to the conclusion that the complication of singing voice separation problem is affected by the instruments in the music. In addition, in separation from real backing tracks, female singing voices give higher performance than male singing voices. It is important to note that almost all of the previous work only displayed the plotted signals after the separation without using formal assessment metric. We therefore decide to quantita-

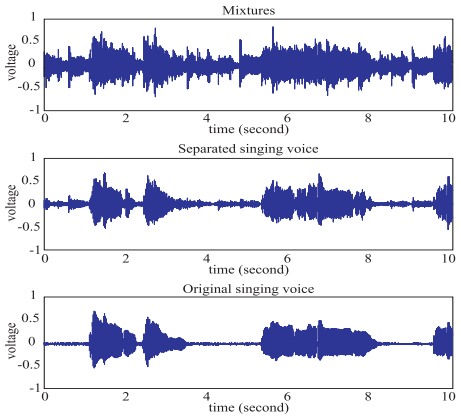


Figure 4 example of separated singing voice

tively measure the raw results regardless of its relatively low figures.

To demonstrate the utility of our approach, we additionally compare the detected pitch of the separated singing voice with that of the original singing voice using Praat [1]. The accuracy shown in Table 2 corresponds to the previous table, as larger SNR values contribute to lower mean errors.

Artificial Backing Track

As mentioned earlier, we divide the instrumental sounds into groups and remix with the singing voice to analyze the performance of our separation method on different mixtures. We make an assumption that the performance of separation algorithm is directly related to pitch or periodicity of instrument sound and singing voice. Consequently, we divide the instruments into three groups; non-periodic percussion instruments (e.g., drum, drum set, and cymbal), bass instruments, and high-pitch instruments (e.g. string, bell, piano).

The result including the full mixture of all instruments is shown in Table 3. It reveals that the SNR separation result of percussion and bass instrument are the highest for male and female singing voice, respectively. Unexpectedly, the result of high-pitch instruments is quite low, even for male singing voice. Although the male singing voice pitch is lower which should have good contrast and give a higher SNR by our assumption, it occurs that there are many instruments of this group within the music while the percussion component was previously removed, making this separation much more different and reduces the performance of the algorithm.

V. Conclusion and Discussion

In this work, we propose the singing voice separation method that utilizes Non-negative Matrix Factorization (NMF) algorithm. The NMF has distinctive ability of decomposing the spectrum of mixtures. With our proposed method, the singing voice can be separated from the mono-channel music mixtures. The SNR values refer to the satisfying performance of our singing voice separation algorithm. In addition, we perform a pitch detection with the separated singing voice output. The mean errors between our separated singing voice and the pure singing voice are quite low, illustrating the utility and good possibility in extending and incorporating this into other research components in computer music, including Music Information Retrieval systems. In future work, we plan to study the effects of different instruments to the method.

TABLE 1 AVERAGE SNR OF THE SEPARATED SINGING VOICE FROM REAL BACKING TRACKS

	Average SNR (dB)
All	7.11
Male	4.23
Female	9.99

TABLE 2 PITCH DETECTION ACCURACY COMPARISONS OF SEPARATED SINGING VOICE AND FULL MIXTURE

	Mean error (Hz)	
	Separated singing voice	Full mixture
All	28	114
Male	37	94
Female	20	133

TABLE 3 AVERAGE SNR OF THE SEPARATED SINGING VOICE FROM ARTIFICIAL BACKING TRACKS

	Average SNR (dB)	
	Male	Female
Full mix	1.14	4.66
Percussion	**11.63**	9.61
Bass	4.73	**17.32**
High-pitch instruments	0.31	3.80

References

[1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.5.12) " [Computer program], Retrieved Oct 25, 2007, from http://www.praat.org/.

[2] Y. Feng, Y. Zhuang, and Y. Pan, "Popular song retrieval based on singing matching," in Proceedings of 3rd IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, 2002, pp. 639-646.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, 2001, pp. 556-562.

[4] Y. Li and D. L. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1475-1487, 2007.

[5] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 2005.

[6] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," Journal of the Acoustical Society of America, vol. 114, pp. 2236-2252, 2003.

[7] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 330-341, 2006.

[8] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in 6th International Conference on Music Information Retrieval, London, UK, 2005, pp. 337-334.

[9] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in Proceedings of the DMRN Summer Conference, Glasgow, 2005.

[10] Y.-G. Zhang and C.-S. Zhang, "Separation of music signals by harmonic structure modeling," in Advances in Neural Information Processing Systems, 2006, pp. 1617-1624.

Singing Voice Separation in Mono-Channel Music

Angkana Chanrungutai and Chotirat Ann Ratanamahatana
Department of Computer Engineering
Chulalongkorn University
Phayathai Road, Pathumwan, Bangkok 10330 Thailand
E-mail: {g49ach, ann}@cp.eng.chula.ac.th

***Abstract*— With predominant communication technology in our digital world, music has turned digital, and much research has been shifted toward digital music processing. Singing voice separation is one of the active research areas since the singing voice itself contains abundant information within, such as melody, singer's characteristic, lyrics, language, emotion, etc. This wide variety of resources is very useful for diverse areas of multimedia research, including Music Information Retrieval (MIR), singer identification, or even karaoke systems. However, this singing voice separation, especially in mono-channel environment, is very challenging problem, whose existing methods are still impractical for real world music. Our main contribution of this work is to propose a singing voice separation algorithm based on Non-negative Matrix Factorization (NMF) method that can effectively extract the singing voice from the mono-channel music. We demonstrate its utility and effectiveness by experiments in both quantitative evaluation and its pitch detection accuracy from the separated singing voice.**

I. INTRODUCTION

With the widespread of digital music over the internet and other sources at present, music processing has increasingly become part of many people's daily routines, e.g., Music Information Retrieval (MIR) system, karaoke system using lyrics and singing voice alignment, singer identification, music enhancement, etc. Some MIR systems such as a popular query-by-humming system are quite viable only for specific type of music, i.e., MIDI music format, since the melody of the song is separated in an individual channel and can readily be extracted for use. Unfortunately, most of the systems are still impractical for common digital music formats such as MP3, WMA, or WAV. Although it can be argued that these melodies can be extracted by a pitch detection algorithm, the instrumental sounds within the music, however, always cause high errors during the detection. Therefore, singing voice separation must be performed in advance. Other than the usefulness for a query-by-humming system, this singing voice actually contains abundant information, e.g., lyrics, language, singer characteristics, emotion, etc. that could be useful for wide variety of research.

The singing voice separation is one of the challenging aspects of the sound source separation problem. The major complications lie in the fact that it is extremely difficult to determine the relationship between the number of observed signals and that of sources. In addition, the type of sources in the music mixture must also be discovered. Theoretically, singing voice is produced by vocal tract, which always varies by organ movements of each unique individual, making this singing voice separation problem more challenging comparing to instrumental sound that is much more stable. This perhaps explains the relatively few research works that are directed toward singing voice separation, regardless of the dire need in many digital music processing systems; existing approaches are still impractical for many genres of music.

One of the direct approaches for source separation, such as Independent Component Analysis (ICA) [2, 8], requires that the number of observed mixtures must be larger or equal to the number of sources. However, for singing voice separation problem in typical music, the number of sources or instruments cannot be easily determined or specified, and the number of sensors is also generally limited to one or two channels (mono or stereo), always being less than the number of sources, which clearly does not satisfy the ICA's fundamental requirement.

Other common approaches are statistical modeling [5, 7], and Computational Auditory Scene Analysis (CASA) [4, 10]. For statistical modeling, models have to be learned from pure singing voice and instrumental sounds in advance. Then the separation is done from the learned model. However, the acquisition of pure singing voice is very difficult in reality. Although CASA approach does not need pure singing voice for learning, it requires some cues to help the separation, such as harmonic structure of instrumental sounds [10] or estimated pitch of the singing voice itself [4]. Nonetheless, it is still infeasible in real world music to identify the exact number of sources or to accurately estimate the pitch of singing voice while there are many other sounds producing pitch, including music instruments or even percussions.

Another attractive algorithm called Non-negative Matrix Factorization (NMF) [3], has been used for spectral data analysis and other areas such as digital image processing and text mining. Recently, NMF has been applied to the musical source separation problem in an attempt to separate musical instrument sounds [9], and has been shown to give good result without needing any prior information about each instrument sound. These promising results are part of our main motivation to apply NMF to the singing voice separation problem.

To generalize the problem, we therefore propose a singing voice separation method for mono-channel music which can essentially be applied to any type and format of digital music. We use NMF as a tool to decompose the music spectrum. The contribution of this work is to propose an NMF's component selection algorithm which is the crucial step in singing voice

978-1-4244-2336-1/08/$25.00 © 2008 IEEE

separation. In addition, we have analyzed the factor that can affect the performance of our algorithm.

The rest of the paper is organized as follows. The NMF algorithm and our proposed singing voice separation method are explained in sections 2 and 3, respectively. The experimental results are given in section 4. Finally, we conclude and discuss the direction for future work in section 5.

II. NON-NEGATIVE MATRIX FACTORIZATION

A. The NMF Principles

Non-Negative Matrix Factorization (NMF), developed by Lee and Seung [3], is an algorithm for multivariate data analysis where a matrix V is decomposed into a product of two matrices W and H as shown in eq. (1), where all elements in the matrices are non-negative, and f, t, and r are positive integer.

$$V_{f\times t} \approx W_{f\times r} H_{r\times t} \tag{1}$$

To find W and H, they have designed the multiplicative update rules to minimize the cost function between V and WH. They also have proposed two cost functions, i.e., the square of the Euclidean distance shown in eq. (2) and the divergence of V and WH shown in eq. (3).

$$\|V - WH\|^2 = \sum_{ij} \left(V_{ij} - (WH)_{ij}\right)^2 \tag{2}$$

$$D(V\|WH) = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right) \tag{3}$$

Due to space limitations, more information about NMF and the proofs of convergence can be found in [2].

B. NMF for Singing Voice Separation

For singing voice separation, we can use the spectra of music signal as an input matrix of NMF, V (Fig. 1A), whose non-negative value at any entry $V_{f,t}$ is the amplitude of the signal at frequency-bin f and frame t.

In the aspect of linear algebra, the spectra or matrix V can be considered as t column vectors of f elements. After NMF decomposition, we will get two matrices; W which consists of r basis column vectors of f elements, w_i, and H which consists of r row vectors of t elements, h_i, representing the coefficients of each basis vector along the time frame t, i.e., $W = \{w_1, w_2,\ldots, w_r\}$ and $H = \{h_1, h_2,\ldots, h_r\}^T$. Using these notation, the equation (1) can be written as

$$V \approx \sum_{i=1}^{r} w_i h_i \tag{4}$$

In addition, the interesting part of NMF is the non-negative constraint that gives the sparseness to the basis spectra or column vectors in matrix W (for clarity, we provide a sample basis spectra in section IIIC). Because the only operation involved among $w_i h_i$ in eq. (4) is addition operator with a non-negative constraint, the basis spectra therefore contain sparse

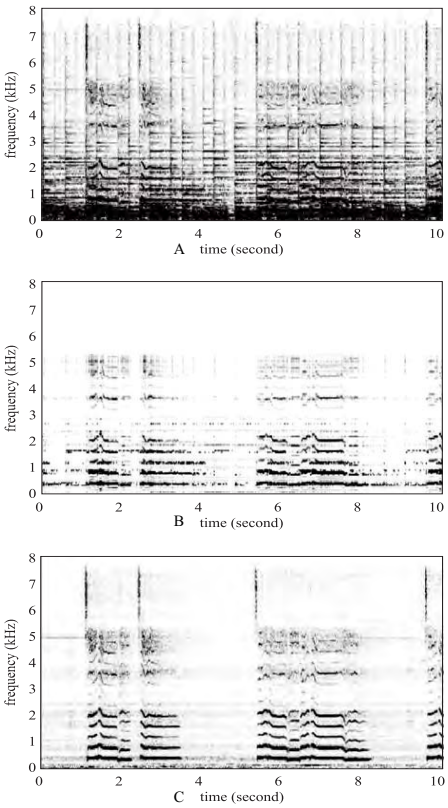


Fig. 1. Spectra of A) A sample music clip as the input of NMF, V; B) $W'H'$ after component selection (this can also be considered as the separated singing voice); and C) Ideal singing voice. Black represents high amplitude, and white represents low amplitude. The details of figures are given in section V.

frequencies such that the non-negative values can be distributed to all basis spectra. Specifically, the frequencies usually produced simultaneously are typically used for the frequency distribution. In other words, the basis spectra can reflect the harmonics or timbres of each note for the instrumental sounds. In this case, the integer r for NMF decomposition can be set to the total number of notes of each instrument as mentioned in [9].

However, the singing voice separation problem is much more complicated; produced frequencies of singing voices always vary across different utterances and persons giving intractable amount of variation of the synchronous frequencies, unlike the sounds from instruments which are much more consistent. Therefore, in this work, we select the basis vectors based on the instrumental sounds instead of the singing voice to achieve better separation.

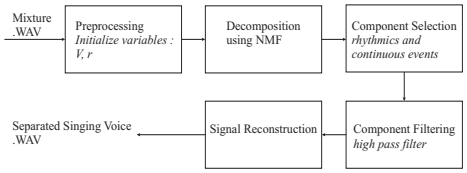


Fig. 2. Block diagram of the proposed singing voice separation algorithm

III. SINGING VOICE SEPARATION

Our proposed singing voice separation method using NMF are fivefold, as shown by a block diagram in Fig. 2. First, input music is preprocessed to the form that can be decomposed by NMF, and some parameters are prepared. Next, NMF decomposition is performed before the component selection process. In the case that the selected components contain some frequencies of the instrumental sounds, we remove them with filter in the component filtering stage. Finally, the singing voice is reconstructed. The details of our method are provided as follows.

A. Preprocessing

In this stage, we transpose the mono-channel music clip into frames and pass it through the Hamming window. Then, Discrete Fourier Transform (DFT) is calculated to construct the spectrum for each frame. Then, the amplitude of each frequency-bin in the spectrum is calculated to initialize the NMF input, i.e., the non-negative matrix V. The output from this stage can be illustrated as a spectrogram shown in Fig. 1A.

In addition, we initialize the number of basis components, r, as mentioned in section II. Our preliminary results reveal that they are not affected much by r values, as long as it is a positive integer that is not too small. According to our preliminary testing, r value of 64 deems to be a generally good number.

B. Decomposition using NMF

The input matrix V and the number of basis components, r from previous stage is now passed through the NMF in order to decompose into the result matrices W and H that are considered as the basis spectra and their coefficients, respectively. Examples of basis spectra of W and the coefficients in H along the time axis are shown in Fig. 3.

As mentioned in section IIB that the basis spectra are sparse, that is, from Fig. 3, most w_i have small amount of active frequencies. It is obvious that the active frequencies of w_1, for example, are harmonics which appear in the sound according to h_1. However, there are some cases that many frequencies are active, e.g., w_2 and w_3. This situation could be a result of noise that may be played by many types of instruments, such as some percussion and wind instrument, or even by human. For this reason, the coefficients of the basis spectrum have been used in the next stage.

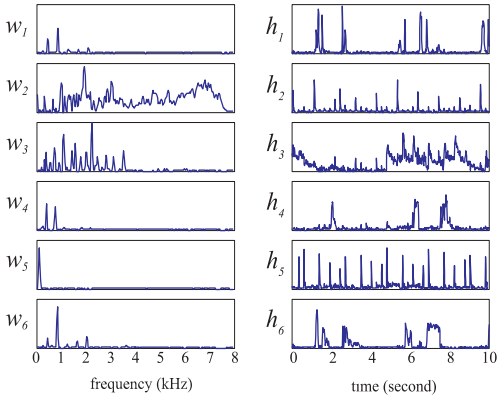


Fig. 3. An example of components decomposed by NMF into matrices W (left) and H (right)

C. Component Selection

In this stage, we remove the components that are not of the singing voice or are expected to be instrumental sounds. To achieve this, we employ two general ideas, i.e. rhythmic and continuous events. Firstly, the main instrument in typical music is percussion that is always played rhythmically (except in solo music or singing voice accompanied by some instruments). Other instruments are composed to be played concurrently with the music. Some instruments are played simultaneously, so the components decomposed by NMF might not separate each note of each instrumental sound correctly. However, we could assume that most instruments except for percussion tend to be played continuously in time.

According to these ideas, we consider the coefficients in matrix H, which tells us how each component varies along the time axis. For instance, all the graphs in the right-hand column of Fig. 3 are the coefficients of basis spectra of the left column. We can see that h_2 and h_5 are rhythmic; in other words, the sound of this component is played rhythmically, so we remove this component out of the singing voice basis list.

As mentioned above, each basis spectrum reflects harmonic or noise. For human singing voice, the harmonic appears according to a vowel which usually does not last for long period in general music (unless it is a tiding note). In the case of noise, the human produces noise only in short period of consonant uttering. In the continuous event case, we can conclude for general music clips that if the basis spectrum is lasting too long, it is most definitely not the spectrum of the singing voice. Based on our experiments, the time threshold is determined to be four seconds, i.e., the period of time threshold that the basis spectrum does not belong to the spectrum of the singing voice. However, this parameter can be adjusted to suit various types of music. After this stage, the three components in Fig. 3; the second and fifth are removed by the rhythmic criterion, and the third is removed by the continuous event criterion.

The output of this stage are *W'*, denoted as the selected basis spectra, and *H'*, denoted as the coefficients corresponding to the basis spectra in *W'*.

Although the components are manually selected by part in our preliminary tests, these criteria can be simply developed to be an automatic selection process.

D. Component Filtering

From the result of the previous stage, there are some frequencies of the selected basis spectra that are to be removed. Accordingly, we refine each component using the fact that human voice frequencies are never below 40 Hz and never over 1.4 kHz [4]. However, there exists some types of human sound that could produce frequencies over this threshold, e.g., fricative sound like 's' or 'f' whose frequencies can be above 1.4 kHz. Therefore, we only use lower bound to filter the basis spectra of *W'*.

E. Signal Reconstruction

Once all the components or frequency sets of the singing voice are selected, we multiply matrices *W'* to *H'* whose product can be shown as a spectrogram in Fig. 1B. However, the elements of this matrix are still non-negative and lack of phase information. To fix this, we simply filter the original spectra by multiplying each individual element with the element at the same position of *W'H'* in the frequency domain. For example, the element (2, 3) of the resultant matrix is the product of the element (2, 3) of the original spectra and of the (*W'H'*) matrix.

Finally, we calculate the Inverse Discrete Fourier Transform (IDFT) from the output spectra. Then the singing voice signal is re-synthesized.

IV. Experiments and Evaluation

We design the experiments for evaluation in two aspects, i.e., the quality of separated singing voice, and the accuracy of pitch detection from the singing voice separated by our proposed method. In this section, we give the details of experimental setup, metric for evaluation in these aspects, and the methodology of the experiments.

A. Experimental Setup

To make quantitative evaluation, we test our proposed method on various mixtures of music test set, which contains 16 music clips sung by various male and female singers accompanied by various types of instruments. The average length of clips is 9.60 second. The input signal is sampled at 16 kHz, 16-bit resolution, with 32 ms frame size and 16 ms overlap for the experiments. For a 10-second clip, we will obtain 624 column vectors of length 512 which is then transformed to 624 spectra of length 257 after the preprocessing stage. In addition, fourteen music clips are sung on randomly selected instrument backing track, while the rest are sung on real backing track. All backing tracks are in MIDI format, only for evaluation purpose.

The advantage of MIDI backing track usage in our experiment is that we can easily extract each channel of pure instrumental sound and mix with singing voice for the experiments, including a convenient evaluation on which type of instrumental sounds and how they affect the performance of the separation method.

B. Evaluation

To evaluate the quality of singing voice separated by our proposed method, we take the signal-to-noise ratio (SNR) used by [6, 9] as shown in eq. (5), which considers the separated singing voice, *s*, as a signal, and least square error between the separated singing voice and the original one, *x*, as noise.

$$SNR = 10\log\frac{\sum_{i=1}^{n} s(i)^2}{\sum_{i=1}^{n}\left[s(i)-x(i)\right]^2} \tag{5}$$

where *n* is the length of the singing voice signals, and *i* is the sampling point of each signal.

The higher performance of separation method can be explained as the higher SNR. In an ideal case, the SNR should reach its upper bound, which is defined in the next section.

For pitch detection accuracy evaluation, we compare the pitch detected from our separated singing voice with the pitch detected from the original pure singing voice using Pratt detection software [1]. We measure the error between pitch of the separated and original singing voice pair using Manhattan distance metric, such that we can obviously notice the distance in Hertz unit.

C. Methodology

Our experiments consist of two main components. First, the testing objective is to study the results of our singing voice separation method under three scenarios; single-instrument, mixed-instrument, and real backing tracks. For single-instrument backing track, we are interested in the effect of the type of instrument to the performance of separation algorithm; we randomly mix female and male singing voice with various types of instruments that can be grouped as *percussion*, such as drum, tambourine, and cymbals; *low-pitch instrument*, such as guitar-bass and trumpet-bass; and *high-pitch instrument*, such as piano, guitar, and saxophone. In the second scenario, we are interested in the singing voice among multiple instruments mixed together, which is closer to the real music. Finally, we test our method with real world music, whose singing voice and accompaniment are highly correlated. We mixed random male and female singing voice tracks with their real backing tracks, which can consist of up to seven instruments.

The second part of our evaluation is to measure the accuracy of the singing pitch extracted from the separated singing voice of the music from the third scenario, the real world music. The results are shown and discussed section V.

V. Results and Discussion

As described in previous section, the experiments are performed in two phases under three scenarios using different types of backing tracks. The details of the results and discussion are as follows.

TABLE 1.
SNR OF THE SEPARATED SINGING VOICE
FROM SINGLE-INSTRUMENT BACKING TRACKS

Music set		SNR (dB)	
Singer	Instrument	Result[a]	Ideal[b]
Male	P[c]	13.68	21.52
	L[d]	15.22	25.66
	H[e]	21.75	25.42
Female	P	19.26	24.84
	L	22.27	26.60
	H	17.99	23.97
Average		18.36	24.67

[a] SNR between the separated singing voice and the original singing voice,
[b] SNR between the approximated mixture and the input mixture,
[c] Percussion instrument, [d] Low-pitch instrument, [e] High-pitch instrument

TABLE 2.
SNR OF THE SEPARATED SINGING VOICE
FROM MIXED-INSTRUMENT BACKING TRACKS

Music set		SNR (dB)	
Singer	Instrument	Result	Ideal
Male	P + L	14.87	24.08
	P + H	16.88	26.53
	L + H	16.13	23.79
	P + L + H	15.29	23.87
Female	P + L	24.81	26.38
	P + H	17.57	27.99
	L + H	17.34	25.85
	P + L + H	15.32	25.22
Average		17.28	25.46

TABLE 3.
SNR OF THE SEPARATED SINGING VOICE
FROM REAL BACKING TRACKS

Music set		SNR (dB)	
Singer	Instrument	Result	Ideal
Male	Real BKT	17.02	20.73
Female	Real BKT	8.12	15.99
Average		12.57	18.36

A. Singing Voice Separation Results

Single-Instrument Backing Track

The SNR results from this experiment are shown in Table 1; P, L, and H represent percussion, low-pitch instrument, and high-pitch instrument, respectively. We also add the ideal SNR in the last column of result tables to give some information about the SNR value in an ideal case. This ideal SNR is the ratio between the approximated mixture from NMF and the original input mixture. Although this value is the ratio between the approximated and the real mixture, which yields some error, it is the best SNR achievable from NMF approach. In addition, these two mixtures are essentially indistinguishable for human auditory system. These ideal SNR values do vary depending on the type of instruments as well as singers.

From Table 1, the female singing voice, which produces higher pitch compare to the male's, accompanied by randomly selected low-pitch and percussion instruments as the backing track, results in higher SNR. In contrast to the male singing voice, the highest SNR is from the high-pitch instruments. This experiment does reconfirm our hypothesis that the performance of the singing voice separation is higher if the sound sources are different in aspect of pitch.

Mixed-Instrument Backing Track

In the previous experiment, we are interested in the effect of only one single instrument which is still far from the music reality. Hence, in this phase, we start mixing instruments together as backing track. The results are shown in Table 2.

The average result does degrade from the first scenario as expected, which is caused by the higher number of instrumental sounds. However, we also get similar trend in the results, where the female singing voice accompanied by percussion and low-pitch instruments does perform better.

Real Backing Track

In the previous two experiments, we randomly select the singing voice and the instruments for the mixture, whose elements may not be entirely correlated. So, in this final phase, we test our on the real scenario where the singing voice is accompanied by its real backing track which contains high pitch between the singing voice and the real backing track. The results are shown in Table 3.

As expected, the average results from Table 3 are inferior to the two previous scenarios due to the diversity of accompanied instruments in real backing track. Nevertheless, when we analyze each individual case, i.e., male and female singing voice, the result of separated male singing voice is relatively high comparing to the ideal SNR; the separated female singing voice result has much lower SNR. By looking at the music we test, we found that a large number of instruments (up to seven) in this real backing track are mostly high-pitch (high-frequency) instruments, thus interfering with the female singing voice.

The spectra in Fig. 1 are of this female mixture. Fig. 1A is spectra of the mixture shown in grayscale where black and white colors represent high and low amplitude of the frequency, respectively. Fig. 1B and 1C are spectra of the separated singing voice and pure original singing voice, respectively. As we can see from the Fig. 1A, there exists the frequency pattern mixed with other chaotic pattern of frequencies, including the rhythmic along the time and other instruments that produce the same range of singing voice frequency. In addition, we can notice that in Fig. 1B, the original singing voice still contains some noise at about 5 kHz, 2.4 kHz, and

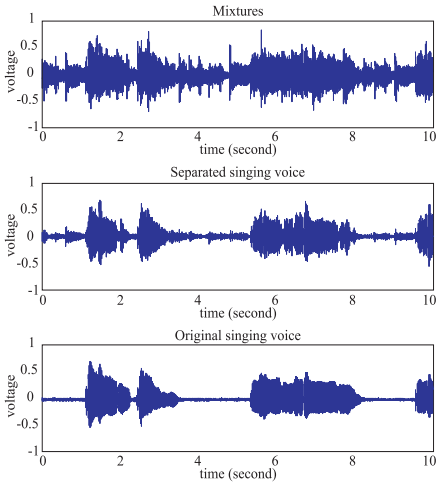


Fig. 4. Example of separated singing voice

other lower frequencies for the whole period of time. For these reasons, the separated singing voice gives quite low SNR. However, by comparing Fig. 1B with the original singing voice in Fig. 1C, the shape of the spectra are well matched and most of instrumental sounds can be removed.

Nonetheless, from the human perspective, using pure SNR to evaluate the method can be seen as an extreme case since the noise produced in many cases that can be clearly reflect in SNR value is actually undetectable in human ears. Fig. 4 illustrates the mixture, separated singing voice, and the original singing voice of this female singing voice accompanied by real backing track in wave form, in other words, it is another visualization of the Fig. 1A, 1B, and 1C, respectively. We can also see from Fig. 4 that the separated singing voice extracted by our proposed method is highly similar to the pure original singing voice.

B. Pitch Detection Results

To demonstrate the utility of our approach, we do the another experiment to test the pitch detection of the separated singing voice compared with the original singing voice using Praat [1]. The accuracy of pitch detection is shown in Table 4.

We can see that the separated singing voice has much lower pitch errors (in Hertz), comparing with the errors from the full mixture of the song. These errors are considered extremely low, considering the pitch difference between any two adjacent notes differ in relatively larger scale in Hertz.

VI. Conclusion and Discussion

In this work, we propose the singing voice separation method that utilizes Non-negative Matrix Factorization (NMF) algorithm. The NMF has distinctive ability of decomposing the spectrum of mixtures. With our proposed method, the singing voice can be separated from the mono-channel music mixtures. The results from the experiments including the satisfying signal and spectrum demonstrate the effectiveness of our method in singing voice separation. Moreover, from our experiments, we can conclude that the performance of the proposed method is affected by the type of instrument and singer. Pitch differences between singing voice and the musical instrument also yield higher result than the more similar ones. In addition, the number of instruments also affects the performance. This supports the prior knowledge about the sound source separation work. In future work, we plan to study the effects of different instruments to the method and develop the algorithm to relieve the effect of factors that cause the reduction of SNR.

TABLE 4.
PITCH DETECTION ACCURACY COMPARISONS OF SEPARATED SINGING VOICE AND FULL MIXTURE

	Mean error (Hz)	
	Separated singing voice	Full mixture
All	28	114
Male	37	94
Female	20	133

References

[1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.5.12) [Computer program]," 2007.
[2] Y. Feng, Y. Zhuang, and Y. Pan, "Popular song retrieval based on singing matching," in Proceedings of 3rd IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, 2002, pp. 639-646.
[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, 2001, pp. 556-562.
[4] Y. Li and D. L. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1475-1487, 2007.
[5] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 2005.
[6] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," Journal of the Acoustical Society of America, vol. 114, pp. 2236-2252, 2003.
[7] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 330-341, 2006.
[8] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in 6th International Conference on Music Information Retrieval, London, UK, 2005, pp. 337-334.
[9] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in Proceedings of the DMRN Summer Conference, Glasgow, 2005.
[10] Y.-G. Zhang and C.-S. Zhang, "Separation of music signals by harmonic structure modeling," in Advances in Neural Information Processing Systems, 2006, pp. 1617-1624.

ภาคผนวก ข

ผลงานตีพิมพ์อื่น ๆ

บทความทางวิชาการเรื่อง “การแปลผลและการบรรยายรูปภาพตารางสำหรับผู้พิการทางสายตา” โดยวงศ์ยศ เกิดศรี อังคนา จันทร์รุ่งอุทัย และโชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการ “11th National Computer Science and Engineering Conference (NCSEC 2007)” ซึ่งจัดขึ้น ณ กรุงเทพมหานคร ประเทศไทย ระหว่างวันที่ 19-21 พฤศจิกายน 2550

National Computer Science and Engineering Conference (NCSEC2007)
19 - 21 November 2007

P123

การแปลผลและการบรรยายรูปภาพตารางสำหรับผู้พิการทางสายตา
Tabular Image Translation and Description for Visually Impaired People

วงศ์ยศ เกิดศรี อังคนา จันทร์รุ่งอุทัย และ โชติรัตน์ รัตนามหัทธนะ
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ถนนพญาไท แขวงวังใหม่ เขตปทุมวัน กรุงเทพมหานคร 10330
อีเมล: Wongyos.K@student.chula.ac.th, {g49ach, ann}@cp.eng.chula.ac.th

บทคัดย่อ

ปัจจุบันผู้พิการทางสายตาสามารถอ่านเอกสารอิเล็กทรอนิกส์บนเครื่องคอมพิวเตอร์ได้โดยใช้โปรแกรมอ่านหน้าจอ ซึ่งอ่านออกเสียงได้เฉพาะข้อมูลที่เป็นตัวอักษรเท่านั้น ไม่สามารถอ่านข้อมูลที่เป็นรูปภาพได้ ทำให้ผู้พิการทางสายตาไม่อาจทราบถึงรายละเอียดของข้อมูลรูปภาพเหล่านั้น โดยปกติแล้วรูปภาพส่วนใหญ่จะถูกบรรยายให้ผู้พิการทางสายตาเข้าใจผ่านการใช้มือสัมผัสโดยใช้เครื่องพิมพ์อักษรเบรลล์หรือเครื่องแสดงผลเบรลล์ในการแสดงรายละเอียดของรูปภาพ แต่วิธีการดังกล่าวมีค่าใช้จ่ายสูง ใช้เวลานาน และไม่สะดวกกับการใช้งาน งานวิจัยนี้จึงได้นำเสนอวิธีการแปลผลและบรรยายรูปภาพแบบใหม่ด้วยการแปลรูปภาพให้เป็นข้อความบรรยาย แล้วใช้โปรแกรมอ่านหน้าจออ่านข้อความเหล่านั้นขึ้นมาแทน โดยใช้หลักการของการประมวลผลภาพดิจิทัลร่วมกับการรู้จำตัวอักษรในการประมวลผล ซึ่งใช้ตัวอย่างรูปภาพตารางเป็นกรณีศึกษา

คำสำคัญ การแปลความหมายรูปภาพ การบรรยายรูปภาพ การประมวลผลภาพ ผู้พิการทางสายตา โปรแกรมอ่านหน้าจอ เทคโนโลยีอำนวยความสะดวก

Abstract

Nowadays, visually impaired people can read electronic documents on the computer by using some screen reader software. Nevertheless, these software products only work on the textual parts, but they are unable to read nor describe images. Consequently, the visually impaired people may not understand the detail of the images. The task of translating and describing images is usually achieved by printing the images with the Braille printer or showing them via the Braille display; as a result, the visually impaired people can touch the images by their hands. However, the process is very time consuming, costly, and inconvenient. This paper proposes a new concept based on Digital Image Processing and the Optical Character Recognition (OCR) techniques for describing and translating the images into textual description which is readable from the screen reader software. We demonstrate the utility of our approach by testing on a set of tabular images.

Key Words: Image Translation, Image Description, Image Processing, Visually Impaired People, Screen Reader, Assistive Technology

1. บทนำ

ในปัจจุบันเทคโนโลยีอำนวยความสะดวก หรือเทคโนโลยีเพื่อช่วยเหลือ (Assistive Technology: AT) ได้กลายเป็นเทคโนโลยีสำคัญสำหรับผู้พิการ ที่จะคอยช่วยเหลือและเติมเต็มให้ผู้พิการสามารถทำงานและใช้ชีวิตประจำวันได้อย่างปกติสุขเช่นเดียวกับคนทั่วไป อย่างไรก็ตาม ยังมีประเด็นที่ท้าทายอีกมากที่ต้องการการพัฒนาอย่างต่อเนื่อง ซึ่งหนึ่งในนั้นคือ การแปลผลและการบรรยายรูปภาพให้กับผู้พิการทางสายตา

การแปลผลและการบรรยายรูปภาพให้กับผู้พิการทางสายตานั้น เป็นเรื่องที่มีการวิจัยและพัฒนาอย่างต่อเนื่องมาโดยตลอด ซึ่งเริ่มจากการแปลงรูปภาพจากหนังสือหรือจากแผ่นกระดาษ ให้เป็นรูปภาพที่ประกอบด้วยสัญลักษณ์ของเบรลล์โดยใช้หลักการพิมพ์ภาพด้วยเครื่องพิมพ์เบรลล์ (Braille printer) หรือใช้การแสดงผลภาพออกทางเครื่องแสดงผลเบรลล์ (Braille display) เพื่อให้ผู้พิการทางสายตาใช้มือสัมผัสและเข้าใจถึงลักษณะต่างๆ ของรูปภาพได้ แต่วิธีการดังกล่าวต้องใช้เวลานานในการสร้างแผนภาพเบรลล์ มีค่าใช้จ่ายสูง และไม่สะดวกในการใช้งานจริง

รูปภาพส่วนใหญ่ในปัจจุบันถูกจัดเก็บไว้ในรูปแบบของรูปภาพดิจิทัลซึ่งอาจปรากฏบนเอกสารอิเล็คทรอนิกส์ต่างๆ ในเครื่องคอมพิวเตอร์ ซึ่งโดยทั่วไปผู้พิการทางสายตาสามารถเข้าถึงเอกสารเหล่านี้ได้โดยใช้โปรแกรมอ่านหน้าจอ (Screen reader) [1] เช่น JAWS [2] Apple VoiceOver [3] Adobe Read Out Loud [4] พีพีเอตาทิพย์ (PPA Tatip) [5] เป็นต้น อย่างไรก็ตามโปรแกรมเหล่านี้สามารถอ่านออกเสียงได้เฉพาะในส่วนของข้อความหรือตัวหนังสือ ไม่สามารถอ่านออกเสียงและแสดงรายละเอียดของรูปภาพที่ปรากฏอยู่บนหน้าจอ ทำให้ผู้พิการทางสายตาไม่สามารถเข้าใจ และทราบถึงความหมายของรูปภาพเหล่านั้นได้ ข้อจำกัดดังกล่าวจึงเป็นจุดเริ่มต้นของงานวิจัยนี้ในการแปลผลและบรรยายรูปภาพให้กับผู้พิการทางสายตาได้เข้าใจถึงความหมายและรายละเอียดต่างๆ ของรูปภาพ อย่างไรก็ตามรูปภาพที่ปรากฏอยู่บนเอกสารอิเล็กทรอนิกส์ในปัจจุบันนั้นมีหลายประเภท เช่น ภาพวาด ภาพเขียน ภาพถ่าย ภาพกราฟิกส์ เป็นต้น ซึ่งแต่ละประเภทมีโครงสร้างและลักษณะของภาพที่แตกต่างกัน จึงทำให้มีวิธีการแปลผลและบรรยายภาพที่แตกต่างกัน รวมทั้งต้องใช้วิธีที่ซับซ้อนในการประมวลผล ซึ่งยังเป็นเรื่องยากในการคิดหาวิธีที่มีประสิทธิภาพในการทำงาน งานวิจัยนี้จึงเริ่มต้นด้วยวิธีการประมวลผลที่ไม่ซับซ้อนเกินไปโดยเลือกรูปภาพตาราง (Tabular Image) เป็นกรณีศึกษา ซึ่งใช้หลักการทำงานร่วมระหว่างโปรแกรมแปลความหมายรูปภาพตารางให้อยู่ในรูปของข้อความบรรยาย และโปรแกรมอ่านหน้าจอในการอ่านออกเสียงข้อความที่ได้ให้กับผู้พิการทางสายตาฟัง

เหตุผลที่แปลความหมายรูปภาพให้อยู่ในรูปแบบของข้อความบรรยาย แต่ไม่แปลให้อยู่ในรูปแบบของเสียงพูดเนื่องด้วย โดยปกติแล้วโปรแกรมอ่านหน้าจอสามารถอ่านออกเสียงข้อความที่ปรากฏอยู่บนจอภาพได้ในทุกกรณี ดังนั้นการแปลให้อยู่ในรูปแบบของข้อความก็เพียงพอแล้วเพื่อให้โปรแกรมอ่านหน้าจอนำข้อความที่ได้นั้นไปประมวลผลและอ่านออกเสียงต่อไป ซึ่งจะช่วยลดความซับซ้อนของระบบลง และลดค่าใช้จ่ายในการพัฒนาระบบได้ โดยวิธีการทำงานดังกล่าวใช้กระบวนการประมวลผลภาพดิจิทัล (Digital Image Processing) และกระบวนการรู้จำตัวอักษร (Optical Character Recognition: OCR) เข้ามาประมวลผล

2. งานวิจัยที่เกี่ยวข้อง

การแปลผลและการบรรยายรูปภาพให้กับผู้พิการทางสายตาเป็นเรื่องที่ได้รับความสนใจ โดยมีงานวิจัยจำนวนหนึ่งที่พยายามหาวิธีการและแนวทางในการแปลผลและบรรยายรูปภาพให้อยู่ในรูปแบบต่างๆ โดยเมื่อปี ค.ศ. 2001 Hesham M. Kame และคณะ [6] ได้นำเสนอโปรแกรมวาดภาพไอซีทูดี (IC2D) ที่ช่วยให้ผู้พิการทางสายตาสามารถเข้าถึงรูปภาพได้ด้วยวิธีการติดฉลาก (Labeling Method) ร่วมกับการจัดกลุ่มของจุด เส้นตรง และรูปเรขาคณิต ซึ่ง

โปรแกรมนี้ช่วยให้ผู้พิการทางสายตาสามารถสื่อสารกับคนปกติโดยการวาดและการอ่านรูปภาพที่วาดด้วยโปรแกรมไอซีทูดีได้ แต่ยังมีข้อจำกัดสำหรับการอ่านรูปภาพที่ต้องการรายละเอียดสูงหรือรูปภาพที่ใช้สำหรับการอธิบายผล เช่น กราฟ ตาราง แผนผัง เป็นต้น ต่อมาในปี ค.ศ. 2004 R. Iglesias และคณะ [7] ได้เสนอวิธีการใหม่สำหรับการเข้าถึงรูปภาพกราฟิกส์คอมพิวเตอร์สามมิติด้วยการสัมผัสอุปกรณ์แสดงผลพร้อมกับการฟังเสียง ซึ่งวิธีการดังกล่าวทำให้ผู้พิการทางสายตาสามารถเข้าใจความหมายของรูปภาพได้ดีและเร็วยิ่งขึ้น เช่นเดียวกับงานวิจัยของ Stephen E. Krufka และคณะ [8] ที่ได้นำเสนอหลักการของการทำกราฟิกส์เวกเตอร์แบบปรับขนาดได้ (Scalable Vector Graphics) ในการสร้างรูปภาพแบบสัมผัส (Tactile Image) เพื่อให้ผู้พิการทางสายตาได้สัมผัสและเข้าใจถึงรายละเอียดของรูปภาพได้ สอดคล้องกับงานวิจัยของ Richard E. Ladner และคณะ [9] ที่พยายามสร้างรูปแบบการบรรยายรูปภาพด้วยการสัมผัสแบบอัตโนมัติซึ่งทำให้ผู้พิการทางสายตาสามารถเข้าใจรูปภาพที่กำลังเข้าถึง ณ เวลานั้นได้ดียิ่งขึ้น อีกทั้ง David K. McGookin และคณะ [10] ได้ออกแบบและพัฒนาเครื่องมือบรรยายกราฟแท่งหรือซาวด์บาร์ (SoundBar) เพื่ออ่านออกเสียงผลลัพธ์ ค่าระดับและตำแหน่งของกราฟให้ผู้พิการทางสายตาฟังได้ อย่างไรก็ตามงานวิจัยเหล่านี้ [7-10] ต้องใช้อุปกรณ์และเครื่องมือเสริมในการทำงานซึ่งเป็นอุปกรณ์เฉพาะที่มีข้อจำกัดในการใช้งานและมีราคาแพง ทำให้ผู้พิการทางสายตาส่วนใหญ่ไม่มีโอกาสได้ใช้งานอุปกรณ์เหล่านี้ เมื่อปี ค.ศ. 2005 Steve Murphy [11] จากบริษัทไอบีเอ็ม (IBM) ได้มีแนวคิดที่จะเพิ่มความสามารถของผู้พิการทางสายตาในการเข้าถึงรูปภาพในหน้าเว็บเพจหรือเอกสารเอชทีเอ็มแอล (HTML) ด้วยการสร้างป้ายระบุหรือแทก (Tag) พิเศษเพื่อช่วยในการบรรยายรูปภาพที่ประกอบอยู่ในหน้าเว็บเพจ ซึ่งทำให้ผู้พิการทางสายตาสามารถเข้าใจรายละเอียดของรูปภาพได้มากยิ่งขึ้น แต่สามารถประมวลผลกับรูปภาพที่ปรากฏอยู่บนหน้าเว็บเพจได้เพียงอย่างเดียวเท่านั้น

จากงานวิจัยที่กล่าวมาข้างต้นยังมีข้อจำกัดคือ ใช้เวลานาน เสียค่าใช้จ่ายสูง และไม่สะดวกต่อการใช้งาน จึงทำให้เกิดงานวิจัยนี้ขึ้น

3. สถาปัตยกรรมของระบบ

โดยปกติแล้วเมื่อโปรแกรมอ่านหน้าจอ [2-5] อ่านพบรูปภาพในเอกสารอิเล็กทรอนิกส์ ก็จะข้ามการทำงานในส่วนของรูปภาพนั้นไป หรืออาจแจ้งให้ผู้ใช้ทราบแต่เพียงว่า ณ ตำแหน่งที่อ่านในเวลานั้นคือรูปภาพ แต่ไม่สามารถอธิบายหรือบรรยายถึงรายละเอียดภายในรูปภาพเหล่านั้นออกมาได้ งานวิจัยนี้จึงมีแนวคิดที่จะออกแบบโปรแกรมย่อยสำหรับแปลผลและบรรยายรูปภาพเหล่านั้น เพื่อให้ผู้พิการทางสายตาได้อ่านและสามารถเข้าใจถึงรายละเอียดของเอกสารซึ่งมีรูปภาพเป็นองค์ประกอบได้ดียิ่งขึ้น

ตามโครงสร้างและสถาปัตยกรรมของงานวิจัยนี้ เมื่อโปรแกรมอ่านหน้าจออ่านพบรูปภาพจากเอกสารอิเล็กทรอนิกส์ใดๆ จะทำการสลับการทำงานมาที่โปรแกรมย่อยที่ทำหน้าที่แปลผลและบรรยายรูปภาพเพื่อประมวลผลกับรูปภาพที่ตรวจพบ แล้วจึงย้อนกลับมาทำงานกับเอกสารอิเล็กทรอนิกส์เดิมอีกครั้งหนึ่ง โดยขั้นตอนของสถาปัตยกรรมระบบได้แสดงไว้ตามรูปที่ 1 ดังรายละเอียดต่อไปนี้

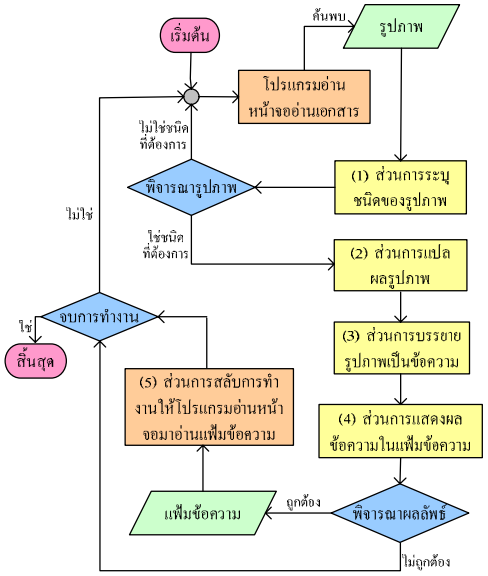
(1) *ส่วนการระบุชนิดของรูปภาพ* ทำหน้าที่บอกชนิดของรูปภาพที่ตรวจพบว่าเป็นรูปภาพชนิดใด เช่น ตาราง กราฟ แผนภูมิ แผนผัง หรือภาพวาด เป็นต้น เพื่อตัดสินใจว่าการแปลผลสามารถทำต่อไปได้หรือไม่

(2) *ส่วนการแปลผลรูปภาพ* ทำหน้าที่แปลผลรูปภาพตามอัลกอริทึมสำหรับรูปภาพแต่ละชนิด

(3) *ส่วนการบรรยายรูปภาพเป็นข้อความ* ทำหน้าที่บรรยายรูปภาพที่ได้จากส่วนของขั้นตอนที่ 2 ออกมาเป็นตัวหนังสือหรือข้อความบรรยาย

(4) *ส่วนการแสดงผลข้อความในแฟ้มข้อความ* จะนำข้อความที่ได้มาแสดงผลในแฟ้มข้อความ (Text File) ใหม่ที่อยู่ภายนอกเอกสารเดิม

(5) *ส่วนการสลับการทำงานให้โปรแกรมอ่านหน้าจอมาอ่านแฟ้มข้อความ* เป็นการสลับการทำงานของโปรแกรมอ่านหน้าจอเพื่ออ่านแฟ้มข้อความแล้วจึงสลับกลับไปอ่านยังแฟ้มเอกสารเดิม



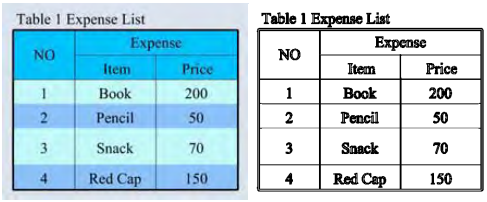
รูปที่ 1 สถาปัตยกรรมแนวคิดของระบบการแปลผลและการบรรยายรูปภาพสำหรับผู้พิการทางสายตา

งานวิจัยนี้ได้เลือกรูปภาพตารางซึ่งมีโครงสร้างที่ไม่ซับซ้อนมาใช้เป็นกรณีศึกษาและรูปภาพทดสอบ โดยเจาะจงในรายละเอียดการทำงานของขั้นตอนที่ (1) (2) และ (3) เป็นหลัก ซึ่งขั้นตอน วิธีการ และผลลัพธ์จากการประมวลต่างๆ ได้แสดงไว้ในหัวข้อถัดไป

4. วิธีการแปลผลและบรรยายรูปภาพ

งานวิจัยนี้ใช้เทคนิคการประมวลผลภาพดิจิทัล และการรู้จำตัวอักษรเพื่อใช้แปลผลและบรรยายรูปภาพ เนื่องจากรูปภาพมีหลายชนิด โดยแต่ละชนิดมีลักษณะเฉพาะที่แตกต่างกัน จึงมีรายละเอียดของขั้นตอนการแปลผลและบรรยายรูปภาพที่แตกต่างกัน งานวิจัยนี้จึงเลือกเฉพาะรูปภาพตารางเพื่อใช้เป็นกรณีศึกษา เนื่องจากรูปภาพตารางเป็นรูปภาพพื้นฐานที่มีการใช้งานกันอย่างกว้างขวาง และมีโครงสร้างที่ไม่ซับซ้อนในการประมวลผล โดยขั้นตอนการแปลผลและการบรรยายรูปภาพสามารถแสดงรายละเอียดได้ดังต่อไปนี้

ก.



ข. ค.

รูปที่ 2 ตัวอย่างรูปภาพตาราง

ก. รายละเอียดของรูปภาพตาราง ซึ่งประกอบด้วย ความกว้างและความสูงของภาพ ความกว้างและความสูงจริงของภาพตาราง และความกว้างและความสูงของตาราง

ข. รูปภาพตารางต้นฉบับ (ก่อนการประมวลผล)

ค. รูปภาพตารางที่ผ่านการหาเส้นขอบ และระบุขอบเขตของรูปภาพ (หลังการประมวลผล)

4.1 การนิยามรูปภาพตาราง

ตาราง ประกอบด้วยแถว (Rows) และหลัก (Columns) ที่มีการจัดเรียงกันอย่างเป็นระเบียบ [12] สำหรับการนิยาม *รูปภาพตาราง* งานวิจัยนี้ได้มีการกำหนดรายละเอียดและคุณลักษณะเฉพาะต่างๆ ของรูปภาพตารางดังนี้

1. เป็นแฟ้มรูปภาพ เช่น แฟ้มนามสกุล .gif .jpg .bmp .png เป็นต้น
2. เป็นรูปตารางที่ประกอบด้วยแถวและหลัก
3. ใช้เส้นตรงในการแบ่งแต่ละแถวและหลัก
4. ไม่เป็นรูปภาพที่มีหนึ่งแถวและหนึ่งหลัก

5. ช่องแต่ละช่องในรูปภาพตาราง เรียกว่าเซลล์ซึ่งสามารถเชื่อมต่อผสานกันได้

4.2 โครงสร้างของรูปภาพตาราง

การศึกษาโครงสร้างของรูปภาพตารางเป็นการวิเคราะห์ถึงชนิดของรูปภาพตารางและลักษณะเฉพาะของรูปภาพตารางโดยใช้นิยามจากหัวข้อ 4.1 ซึ่งโครงสร้างของรูปภาพตารางนั้นประกอบด้วย (1) ชื่อตารางซึ่งอาจปรากฏอยู่ส่วนบนสุดหรือล่างสุดของรูปภาพตาราง และ (2) ตัวตารางซึ่งเป็นส่วนที่เหลือทั้งหมดของรูปภาพตาราง

ส่วนของตัวตารางนั้นจะประกอบไปด้วยเซลล์ต่างๆ โดยที่เซลล์แต่ละเซลล์นั้นเป็นรูปสี่เหลี่ยมที่ถูกแบ่งด้วยเส้นตรงซึ่งมีจุดพิกัดตำแหน่งของมุมทั้งสี่ของเซลล์ และอาจมีตัวหนังสืออยู่ภายในเซลล์เหล่านั้น ดังแสดงตัวอย่างในรูปที่ 2ก.

4.3 การแปลผลรูปภาพตาราง

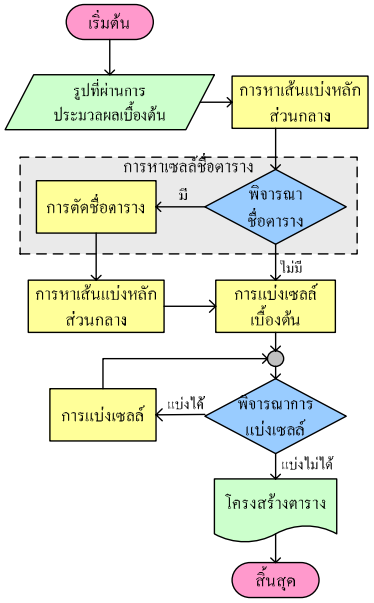
กระบวนการแปลผลรูปภาพตารางเป็นการนำรายละเอียดของโครงสร้างรูปภาพตาราง มาพิจารณาเพื่อตรวจสอบว่ารูปภาพใดบ้างที่มีลักษณะเป็นรูปภาพตารางและไม่ใช่รูปภาพตาราง ซึ่งมีขั้นตอนดังต่อไปนี้

4.3.1 การประมวลผลเบื้องต้น

โดยทั่วไปรูปภาพตารางจะมีส่วนของพื้นที่ว่างบริเวณขอบของรูปภาพก่อนถึงเส้นขอบของรูปตารางที่แท้จริง ซึ่งขั้นตอนนี้เป็นการหาเส้นขอบและระบุขอบเขตของรูปภาพตารางที่แท้จริงโดยใช้เทคนิคการประมวลผลภาพดิจิทัลและแปลงให้เป็นภาพขาวดำเพื่อลดทรัพยากรของระบบและเพื่อง่ายต่อการคำนวณเกี่ยวกับจำนวนจุดภาพ (Pixel) การหาขอบเขตของรูปภาพตารางที่แท้จริงทำได้โดยการหาขอบเขตสี่เหลี่ยมที่เล็กที่สุดที่สามารถบรรจุจุดภาพสีดำไว้ได้ทั้งหมด ผลลัพธ์ของขั้นตอนนี้แสดงไว้ในรูปที่ 2ค.

4.3.2 การหาโครงสร้างของตาราง

การหาโครงสร้างของตารางจะเริ่มต้นจากการหาส่วนของชื่อตาราง และส่วนของเซลล์ที่เป็นเนื้อหาภายในตารางขนาดแถวคูณหลัก โดยใช้วิธีการหาเส้นตรงที่ยาวเพียงพอในแนวนอนและแนวตั้ง เรียกว่าเส้นแบ่งหลักส่วนกลาง เพื่อแบ่งแถวและหลักในเบื้องต้นก่อน หลังจากนั้นแต่ละเซลล์จะถูกพิจารณาว่ามีเส้นแบ่งหลักเฉพาะที่ซึ่งสามารถแบ่งเซลล์นั้นออกได้อีกหรือไม่ โปรแกรมจะแบ่งจนกว่าไม่มีเซลล์ใดสามารถแบ่งได้อีก ขั้นตอนการทำงานแสดงดังรูปที่ 3 ซึ่งรายละเอียดของขั้นตอนย่อยต่างๆ มีดังนี้



รูปที่ 3 ผังงานของการค้นหาโครงสร้างตาราง

(1) การหาเส้นแบ่งหลักส่วนกลาง ทำได้โดยการคำนวณอัตราส่วนของผลรวมของสี (กำหนดให้ค่าสีดำเป็น 1 สีขาวเป็น 0) ในแถวเดียวกันเทียบกับความกว้างของภาพแล้วเลือกเก็บตำแหน่งแถวที่มีค่าอัตราส่วนนี้มากกว่าค่าอัตราส่วนขั้นต่ำของเส้นหลัก (α) สำหรับค่าตำแหน่งของหลักก็สามารถคำนวณได้ในทำนองเดียวกัน กำหนดเซตของตำแหน่งแถว (R) และเซตของตำแหน่งหลัก (C) ดังสมการที่ (1) และ (2) ตามลำดับ

$$R = \left\{ r \mid 1 \le r \le h, \frac{\sum_{c=1}^{w} B(r,c)}{w} \ge \alpha \right\} \qquad (1)$$

$$C = \left\{ c \mid 1 \le c \le w, \frac{\sum_{r=1}^{h} B(r,c)}{h} \ge \alpha \right\} \qquad (2)$$

เมื่อ h และ w คือความสูงและความกว้างของรูปภาพ $B(r,c)$ คือค่าสีในจุดภาพแถวที่ r หลักที่ c และค่าสัดส่วนขั้นต่ำของเส้นหลัก (α) นี้จะไม่เกิน 1.00 และไม่ควรน้อยกว่า 0.90 เพื่อให้มั่นใจได้ว่า ตำแหน่งที่เลือกมาเป็นเส้นตรง ไม่ใช่ส่วนของตัวอักษรที่มีความสูงใกล้เคียงกับความสูงของเซลล์ที่อักษรนั้นอยู่

สมาชิกในแต่ละเซตจะถูกนำมาเรียงลำดับ และหาผลต่างของคู่ตัวเลขที่ติดกัน คู่ที่มีผลต่างมากกว่าค่าขั้นต่ำของเซลล์ (β) หรือค่าจำนวนจุดภาพขั้นต่ำของความกว้างและความยาวของเซลล์ จะถูกเลือกไปใช้ในการแบ่งเซลล์เบื้องต้น

(2) การหาชื่อตาราง และการแบ่งเซลล์เบื้องต้น ชื่อตารางจะอยู่ได้ทั้งบริเวณเหนือหรือใต้ของตาราง ซึ่งสามารถหาได้โดยพิจารณาขอบเขตของรูปภาพที่ตำแหน่งบนสุดหรือล่างสุดที่ไม่ใช่เส้นตรง (ตำแหน่งของเส้นตรงได้คำนวณไว้แล้วจากสมการที่ (1)) โดยที่ชื่อตารางนี้จะถูกเก็บเป็นส่วนหนึ่งของโครงสร้างตาราง และขณะเดียวกันก็ทำการลดขนาดของภาพลงเป็นเพียงส่วนของตารางที่แท้จริง จากนั้นหาเส้นแบ่งหลักส่วนกลางใหม่อีกครั้งเพื่อบอกแถวและหลักของตาราง และเก็บข้อมูลของเซลล์เริ่มต้นในอาเรย์ของเซลล์

(3) การตรวจสอบเซลล์ ในขั้นตอนนี้จะทำการตรวจสอบทุกๆ เซลล์ว่าสามารถแบ่งออกได้หรือไม่จากเซลล์ทางด้านซ้ายไปขวา และจากด้านบนลงล่าง โดยใช้หลักการเดียวกับการหาเส้นแบ่งหลักส่วนกลาง แต่ในที่นี้จะเป็นการหาเส้นแบ่งหลักเฉพาะเซลล์แทน หากพบว่าเซลล์ใดสามารถแบ่งได้ หรือมีจำนวนแถวหรือหลักที่มากกว่า 1 แล้ว จะแบ่งเซลล์เหล่านั้นอีกครั้งจนกว่าจะไม่มีเซลล์ใดที่สามารถแบ่งได้อีก ซึ่งถือว่าสิ้นสุดในขั้นตอนนี้

(4) การแบ่งเซลล์ ขั้นตอนนี้จะทำงานเมื่อขั้นตอนการตรวจสอบเซลล์ตรวจพบว่ามีเซลล์ที่สามารถแบ่งได้ กรณีการแบ่งเซลล์เพิ่มจำนวนแถว เซลล์ทั้งหมดที่อยู่ด้านใต้เซลล์ที่พิจารณาจะถูกเลื่อนลงตามจำนวนแถวที่เพิ่ม และแบ่งเซลล์ที่พิจารณาออก โดยการใส่ข้อมูลในโครงสร้างตารางตามขอบเขตที่ได้จากการหาเส้นแบ่งเฉพาะที่ ส่วนเซลล์อื่นๆ ที่เคยอยู่ในแถวเดียวกันจะแบ่งเป็นสองด้าน คือ เซลล์ทางด้านซ้ายซึ่งเคยถูกตรวจสอบมาก่อนหน้านี้แล้วว่าไม่สามารถแบ่งได้ซึ่งจะถูกคัดลอกมาทั้งหมดเพื่อใส่ให้กับเซลล์ที่เพิ่มทุกๆ แถว ส่วนเซลล์ทางด้านขวานั้นจะถูกตรวจสอบก่อนว่าแบ่งแถวได้หรือไม่ ถ้าแบ่งไม่ได้ก็จะทำเช่นเดียวกับเซลล์ทางซ้าย แต่ถ้าแบ่งได้ก็จะแบ่งตามจำนวนเดียวกับเซลล์ที่กำลังพิจารณา

ตัวอย่างรูปตารางต้นฉบับดังแสดงในรูปที่ 2ข. นั้น เมื่อผ่านขั้นตอนการหาโครงสร้างของตารางเสร็จสมบูรณ์แล้วจะได้เป็นอาร์เรย์ของเซลล์ตารางดังแสดงในรูปที่ 4

Table 1 Expense List

Title

NO	Expense	Expense
Cell-1-1	Cell-1-2	Cell-1-3
NO	Item	Price
Cell-2-1	Cell-2-2	Cell-2-3
1	Book	200
Cell-3-1	Cell-3-2	Cell-3-3
2	Pencil	50
Cell-4-1	Cell-4-2	Cell-4-3
3	Snack	70
Cell-5-1	Cell-5-2	Cell-5-3
4	Red Cap	150
Cell-6-1	Cell-6-2	Cell-6-3

รูปที่ 4 ตัวอย่างอาเรย์ของเซลล์จากตารางต้นฉบับของรูปที่ 2ข. ที่ผ่านการหาโครงสร้างตาราง

4.3.3 การรู้จำตัวอักษร

เมื่อได้โครงสร้างของตารางที่สามารถอ้างถึงตำแหน่งของแต่ละเซลล์ในรูปภาพตารางที่รับเข้ามาและถูกแบ่งออกเป็นเซลล์แต่ละเซลล์เรียบร้อยแล้ว ขั้นตอนนี้จะเป็นกระบวนการรู้จำตัวอักษร ซึ่งงานวิจัยนี้ได้เลือกใช้โปรแกรมสำเร็จรูป SimpleOCR [13] ในการประมวลผลและเป็นเพียงขั้นตอนอย่างง่ายเท่านั้น โดยเพียงต้องการที่จะเลือกตัวอย่างของอัลกอริทึมรู้จำตัวอักษรในการทดสอบรูปภาพของเซลล์แต่ละเซลล์ที่ได้จากขั้นตอนที่ 4.3.2 ว่า

สามารถรู้จำตัวอักษรได้หรือไม่ ซึ่งความถูกต้องของการรู้จำตัวอักษรจะขึ้นอยู่กับกระบวนการทำงานของอัลกอริทึมรู้จำตัวอักษรที่นำมาใช้

4.4 การบรรยายรูปภาพตาราง

ในขั้นตอนนี้ โครงสร้างของตารางที่ได้จากขั้นตอนที่ 4.3 จะถูกบรรยายเป็นข้อความ ซึ่งใช้การบรรยายเป็นข้อความสั้นๆ เพื่อทำให้ผู้พิการทางสายตาสามารถทำความเข้าใจได้ง่าย และไม่เสียเวลาในการฟัง ดังได้แสดงตัวอย่างคำบรรยายไว้ในรูปที่ 5 ซึ่งข้อความในแต่ละเซลล์จะขึ้นบรรทัดใหม่ ทำให้สามารถกดปุ่มลูกศรขึ้นลงเพื่อเลือกบรรทัดที่จะฟังได้

```
6-Row-3-Column Tabular Image
Title : Table 1 Expense List
Cell-Row1-Column1 : NO
Cell-Row1-Column2 : Expense
Cell-Row1-Column3 : Expense
Cell-Row2-Column1 : NO
Cell-Row2-Column2 : Item
Cell-Row2-Column3 : Price
Cell-Row3-Column1 : 1
Cell-Row3-Column2 : Book
Cell-Row3-Column3 : 200
Cell-Row4-Column1 : 2
Cell-Row4-Column2 : Pencil
Cell-Row4-Column3 : 50
Cell-Row5-Column1 : 3
Cell-Row5-Column2 : Snack
Cell-Row5-Column3 : 70
Cell-Row6-Column1 : 4
Cell-Row6-Column2 : Red Cap
Cell-Row6-Column3 : 150
End of Table
```

รูปที่ 5 ผลลัพธ์การแสดงข้อความคำบรรยายของรูปภาพตารางให้กับผู้พิการทางสายตา

5. การทดลองและการประเมินผล

ในการประเมินผลการแปลผลและการบรรยายรูปภาพตารางนั้น งานวิจัยนี้ได้ใช้ชุดของข้อมูลรูปภาพนามสกุลต่างๆ จำนวน 50 รูปแรกที่เหมาะสมที่ได้จากการค้นหาใน Google ด้วยคำค้น *"Tabular Image"* โดยเกณฑ์การวัดที่ใช้ได้แก่ (1) ความถูกต้องของการระบุได้ว่าเป็นตาราง และ (2) ความถูกต้องของการแบ่งเซลล์ตาราง

ในการทดลอง มีการใส่ค่าพารามิเตอร์สองค่าคือ ค่าสัดส่วนขั้นต่ำของเส้นหลัก (α) และค่าขั้นต่ำของเซลล์ (β) เป็น 0.96 และ 10 จุดภาพ ตามลำดับ ซึ่งเป็นค่าที่ดีที่สุดของค่าต่างๆ ที่ใช้ในการทดสอบ โดยผลการทดสอบแสดงได้ในตารางที่ 1

ตารางที่ 1 ความถูกต้องของอัลกอริทึมของชุดรูปภาพตารางทดสอบจำนวน 50 รูป ด้วยคำค้น *"Tabular Image"* จาก Google

เกณฑ์การวัด	จำนวนรูปภาพ	เปอร์เซ็นต์
ความถูกต้องของการระบุได้ว่าเป็นตาราง	48	96.0%
ความถูกต้องของการแบ่งเซลล์ตาราง	43	89.6%

จากตารางที่ 1 ความถูกต้องของการระบุได้ว่าเป็นรูปตารางจากรูปภาพจำนวน 50 รูป เป็น 96.0 เปอร์เซ็นต์ และความถูกต้องของการแบ่งเซลล์ตารางจากรูปภาพจำนวน 48 รูปที่ระบุได้แล้วว่าเป็นรูปตารางเป็น 89.6 เปอร์เซ็นต์ ซึ่งเมื่อพิจารณาผลการทำงานโดยรวมแล้ว อัลกอริทึมนี้สามารถประมวลผลรูปภาพตารางได้ถูกต้องจำนวน 43 รูป จากรูปภาพทั้งหมด 50 รูป ซึ่งคิดเป็น 86.0 เปอร์เซ็นต์ โดยปัจจัยที่ทำให้เกิดความผิดพลาดสามารถวิเคราะห์ได้เป็นสองกลุ่ม ได้แก่

(1) ความผิดพลาดที่โปรแกรมระบุได้ว่าไม่ใช่รูปภาพตาราง (False Negative) เกิดจากการที่อัลกอริทึมไม่สามารถหาเส้นตรงที่บอกขอบเขตของรูปภาพหรือขอบเขตของเซลล์ได้ ซึ่งอาจเป็นเพราะการใช้สีที่ใกล้เคียงกันมากเป็นตัวแบ่งแถวหรือหลัก รูปภาพมีความไม่สมบูรณ์ รูปภาพมีการหมุน รูปภาพมีจุดรบกวน (Noise) และอาจเกิดกรณีหนึ่งแถวและหนึ่งหลักขึ้น

(2) ความผิดพลาดในการแบ่งแยกเซลล์ เกิดจากในบางเซลล์มีตัวอักษรที่มีความสูงเกือบเท่ากับความสูงของเซลล์นั้นๆ ทำให้อัลกอริทึมประมวลผลได้ว่าตัวอักษรตัวนั้นเป็นเส้นแบ่งหลักเฉพาะของเซลล์ด้วย ทำให้การแบ่งแถวและการแบ่งหลักเกิดความผิดพลาดขึ้น

6. บทสรุปและแนวทางการวิจัยในอนาคต

งานวิจัยนี้มีแนวคิดในการแปลผลและการบรรยายรูปภาพให้ออกมาเป็นตัวหนังสือหรือข้อความคำบรรยาย

เพื่อให้ผู้พิการทางสายตาสามารถใช้โปรแกรมอ่านหน้าจออ่านออกเสียง และเข้าใจถึงรายละเอียดของเอกสารอิเล็กทรอนิกส์เหล่านั้นได้ดียิ่งขึ้น โดยใช้หลักการของการประมวลผลภาพดิจิทัลร่วมกับการรู้จำตัวอักษร งานวิจัยนี้ได้เลือกเฉพาะรูปภาพตารางเป็นกรณีศึกษาเนื่องจากมีการใช้งานอย่างกว้างขวาง และไม่ซับซ้อนต่อการอธิบายและทำการทดลอง ซึ่งต้นแบบอัลกอริทึมการประมวลผลรูปภาพที่นำเสนอได้แก่ การนิยามรูปภาพตาราง การออกแบบโครงสร้างของรูปภาพตาราง การแปลผลรูปภาพตาราง การรู้จำตัวอักษร และการบรรยายรูปภาพตารางเป็นข้อความ ในการประเมินผลการทำงานสามารถแบ่งได้เป็นสองกรณีคือ ความถูกต้องที่ระบุได้ว่าเป็นรูปภาพตาราง และความถูกต้องในการแบ่งแยกเซลล์

งานวิจัยนี้ต้องการนำเสนอแนวคิดใหม่ในการแปลผลและบรรยายรูปภาพ และนำเสนอต้นแบบอัลกอริทึมในการประมวลผลกับรูปภาพตาราง ซึ่งจะเป็นแนวทางหนึ่งในการช่วยเหลือให้ผู้พิการทางสายตาสามารถเข้าถึงรูปภาพและเข้าใจความหมายของรูปภาพได้หลากหลายชนิดต่อไปในอนาคต

แนวทางสำหรับงานวิจัยต่อไป จะมีการนำเสนอวิธีการเพิ่มประสิทธิภาพการทำงานของอัลกอริทึมของงานวิจัยนี้ เพื่อให้สามารถใช้ประโยชน์จากการแปลผลและบรรยายรูปภาพได้ดียิ่งขึ้น เช่น สามารถประมวลผลได้กับรูปภาพชนิดต่างๆ สามารถระบุตำแหน่ง ขอบเขต และขนาดของรูปภาพในเอกสารได้ ซึ่งจะช่วยให้ผู้พิการทางสายตาสามารถเข้าถึงรายละเอียดของรูปภาพได้มากยิ่งขึ้น

7. กิตติกรรมประกาศ

ขอขอบคุณนายคณิตย์ ผามะณี ซึ่งเป็นผู้พิการทางสายตา และเป็นผู้ที่จุดประกายความคิดในการพัฒนางานวิจัยเรื่องนี้ นอกจากนี้ขอขอบคุณคณาจารย์ที่ให้ข้อเสนอแนะที่เป็นประโยชน์ซึ่งประกอบด้วย อ.ดร.โชติ-รัตน์ รัตนามหัทธนะ อ.ดร.อติวงศ์ สุชาโต อ.ดร.พิษณุ คนองชัยยศ และ อ.ดร.อรรถสิทธิ์ สุรฤกษ์

8. เอกสารอ้างอิง

[1] Wikipedia, "Screen reader," Available from: http://en.wiki-pedia.org/wiki/Screen_reader, Access date: July 3, 2007.

[2] JAWS, Available from: http://www.freedomscientific.com, Access date: July 3, 2007.

[3] Apple VoiceOver, Available from: http://www.apple.com/macosx/features/voiceover/, Access date: July 3, 2007.

[4] Adobe Read Out Loud, Available from: http://www.adobe.com/enterprise/accessibility/reader6/sec2.html, Access date: July 3, 2007.

[5] PPA Tatip, Available from: http://www.tabod.net, Access date: July 3, 2007.

[6] H. M. Kamel and J. A. Landay, "The Use of Labeling to Communicate Detailed Graphics in a Non-visual Environment," In Proc. of the Conf. on Human Factors in Computing Systems, Washington, USA, 2001.

[7] R. Iglesias, S. Casado, T. Gutierrez, J. I. Barbero, C. A. Avizzano, S. Marcheschi and M. Bergamasco, "Computer Graphics Access for Blind People through a Haptic and Audio Virtual Environment," In Proc. of 3rd IEEE Int. Workshop on Haptic, Audio and Visual Environments and Their Applications, 2004.

[8] S. E. Krufka and K. E. Barner, "Automatic Production of Tactile Graphics from Scalable Vector Graphics," In Proc. of 7th Int. ACM SIGACCESS Conf. on Computers and Accessibility, USA, 2005.

[9] R. E. Ladner, M. Y. Ivory, R. Rao, S. Burgstahler, D. Comden, S. Hahn, M. Renzelmann, S. Krisnandi, M. Ramasamy, B. Slabosky, A. Martin, A. Lacenski, S. Olsen and D. Groce, "Automating Tactile Graphics Translation," In Proc. of 7th Int. ACM SIGACCESS Conf. on Computers and Accessibility, USA, 2005.

[10] D. K. McGookin and S. A. Brewster, "SoundBar: Exploiting Multiple Views in Multimodal Graph Browsing," In Proc. of 4th Nordic Conf. on Human-Computer Interaction, Oslo, Norway, 2006.

[11] S. Murphy, "Accessibility of Graphics in Technical Documentation for the Cognitive and Visually Impaired," In Proc. of 23rd Annual Int. Conf. on Design of Communication, UK, 2005.

[12] Wikipedia, "Basic description of table," Available from:http://en.wikipedia.org/wiki/Table_(information) Access date: July 3, 2007.

[13] Meta Enterprises, LLC, "SimpleOCR Version 3.1," Available from: http://www.simpleOcr.com, Access date: July 3, 2007.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอังคนา จันทร์รุ่งอุทัย เกิดวันที่ 25 ตุลาคม พ.ศ. 2526 ที่กรุงเทพมหานคร สำเร็จการศึกษาระดับมัธยมศึกษาตอนปลายจากโรงเรียนมหิดลวิทยานุสรณ์ อ.ศาลายา จ. นครปฐม หลังจากนั้น ในปีการศึกษา 2545 ได้รับทุนโครงการพัฒนากำลังคนด้านวิทยาศาสตร์ (ทุนเรียนดีวิทยาศาสตร์แห่งประเทศไทย) และเข้าศึกษาที่คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล และสำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต เกียรตินิยมอันดับหนึ่ง สาขาวิชาคณิตศาสตร์ ในปี พ.ศ. 2549 และในปีเดียวกัน ได้เข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย