การติดตามการสั่นของวัตถุในวิดีโอเงียบเพื่อสกัดเสียง

Object Vibration Tracking in Silent Video for Sound Extraction

ธนัสนี เพียรตระกูล* ภัทราวุธ คุณวิภูษิต อมร มิสรา และ กฤษฎิ์ เจริญสาธิต
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยมหิดล
25/25 ถนนพุทธมณฑลสาย4 ตำบลศาลายา อำเภอพุทธมณฑล จังหวัดนครปฐม 73170
Tanasanee Phienthrakul *, Pattrawut Khunwipusit, Amon Mishra and Kris Charoensatit
Department of Computer Engineering, Faculty of Engineering, Mahidol University
25/25 Phuttamonthon 4 Road, Salaya, Phuttamonthon, NakornPathom Thailand, 73170
*ผู้นิพนธ์ประสานงาน: tanasanee.phi@mahidol.ac.th เบอร์โทรศัพท์ 02-889-2138 ต่อ 6257

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอกระบวนการในการสังเคราะห์เสียงจากภาพเคลื่อนไหวที่มีการสั่นของวัตถุในภาพ โดยจะนำภาพจากกล้องที่มีความละเอียดเหมาะสมมาประมวลผลเพื่อหาวัตถุในภาพที่มีการเคลื่อนไหว จากนั้นทำการขยายสัญญาณเพื่อให้เห็นการเปลี่ยนแปลงของตำแหน่งได้ชัดเจนขึ้น แล้วจึงนำตำแหน่งที่เปลี่ยนไปของวัตถุในแต่ละเฟรมมาแปลงให้ออกมาเป็นเสียง จากการทดสอบเบื้องต้นพบว่า การสั่นสะเทือนของวัตถุในภาพสามารถนำมาแปลงเป็นเสียงที่มีความถี่ใกล้เคียงกับความถี่ของเสียงต้นฉบับ ดังนั้นการสกัดเสียงจากการสั่นของวัตถุ จึงช่วยทำให้สามารถได้ยินเสียง ซึ่งสามารถนำไปใช้ช่วยในการสืบสวนสอบสวนต่อไปได้

คำสำคัญ การสกัดเสียง การขยายสัญญาณ การสั่นสะเทือนของวัตถุ วิดีโอเงียบ การติดตามวัตถุ

Abstract

This research proposed the steps of voice synthesis from video, which contain a moving object. Video images with the suitable resolution are processed to locate the vibrating object. Then, the signals are magnified in order to easily detect the change in position. The moving positions of the object in each frame are converted to sound. The initial testing shows that the vibration of object in a video can be converted to sound with the frequency close to the actual frequency of the source. This process of sound extraction can be used to assist in the investigation of a crime or incident.

Keywords: sound extraction, magnification, object vibration, silent video, object tracking.

1. บทนำ

จากเหตุการณ์ความไม่สงบที่เกิดขึ้นบ่อยครั้งในประเทศอาทิเช่น การลอบวางระเบิดทั้งในกรุงเทพมหานคร และต่างจังหวัด การจราจล การชุมนุมประท้วง การโจรกรรม หรือ การลอบวางเพลิง เหตุการณ์เหล่านี้มักมีการตกลงวางแผนกันล่วงหน้า หรือบางครั้งมีการนัดหมายจ้างวานให้ผู้อื่นช่วยลงมือดำเนินการ ทำให้เมื่อเกิดเหตุการณ์ความไม่สงบขึ้น การติดตามผู้กระทำความผิดมาลงโทษจึงทำได้ยากและไม่สามารถหาหลักฐานที่ชัดเจนได้

ในปัจจุบันกล้อง CCTV (Closed Circuit Television) ได้รับความนิยมมากขึ้น มีการติดตั้งกล้องตามสถานที่สำคัญเพื่อสอดส่องดูแลความสงบเรียบร้อย หรือในร้านค้า ร้านอาหาร สถานที่สาธารณะทั่วไป ก็มีการติดตั้งกล้องกันมากขึ้น เพื่อที่จะใช้บันทึกภาพและตรวจสอบย้อนหลังได้เมื่อต้องการ ด้วยเหตุนี้เมื่อเกิดเหตุความไม่สงบขึ้น ภาพของผู้

ก่อเหตุจึงมักถูกบันทึกไว้ได้ทั้งก่อนและหลังเหตุการณ์ความไม่สงบ

อย่างไรก็ตาม การจะนำภาพของผู้ก่อความไม่สงบ ไปขยายผลเพื่อหาผู้สมรู้ร่วมคิด หรือผู้ร่วมก่อการ อาจจะยังทำได้ไม่สะดวกนัก เนื่องจากกล้องส่วนใหญ่จะไม่สามารถบันทึกเสียงสนทนาที่เกิดขึ้นในระยะไกลได้ หรือ หากบันทึกได้ก็จะมีสัญญาณรบกวนทำให้ได้ยินไม่ชัดเจน จึงเป็นเหตุให้ผู้ต้องสงสัยสามารถปฏิเสธความผิดได้ หรือหากต้องการขยายผลเพื่อสืบหาผู้จ้างวานหรือผู้เกี่ยวข้อง ก็อาจจะทำให้ไม่มีหลักฐานที่จะนำมายืนยันได้ว่าบุคคลนั้นเป็นผู้จ้างวานจริง

ด้วยเหตุนี้จึงเกิดแนวความคิดที่ว่า หากเราสามารถได้ยินเสียงสนทนาเหล่านั้น ก็อาจจะทำให้มีหลักฐานที่ชัดเจนขึ้น ว่าใครเป็นผู้จ้างวาน หรือมีใครบ้างเป็นผู้ร่วมก่อเหตุ การสังเคราะห์เสียงจากภาพเคลื่อนไหวที่บันทึกเหตุการณ์นั้นไว้ หรือ ภาพเหตุการณ์ก่อนและหลังก่อเหตุที่ผู้ต้องสงสัยมีการติดต่อพูดคุยกับบุคคลอื่น อาจทำให้ได้ข้อเท็จจริงที่มากขึ้น และสามารถนำไปขยายผลเพื่อติดตามจับกุมผู้ร่วมกระทำความผิดทั้งหมดมาลงโทษได้

ในงานวิจัยนี้จึงมีแนวความคิดที่จะนำเสนอกระบวนการในการสังเคราะห์เสียงจากภาพเคลื่อนไหวที่มีการสั่นสะเทือนของวัตถุในภาพ ในการใช้งานจริง ภาพจากกล้อง CCTV ที่มีความละเอียดเหมาะสม จะถูกนำมาประมวลผลเพื่อหาวัตถุในภาพที่มีการเคลื่อนที่หรือมีการสั่นสะเทือน จากนั้นจะทำการแปลงการสั่นสะเทือนเหล่านั้นให้ออกมาเป็นเสียง เพื่อใช้ประกอบเป็นหลักฐานร่วมกับภาพเหตุการณ์ที่ได้บันทึกไว้ แต่ในการทดสอบแนวความคิดนี้ เราจะเริ่มจากการนำภาพเคลื่อนไหวที่ไม่มีเสียง ซึ่งถูกบันทึกไว้ในสภาวะควบคุมมาใช้ในการทดสอบ โดยจะเป็นภาพของวัตถุที่เป็นแหล่งกำเนิดเสียง และ เป็นวัตถุที่เสียงตกไปกระทบ บันทึกด้วยความละเอียด และความเร็วสูง เพื่อแสดงให้เห็นว่าแนวความคิดดังกล่าวมีความเป็นไปได้

2. ทฤษฎีและวิธีดำเนินการวิจัย

ในหัวข้อนี้จะกล่าวถึงเสียง การกำเนิดของเสียง การเคลื่อนที่ของเสียง และการได้ยิน เพื่อให้เข้าใจขั้นตอนของการกำเนิดเสียง และงานวิจัยที่พยายามจะสกัดเสียงจากระยะไกล

2.1 เสียงกับการได้ยิน

เสียงเกิดจากการสั่นสะเทือนของวัตถุ เมื่อวัตถุมีการสั่นสะเทือน จะทำให้เกิดการอัดตัวและขยายตัวของคลื่นเสียง และถูกส่งผ่านตัวกลางไปยังหู ทำให้ได้ยินเสียงขึ้น [1] จะเห็นได้ว่าการเคลื่อนที่ของเสียงจำเป็นต้องอาศัยตัวกลาง ซึ่งอาจจะเป็นของแข็ง ของเหลว หรือ ก๊าซ หากไม่มีตัวกลางเหล่านี้เสียงก็จะไม่สามารถเดินทางมาจนถึงหูของเราได้

เมื่อวัตถุหรือแหล่งกำเนิดเสียงมีการสั่นสะเทือน จะทำให้โมเลกุลของอากาศที่อยู่รอบๆ มีการเคลื่อนที่ โมเลกุลของอากาศเหล่านั้นจะเคลื่อนที่จากแหล่งกำเนิดเสียงไปชนกับโมเลกุลของอากาศที่อยู่ถัดออกไป และเกิดการถ่ายโอนโมเมนตัม [1] จากนั้นโมเลกุลที่ชนกันก็จะแยกออกจากกัน โดยโมเลกุลของอากาศที่เคลื่อนที่มาชนจะถูกดึงกลับไปยังตำแหน่งเดิมด้วยแรงปฏิกิริยา และโมเลกุลที่ได้รับการถ่ายโอนพลังงาน ก็จะเคลื่อนที่ต่อไปและไปชนกับโมเลกุลของอากาศที่อยู่ถัดไป เป็นเช่นนี้ไปเรื่อยๆ ทำให้เกิดเป็นคลื่นเสียง ดังแสดงในรูปที่ 1

ส่วนของความดังของเสียง คือปริมาตรของพลังงานเสียงที่มาถึงหูของเรา [1] ปัจจัยที่มีผลทำให้เสียงดังหรือเสียงค่อยได้แก่

1. ระยะทาง ถ้าระยะทางใกล้ๆ จะได้ยินเสียงดังมาก และจะได้ยินเสียงค่อยลงไปเมื่อระยะห่างออกไป

2. ความแรงในการสั่นสะเทือนของวัตถุแหล่งกำเนิดเสียง

3. ชนิดของตัวกลางที่คลื่นเสียงเคลื่อนที่ผ่านไป เช่น ถ้าคลื่นเสียงเคลื่อนที่ไปในน้ำจะมีความดังของเสียงมากกว่าคลื่นเสียงที่เคลื่อนที่ไปในอากาศ

4. ขนาดและรูปร่างของวัตถุที่เป็นแหล่งกำเนิดเสียงสั่นสะเทือน

ด้วยเหตุนี้ หากเราอยู่ไกลจากแหล่งกำเนิดเสียง หรือมีการรบกวนจากเสียงอื่นๆ เช่น มีลมพัด มีวัตถุมากั้นทางเดินของเสียง ก็จะทำให้ไม่สามารถได้ยินเสียงชัดเจน แต่หากเราสามารถสังเกตการสั่นสะเทือนของวัตถุที่อยู่ใกล้กับแหล่งกำเนิดเสียงมากกว่า แล้วนำมาแปลงให้เป็นคลื่นเสียงได้ก็น่าจะทำให้ได้ยินเสียงที่ชัดเจนมากยิ่งขึ้น

2.2 งานวิจัยที่เกี่ยวข้องกับการสังเคราะห์เสียงจากระยะไกล

จากการค้นคว้างานวิจัยที่เกี่ยวข้อง ได้ศึกษางานวิจัยนี้เป็นของ James M. Moses และเพื่อนของเขา K.P. Trout [2] ที่ทำการใช้เลเซอร์วัดแรงสั่นสะเทือนบนพื้นผิวที่มีความ

เรียบสูงและสามารถสะท้อนแสงกลับมาได้ ในงานวิจัยดังกล่าว ได้ทำการขยายสัญญาณและแปลค่าแรงสั่นสะเทือนให้เป็นเสียง โดยใช้คุณสมบัติของเลเซอร์ ซึ่งเป็นการบีบอัดของแสง ทำให้แสงที่ได้นั้น นอกจากจะมีความเข้มสูงแล้ว ยังกระเจิงได้ยากกว่าด้วย สัญญาณที่ได้รับจึงค่อนข้างชัดเจนและไม่มีสัญญาณรบกวนมากนัก พวกเขาได้ยิงเลเซอร์ไปที่กระจกและใช้แผงโซลาเซลล์เป็นตัวรับสัญญาณ จากนั้นจึงใช้เครื่องช่วยขยายสัญญาณ ก่อนที่จะปล่อยเสียงออกทางลำโพง วิธีนี้ถึงแม้จะทำให้ได้เสียงที่ค่อนข้างชัดเจน แต่ก็ยังมีข้อจำกัดอยู่มาก เช่นในกรณีที่ไม่มีกระจก และยังไม่สามารถใช้ได้จริงในการสืบสวนสอบสวน เนื่องจากต้องใช้งานในขณะที่อยู่ในที่เกิดเหตุเท่านั้น จึงไม่สามารถนำมาใช้กับสถานการณ์ที่ผ่านไปแล้วได้

งานวิจัยเรื่องถัดมาเป็นของ Michael Rubinstein และทีมงาน [3] ซึ่งงานวิจัยนี้ได้รับความสนใจและงบสนับสนุนจากบริษัทชื่อดัง อาทิเช่น Microsoft, Adobe และแน่นอนมหาวิทยาลัยที่พวกเขาสังกัดอยู่ MIT นั่นเอง พวกเขาได้ใช้ประโยชน์จากคุณสมบัติของเสียงที่เกิดจากการเคลื่อนไหวและก่อให้เกิดแรงสั่นสะเทือน ทดลองวิเคราะห์จากแรงสั่นสะเทือนของสิ่งของรอบข้างที่สามารถหาได้โดยใช้ Video Processing ซึ่งมีพื้นฐานมาจากการประมวลผลภาพ

พวกเขาใช้ตัวอย่างที่ได้จากกล้องที่มีความไวและความละเอียดสูง (Ultra-speed Camera) โดยความถี่ของการถ่ายทีละเฟรมของกล้องประเภทนี้สูงกว่าความถี่ที่มนุษย์สามารถรับฟังได้มาก ซึ่งถ้าสามารถสังเคราะห์เสียงได้ 100% จะส่งผลให้ได้เสียงที่ค่อนข้างคมชัดในระดับที่น้อยคนจะแยกออกได้ แต่ข้อเสียของงานวิจัยนี้คือไม่สามารถใช้จริงกับกล้องทั่วไปได้ และหากจะเปลี่ยนกล้อง CCTV ทั้งประเทศให้เป็นกล้องชนิดนี้ต้องใช้งบประมาณที่สูงมากจนแทบจะเป็นไปไม่ได้เลย เราจึงได้ค้นคว้าต่อเพื่อหาวิธีที่จะสังเคราะห์เสียงให้ได้ใกล้เคียงที่สุดด้วยงบประมาณที่ต่ำลงมาและสามารถนำไปประยุกต์ใช้ได้จริงในสถานการณ์จริง

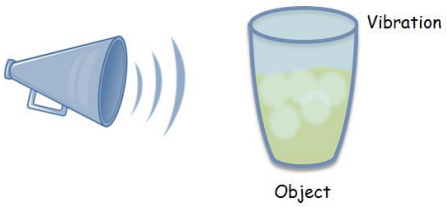
เนื่องจากการใช้กล้องที่มีความละเอียดสูงยังมีข้อจำกัดอยู่อีกมาก จึงมีงานวิจัยที่ได้นำหลักการของชัตเตอร์เวียน (Rolling Shutter) มาประยุกต์ใช้ [4], [5] โดยภาพจะถูกถ่ายและเก็บข้อมูลทีละแถว ดังนั้นหากมีการสั่นสะเทือนของวัตถุในภาพ ก็สามารถที่จะสังเกตได้จากเส้นขอบของวัตถุที่อยู่ในภาพนั้น ซึ่งสามารถนำไปแปลงเป็นความถี่ และสามารถนำมาใช้สกัดเป็นเสียงออกมาได้ด้วยวิธีที่ไม่แตกต่างจากการใช้กล้องความละเอียดสูงมากนัก

3. การสกัดเสียงจากการสั่นสะเทือนของวัตถุ

ในการแปลงการสั่นของวัตถุในภาพเคลื่อนไหวให้เป็นเสียงนั้น ผู้พัฒนาได้ทำการแบ่งขั้นตอนการทำงาน ออกเป็น 4 ขั้นตอน คือ

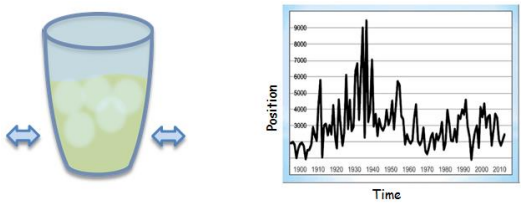
1. จากภาพเคลื่อนไหวที่นำมาวิเคราะห์ จะทำการตรวจหาวัตถุที่อยู่ใกล้กับแหล่งกำเนิดเสียงที่สนใจ และสามารถตรวจจับความการสั่นสะเทือนได้ โดยการใช้วิธีการของการประมวลผลภาพเคลื่อนไหว เปรียบเทียบเฟรมภาพที่อยู่ใกล้กัน เพื่อตรวจหาวัตถุในภาพที่มีการเคลื่อนที่ เช่น แก้วน้ำ กระดาษ ถุงขนม ต้นไม้ ใบไม้ ฯลฯ ที่สามารถตรวจวัดการสั่นสะเทือนและแปลงเป็นเสียงได้

รูปที่ 1 แสดงขั้นตอนแรกในการทำงานที่จะตรวจหาวัตถุที่มีการสั่นสะเทือนในภาพ จากนั้นทำการระบุตำแหน่งของวัตถุเหล่านั้นบนภาพ เพื่อจะนำไปใช้ศึกษาการสั่นของวัตถุนั้นในขั้นตอนถัดไป



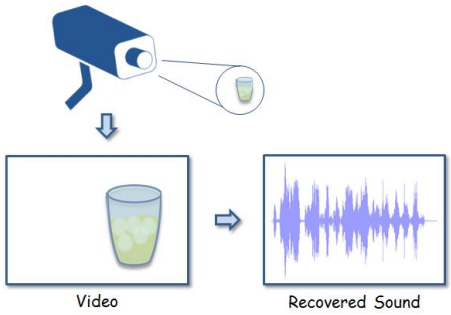
รูปที่ 1 ตรวจหาวัตถุที่มีการสั่นและอยู่ใกล้กับแหล่งกำเนิดเสียง

2. วิเคราะห์การสั่นสะเทือนของวัตถุนั้น โดยทำการกำหนดแกนอ้างอิง และกำหนดจุดสนใจบนวัตถุ เพื่อนำมาสร้างเป็นกราฟที่แสดงตำแหน่งของวัตถุที่มีการสั่น ในขั้นตอนนี้จะทำการบันทึกตำแหน่งที่เปลี่ยนไปของวัตถุที่สนใจ โดยพิจารณาจากตำแหน่งของวัตถุในแต่ละเฟรมจากภาพเคลื่อนไหวที่นำมาพิจารณา แล้วนำมาแสดงในรูปแบบกราฟเพื่อให้เห็นการเปลี่ยนแปลงได้ง่ายขึ้น ดังแสดงตัวอย่างของขั้นตอนนี้ในรูปที่ 2



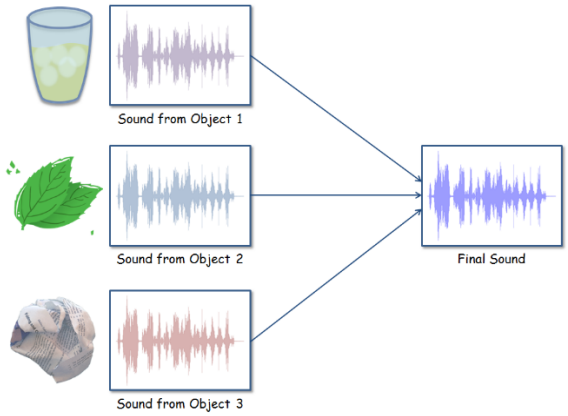
รูปที่ 2 พิจารณาตำแหน่งของวัตถุในแต่ละเฟรม และแสดงผลเป็นกราฟระหว่างตำแหน่งและเวลา

3. จากนั้นจะทำการแปลงข้อมูลจากตำแหน่งของวัตถุที่ได้ให้เป็นคลื่นเสียง ภายหลังจากได้ตำแหน่งของวัตถุในแต่ละเฟรมของภาพเคลื่อนไหวแล้ว สัญญาณที่ได้จะถูกนำมาขยายด้วยวิธีการ Phase-based video processing [6] เพื่อให้เห็นการเปลี่ยนแปลงที่ชัดเจนขึ้น และสามารถนำไปแปลงเป็นเสียงที่มีความดังเหมาะสมได้ นั่นคือเราจะได้เสียงที่สกัดออกมาจากภาพเคลื่อนไหว ดังแสดงในรูปที่ 3



รูปที่ 3 ได้เสียงที่สกัดมาจากวัตถุที่มีการสั่นสะเทือนในภาพ

4. ทำการรวมสัญญาณเสียงที่ได้จากแต่ละวัตถุให้เป็นเสียงเดียวที่มีความชัดเจนมากขึ้น หากการตรวจหาวัตถุในขั้นตอนแรก พบว่ามีวัตถุที่สนใจมากกว่า 1 ชิ้น ในขั้นตอนนี้เราจะทำการรวมสัญญาณเสียงที่ได้จากแต่ละวัตถุที่เราสนใจ ดังแสดงในรูปที่ 4



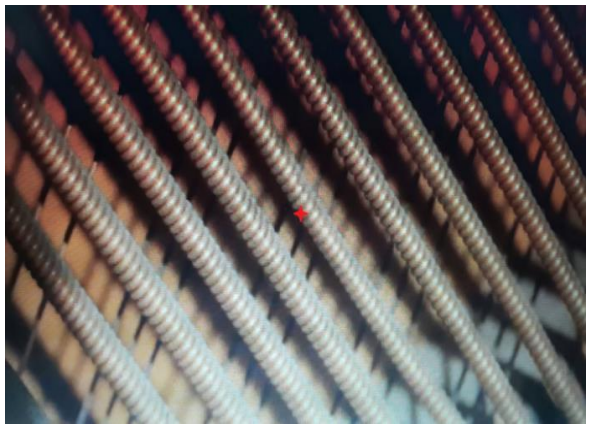
รูปที่ 4 สัญญาณเสียงจากหลายๆ วัตถุจะถูกรวมเข้าด้วยกันเพื่อสังเคราะห์เป็นเสียงที่ต้องการ

ในการรวมสัญญาณเสียงนั้นเราพิจารณาระยะห่างจากแหล่งกำเนิดเสียงของแต่ละวัตถุร่วมด้วย วัตถุที่อยู่ไกลจากแหล่งกำเนิดเสียงมากกว่า อาจจะมีเวลาที่หน่วงกว่า (Delay Time) สำหรับการที่เสียงเดินทางไปถึง นอกจากนั้นยังอาจจะมีสัญญาณรบกวนอื่นๆ ที่ไม่ใช้เสียงหลักที่เราสนใจ สัญญาเหล่านั้นจะถูกกรองออกไปในขั้นตอนนี้ด้วย

4. ผลการวิจัยและอภิปราย

เพื่อที่จะทดสอบแนวคิดที่ได้นำเสนอ ผู้วิจัยได้ทำการเก็บภาพเคลื่อนไหวที่บันทึกด้วยกล้องความเร็วสูงที่ความถี่ 1-20 กิโลเฮิร์ซ พร้อมทั้งบันทึกเสียงจริงเพื่อนำมาใช้ในการเปรียบเทียบในภายหลังด้วย และพยายามบันทึกภาพและเสียงให้มีสัญญารบกวนน้อยที่สุด เพื่อลดความผิดพลาดของสัญญาณที่สกัดออกมา

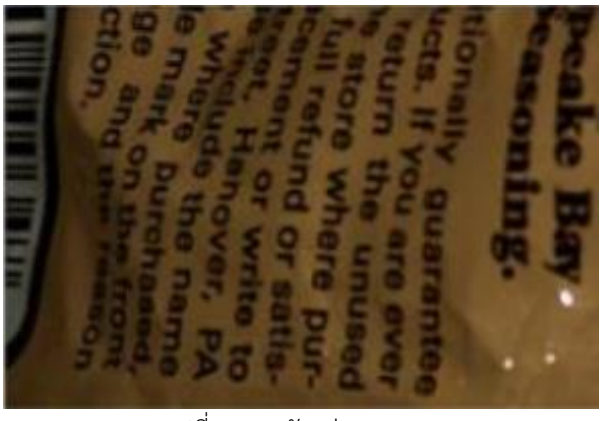
สำหรับภาพของวัตถุที่นำมาทดสอบสกัดเสียง ผู้พัฒนาต้องการทดสอบวิธีการที่นำเสนอ โดยการนำภาพเคลื่อนไหวของวัตถุ มาหาตำแหน่งที่เปลี่ยนแปลงไปในแต่ละเฟรม ในการศึกษาเบื้องต้นพบว่าหากสามารถสกัดเสียงจากการสั่นของแหล่งกำเนิดเสียงได้ จะเห็นการสั่นสะเทือนได้ชัดเจนกว่า จึงเลือกทดสอบการสกัดเสียงจากภาพวัตถุที่เป็นแหล่งกำเนิดเสียง เช่น สายเปียโน สายกีต้าร์ สายไวโอลิน ตัวอย่างภาพของสายเปียโนแสดงในรูปที่ 5



รูปที่ 5 ภาพตัวอย่างสายเปียโนที่มีการสั่น

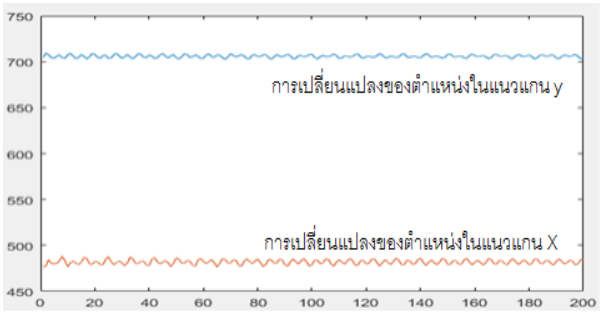
นอกจากนั้นยังเลือกภาพเคลื่อนไหวของวัตถุที่ไม่ได้เป็นแหล่งกำเนิดเสียง แต่ทำหน้าที่เป็นตัวกลาง มาใช้ในการทดสอบด้วย โดยเลือกใช้ภาพถุงขนมที่อยู่ใกล้แหล่งกำเนิดเสียง เพราะเป็นวัตถุที่มีน้ำหนักเบาสามารถเห็นการเคลื่อนไหวได้ง่าย ตัวอย่างภาพถุงขนมแสดงในภาพที่ 6 และ

ยังมีภาพที่บันทึกการเคลื่อนไหวของผิวหน้าลำโพง ซึ่งเป็นตัวกลางที่อยู่ใกล้แหล่งกำเนิดเสียงมากๆ นำมาใช้ในการทดสอบสกัดเสียงด้วย



รูปที่ 6 ภาพตัวอย่างถุงขนม

ผู้พัฒนาทำการกำหนดจุดที่ต้องการพิจารณา โดยเลือกจากบริเวณที่สามารถสังเกตเห็นการเปลี่ยนแปลงได้อย่างชัดเจน และบันทึกการเปลี่ยนแปลงตำแหน่งของจุดดังกล่าวทั้งในแนวแกน X และแกน Y ดังแสดงในรูปที่ 7 จากกราฟแกนนอนเป็นแกนของเวลาที่เปลี่ยนไป ส่วนแกนตั้งเป็นตำแหน่งของจุดที่พิจารณา จะเห็นว่าตำแหน่งของจุดที่สนใจมีการเคลื่อนที่แกว่งในช่วงที่ไม่สูงมากนัก



รูปที่ 7 กราฟแสดงการเปลี่ยนแปลงของจุดที่สนใจในแนวแกน X และแกน Y

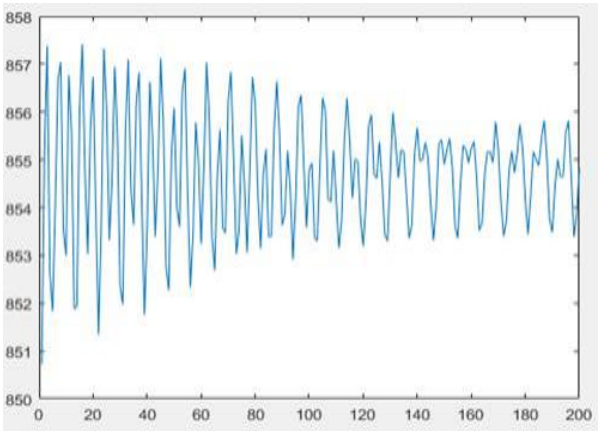
จากนั้นทำการรวมการเปลี่ยนแปลงในแนวแกน X และแกน Y เข้าด้วยกัน โดยใช้สมการพีทาโกรัส (Pythagoras Equation)

$$s_i = \sqrt{x_i^2 + y_i^2} \qquad (1)$$

เมื่อ s_i เป็นค่าที่ได้จากการรวมสัญญาณที่เวลา i

x_i^2 และ y_i^2 เป็นการเปลี่ยนแปลงของตำแหน่งในแนวแกน x และ y ที่เวลา i ตามลำดับ

ทำการขยายสัญญาณเพื่อให้เห็นการเปลี่ยนแปลงได้ง่ายขึ้น โดยใช้วิธีการขยายสัญญาณแบบ Phase-based ที่ให้ผลเป็นกราฟรูปร่างเดิม ผลการรวมและขยายสัญญาณแสดงในรูปที่ 8



รูปที่ 8 กราฟจากการรวมการเปลี่ยนแปลงและขยายสัญญาณ

การทดสอบความสามารถในการสกัดเสียงจากวิดีโอเงียบได้ใช้ตัวอย่างวิดีโอ จำนวน 10 ตัวอย่าง ซึ่งเป็นวิดีโอที่แท้จริงแล้วมีเสียง แต่ผู้วิจัยได้ทำการแยกเสียงไว้ เพื่อนำมาใช้ประเมินผลเทียบกับเสียงที่สกัดขึ้นมาใหม่ โดยจะทำการปรับแอมปิจูดของเสียงให้อยู่ในช่วงเดียวกันก่อน (Normalize) แล้วจึงวัดค่าความแตกต่างกำลังสองเฉลี่ย (Mean Square Error: MSE) ตามสมการที่ 2

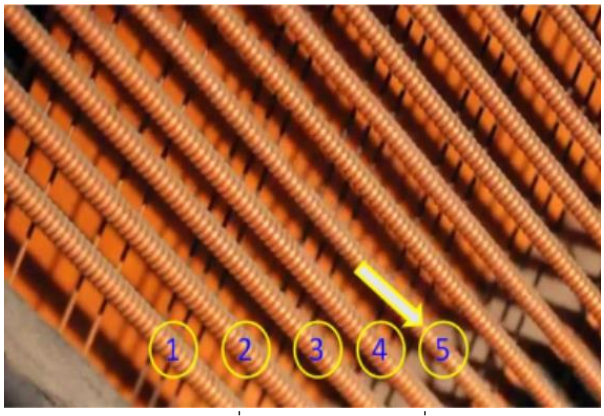
$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}(O(i) - S(i))^2 \qquad (2)$$

โดยที่ n เป็นความยาวของจุดข้อมูล $O(i)$ เป็นแอมปิจูดของเสียงจริง ที่เวลา i และ $S(i)$ เป็นแอมปิจูดของเสียงสังเคราะห์ที่เวลา i ผลการทดสอบแสดงในตารางที่ 1

ตารางที่ 1 ผลการทดสอบวัดค่า MSE

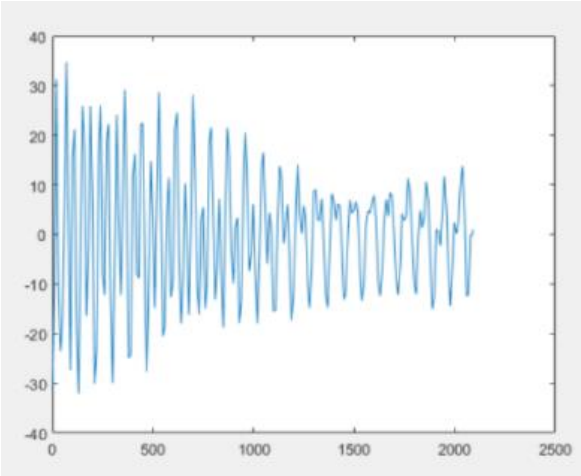
วิดีโอที่	แหล่งกำเนิดเสียง / ชนิดของเสียง	MSE
1	Piano_5th String-C	11.32270
2	Piano_7th String-D	10.7950
3	Piano_12th String-G	2.19270
4	Violin String-G	1.58890
5	Guitar String-E	1.58441
6	Potato chip	**0.26655**
7	Marry in lamb	**0.14003**
8	Musical anthem	**0.75559**
9	Musical anthem slow	1.05190
10	National anthem slow	**0.55879**

จากตาราง ค่า MSE ที่สูงแสดงให้เห็นว่ามีความแตกต่างจากเสียงต้นฉบับอยู่มาก โดยทั่วไปแล้วค่า MSE ที่เหมาะสมควรมีค่าไม่เกิน 3 และถือว่าให้ผลที่ดี ถ้า MSE มีค่าไม่เกิน 1 จากตารางที่ 1 จะเห็นว่า มี 4 วิดีโอที่ได้ผลอยู่ในเกณฑ์ดี คือ Potato chip, Marry in lamb, Musical anthem, และ National anthem slow และอีก 4 วิดีโอ ให้ผลอยู่ในระดับที่พอใช้ได้ ส่วนเปียโนสายที่ 5 และ สายที่ 7 จะเห็นว่ายังมีความคลาดเคลื่อนอยู่สูง ส่วนหนึ่งเป็นอาจเป็นสาเหตุมาจากการเลือกจุดที่จะพิจารณาการสั่นสะเทือน ซึ่งด้วยลักษณะของสายเปียโนทำให้การติดตามจุดดังกล่าวมีโอกาสคลาดเคลื่อนได้ง่าย รูปที่ 9 ชี้ให้เห็นสายเปียโนเส้นที่ 5

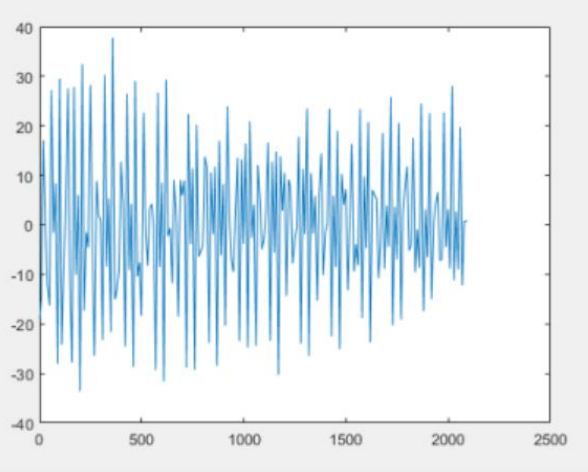


รูปที่ 9 สายเปียโนเส้นที่ 5

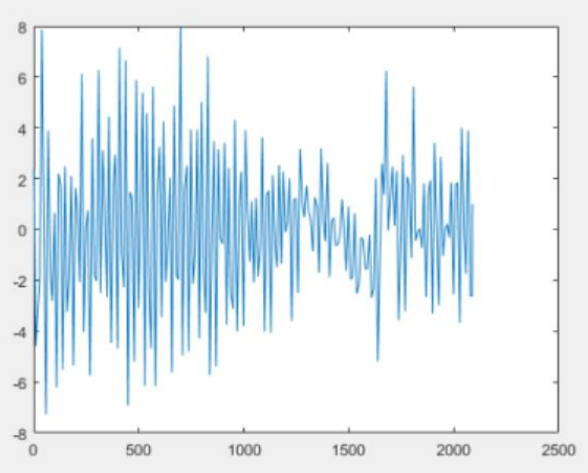
ตัวอย่างของสัญญาณเสียงที่สกัดออกมาได้จากสายเปียโนเส้นที่ 5 เส้นที่ 7 และ เส้นที่ 12 แสดงในรูปที่ 10 – รูปที่ 12 ตามลำดับ สัญญาณที่สกัดมาได้จากคลิปวิดีโอไวโอลินและกีตาร์ แสดงในรูปที่ 13 และ รูปที่ 14 โดยเลือกจุดพิจารณาอยู่บนสายไวโอลินและสายกีตาร์โดยตรง



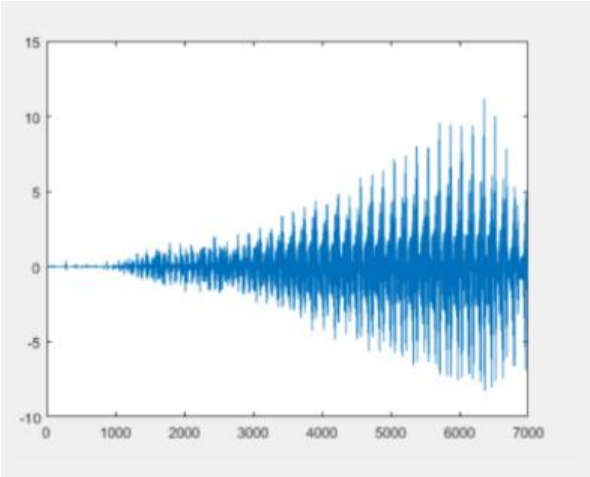
รูปที่ 10 สัญญาณเสียงที่ได้จากสายเปียโนเส้นที่ 5 โน๊ต C#



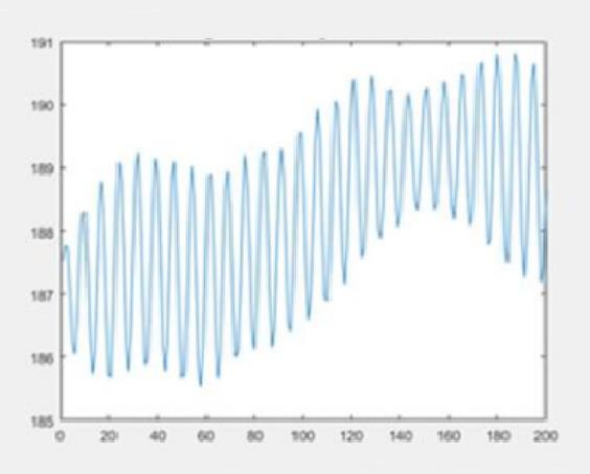
รูปที่ 11 สัญญาณเสียงที่ได้จากสายเปียโนเส้นที่ 7 โน๊ต D



รูปที่ 12 สัญญาณเสียงที่ได้จากสายเปียโนเส้นที่ 12 โน๊ต G



รูปที่ 13 สัญญาณเสียงที่สกัดได้จากสายไวโอลิน โน๊ต G

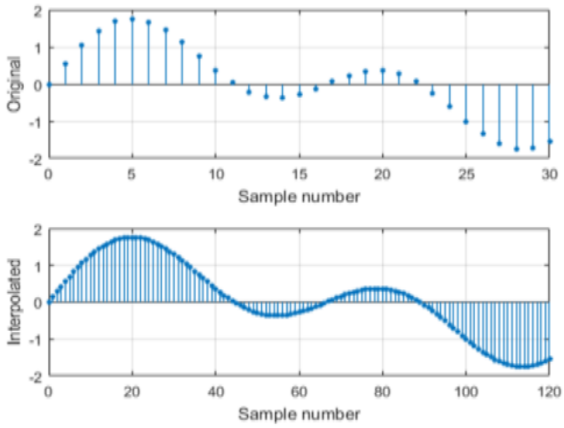


รูปที่ 14 สัญญาณเสียงที่สกัดได้จากสายกีต้าร์ โน๊ต E

ในส่วนของการสกัดเสียงจากวัตถุที่ไม่ใช่แหล่งกำเนิดเสียงโดยตรง เราพบว่าถ้าความถี่ของภาพไม่สูงพอ จะไม่สามารถได้สัญญาณเสียงที่ครบถ้วน ดังนั้นจึงได้เพิ่มในส่วนของการสังเคราะห์สัญญาณเสียงด้วยวิธีการ interpolation ตามสมการที่ 3 เมื่อกำหนดให้ (x_1, y_1) และ (x_2, y_2) เป็นพิกัดของจุดใดๆ ที่อยู่ติดกัน จุดที่สังเคราะห์ขึ้นมาใหม่แทนด้วย (x_n, y_n) ที่อยู่ระหว่าง 2 จุดดังกล่าว สามารถหาได้จาก

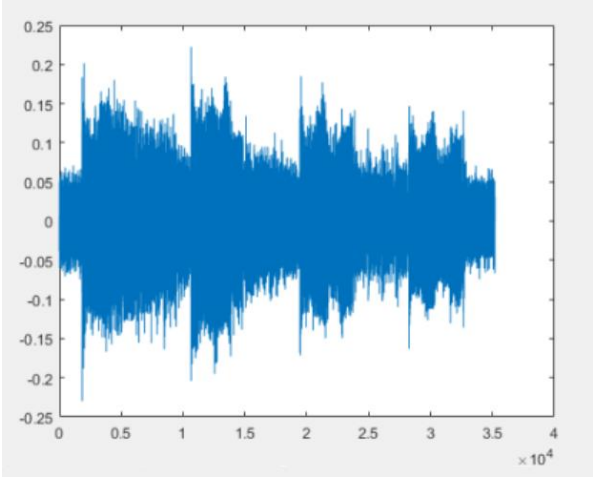
$$y_n = \frac{(x_n - x_1)(y_2 - y_1)}{(x_2 - x_1)} + y_1 \qquad (3)$$

ตัวอย่างผลการทำ interpolation แสดงในรูปที่ 15 ซึ่งการทำ interpolation นี้มีส่วนช่วยให้สัญญาณเสียงที่สกัดได้มีความใกล้เคียงกับสัญญาณจริงมาขึ้น และส่งผลให้ค่า MSE ในตารางที่ 1 ของวิดีโอที่ 6-10 มีค่าต่ำลงเมื่อเทียบกับวิดีโอที่ไม่ได้ผ่านการทำ interpolation ด้วย

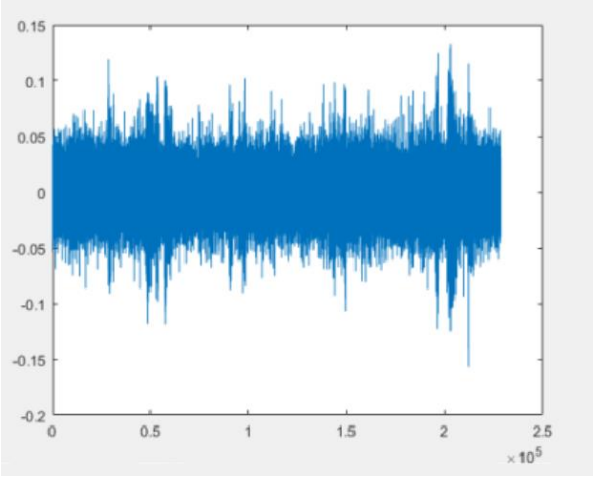


รูปที่ 15 ตัวอย่างผลการทำ interpolation

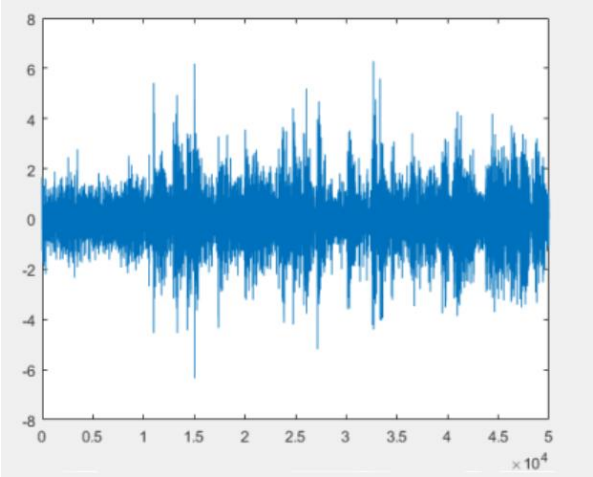
ภาพสัญญาณเสียงที่ได้จากการสกัดผ่านถุงขนม แสดงในรูปที่ 16 ในรูปที่ 17 แสดงสัญญาณเสียงพูด “Marry had a little lamb.” ที่สกัดจากภาพผิวหน้าของลำโพง รูปที่ 18 - รูปที่ 20 เป็นเพลงชาติไทยที่สกัดจากผิวหน้าของลำโพง เพื่อให้สามารถเห็นการเคลื่อนไหวของวัตถุในภาพได้ชัดเจน โดยในรูปที่ 18 เป็นทำนองของเพลงชาติที่เล่นในจังหวะปรกติ รูปที่ 19 เป็นทำนองเพลงชาติที่เล่นให้ช้าลง และในรูปที่ 20 เป็นเสียงร้องเพลงชาติอย่างช้าๆ



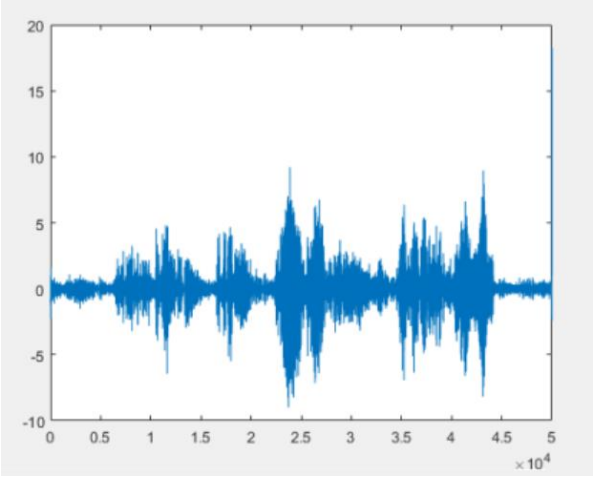
รูปที่ 16 สัญญาณเสียงจากถุงขนม



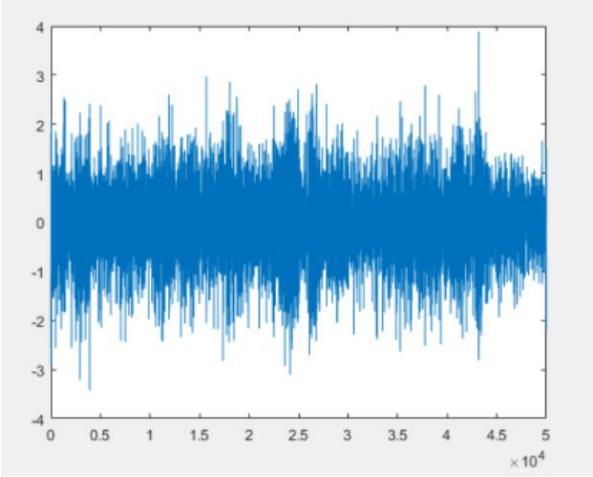
รูปที่ 17 สัญญาณเสียงพูด “Marry had a little lamb.” สกัดจากภาพผิวหน้าของลำโพง



รูปที่ 18 สัญญาณเสียงทำนองเพลงชาติไทย สกัดจากภาพผิวหน้าของลำโพง



รูปที่ 19 สัญญาณเสียงทำนองเพลงชาติไทยโดยเล่นอย่างช้า สกัดจากภาพผิวหน้าของลำโพง



รูปที่ 20 สัญญาณเสียงร้องเพลงชาติไทยโดยเล่นอย่างช้า สกัดจากภาพผิวหน้าของลำโพง

นอกจากนั้นแล้ว ยังมีการสำรวจโดยการให้ผู้ทดสอบทดลองฟังแล้วเลือกว่าเสียงไหนตรงกับต้นฉบับมากกว่า เสียงที่นำมาให้ผู้ทดสอบเลือกหลังจากฟังต้นฉบับแล้ว คือเสียงที่ได้ทำการสังเคราะห์ขึ้นจากวิธีการที่นำเสนอ และเสียงที่นำเอาเสียงที่สังเคราะห์ขึ้นมาได้มาเพิ่มหรือลดสัญญาณ หรือทำการสลับตำแหน่งของสัญญาณ

ในแบบทดสอบจะมีตัวเลือกให้ผู้รับการทดสอบเลือก 3 ทางเลือก คือ ก. เลือกเสียงที่ 1 ข. เลือกเสียงที่ 2 และ ค. ไม่สามารถระบุได้ว่าเสียงใดตรงกับต้นฉบับ โดยเสียงที่สังเคราะห์จากวิธีการที่นำเสนอจะสลับเป็นเสียงที่ 1 หรือเสียงที่ 2 อย่างสุ่ม ทำการทดสอบกับผู้รับการทดสอบจำนวน 10 คน ผลการทดสอบแสดงในตารางที่ 2

ตัวเลขที่อยู่ในช่องสีเทาคือจำนวนคนที่ตอบได้ถูกต้องในข้อนั้น ซึ่งสามารถสรุปเป็นเปอร์เซ็นต์ของคนที่ตอบถูกในคอลัมน์สุดท้าย ซึ่งจากตารางพบว่าการสกัดเสียงจากถุงขนมมีผู้ตอบถูกสูงสุด คือทั้ง 10 คนที่เข้ารับการทดสอบตอบได้ถูกต้อง รองลงมาคือเสียงไวโอลินที่มีผู้ตอบถูก 7 คน สายเปียโนเส้นที่ 5 และ ทำนองเพลงชาติไทย มีผู้ตอบถูกรองลงมาคือ 6 คน

สำหรับเสียงจากสายกีต้าร์ที่ไม่มีผู้ทายถูกเลย หรือเสียงจากเปียโนสายที่ 12 ที่มีผู้ทายถูกเพียงแค่คนเดียว อาจเป็นเพราะเป็นตัวโน๊ต จึงฟังค่อยข้างยากและไม่สามารถแยกความต่างของเสียงที่สกัดได้และเสียงที่ถูกดัดแปลง สำหรับเสียงพูดหรือเสียงร้องอย่างช้าๆ ผู้ฟังอาจจะไม่ค่อยชิน จึงทำ

ให้ยังมีผู้ตอบผิดอยู่บ้าง เสียงเหล่านี้ที่สกัดออกมาได้ยังคงมีสัญญาณรบกวนอยู่อีกมาก ทำให้ผู้ฟังต้องใจฟังเป็นพิเศษจึงจะสามารถแยกแยะความแตกต่างได้

ตารางที่ 2 ผลการทดสอบฟังเสียงที่สกัดออกมาได้

วิดีโอที่	แหล่งกำเนิดเสียง	ตัวเลือกที่ 1	ตัวเลือกที่ 2	ตัวเลือกที่ 3	% ตอบถูก
1	Piano_5th String-C	**6**	1	3	60 %
2	Piano_7th String-D	**4**	3	3	40 %
3	Piano_12th String-G	6	**1**	3	10 %
4	Violin String-G	**7**	3	0	70 %
5	Guitar String-E	9	**0**	1	0 %
6	Potato chip	**10**	**0**	**0**	100 %
7	Marry in lamb	7	**2**	1	20 %
8	Musical anthem	4	**6**	0	60 %
9	Musical anthem slow	**4**	4	2	40 %
10	National anthem slow	**4**	5	1	40 %

สำหรับเสียงพูดที่ไม่ได้มีจังหวะหรือทำนอง ผู้ฟังยังคงแยกแยะเสียงได้ไม่ดีนัก จากวิดีโอที่ 7 ซึ่งเป็นเสียงพูดจากการอ่านเนื้อเพลง แต่ไม่ได้มีจังหวะหรือเครื่องดนตรีประกอบ สังเกตว่าผู้ฟังแยกแยะเสียงได้ถูกต้องเพียงแค่ 20% เท่านั้น ซึ่งหากจะนำวิธีการนี้ไปใช้จริงกับเสียงสนทนา อาจจะต้องการการปรับปรุงให้มีประสิทธิภาพดีขึ้น

5. บทสรุป

งานวิจัยนี้ได้นำเสนอการสกัดเสียงจากการสั่นของวัตถุ ซึ่งจะช่วยทำให้เราสามารถได้ยินเสียงที่ไม่สามารถได้ยินอย่างชัดเจนเมื่อทำการบันทึกภาพเคลื่อนไหว วิธีการที่นำเสนอสามารถสกัดเสียงได้จากภาพเคลื่อนไหวของแหล่งกำเนิดเสียงโดยตรง และภาพเคลื่อนไหวของวัตถุที่ทำหน้าที่เป็นตัวกลางเสียงที่สกัดออกมาได้ยังคงมีสัญญาณรบกวนติดมาด้วย ทำให้ยังไม่เหมือนต้นฉบับอย่างแท้จริง และหากไม่เคยฟังเสียงต้นฉบับมาก่อน อาจต้องตั้งใจฟังเพื่อจับใจความ

ในการทดสอบได้ทดลองสกัดเสียงจากวิดีโอตัวอย่างจำนวน 10 วิดีโอ และนำเสียงที่สกัดได้เป็นเปรียบเทียบกับเสียงจริงเพื่อวัดค่าคลาดเคลื่อน หรือ MSE และทดสอบโดยให้ผู้ฟังแยกแยะเสียงที่สกัดได้กับเสียงอื่นที่ใกล้เคียง ผลปรากฏว่า การสกัดเสียงจากภาพเคลื่อนไหวของวัตถุที่ไม่ได้เป็นแหล่งกำเนิดเสียง แต่เป็นวัตถุเบา สามารถให้เสียงสกัดที่ฟังได้ง่ายกว่า และให้ค่า MSE ที่ต่ำกว่า ซึ่งหมายความว่าใกล้เคียงกับเสียงจริง

ปัจจัยหนึ่งที่มีผลต่อเสียงที่สกัดได้คือ เสียงต้นฉบับ ถ้าเป็นข้อความที่ผู้ฟังคุ้นเคย ก็จะสามารถคาดเดาจากเสียงที่สกัดออกมาได้ง่ายกว่า แต่ถ้าไม่เคยได้ยินเสียงนั้นมาก่อน การคาดเดาจากเสียงที่มีสัญญาณรบกวนจะทำได้ยากกว่า ดังนั้นหากต้องการนำไปใช้กับการสกัดเสียงสนทนาที่ไม่ทราบว่าผู้พูดจะกล่าวอะไรออกมาบ้าง จะยังมีอุปสรรค์อยู่ค่อนข้างมาก ในเรื่องของการเลือกวัตถุที่เหมาะสมในภาพ เพื่อจะนำมาศึกษาการเคลื่อนไหว ระยะห่างจากแหล่งกำเนิดเสียง และสัญญาณรบกวนต่างๆ ที่อาจเกิดในบริเวณรอบๆ ในกรณีเช่นนั้นอาจใช้การอ่านริมฝีปากร่วมด้วยเพื่อช่วยเพิ่มความเชื่อมั่นให้มากยิ่งขึ้น

หากเปรียบเทียบวิธีการที่นำเสนอกับวิธีการสกัดเสียงแบบอื่น พบว่ามีเป้าหมายในการใช้งานที่แตกต่างกันออกไป เช่น การยิงเลเซอร์เพื่อวัดแรงสั่นสะเทือนบนพื้นผิวจะเหมาะกับกรณีที่ต้องการดังฟังเสียงในขณะที่ผู้สนทนากำลังพูดคุยอยู่ ซึ่งหากเราไม่ทราบสถานที่หรือไม่มีการเตรียมการก่อน ก็จะไม่สามารถใช้วิธีการเลเซอร์นี้ได้ ส่วนงานวิจัยอื่นๆที่ศึกษาการสั่นสะเทือนของวัตถุในภาพเช่นเดียวกัน โดยส่วนใหญ่จะใช้วิดีโอที่มีความละเอียดสูง ทำให้ใช้พื้นที่หน่วยความจำมากและใช้เวลาในการประมวลผลสูงตามไปด้วย นอกจากนั้นยังมีต้นทุนในเรื่องของอุปกรณ์ที่มีราคาแพง ทำให้ยังไม่สามารถนำมาใช้งานจริงได้

ในงานวิจัยนี้จึงได้ปรับปรุงวิธีการสกัดเสียงจากการสั่นของวัตถุในวิดีโอเงียบ โดยใช้การติดตามวัตถุแบบสุ่มในช่วงที่เหมาะสมและเติมส่วนที่ขาดพร้อมทั้งขยายสัญญาณเสียง ทำให้สามารถสกัดเสียงกลับคืนได้ ผู้วิจัยคาดหวังว่าเมื่อนำ

วิธีการที่นำเสนอไปใช้ในการสืบสวนสอบสวน จะช่วยให้เจ้าหน้าที่มีข้อมูลมากขึ้นและข้อมูลเหล่านี้อาจนำมาประกอบกับภาพเคลื่อนไหว เพื่อใช้เป็นหลักฐานในการพิจารณาคดี หรืออาจมีส่วนช่วยให้การติดตามผู้ต้องสงสัยหรือผู้ร่วมดำเนินการ มาดำเนินคดีตามกฎหมายเป็นไปได้รวดเร็วยิ่งขึ้น

6. เอกสารอ้างอิง

[1] AMFINEWELL, เสียงกับการได้ยิน, Available from: https://amfinewell.wordpress.com/2013/01/22/เสียงกับการได้ยิน-3/ [Accessed 1st March 2016).

[2] Moses J.M. and Trout K.P., A Simple Laser Microphone for Classroom Demonstration, The Physics Teacher, Vol. 44, December 2006.

[3] The visual microphone: Passive recovery of sound from video, Washington Post, 2014. [Online]. Available: https://www.washingtonpost.com/video/c/embed/098665de-1c0c-11e4-9b6c-12e30cbe86a3. [Accessed 8th August 2016].

[4] Ait-aider O., Bartoli A., and Andreff N., Kinematics from lines in a Single rolling shutter image. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 1–6.

[5] Meingast M., Geyer C., and Sastry S., Geometric models of rolling-shutter cameras. arXiv preprint cs/0503076. NAKAMURA, J. 2005. Image sensors and signal processing for digital still cameras. CRC Press.

[6] Wadhwa N., Rubinstein M., Durand F., and Freeman W.T., Phase-based video motion processing, ACM Transactions on Graphics (TOG), Vol. 32, No. 4, 2013.