doi:10.4186/ejth.2011.3.2.33

ไทยลิปซิงค์ : การจับคู่การเคลื่อนไหวของริมฝีปากตามเสียงพูดภาษาไทย

ทวีศักดิ์ ชื่นสายชล[1]*, พิษณุ คนองชัยยศ[1] และ ชัย วุฒิวิวัฒน์ชัย[2]

[1]ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพฯ ประเทศไทย 10330
[2]ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ กระทรวงวิทยาศาสตร์และเทคโนโลยี ปทุมธานี ประเทศไทย 12120
E-mail : me@zaedgebz.net*, pizzanu.k@chula.ac.th, chai.wutiwiwatchai@nectec.or.th

บทคัดย่อ

ในปัจจุบันการสร้างแอนิเมชันได้ก้าวหน้าขึ้นเป็นอย่างมาก งานวิจัยนี้นำเสนอการเพิ่มประสิทธิภาพและการลดเวลาและค่าใช้จ่ายในการทำงานของขั้นตอนการสร้างการเคลื่อนไหวริมฝีปากของตัวละครให้สัมพันธ์กับเสียงที่พูด (Lip-sync) ซึ่งในการทำแอนิเมชันจะใช้เวลาในการสร้างและค่าใช้จ่ายในการจ้างศิลปิน งานวิจัยนี้จึงนำเสนอแอปพลิเคชั่นในการสร้างการเคลื่อนไหวของริมฝีปากตามเสียงพูด โดยพิจารณาแยกเป็นส่วนงานสำคัญได้สองส่วนคือ ส่วนแรกคือการระบุระยะเวลาของแต่ละหน่วยเสียง (force-alignment) ซึ่งจะได้ค่าเป็นหน่วยเสียง (phoneme) และเวลาเริ่มต้นและสิ้นสุดของหน่วยเสียง ส่วนที่สองคือส่วนของการสร้างการเคลื่อนไหวต่อเนื่องให้กับริมฝีปาก ซึ่งจะนำผลลัพธ์ที่ได้จากส่วนที่หนึ่งมาประมวลผลต่อ ซึ่งในส่วนนี้จะทำการสร้างโมเดลและทำการเชื่อมโยงกระดูกตามหลักการของการสร้างแบบจำลองหน้า (Facial animation) ซึ่งผลลัพธ์ของงานวิจัยคือ ตัวละครแอนิเมชันที่เคลื่อนไหวริมฝีปากได้สัมพันธ์กับเสียงพูด และสามารถนำไปพัฒนาใช้ในการสร้างแอนิเมชันต่อไปได้

คำสืบค้น

ลิปซิงค์ เสียงพูดภาษาไทย การเคลื่อนไหวริมฝีปากตามเสียงพูด การระบุระยะเวลาหน่วยเสียง ฐานข้อมูลการเคลื่อนไหวของริมฝีปาก หน่วยเสียงภาษาไทย

doi:10.4186/ejth.2011.3.2.33

Thai Lip-Sync : Mapping Lip Movement to Thai Speech

Thavesak Chuensaichol[1], Pizzanu Kanongchaiyos[1] and Chai Wutiwiwatchai[2]

[1]Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University,
Bangkok, Thailand 10330
[2]National Electronics and Computer Technology,
Center National Science and Technology Development Agency,
Ministry of Science and Technology, Pathumthaini 10120
E-mail : me@zaedgebz.net* , pizzanu.k@chula.ac.th,
chai.wutiwiwatchai@nectec.or.th

ABSTRACT

Nowadays, the animation industry is growing drastically. This results in increasing demand for better performance and reduction of process time and cost. One of the most important processes is Lip-synchronization. Generally, in creating an animation, this process can cost a lot of time and expenses. In this project, we will research about how to create the Lips Synchronization model. There are two main parts that we will focus. First, the force-alignment which is used to analyst wave files and generate them with the length of time used in each phoneme separately. Second, creating the animation, the label output from force-alignment will be collected and used as an input to this part. In this process, we will develop an object model, link armatures to it by using the facial animation method as a reference. The output result of this project is the animation model that the viseme can move synchronously with the sound waves and the length of time used in each phoneme. Therefore, this model can be used to develop the Lip-synchronization process and will also help improving the animation industry in the future.

KEYWORDS

Lip-sync, Thai speech, lip movement, force alignment, Viseme database, Thai phoneme

doi:10.4186/ejth.2011.3.2.33

I. บทนำ

เนื่องจากการสร้างภาพยนตร์แอนิเมชันในปัจจุบันมีขั้นตอนในการสร้างการเคลื่อนไหวของริมฝีปากตามเสียงพูด ด้วยวิธีการกำหนดการเคลื่อนไหวให้กับตัวละครแอนิเมชันโดยใช้ศิลปินเป็นผู้กำหนด โดยศิลปินจะทำการกำหนดคีย์เฟรมให้กับหน่วยเสียงแต่ละหน่วยในการพูดของตัวละครแอนิเมชันตัวนั้น ซึ่งระยะเวลาในการสร้างการเคลื่อนไหวของริมฝีปากจะแปรผันตามกับความยาวของประโยคที่ตัวละครพูดและปริมาณคำที่ใช้พูดว่ามีจำนวนคำซ้ำกันมากน้อยเพียงใด

งานวิจัยนี้นำเสนอการแก้ปัญหาข้างต้นด้วยการจับคู่การเคลื่อนไหวของหน่วยเสียงการเคลื่อนไหวของริมฝีปากตามหน่วยเสียง โดยผู้วิจัยจะใช้วิธีระบุเวลาของหน่วยเสียงแต่ละหน่วยเสียงว่าเวลาเริ่มต้นจนกระทั่งสิ้นสุดเสียงของหน่วยเสียงนั้นในประโยคที่ต้องการสร้างการเคลื่อนไหวตามเสียงพูดของตัวละคร จากนั้นจะทำการจับคู่ระหว่างเวลาของหน่วยเสียงกับการเคลื่อนไหวในฐานข้อมูลการเคลื่อนไหวของริมฝีปากในแต่ละหน่วยเสียงและทำการเรียบเรียงการเคลื่อนไหวของริมฝีปากทั้งประโยคเข้าด้วยกัน

ในการทดลองเพื่อประเมินผลซอฟต์แวร์ที่สร้างขึ้น จะสร้างฐานข้อมูลการเคลื่อนไหวของริมฝีปากในแต่ละหน่วยเสียงจะสร้างจากการเก็บข้อมูลจากเครื่องจับการเคลื่อนไหวในตำแหน่งริมฝีปากและรอบบริเวณใบหน้าที่กำหนดตามโครงกระดูกกำกับการเคลื่อนไหวของริมฝีปาก โดยจะเก็บจากผู้ใช้ภาษาไทยเพื่อให้ได้ข้อมูลที่มีความใกล้เคียงกับการพูดของผู้ใช้ภาษาไทย

ผลการวิจัยแสดงให้เห็นว่างานวิจัยนี้สามารถแก้ไขปัญหาและปรับปรุงขั้นตอนในการสร้างการเคลื่อนไหวของริมฝีปากตามเสียงพูดภาษาไทยให้มีความเที่ยงตรงและแม่นยำกว่าวิธีในปัจจุบัน และลดระยะเวลาและค่าใช้จ่ายในการจ้างศิลปินเพื่อการสร้างการเคลื่อนไหวให้กับตัวละครส่วนการพูดได้ และสามารถนำไปเป็นสื่อการเรียนรู้เพื่อใช้สอนให้กับผู้ที่ต้องการเรียนรู้ภาษาไทยให้ผู้เรียนสามารถเลียนลักษณะการออกเสียงจากผู้ใช้ภาษาไทย

II. งานวิจัยที่เกี่ยวข้อง

แอปพลิเคชันที่ประยุกต์ใช้คอมพิวเตอร์สำหรับการรู้จำสียงในการฝึกหัดการออกเสียงหรือพูดนั้นได้ถูกพัฒนาขึ้นอย่างต่อเนื่องตัวอย่างเช่น โปรแกรม ‘สปีชวิวเวอร์’ (*Speech Viewer*) ซึ่งมีความสามารถในการแสดงผลของการออกเสียงด้วยภาพทางหน้าจอซึ่งช่วยพัฒนาความเข้าใจในการออกเสียงของผู้ใช้ โดยเน้นการออกเสียงคำในภาษาอังกฤษ เช่นเดียวกับเครื่องมือส่วนมากที่มักจะเป็นภาษาตะวันตก [1] แม้ว่าจะมีการพัฒนาในภาษาจีน [2] บ้างก็ตาม ในการออกเสียงแต่ละภาษายังมีความแตกต่างในเรื่องของหน่วยเสียงและการเคลื่อนไหวของริมฝีปากซึ่งทำให้บางครั้งการใช้โปรแกรมที่ออกแบบสำหรับภาษาต่างประเทศนั้นยังไม่สามารถใช้งานได้อย่างสมบูรณ์แบบ เนื่องจากเทคนิคที่มีอยู่จะใช้การจับคู่ลักษณะเด่นของเสียงผู้ใช้กับคำพูดในฐานข้อมูลโดยแบบจำลองเอชเอ็มเอ็ม แต่เสียงในภาษาที่มีรากของภาษาต่างกัน จะมีสมบัติต่าง ๆ กัน เช่น หน่วยเสียง ลักษณะของรูปปาก และอวัยวะในการออกเสียง เป็นต้น ดังนั้นการสร้างภาพเคลื่อนไหวสำหรับเสียงภาษาไทยจึงจำเป็นต้องปรับปรุงวิธีการการจับคู่ รวมทั้งเพิ่มเติมรูปแบบของหน่วยเสีย รูปปาก รวมถึงปรับค่าแบบจำลองในการวิเคราะห์เสียงให้เหมาะสมกับภาษาไทยมากขึ้นโดย คำนึงถึงปัจจัยในการสร้างภาพเช่น เวลาในการคำนวณ หรือความสมจริงของผลลัพธ์โดยผู้ใช้งานมากขึ้น

ในปัจจุบันมีโมเดลการรู้จำเสียงที่ถูกพัฒนาเพื่อใช้ในภาษาไทย [3] สำหรับการระบุข้อความจากเสียงพูด ในงานวิจัยนี้จะใช้โมเดลเสียงที่ใช้ในภาษาไทยดังกล่าวมาทำการปรับปรุงโมเดลเสียงเพื่อใช้ในการระบุระยะเวลา

doi:10.4186/ejth.2011.3.2.33

ของหน่วยเสียงจากเสียงพูดและนำหน่วยเสียงที่ระบุระยะเวลาได้ประกอบกับฐานข้อมูลการเคลื่อนไหวของริมฝีปากและทำการเรียบเรียงเป็นการเคลื่อนไหวของริมฝีปากในเสียงพูดตั้งต้นได้ ขั้นตอนดังกล่างเป็นขั้นตอนที่ได้มาจากการดัดแปลงโครงสร้างซอฟต์แวร์ในการสร้างการเคลื่อนไหวของตัวละครสามมิติโดยเสียงพูด [4] และการสร้างฐานข้อมูลการเคลื่อนไหวของริมฝีปาก [5] และการเชื่อมการเคลื่อนไหวระหว่างหน่วยเสียง [6]

III. ขั้นตอนดำเนินการวิจัย

ขั้นตอนการวิจัยจะประกอบด้วยเจ็ดขั้นตอนซึ่งทำการประยุกต์วิธีจากการจับคู่เสียงกับภาพของภาษาต่างประเทศและปรับให้เข้ากับภาษาไทยที่ลักษณะเฉพาะไม่เหมือนกัน

3.1 การระบุระยะเวลาของหน่วยเสียงภาษาไทย

โมเดลเสียงที่สร้างขึ้นจะสร้างจากหน่วยเสียงบนเสียงภาษาไทยซึ่งมีจำนวน 74 หน่วยเสียงรวมกับเสียงเงียบอีก 1 เสียง (sil) เป็น 75 หน่วยเสียงตามมาตรฐานของหน่วยเสียงภาษาไทย (CMU Thai phoneme)[7] โดยโมเดลเสียงจะใช้วิธีการ แบบจำลองฮิดเด้นมาคอฟหรือเอชเอ็มเอ็ม (Hidden Markov Model: HMM) ในการประมวลผล โมเดลเสียงจะใช้ระบุระยะเวลาของแต่ละหน่วยเสียงโดยใช้วิธีระบุระยะเวลาของหน่วยเสียงในเสียงพูด (Force-alignment) ซึ่งเป็นฟังก์ชันหนึ่งใน เครื่องมือเอชทีเค มีข้อมูลนำเข้าเป็นข้อมูลเสียง (wave file) และข้อมูลลำดับของหน่วยเสียงจากไฟล์เสียง

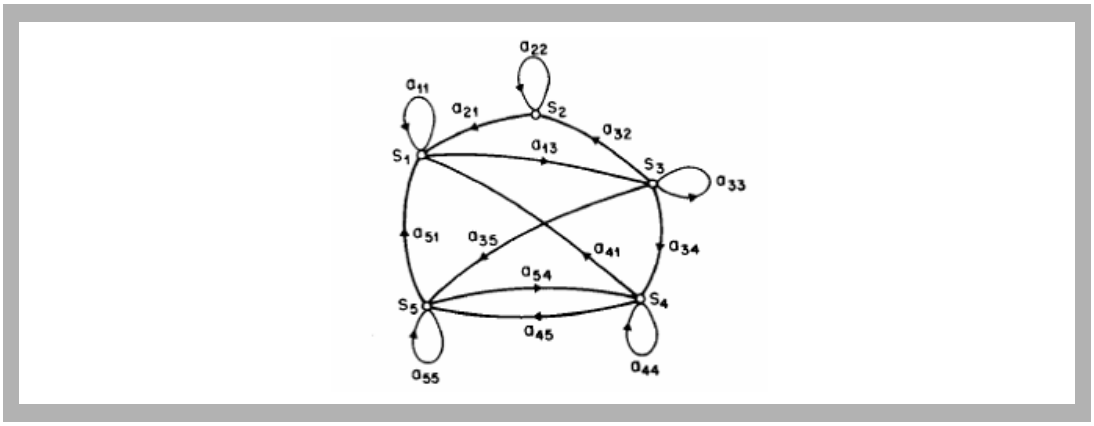
3.2 แบบจำลองฮิดเด้นมาคอฟ (Hidden Markov Model)

หลักการทำงานของ เอชเอ็มเอ็ม คือการแปลงตัวอย่างหรือลำดับ ของคำใน ข้อมูลหรือสัญญาณเสียง มาคำนวณความน่าจะเป็นของข้อมูลเหล่านั้น โดย เอชเอ็มเอ็ม จะประกอบไปด้วย สถานะต่างๆ ประกอบไปด้วยความน่าจะเป็นสองชุด คือ ความน่าจะเป็นในการสังเกตสถานะ และ ความน่าจะเป็นในการเปลี่ยนสถานะปัจจุบันไปยังสถานะอื่น โดยการทำงานของเอชเอ็มเอ็มมีขั้นตอนดังนี้

```
01  วนซ้ำ {
02      สุ่มค่า ความน่าจะเป็นในการสังเกตสถานะปัจจุบัน
03      สุ่มค่า ความน่าจะเป็นการเปลี่ยนสถานะปัจจุบัน ไปยังสถานะอื่น
04      กำหนด สถานะที่สุ่มใหม่มาแทนที่เป็นสถานะปัจจุบัน
05  } จนกระทั่งพบเงื่อนไขในการสุ่ม (ความยาวหรือขนาดที่ต้องการ)
06  พบสถานะจบ
07  สุ่มค่าจนได้ค่าเป็น สถานะจบ
```

กระบวนการมาคอฟจะพิจารณาระบบที่อธิบายถึงช่วงเวลาของกลุ่มสถานะจำนวน N สถานะ เขียนแทนด้วย S_1 ไปจนถึง S_n จากรูปด้านล่างเป็นการกำหนดสถานะ 5 สถานะ และให้ i เป็นสถานะต้น และ j เป็นสถานะปลาย

doi:10.4186/ejth.2011.3.2.33

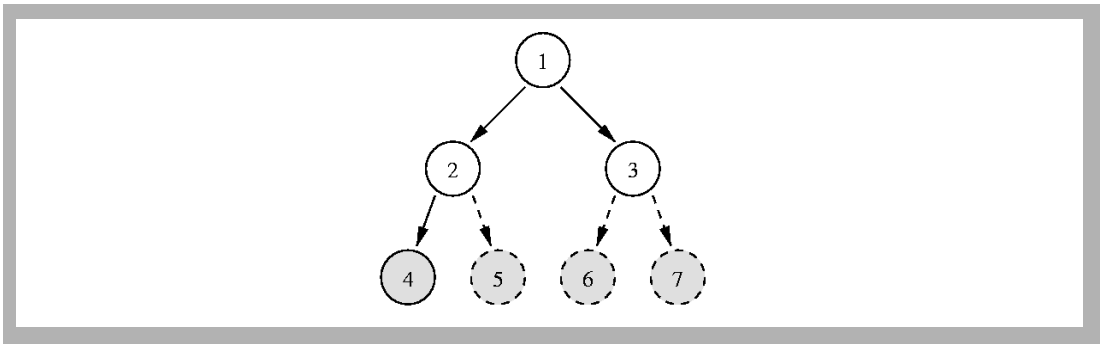


รูปที่ 1
แผนภาพแสดงการทำงานของ Markov process

3.3 การปรับปรุงโมเดลเสียง (Model adaptation)

การปรับปรุงและพัฒนาโมเดลเสียงจะทำให้โมเดลเสียงเดิมมีประสิทธิภาพในการระบุระยะเวลาของหน่วยเสียงสูงขึ้นเมื่อนำไปใช้กับผู้พูดใหม่หรือกลุ่มเฉพาะของผู้พูดมากกว่าโมเดลเสียงดั้งต้นทั่วไปมีรูปแบบการปรับปรุงโมเดลเสียงที่ผู้วิจัยเลือกใช้มาสองวิธี

วิธีเอ็มแอลแอลอาร์ (Maximum likelihood linear regression: MLLR) [8]
การแปลงรูปแบบของโมเดลเพื่อเตรียมทำการปรับปรุงโมเดลเสียงโดยจะยึดความเป็นไปได้ที่ใกล้เคียงโดยการสร้างต้นไม้ถดถอย โดยจะเลือกจำนวนและประเภทของข้อมูล เพื่อหาวิธีการปรับปรุงที่ใกล้เคียงที่สุดโดยจะคำนึงค่าความสูงของต้นไม้เพื่อใช้ในการหาโมเดลเสียงที่ใกล้เคียงมากที่สุด



รูปที่ 2
แผนภาพแสดงต้นไม้ถดถอย

วิธี เอ็มเอพี (Maximum a posteriori estimation: MAP) [8]
การปรับปรุงโดยใช้วิธีของเบยย์ (Bayesian) เป็นการปรับโมเดลโดยใช้ตัวแปรจากโมเดลเดิมมาเป็นขั้นตอนก่อนหน้าเพื่อที่จะหาแนวโน้มความใกล้เคียงค่าใหม่โดย การหาค่าสูงสุดของความหนาแน่นในโมเดลเอชเอ็มเอ็ม

3.4 การแปลงค่ามุมออยเลอร์และควอเทอเนียน

โดยทั่วไปการหมุนวัตถุจะนิยมใช้มุมออยเลอร์เนื่องจากค่าของตัวแปรในการกำหนดเป็นค่าที่มนุษย์สามารถนำไปใช้ได้ง่ายกว่าเพราะเป็นค่ามุมบนระนาบสามมิติ และมุมควอเทอเนียนก็เป็นทางเลือกของการหมุนเช่นเดียวกันโดยแอปพลิเคชัน 'เบลนเดอร์' (*Blender*) จะใช้การหมุนโดยใช้มุมควอเทอเนียนในการคำนวณการหมุนกระดูกของโมเดล ซึ่งมุมควอเทอเนียนมีพื้นฐานในการคิดมาจากจำนวนเชิงซ้อนและมีข้อดีคือสามารถสร้างและปรับค่าของการให้มีความลื่นไหลมากกว่าวิธีการใช้การหมุนด้วยการปรับค่าบนมุมออยเลอร์ การใช้มุมควอเทอเนียนสามารถประยุกต์นำมาใช้ในการคำนวณเมตริกซ์ขนาด 4x4 เมตริกซ์ซึ่งเป็นเมตริกซ์ที่นิยมใช้ในการคำนวณการหมุนของวัตถุได้ง่ายกว่า

doi:10.4186/ejth.2011.3.2.33

มุมควอเทอเนียนมีรูปแบบการเขียนสมการดังนี้

$$Quarternion's\ angle = [a,b,c,d]$$

สมการแปลงมุมออยเลอร์ ให้เป็นมุมควอเทอเนียนโดยกำหนดให้ตัวแปรในมุมออยเลอร์เป็น(X,Y,Z) และตัวแปรในมุมควอเทอเนียนเป็น (a,b,c,d) ตามลำดับ จะได้วิธีการแปลงสมการจากมุมออยเลอร์เป็นมุมควอเทอเนียนดังนี้

$$a = \left(cos\left(\frac{x}{2}\right) \times cos\left(\frac{y}{2}\right) \times cos\left(\frac{z}{2}\right)\right) + \left(sin\left(\frac{x}{2}\right) \times sin\left(\frac{y}{2}\right) \times sin\left(\frac{z}{2}\right)\right)$$

$$b = \left(sin\left(\frac{x}{2}\right) \times cos\left(\frac{y}{2}\right) \times cos\left(\frac{z}{2}\right)\right) + \left(cos\left(\frac{x}{2}\right) \times sin\left(\frac{y}{2}\right) \times sin\left(\frac{z}{2}\right)\right)$$

$$c = \left(cos\left(\frac{x}{2}\right) \times sin\left(\frac{y}{2}\right) \times cos\left(\frac{z}{2}\right)\right) + \left(sin\left(\frac{x}{2}\right) \times sin\left(\frac{y}{2}\right) \times sin\left(\frac{z}{2}\right)\right)$$

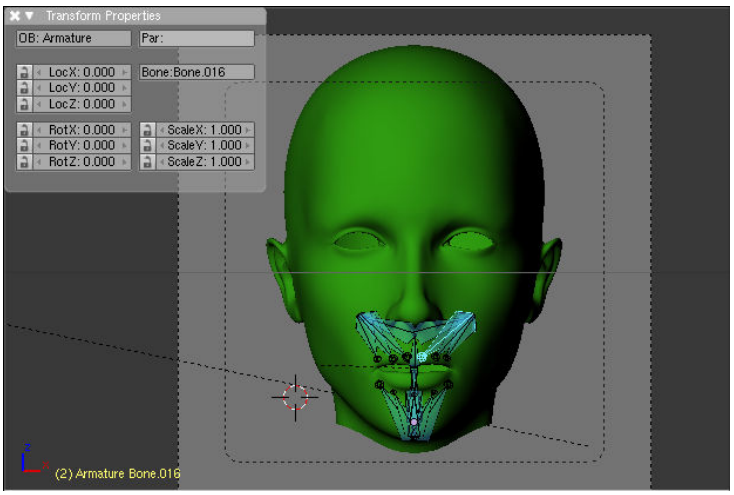
$$d = \left(cos\left(\frac{x}{2}\right) \times cos\left(\frac{y}{2}\right) \times sin\left(\frac{z}{2}\right)\right) + \left(sin\left(\frac{x}{2}\right) \times sin\left(\frac{y}{2}\right) \times cos\left(\frac{z}{2}\right)\right)$$

3.5 การสร้างแบบจำลองใบหน้า

ในงานวิจัยนี้จะใช้แบบจำลองใบหน้าที่จะลองกล้ามเนื้อมนุษย์ที่ใช้เคลื่อนไหวบนใบหน้ามาใช้ โดยจะทำการวางโครงกระดูกให้กับใบหน้า (armature) โดยเราจะใช้กระดูกในการขยับริมฝีปากด้านบนซ้ายจำนวน 3 ท่อนและด้านบนขวา 3 ท่อน และทำการวางกระดูกด้านล่างทั้งหมด 5 กระดูกตามรูปกล้ามเนื้อที่ใช้ในการเคลื่อนที่บนใบหน้า ซึ่งในการเคลื่อนที่ระหว่างเฟรมที่เราได้ทำการตั้งค่า (Key frame) ไว้

3.6 การสร้างฐานข้อมูลการเคลื่อนไหวของริมฝีปากตามเสียงพูด

ฐานข้อมูลการเคลื่อนไหวของริมฝีปากตามเสียงพูดจะใช้วิธีการเก็บข้อมูลจากการนำวีดีโอการพูดของผู้พูดมาทำการระบุระยะเวลาของแต่ละหน่วยเสียงจากในวีดีโอ จากนั้นจะทำการเก็บข้อมูลของตำแหน่งใบหน้าวีดีโอขณะที่สิ้นสุดหน่วยเสียง โดยจะเก็บค่าของตำแหน่งบนใบหน้าให้กับทุกหน่วยเสียงในภาษาไทย



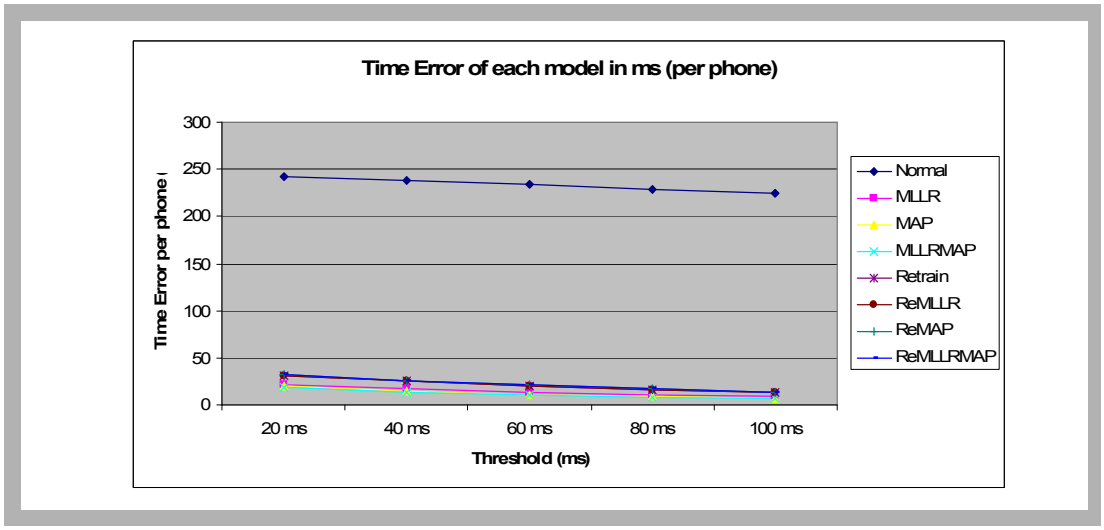
รูปที่ 3
ภาพแสดงใบหน้าของตัวละครต้นแบบขณะออกเสียง sil

doi:10.4186/ejth.2011.3.2.33

3.7 การประสานเวลาระหว่างการเคลื่อนไหวของริมฝีปากและเสียงพูด

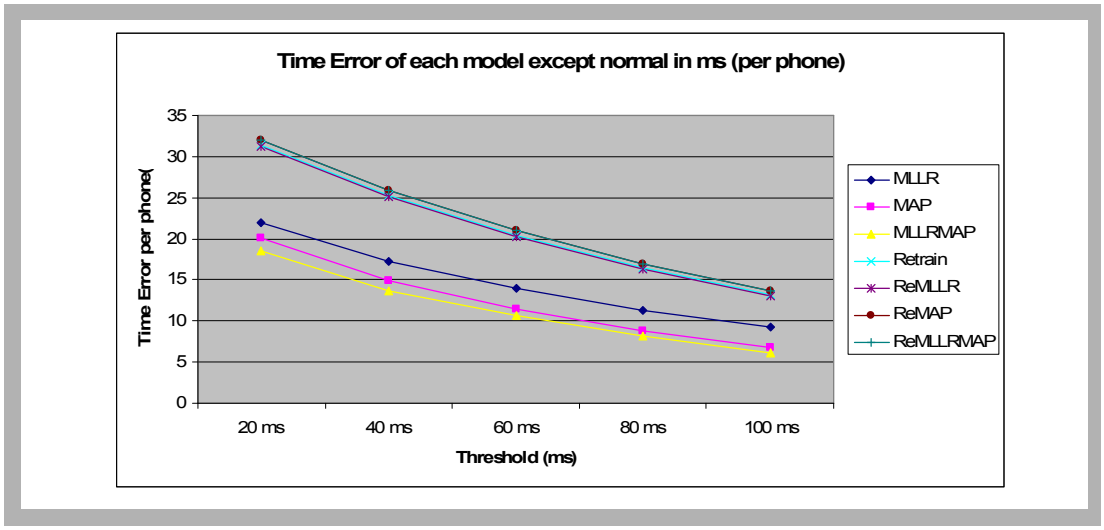
ขั้นตอนนี้จะทำการกำหนดค่าของคีย์เฟรมในการเคลื่อนไหวของริมฝีปากด้วยเวลาสิ้นสุดของหน่วยเสียงในการพูด โดยจะใช้ฐานข้อมูลของการเคลื่อนไหวของริมฝีปากเป็นตัวกำหนดตำแหน่งบนใบหน้าที่มีการขยับริมฝีปากไปตามหน่วยเสียงต่างๆในประโยค

IV. ผลการทดลอง



รูปที่ 4 กราฟแสดงร้อยละความผิดพลาดของการระบุระยะเวลาหน่วยเสียง ทุกโมเดลเสียง

การคัดเลือกวิธีในการปรับแต่งโมเดลเสียงที่ได้ผลลัพธ์ในการระบุระยะเวลาเสียงที่แม่นยำมากที่สุดจากรูป 4 เป็นการวัดค่าความผิดพลาดของการระบุระยะเวลาหน่วยเสียงในแต่ละวิธีการปรับปรุงโมเดลเสียงทุกรูปแบบในการทดลองซึ่งสามารถสังเกตได้ว่าความผิดพลาดในรูปแบบที่ไม่มีการปรับปรุงโมเดลเสียงมีความผิดพลาดมากกว่าการปรับปรุงด้วยวิธีต่างๆอย่างชัดเจน ซึ่งจะแยกพิจารณาในรูปที่ 5 เป็นการวัดค่าความผิดพลาดของการระบุระยะเวลาหน่วยเสียงในแต่ละวิธีการปรับปรุงโมเดลเสียง (ไม่รวมการใช้โมเดลที่ไม่มีการปรับปรุง) เพื่อเปรียบเทียบระหว่างโมเดลเสียงที่ทำการปรับปรุงแล้วเท่านั้น โดยแบ่งการพิจารณาตามค่าความผิดพลาดที่ยอมรับได้แบ่งออกเป็นห้าช่วง โดยจะนับความผิดพลาดจากเวลาที่หน่วยเสียงนั้นเกิน ค่าความผิดพลาดที่ยอมรับได้ไปในหน่วยเวลา มิลลิวินาที (ms) ได้ผลลัพธ์ว่าการปรับปรุงโมเดลเสียงแบบ เอ็มแอลแอลอาร์ (MLLR) เอ็มเอพี (MAP) และ เอ็มแอลแอลอาร์เอ็มเอพี (MLLRMAP) ดีที่สุดตามลำดับ โดยโมเดลเสียงที่ใช้การเรียนรู้ซ้ำกับโมเดลเสียง (Retrain) จะได้ผลลัพธ์ที่มีความแม่นยำต่ำกว่า



รูปที่ 5 กราฟแสดงร้อยละความผิดพลาดของการระบุระยะเวลาหน่วยเสียง ในโมเดลที่ทำการปรับปรุงแล้ว

doi:10.4186/ejth.2011.3.2.33

การปรับปรุงโมเดลเสียงด้วยวิธีเอ็มแอลแอลอาร์ (MLLR) ได้ผลดีกว่าวิธีอื่น เนื่องจากจำนวนหน่วยเสียงในคำพูดมีจำนวนพยางค์ไม่มาก ซึ่งวิธีเอ็มแอลแอลอาร์จะสามารถปรับปรุงโมเดลเสียงที่มีความสูงของต้นไม้ถดถอยได้รวดเร็วและได้ผลลัพธ์แม่นยำ ส่วนวิธีเอ็มเอพี (MAP) ได้ผลลัพธ์ดีเช่นกันเนื่องจากจำนวนข้อมูลที่ใช้ทดสอบและปรับปรุงโมเดลเสียงมีจำนวนมากซึ่งส่งผลให้การปรับปรุงในลักษณะนี้มีความแม่นยำสูงด้วย ส่วนวิธีปรับปรุงแบบเอ็มแอลแอลอาร์และปรับปรุงต่อด้วยเอ็มเอพี (MLLRMAP) จะได้ความแม่นยำมากที่สุดเพราะได้รับข้อดีในการปรับปรุงจากทั้งสองรูปแบบข้างต้น

การทดลองจากผู้ใช้จำนวนห้าคนวัดระดับความเหมือนระหว่างการเคลื่อนไหวของตัวละครแอนิเมชันกับการเคลื่อนไหวของมนุษย์ โดยทำการทดลองกับวีดีโอที่สร้างขึ้นจากเสียงพูดเจ็ดประโยค ได้ผลดังนี้

- Lab_001.avi (“ นิสิตจุฬา ”) ผลที่ได้คือ 88.89%
- Lab_003.avi (“ ป้ายรถเมล์ “) ผลที่ได้คือ 75.75%
- Lab_004.avi (“ เจ้าขุนทองมาแล้วจ้า ”) ผลที่ได้คือ 82.46%
- Lab_005.avi (“ หิวข้าวจังเลย ”) ผลที่ได้คือ 87.18%
- Lab_006.avi (“ คุณสบายดีมั้ย ”) ผลที่ได้คือ 83.33%
- Lab_007.avi (“ น้องมาลีมีลูกแมวเหมียว ”) ผลที่ได้คือ 87.04%
- Lab_009.avi (“ ขอเกรดเอด้วยนะครับอาจารย์ ”) ผลที่ได้คือ 92.08%

จากการพิจารณาระดับคะแนนที่ได้ของแต่ละวีดีโอพบว่าคะแนนส่วนใหญ่ในการประเมินความเหมือนจะอยู่ในระดับคะแนน 75 เปอร์เซ็นต์ขึ้นไป จนถึงระดับ 92 เปอร์เซ็นต์ ซึ่งผลลัพธ์จัดว่าอยู่ในระดับที่น่าพอใจ ซึ่งสามารถอธิบายว่าการเคลื่อนไหวของตัวละครแอนิเมชันที่เป็นผลลัพธ์ของแอปพลิเคชันมีความเหมือนการออกเสียงจริงของมนุษย์ค่อนข้างสูง ส่วนหน่วยเสียงบางหน่วยในข้อมูลเสียงที่ได้ผลการทดลองที่ไม่ดีนักส่วนใหญ่จะเกิดในข้อมูลเสียงที่มีสระเสียงสั้นหรือมีการออกเสียงเร็วทำให้เกิดการระบุระยะเวลาของหน่วยเสียงมีความคลาดเคลื่อนสูง และส่งผลให้การเปลี่ยนของหน่วยเสียงเกิดขึ้นบ่อยจึงทำให้การเคลื่อนไหวของริมฝีปากของตัวละครแอนิเมชันไม่ลื่นไหลหรือเกิดขึ้นเร็วมากเกินไป จึงทำให้ผู้ประเมินมองว่าการเคลื่อนไหวของริมฝีปากผิดจากความเป็นจริงหรือไม่ลื่นไหลเท่าที่ควร แต่เมื่อเปรียบเทียบกับวิธีการใช้แบบจำลองเอชเอ็มเอ็มกับหน่วยเสียงภาษาต่างประเทศพบว่าสามารถให้ความแม่นยำซึ่งวัดจากเวลาในการออกหน่วยเสียงสูงกว่า

V. สรุปผล

งานวิจัยจับคู่การเคลื่อนไหวของหน่วยเสียงการเคลื่อนไหวของริมฝีปากตามหน่วยเสียง โดยใช้วิธีระบุเวลาของหน่วยเสียงแต่ละหน่วยจับคู่ระหว่างเวลาของหน่วยเสียงกับการเคลื่อนไหวในฐานข้อมูลการเคลื่อนไหวของริมฝีปากในแต่ละหน่วยเสียงและทำการเรียบเรียงการเคลื่อนไหวของริมฝีปากทั้งประโยคเข้าด้วยกัน โดยขั้นตอนการสร้างการเคลื่อนไหวของริมฝีปากตามเสียงพูดในภาษาไทย ทำให้มีประสิทธิภาพในด้านความเร็วในการสร้าง และความถูกต้องเพิ่มขึ้น อีกทั้งยังสร้างฐานข้อมูลการเคลื่อนไหวของริมฝีปากในแต่ละหน่วยเสียงซึ่งสามารถนำไปประยุกต์ใช้กับตัวละครแอนิเมชันที่ต้องการพูดภาษาไทย และสร้างโครงกระดูกกำกับการเคลื่อนไหวของริมฝีปาก ที่สอดคล้องกับการใช้ฐานข้อมูลการเคลื่อนไหวของริมฝีปากซึ่งสามารถปรับเปลี่ยนไปใช้กับตัวละครที่มีลักษณะใกล้เคียงกับตัวละครต้นแบบได้

doi:10.4186/ejth.2011.3.2.33

ในกรณีที่ที่ข้อมูลเสียงที่มีสระเสียงสั้นหรือมีการออกเสียงเร็วทำให้เกิดการระบุระยะเวลาของหน่วยเสียงมีความคลาดเคลื่อนสูง และส่งผลให้การเปลี่ยนของหน่วยเสียงเกิดขึ้นบ่อยจึงทำให้การเคลื่อนไหวของริมฝีปากเร็วเกินไป ซึ่งสามารถพัฒนาเพิ่มเติมโดยปรับปรุงวิธีการสร้างภาพเคลื่อนไหวที่มีประสิทธิภาพสูงขึ้น

กิตติกรรมประกาศ

งานวิจัยชิ้นนี้ได้รับทุนสนับสนุนจาก โครงการทุนสถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยีไทย (ทุน TGIST) โดยสำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.) รหัสทุน TG-44-09-53-063M

doi:10.4186/ejth.2011.3.2.33

บรรณานุกรม

[1] D.W. Massaro, “A computer-animated tutor for spoken and written language learning,” *Proceedings of the 5th international conference on Multimodal interfaces*, New York, New York, USA: ACM, 2003, p. 172–175.

[2] X. Jiang, Y. Wang, and F. Zhang, “Visual speech analysis and synthesis with application to Mandarin speech training,” *Proceedings of the ACM symposium on Virtual reality software and technology*, New York, New York, USA: ACM, 1999, p. 111–115.

[3] N. Thatphithakkul and B. Kruatrachue, “Denoise speech recognition based on wavelet transform using threshold estimation,” *Electrical Engineering Conference (EECON), Thailand (in Thai)*, 2004, pp. 2-5.

[4] K. Waters, T.M. Levergood, and D.E.C.C.R. Laboratory, *DECface: An automatic lip-synchronization algorithm for synthetic faces*, New York, New York, USA: Citeseer, 1993.

[5] P. Kakumanu, R. Gutierrez-Osuna, A. Esposito, R. Bryll, A. Goshtasby, and O. Garcia, “Speech driven facial animation,” *Proceedings of the 2001 workshop on Perceptive user interfaces*, ACM, 2001, p. 1–5.

[6] others, “Expressive Facial Animation Synthesis by Learning Speech Co-Articulation and Expression,” *Space, IEEE Transaction on Visualization and Computer Graphics*, Citeseer, pp. 1-12, 2006.

[7] S. Suebvisai, *Thai automatic speech recognition*, 2005.

[8] H.V. Sharma and M. Hasegawa-johnson, “State-Transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition,” *Computational Linguistics*, pp. 72-79, 2010.