# ระบบสังเคราะห์เสียงร้องเพลงภาษาไทยโดยใช้แบบจำลองฮิดเดนมาร์คอฟ

นายลัทธพล จีระประดิษฐ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

# HMM-BASED THAI SINGING VOICE SYNTHESIS SYSTEM

Mr. Lattapon Jeerapradit



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | ระบบสังเคราะห์เสียงร้องเพลงภาษาไทยโดยใช้แบบจำลองฮิดเดนมาร์คอฟ |
| โดย | นายลัทธพล จีระประดิษฐ |
| สาขาวิชา | วิศวกรรมคอมพิวเตอร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม | ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ |

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ธนารัตน์ ชลิดาพงศ์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.ศิรินาถ ตั้งรวมทรัพย์)

ลัทธพล จีระประดิษฐ : ระบบสังเคราะห์เสียงร้องเพลงภาษาไทยโดยใช้แบบจำลองฮิดเดนมาร์คอฟ. ( HMM-BASED THAI SINGING VOICE SYNTHESIS SYSTEM) อ.ที่ปรึกษาหลัก : รศ. ดร.อติวงศ์ สุชาโต, อ.ที่ปรึกษาร่วม : ผศ. ดร.โปรดปราน บุณยพุกกณะ

การร้องเพลงในแต่ละภาษานั้นมีเอกลักษณ์เฉพาะตัวบางอย่างซึ่งส่งผลให้การพัฒนาความเป็นธรรมชาติของเสียงร้องเพลงสังเคราะห์ในแต่ละภาษานั้นมีความท้าทายแตกต่างกัน เสียงวรรณยุกต์เป็นส่วนที่มีผลมากกับการสื่อสารในภาษาที่มีเสียงวรรณยุกต์ แต่ระบบสังเคราะห์เสียงร้องเพลงในปัจจุบันไม่ได้คำนึงถึงเสียงวรรณยุกต์ นอกจากนี้เมลิสมาเป็นอีกสถานการณ์หนึ่งที่พบได้บ่อยครั้งในการร้องเพลงป็อปไทยซึ่งต้องมีการจัดการเพื่อจำลองการร้องเมลิสมา เป้าหมายของวิทยานิพนธ์นี้จึงมุ่งเน้นที่การปรับระบบสังเคราะห์เสียงร้องเพลงให้รองรับการจำลองเสียงในสถานการณ์เมลิสมาและผลกระทบของเสียงวรรณยุกต์

งานวิทยานิพนธ์นี้เสนอ 1) ปัจจัยบริบทที่ใช้ในระบบสังเคราะห์เสียงร้องเพลงสำหรับภาษาที่วรรณยุกต์มีผลต่อเสียงร้องเพลงและคำนึงถึงเมลิสมา 2) วิธีการทำสำเนารูปเขียน จากการประเมินผลพบว่า วิธีการทำสำเนารูปเขียนที่เสนอทั้งสองแบบนั้นส่งผลให้ระบบสังเคราะห์เสียงร้องเพลงรองรับเมลิสมา โดยวิธีการทำสำเนารูปเขียนที่คำนึงถึงสระเสียงสั้น-ยาวและตัวสะกดนั้นมีรูปคลื่นของเสียงร้องเพลงสังเคราะห์ใกล้เคียงกับรูปคลื่นของเสียงร้องเพลงจริงมากกว่า รวมถึงมีความเป็นธรรมชาติมากกว่าโดยใช้มาตรวัดเอ็มโอเอส อีกทั้งเมื่อมีปัจจัยบริบทที่เกี่ยวข้องกับเสียงวรรณยุกต์ เค้ารูปของความถี่มูลฐานที่สังเคราะห์ได้นั้นมีความใกล้เคียงเสียงร้องเพลงจริงมากกว่าในระบบที่ไม่มีปัจจัยบริบทที่เกี่ยวข้องกับเสียงวรรณยุกต์ และมีความเป็นธรรมชาติมากขึ้นโดยใช้มาตรวัดเอ็มโอเอส นอกจากนี้เพื่อเพิ่มความเป็นธรรมชาติให้เสียงร้องเพลงสังเคราะห์จึงมีการทดลองเกี่ยวกับจำนวนสถานะของแบบจำลองเสียงพบว่า เมื่อจำนวนสถานะเพิ่มขึ้น ความเป็นธรรมชาติของเสียงร้องเพลงสังเคราะห์ก็มากขึ้น แต่เมื่อถึงจุดหนึ่งเสียงร้องเพลงสังเคราะห์ที่ได้จะมีความเป็นธรรมชาติลดลง

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

# # 5870234121 : MAJOR COMPUTER ENGINEERING

KEYWORD: SINGING VOICE SYNTHESIS, HMM

Lattapon Jeerapradit : HMM-BASED THAI SINGING VOICE SYNTHESIS SYSTEM. Advisor: Assoc. Prof. Atiwong Suchato, Ph.D. Co-advisor: Asst. Prof. Proadpran Punyabukkana, Ph.D.

Singing synthesis in each language has its unique characteristics and challenges aiming to improve its naturalness. The effort regarding naturalness becomes more complicated for tonal languages. One of the reasons is due to the fact that the same word uttered in different tone yields different meaning. Nonetheless, no known research has attempted to include tone consideration into their singing synthesis models. Another challenge the tonal language singing synthesis faces is melisma for the same reason. Therefore, this research offers a tonal-melisma-compatible singing voice synthesis system. To do so, we propose 1) a contextual factors design which includes tone and melisma contexts, and 2) phoneme duplication methods. The results showed that the proposed phoneme duplication methods made the system compatible with melisma, where short vowels and final consonants constructed a favorable waveform closer to real singing voice and have a higher naturalness in MOS evaluation. Furthermore, a system with a tone context outperformed the baseline due to similarity of the generated F0 contour. Finally, in order to improve naturalness in the synthesized singing voice, an experiment with HMM state numbers was conducted. The outcome demonstrated that the naturalness increased as the state numbers grew to a certain point.

Field of Study: Computer Engineering
Academic Year: 2018

Student's Signature .....
Advisor's Signature .....
Co-advisor's Signature .....

# กิตติกรรมประกาศ

ข้าพเจ้าขอขอบคุณ รศ.ดร. อติวงศ์ สุชาโต อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ ผศ.ดร. โปรด ปราน บุณยพุกกณะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ท่านทั้งสองให้ความช่วยเหลือ คำแนะนำ และข้อคิดที่เป็นประโยชน์ อันเป็นส่วนสำคัญที่ทำให้วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้

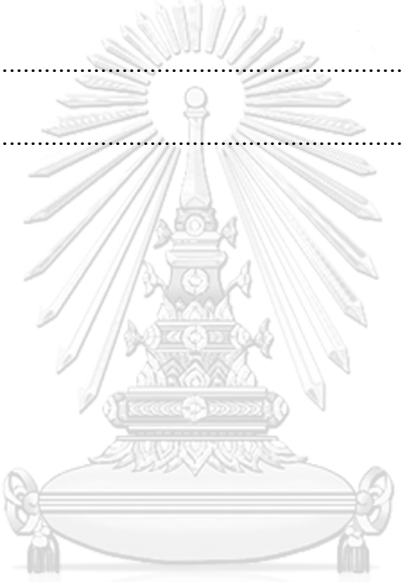
ขอขอบคุณคณะกรรมการสอบ ผศ.ดร. ธนารัตน์ ชลิดาพงศ์ และ ดร. ศิรินาถ ตั้งรวมทรัพย์ ที่สละเวลามานำเดินการสอบวิทยานิพนธ์ให้กับข้าพเจ้า รวมทั้งให้คำแนะนำและข้อคิดต่าง ๆ ที่เป็นประโยชน์ในการทำวิทยานิพนธ์

สุดท้ายนี้ขอขอบคุณบิดา มารดา รวมถึงพี่ ๆ เพื่อน ๆ และน้อง ๆ ที่ให้ความช่วยเหลือจนกระทั่งวิทยานิพนธ์นี้สำเร็จได้

ลัทธพล จีระประดิษฐ

# สารบัญ

**หน้า**



# บทที่ 1
# บทนำ

## 1.1 ที่มาและความสำคัญของปัญหา

ในช่วงศตวรรษที่ผ่านมาเทคโนโลยีต่าง ๆ ถูกพัฒนาอย่างต่อเนื่อง กระทั่งปัจจุบันคอมพิวเตอร์สามารถสังเคราะห์เสียงพูดของมนุษย์ได้ และในบางภาษามีคุณภาพเสียงของเสียงสังเคราะห์อยู่ในระดับที่ผู้ฟังนั้นพึงพอใจในความเป็นธรรมชาติ [1] หลังจากระบบสังเคราะห์เสียงพูดของมนุษย์ถูกพัฒนาขึ้นมา ระบบนั้นก็ถูกนำมาประยุกต์ในการสังเคราะห์เสียงร้องเพลงเช่นกัน เสียงร้องเพลงสังเคราะห์นั้นเริ่มได้รับความนิยมในญี่ปุ่นเป็นพิเศษ [2] เมื่อโปรแกรมสังเคราะห์เสียงร้องเพลงโวคาลอยด์ (VOCALOID) [3] ถูกพัฒนาขึ้นโดยบริษัทยามาฮ่า (YAMAHA) โปรแกรมสังเคราะห์เสียงร้องเพลงโวคาลอยด์นั้น เป็นหนึ่งในไม่กี่ตัวเลือกที่ผู้ใช้ทั่วไปสามารถเลือกใช้ได้และถูกเลือกใช้มากที่สุด เพราะโปรแกรมโวคาลอยด์ไม่ได้มีเพียงเครื่องประมวลผลสังเคราะห์ (Synthesis Engine) เท่านั้น แต่มาพร้อมกับสภาพแวดล้อมที่ใช้งานง่าย เหมาะกับผู้ใช้ทั่วไป

ปัจจุบันระบบสังเคราะห์เสียงร้องเพลงถูกพัฒนาขึ้นในหลากหลายภาษา เช่น ภาษาอังกฤษ [4] ภาษาจีน [5] และภาษาไต้หวัน [6] เป็นต้น ซึ่งในการพัฒนาเสียงร้องเพลงของแต่ละภาษาเพื่อให้มีความเป็นธรรมชาตินั้น มีความท้าทายแตกต่างกันไป เช่น การนำระบบสังเคราะห์เสียงร้องเพลงภาษาญี่ปุ่นมาใช้กับภาษาอังกฤษ ต้องมีการปรับแต่งระบบเพิ่มเติมเพื่อให้รองรับธรรมชาติของภาษานั้น ๆ เช่น ระบบสังเคราะห์เสียงร้องเพลงในภาษาจีนซึ่งเป็นภาษาที่มีเสียงวรรณยุกต์และหน่วยเสียงย่อยที่ไม่เหมือนภาษาญี่ปุ่น [7] เป็นต้น เมื่อกล่าวถึงงานด้านระบบสังเคราะห์เสียงมนุษย์ของภาษาไทยนั้น มีงานที่เกี่ยวกับระบบสังเคราะห์เสียงพูด [8, 9] ระบบสังเคราะห์เสียงพูด VAJA [10] ซึ่งพัฒนาโดย NECTEC เป็นหนึ่งในระบบสังเคราะห์เสียงพูดภาษาไทยที่พัฒนาขึ้น แต่จากการสำรวจยังไม่พบว่ามีงานวิจัยเกี่ยวกับระบบสังเคราะห์เสียงร้องเพลงภาษาไทย

ในงานวิจัยนี้จึงเสนอระบบสังเคราะห์เสียงร้องเพลงภาษาไทยโดยใช้แบบจำลองฮิดเดนมาร์คอฟ

## 1.2 วัตถุประสงค์งานวิจัย

งานวิจัยนี้มีขึ้นเพื่อเสนอระบบสังเคราะห์เสียงร้องเพลงภาษาไทยและพัฒนาเสียงร้องเพลงสังเคราะห์ให้มีความเป็นธรรมชาติมากขึ้น

## 1.3 ขอบเขตงานวิจัย

- กลุ่มเป้าหมายเพลงที่สนใจศึกษาคือ เพลงป็อปไทย (Thai pop songs)
- ชุดข้อมูลเสียงร้องเพลงเป็นเสียงนักร้องชายจากวงประสานเสียง 1 คน อัดเสียงโดยใช้ไมค์คอนเดนเซอร์ในสภาพแวดล้อมที่ไม่มีเสียงรบกวน

- ข้อมูลนำเข้าของระบบนี้คือ โน้ตเพลงและลำดับของรูปเขียนของเนื้อเพลง
- งานวิจัยนี้เป็นชุดเครื่องมือที่พัฒนาบนระบบปฏิบัติการลินุกซ์ (Linux)
- งานวิจัยนี้มุ่งความสนใจที่การพัฒนาความถี่มูลฐานและระยะเวลาของคำร้องที่สังเคราะห์ได้ เพื่อมีความเป็นธรรมชาติมากขึ้น

**1.4 แผนดำเนินงานวิจัย**

- ศึกษาและค้นคว้างานวิจัยที่เกี่ยวข้อง

ศึกษางานวิจัยเรื่องการสังเคราะห์เสียงร้องเพลงในภาษาต่าง ๆ รวมถึงเทคนิคต่าง ๆ ที่ใช้พัฒนาการฝึกฝนแบบจำลองเสียงร้องเพลง และเครื่องมือที่เกี่ยวข้องกับการสังเคราะห์เสียงร้องเพลง

- เตรียมชุดข้อมูลที่ใช้ฝึกฝนและทดลองแบบจำลองเสียง

ระบบสังเคราะห์เสียงร้องเพลงนั้นจำเป็นต้องใช้ข้อมูลเสียง ดังนั้นหลังจากศึกษางานวิจัยที่เกี่ยวข้อง จึงกำหนดชุดข้อมูลที่จะใช้ในระบบทั้งในส่วนฝึกฝนแบบจำลองเสียงและสำหรับการประเมินผล

- พัฒนาระบบสังเคราะห์เสียงร้องเพลงภาษาไทย

ภายหลังการเตรียมชุดข้อมูลเสร็จสิ้น จึงเริ่มพัฒนาระบบสังเคราะห์เสียงร้องเพลงภาษาไทยขึ้นมา พร้อมทั้งปรับเทคนิคต่าง ๆ เพื่อให้เหมาะสมกับระบบสังเคราะห์เสียงร้องเพลงภาษาไทย รวมถึงเสนอปัจจัยหรือเทคนิคอื่น ๆ ที่เกี่ยวข้องเพื่อพัฒนาให้เสียงร้องเพลงสังเคราะห์ภาษาไทยมีความเป็นธรรมชาติมากขึ้น

- ออกแบบการทดลอง

ตัวแปรในการทดลองคือปัจจัยและเทคนิคที่เสนอขึ้นมา เพื่อศึกษาผลกระทบของตัวแปรเหล่านั้นกับระบบสังเคราะห์เสียงร้องเพลงสังเคราะห์ภาษาไทย รวมถึงออกแบบการประเมินผลเพื่อชี้วัดตัวแปรต่าง ๆ ในการทดลอง



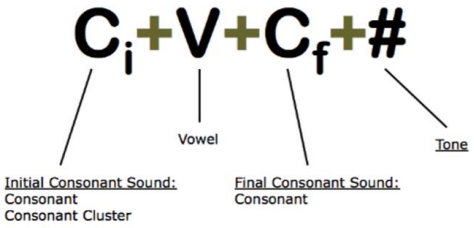
- รวบรวมผลการทดลองและสรุปผลการทดลอง

เมื่อกำหนดการทดลองได้แล้วจึงทำตามแผนการทดลองเพื่อเก็บผลการประเมิน จากนั้นจึงนำมาวิเคราะห์และสรุปผลการทดลอง

# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

## 2.1 เสียงในภาษาไทย

ภาษาไทยเป็นภาษาที่มีเสียงวรรณยุกต์ [11] ในหนึ่งคำประกอบไปด้วยพยางค์ตั้งแต่หนึ่งพยางค์ขึ้นไป แต่ละพยางค์จะมีเสียงอย่างน้อย 3 เสียง คือ เสียงพยัญชนะต้น เสียงสระ และเสียงวรรณยุกต์ สำหรับเสียงพยัญชนะสะกดนั้นในบางพยางค์อาจจะมีหรือไม่มีก็ได้ ดังนั้นโครงสร้างของพยางค์ในเสียงภาษาไทยจึงเป็นดังรูปที่ 1




รูปที่ 1 โครงสร้างของพยางค์ในภาษาไทย [12]

| พยัญชนะต้น | | | |
|---|---|---|---|
| เดี่ยว | ตัวอย่าง | ผสม | ตัวอย่าง |
| p | ปาก | pr | ประสาน |
| t | เต่า, กฎ | phr | พราน |
| c | จะ | tr | เตรียม |
| k | กอน | kr | กราบ |
| z | อาน | khr | คร่ำ |
| ph | พบ, ภัย, ผ่าน | pl | ปลา |
| th | ทึ่ง, ธง, เฒ่า,ฐาน, มณโฑ | phl | พลาด |
| ch | ช้อบ, เฌอ | thr | จันทรา |
| kh | คน, เข็น, ฆ่า | kl | เกลอ |
| b | บอก | khl | เคลื่อน |
| d | ด้าน, ชฎา | kw | กว้าง |
| m | ไม้ | khw | ขวา |
| n | นาน, เณร | เสียงทับศัพท์ | |
| ng | เงิน | br | เบรน |
| l | เล่น, ก็ฬำ | bl | บลู |
| r | รอ, ฤทัย | fr | ฟราย |
| f | ฝน, ฟัน | fl | เฟลม |
| s | สาย, ศิลา, รักษา, ซ่อน | dr | ดรากอน |
| h | โหน, เฮฮา | 17 หน่วย | |
| w | ว่า | | |
| j | ย้อน, หญิง | | |
| 21 หน่วย | | | |

| สระ | | | |
|---|---|---|---|
| เดี่ยว | ตัวอย่าง | ผสม | ตัวอย่าง |
| a | อะ | ia | เอียะ |
| aa | อา | iia | เอีย |
| i | อิ | va | เอือะ |
| ii | อี | vva | เอือ |
| v | อึ | ua | อัวะ |
| vv | อื | uua | อัว |
| u | อุ | 6 หน่วย | |
| uu | อู | | |
| e | เอะ | | |
| ee | เอ | | |
| x | แอะ | | |
| xx | แอ | | |
| o | โอะ | | |
| oo | โอ | | |
| @ | เอาะ | | |
| @@ | ออ | | |
| q | เออะ | | |
| qq | เออ | | |
| 18 หน่วย | | | |

| ตัวสะกด | |
|---|---|
| เดี่ยว | ตัวอย่าง |
| p^ | พบ |
| t^ | เกร็ด |
| k^ | ปาก |
| n^ | หาร |
| m^ | ลม |
| ng^ | ฟาง |
| j^ | ยาย |
| w^ | กาว |
| เสียงทับศัพท์ | |
| f^ | กราฟ |
| l^ | แอล |
| s^ | เอส |
| ch^ | คลัช |
| 12 หน่วย | |

รูปที่ 2 ตารางหน่วยเสียงรูปอ่านในภาษาไทย [13]

เสียงพยัญชนะต้น (Initial Consonant Sound) เสียงพยัญชนะต้นเดี่ยวมีทั้งหมด 21 เสียง เสียงพยัญชนะผสม 12 เสียง เมื่อรวมเสียงพยัญชนะผสมจากคำทับศัพท์อีก 5 เสียง จะได้เสียงพยัญชนะผสม 17 เสียง ดังนั้นเสียงพยัญชนะต้นรวมแล้วมีทั้งหมด 38 เสียง

เสียงสระ (Vowel) เสียงสระเดี่ยวมีทั้งหมด 18 เสียง และเสียงสระประสมอีก 6 เสียง รวมแล้วมีทั้งหมด 24 เสียง สำหรับสระอำ ไอ ใอ และเอา นั้นเรียกว่า สระเกิน คือ เป็นสระที่มีเสียงพยัญชนะผสมอยู่ด้วย

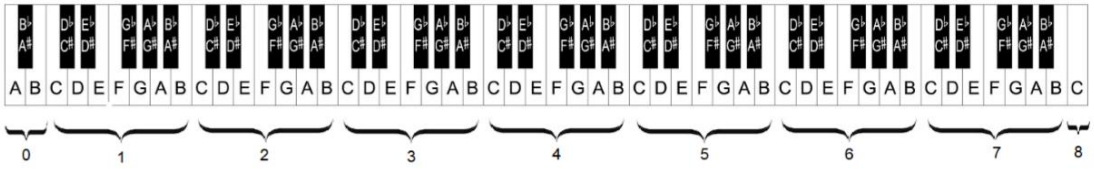
เสียงพยัญชนะสะกด (Final Consonant Sound) เสียงพยัญชนะสะกดมีทั้งหมด 8 เสียง เมื่อรวมเสียงพยัญชนะสะกดจากคำทับศัพท์อีก 4 เสียง จะได้เสียงพยัญชนะสะกดทั้งหมด 12 เสียง

เสียงวรรณยุกต์ (Tone) เสียงวรรณยุกต์มีทั้งหมด 5 เสียง [14]

**2.2 ความถี่ของระดับเสียงทางดนตรี (Musical Pitch Frequency)**

2.2.1 ระดับเสียง (Pitch)

มนุษย์สามารถรับรู้ความความสูงต่ำของเสียงได้จากความถี่ของเสียงนั้น ๆ เมื่อความถี่ยิ่งมาก เสียงที่เรารับรู้ก็จะยิ่งสูง ความสูงต่ำของเสียงที่รับรู้ได้นี้เรียกว่า ระดับเสียง (Pitch) ระดับเสียงในทางดนตรีนั้นมีการแทนด้วยสัญลักษณ์หลายรูปแบบ หนึ่งในรูปแบบที่นิยมมากที่สุดคือ การใช้ตัวอักษรภาษาอังกฤษ A ถึง G ตามด้วยเครื่องหมายแปลงเสียง (หากมี) และตามด้วยตัวเลข เช่น Bb4 หมายถึงระดับเสียง Bb ในอ็อกเทฟที่ 4 นอกจากนั้นยังมีการกำหนดระดับเสียงมาตรฐาน (Standard Pitch) เพื่อใช้ในการปรับแต่งเครื่องดนตรี ระดับเสียงมาตรฐานนี้แตกต่างกันไปตามยุคสมัย แต่ในปัจจุบันระดับเสียงมาตรฐานที่นิยมใช้กันคือ A4 = 440 Hz ตามมาตรฐานไอเอสโอ 16 ปี 1975




รูปที่ 3 ภาพแสดงตำแหน่งของระดับเสียงบนลิ่มเปียโน

2.2.2 การปรับแต่งความถี่ของระดับเสียงทางดนตรี (Musical Pitch Frequency Tuning)

การปรับแต่งความถี่ของระดับเสียงทางดนตรีนั้นมีการพัฒนามาเรื่อย ๆ หากกล่าวถึงเฉพาะการปรับแต่งความถี่ของระดับเสียงบนเปียโน การปรับแต่งระดับเสียงเสียงแบบเท่ากัน (Equal Tempered Tuning) เป็นวิธีที่ใช้ในปัจจุบัน [15] การปรับแต่งระดับเสียงแบบเท่ากันนั้นใช้วิธีแบ่งระยะห่างของแต่ละระดับเสียงในหนึ่งอ็อกเทฟโดยที่ ระยะห่างของแต่ละระดับเสียงมีอัตราส่วนของความถี่เท่ากัน วิธีนี้นำมาใช้ปรับแต่งระดับเสียงบนเปียโนเพื่อรักษาระยะห่างของระดับเสียงเดียวกันในแต่ละอ็อกเทฟให้เป็นขั้นคู่แปดเพอร์เฟกต์

$$f_n = f_{n-1} \times 2^{\frac{1}{12}} \qquad (1)$$

สมการ (1) เป็นความสัมพันธ์ของความถี่ของระดับเสียงที่อยู่ติดกันเมื่อใช้วิธีการปรับแต่งระดับเสียงแบบเท่ากัน เมื่อ $f_n$ คือความถี่ของระดับเสียงใด ๆ และ $f_{n-1}$ คือความถี่ของระดับเสียงที่ต่ำกว่าหนึ่งระดับเสียง หากกำหนดให้ระดับเสียงมาตรฐานคือ A4 = 440 Hz เราสามารถคำนวณหาความถี่ของระดับเสียงอื่น ๆ ได้ดัง แสดงในตารางที่ 1

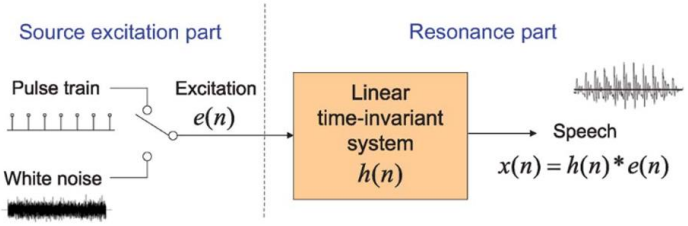
| ระดับเสียง | ความถี่ (Hz) | ระดับเสียง | ความถี่ (Hz) |
|---|---|---|---|
| A3 | 220 | A4 | 440 |
| A#3/Bb3 | 233.08 | A#4/Bb4 | 466.16 |
| B3 | 246.94 | B4 | 493.88 |
| C4 | 261.63 | C5 | 523.25 |
| C#4/Db4 | 277.18 | C#5/Db5 | 554.37 |
| D4 | 293.66 | D5 | 587.33 |
| D#4/Eb4 | 311.13 | D#5/Eb5 | 622.25 |
| E4 | 329.63 | E5 | 659.26 |
| F4 | 349.23 | F5 | 698.46 |
| F#4/Gb4 | 369.99 | F#5/Gb5 | 739.99 |
| G4 | 392 | G5 | 783.99 |
| G#4/Ab4 | 415.3 | G#5/Ab5 | 830.61 |
| A4 | 440 | A5 | 880 |

ตารางที่ 1 ตัวอย่างความถี่ของระดับเสียงที่คำนวณได้จากระดับเสียงมาตรฐาน A4 = 440 Hz โดยใช้การปรับแต่งระดับเสียงแบบเท่ากัน



**2.3 ตัวเข้ารหัสเสียงพูด (Vocal encoder)**

ตัวเข้ารหัสเสียงพูด (Vocal encoder) หรือมีอีกชื่อหนึ่งว่า โวโคดเดอร์ (Vocoder) คือเทคนิคการสังเคราะห์ และวิเคราะห์เสียงมนุษย์โดยใช้แบบจำลองแหล่งกำเนิด-ตัวกรอง (Source-filter model) [16] ตัวกรองช่องเสียง (Vocal tract filter) ประมาณค่าได้จากการแจกแจงของพลังงานบนสเปกตรัม (Spectrum) ซึ่งคำนวณได้ด้วยแถวลำดับของตัวกรองแบนด์พาส (Bandpass filter) นอกจากนี้ยังมีตัวตรวจหาระดับเสียง (Pitch detector) เพื่อตรวจหาว่าเป็นเสียงโฆษะ (Voiced) หรืออโฆษะ (Unvoiced) และประมาณความถี่มูลฐานสำหรับเสียงโฆษะ




รูปที่ 4 แผนภาพแสดงแบบจำลองแหล่งกำเนิด-ตัวกรอง [17]

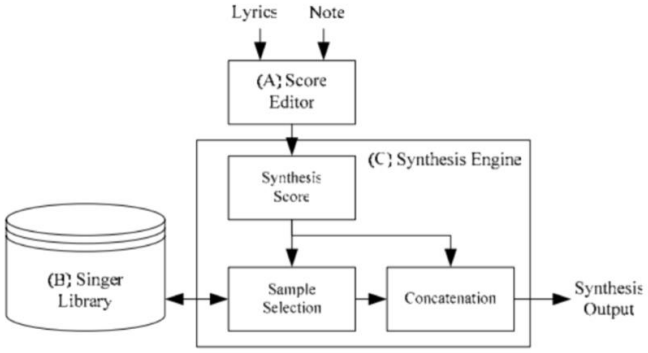
เนื่องจากความแปรผันบนพารามิเตอร์ของแบบจำลองเสียงน้อยกว่าความแปรผันของเสียงจริงมาก ส่งผลให้ความกว้างแถบความถี่ (Bandwidth) ที่ใช้ในการสื่อสัญญาณ (Transmission) ลดลง [18]

**2.4 การสังเคราะห์เสียงร้องเพลง (Singing Voice Synthesis)**

ในการสังเคราะห์เสียงร้องเพลงนั้นแบ่งวิธีการสังเคราะห์ได้หลัก ๆ เป็น 2 วิธี คือ

2.4.1 สังเคราะห์เสียงร้องเพลงด้วยการต่อกัน (Concatenative Synthesis)

แนวคิดของการสังเคราะห์เสียงร้องเพลงวิธีนี้คือ เมื่อได้รับข้อมูลนำเข้าที่ต้องการแล้ว ระบบจะเลือกหน่วยเสียงที่มีความสอดคล้องกัน และปรับแต่งเสียงโดยใช้เทคนิคด้านประมวลผลสัญญาณ หลังจากนั้นจึงนำเสียงที่ได้มาต่อกัน โวคาลอยด์เป็นระบบสังเคราะห์เสียงร้องเพลงระบบหนึ่งซึ่งใช้การสังเคราะห์เสียงร้องเพลงด้วยการต่อกัน เสียงที่ได้จากการสังเคราะห์ด้วยวิธีนี้จะมีปัญหาเกี่ยวกับความไม่ต่อเนื่องบริเวณรอยต่อของหน่วยเสียงที่เลือกมาต่อกัน และต้องใช้ฐานข้อมูลเสียงร้องจำนวนมากเพื่อให้เสียงตัวอย่างมีความครอบคลุม รายการอ้างอิง [3, 19] เป็นตัวอย่างของระบบที่ใช้วิธีนี้



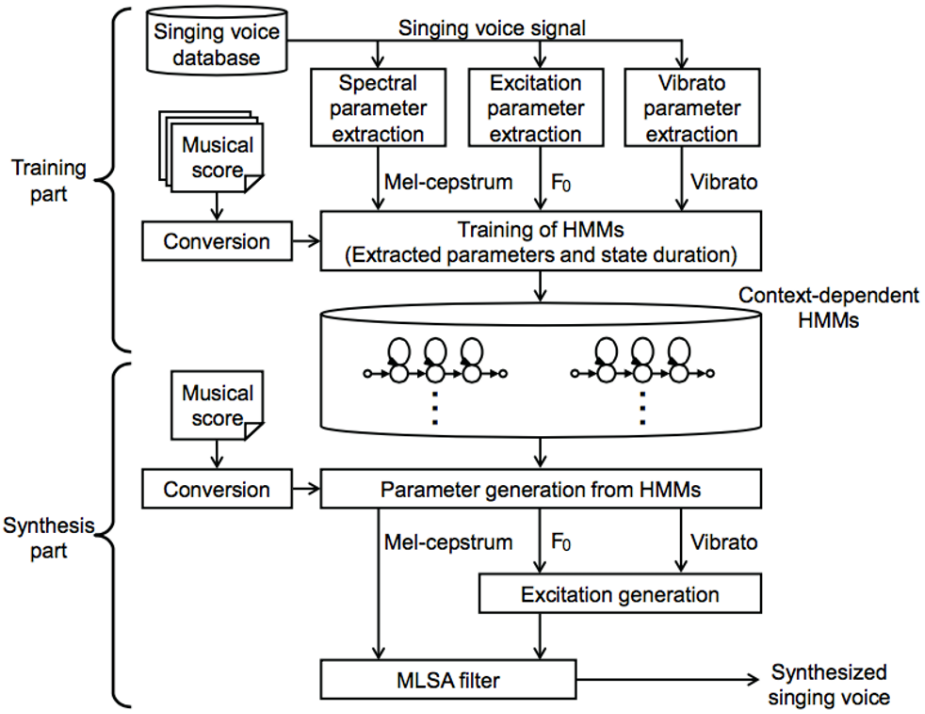

รูปที่ 5 แผนภาพระบบสังเคราะห์เสียงร้องของโวคาลอยด์ [3]

2.4.2 สังเคราะห์เสียงร้องด้วยพารามิเตอร์ (Parametric Synthesis)

แนวคิดของการสังเคราะห์เสียงวิธีนี้คือ ตัวอย่างเสียงจะไม่ถูกเก็บในรูปคลื่นสัญญาณ แต่จะเก็บเป็นพารามิเตอร์ซึ่งได้จากการสกัดคุณลักษณะสำคัญของหน่วยเสียง เมื่อระบบรับข้อมูลนำเข้าที่ต้องการแล้ว พารามิเตอร์จะถูกสร้างขึ้นจากข้อมูลนำเข้า และระบบจะสร้างเสียงขึ้นมาจากพารามิเตอร์นั้น ตัวอย่างหนึ่งของการสังเคราะห์เสียงร้องด้วยพารามิเตอร์คือ การสังเคราะห์เสียงร้องโดยใช้แบบจำลองกายภาพพัฒนา (Singing Synthesis with an Evolved Physical Model) [20] ในงานนี้แบบจำลองสองมิติถูกสร้างขึ้นมาเพื่อเลียนแบบช่องเสียงของมนุษย์โดยใช้เจเนติกอัลกอริทึม (Genetic Algorithm) เสียงที่ได้จากการทดลองนั้นเมื่อนำมาวิเคราะห์พบว่ามีความใกล้เคียงกับเสียงของมนุษย์ แต่แบบจำลองที่ได้จากการทดลองกับช่องเสียงของมนุษย์ยังไม่มีความใกล้เคียงกัน อีกตัวอย่างหนึ่งที่เป็นการสังเคราะห์เสียงร้องด้วยพารามิเตอร์คือ การสังเคราะห์เสียงร้องเพลงโดยใช้แบบจำลองฮิดเดนมาร์คอฟ [21]

**2.5 ระบบสังเคราะห์เสียงร้องเพลงโดยใช้แบบจำลองฮิดเดนมาร์คอฟ (HMM-based Singing Voice Synthesis)**

การสังเคราะห์เสียงร้องโดยใช้แบบจำลองฮิดเดนมาร์คอฟนั้นแบ่งออกเป็น 2 ส่วนคือ ส่วนฝึกฝน (Training Part) และส่วนสังเคราะห์ (Synthesis Part) ในส่วนฝึกฝนนั้นเสียงร้องในฐานข้อมูลจะถูกสกัดคุณลักษณะสำคัญออกมาแล้วนำไปฝึกฝนแบบจำลอง และในส่วนสังเคราะห์จะรับข้อมูลนำเข้ามาเพื่อสร้างพารามิเตอร์สำหรับเสียงผลลัพธ์ [22] และเสียงร้องจะถูกสังเคราะห์ขึ้นโดยตรงจากพารามิเตอร์ที่สร้างขึ้นมา ข้อดีของการสังเคราะห์เสียงด้วยพารามิเตอร์คือ เสียงร้องที่ได้มีความต่อเนื่องและไม่จำเป็นต้องใช้ฐานข้อมูลเสียงร้องขนาดใหญ่ [6] แต่ยังมีส่วนที่ต้องพัฒนาคือสัญญาณรบกวนในเสียงบางเสียง [7]

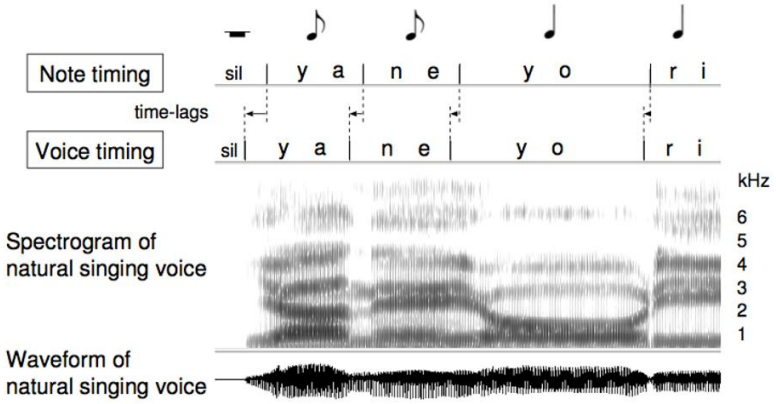



รูปที่ 6 แผนภาพระบบสังเคราะห์เสียงร้องโดยใช้แบบจำลองฮิดเดนมาร์คอฟ [4]

ต่อมาการสังเคราะห์เสียงร้องโดยใช้แบบจำลองฮิดเดนมาร์คอฟถูกพัฒนาเพิ่มเติมแบบจำลองไทม์แล็ก (Time-lag model) [21] ถูกเสนอขึ้นเพื่อจำลองความคลาดเคลื่อนของเวลาที่มนุษย์ใช้ในการเปล่งเสียงร้องเทียบกับเวลาตามโน้ตดนตรีจริง จากสรุปผลการทดลองพบว่าเมื่อเพิ่มการจำลองไทม์แล็กเข้าไปในแบบจำลองเสียง เสียงที่ได้จากระบบนี้มีความเป็นธรรมชาติมากขึ้นจากระบบธรรมดาที่ไม่มีแบบจำลองไทม์แล็ก

คุณภาพของเสียงร้องเพลงสังเคราะห์ที่ใช้ฐานข้อมูลนั้นขึ้นอยู่กับปริมาณและความครอบคลุมของชุดข้อมูลฝึกฝน โดยเฉพาะการเตรียมชุดข้อมูลสำหรับการสังเคราะห์เสียงร้องเพลงซึ่งมีความหลากหลายมากกว่าเสียงพูดทั้งในเรื่องของระดับเสียงและความสั้นยาว เพื่อให้เสียงที่ได้จากการ

สังเคราะห์มีความชัดเจน การฝึกฝนแบบปรับตัวระดับเสียงได้ (Pitch Adaptive Training) [23] จึงถูกเสนอขึ้นเพื่อปรับปรุงกระบวนการฝึกฝนแบบจำลองเสียง การฝึกฝนนี้ใช้แนวความคิดจากการฝึกฝนแบบจำลองเสียงพูดที่ใช้ชุดข้อมูลฝึกฝนของผู้พูดหลายคน เรียกว่า การฝึกฝนแบบปรับตัวผู้พูดได้ (Speaker Adaptive Training) [24]




รูปที่ 8 ตัวอย่างของไทม์แล็ก (Time-lag) [21]

การฝึกฝนแบบปรับตัวระดับเสียงได้นั้นสามารถมองในรูปของการทำข้อมูลระดับเสียงให้เป็นกลาง (Pitch Normalization) กล่าวคือ ในการร้องเพลงนั้นความถี่ของเสียงร้องจะแกว่งอยู่รอบค่าความถี่ของระดับเสียงทางดนตรีหนึ่ง ๆ ส่วนที่เราสนใจจึงเป็นแค่ส่วนต่างของความถี่ของเสียงร้องกับความถี่ของระดับเสียงทางดนตรีนั้น ๆ การสังเคราะห์เสียงร้องเพลงจึงสามารถสังเคราะห์ได้ทุกระดับเสียงแม้ไม่มีระดับเสียงนั้นในชุดข้อมูลฝึกฝน

นอกจากนี้ยังมีการจำลองสไตล์การร้องเพลง [17] การสร้างแบบจำลองเสียงที่ใช้สำหรับแปลงไปเนื้อเสียงอื่น [25] ซึ่งแสดงให้เห็นว่าการสังเคราะห์เสียงร้องเพลงด้วยพารามิเตอร์นั้นมีความยืดหยุ่นสูงในการปรับแต่งพารามิเตอร์เพื่อให้ได้เสียงแบบต่าง ๆ รายการอ้างอิง [4, 5, 26] เป็นตัวอย่างของระบบสังเคราะห์เสียงโดยใช้แบบจำลองฮิดเดนมาร์คอฟ

การนำระบบสังเคราะห์เสียงร้องเพลงของภาษาหนึ่งไปใช้กับอีกภาษาหนึ่งมักต้องมีการปรับแต่งเพิ่มเติม เพื่อให้เข้ากับลักษณะเฉพาะของภาษานั้น ๆ เช่น การออกแบบปัจจัยบริบทของภาษาญี่ปุ่นให้รองรับบริบทเฉพาะของภาษาอังกฤษ [4] การจัดสรรพยางค์และการทำสำเนาพยางค์เมื่อจำนวนโน้ตดนตรีไม่เท่ากับจำนวนพยางค์เสียงร้อง [6]

2.5.1 การออกแบบปัจจัยบริบท (Contextual Factors Design)

ปัจจัยบริบทที่ใช้ในงานสังเคราะห์เสียงร้องเพลงนั้นต่างจากงานสังเคราะห์เสียงพูดเนื่องจากมีปัจจัยที่เกี่ยวของเพิ่มขึ้นจากข้อมูลทางด้านดนตรี เช่น ระยะเวลา ระดับเสียง ตำแหน่งปัจจุบันของตัวโน้ต เป็นต้น

| | |
|---|---|
| **Initial/final level** | Triphone (Current initial/final with preceding and preceding initial/final) |
| **Note level (or Syllable level)** | Absolute and relative pitch of current note |
| | Pitch difference with preceding and succeeding note |
| | Length of preceding, current and succeeding note (in 32th note, millisecond) |
| | Position of current note in musical bar and phrase (in 32th note, millisecond) |
| | *Melisma or not, absolute pitch of the last note in melisma, difference between the pitch of the first note and the last note in melisma* |
| **Phrase** | Number of notes in current phrase |
| | Length of current phrase (in 32th note, millisecond) |
| **Song** | Key, time signature and tempo |
| | Number of notes in current song |
| | Length of current song (in 32th note, millisecond) |

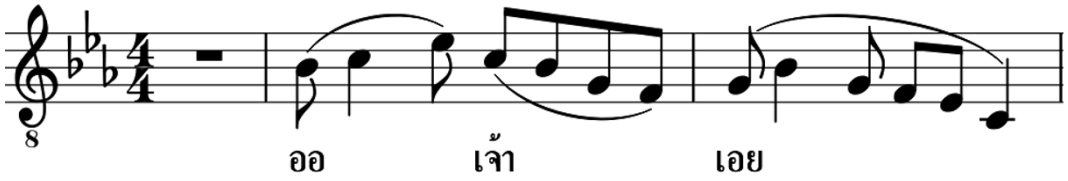
ตารางที่ 2 ตัวอย่างปัจจัยบริบทของระบบภาษาจีน [5]

| | |
|---|---|
| **Phoneme** | Quinphone (Phoneme within the context of two immediately preceding and succeeding phonemes) |
| **Syllable (Mora)** | Number of phonemes in {previous, current, next} syllable |
| | Position of {previous, current, next} syllable in note |
| | Language dependent context in {previous, current, next} syllable (English: with of without {accent, stress}, Japanese: undefined) |
| **Note** | Musical {tone, key, beat, tempo and length} of {previous, current, next} note |
| | Position of current note in {measure, phrase} |
| | With of without a slur between current and {previous, next} note |
| | Dynamics to witch current note belongs |
| | Difference in pitch between current note and {previous, next} note |
| | Distance between current note and {next, previous} {accent, staccato} |
| | Position of current note in current {crescendo, decrescendo} |
| **Phrase** | Number of {syllables, notes} in {previous, current, next} phrase |
| **Song** | Number of {syllables, notes} / Number of measures |
| | Number of phrases |

ตารางที่ 3 ตัวอย่างปัจจัยบริบทของระบบภาษาญี่ปุ่นและรองรับภาษาอังกฤษ [4]

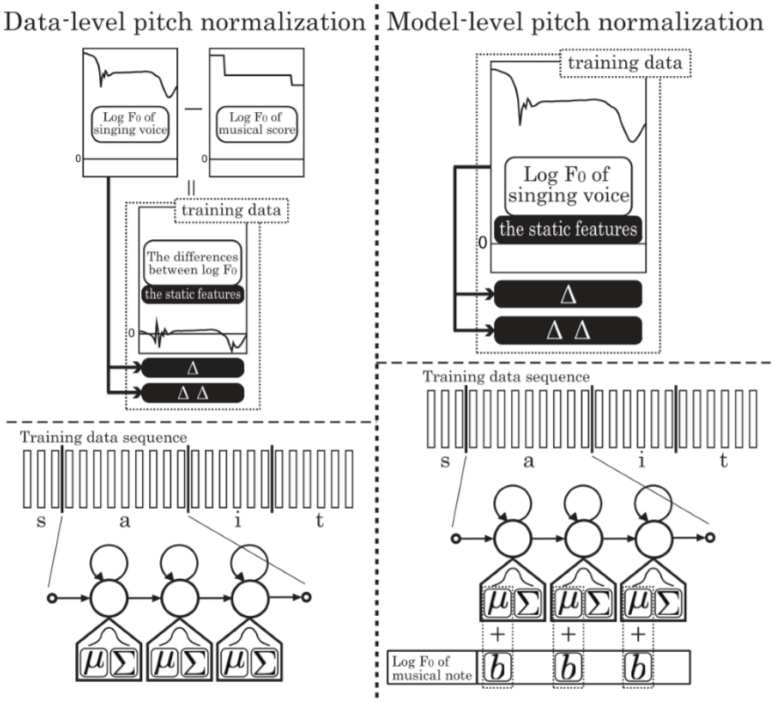
2.5.2 เมลิสมากับการออกแบบระบบสังเคราะห์เสียงร้องเพลง (Melisma and singing voice synthesis system design)

ในการร้องเพลงนั้นหลายครั้งมักจะพบโน้ตดนตรีที่มีการไล่หลายระดับเสียงในหนึ่งพยางค์ เหตุการณ์นี้ เรียกว่า เมลิสมา (Melisma) จากตัวอย่างงานการออกแบบปัจจัยบริบท พบว่าปัจจัยที่เกี่ยวข้องกับระดับเสียงของโน้ตดนตรีนั้นอยู่ในระดับที่สูงกว่ารูปเขียน เมื่อพิจารณาถึงสถานการณ์ที่มีเมลิสมาซึ่งการเปลี่ยนแปลงของระดับเสียงอยู่ในระดับต่ำกว่าพยางค์ หากไม่มีปัจจัยบริบทที่เกี่ยวข้องกับเมลิสมาหรือการจัดการกับสถานการณ์ที่พบเมลิสมา ข้อมูลเหล่านี้จะสูญหายไปในการฝึกฝนแบบจำลองเสียงและการสังเคราะห์เสียง ซึ่งส่งผลกระทบต่อคุณภาพของแบบจำลองเสียงที่ได้ โดยเฉพาะในการสังเคราะห์เค้ารูปความถี่มูลฐานในกรณีที่พบเมลิสมา




รูปที่ 9 ตัวอย่างเมลิสมา

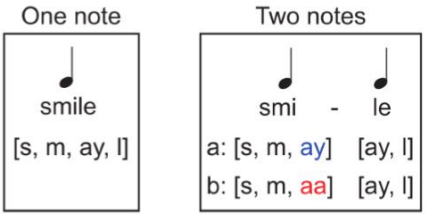
เนื่องจากในงานวิจัยที่เสนอการฝึกฝนแบบปรับตัวระดับเสียงได้นั้นไม่ได้พิจารณาในสถานการณ์ที่พบเมลิสมา หนึ่งในวิธีแก้ไขที่เสนอขึ้นมาคือ การจำลองความถี่มูลฐานโดยทำระดับเสียงให้เป็นกลางในระดับข้อมูล (Data-level pitch normalization) พร้อมกับการเพิ่มปัจจัยบริบทที่เกี่ยวข้องกับเมลิสมา [5]




รูปที่ 10 ความแตกต่างของการทำระดับเสียงให้เป็นกลางในระดับข้อมูลและระดับแบบจำลอง [23]

เหตุที่เรียกว่าการทำระดับเสียงให้เป็นกลางในระดับข้อมูล เนื่องจากความแตกต่างของความถี่มูลฐานที่ได้จากเสียงมนุษย์ และโน้ตเพลงนั้นได้เตรียมไว้ก่อนจะนำไปฝึกฝนแบบจำลองเสียงที่เกี่ยวข้อง แตกต่างกับการทำระดับเสียงให้เป็นกลางในระดับแบบจำลอง [23] ซึ่งจำลองความแตกต่างของระดับเสียงโดยไม่ต้องเตรียมคำนวณความแตกต่างไว้ล่วงหน้า นอกจากนั้นการจำลองความแตกต่างที่ระดับแบบจำลองทำให้ไม่ต้องยึดการวางแนวในการฝึกฝนแบบจำลอง ส่งผลให้การประมาณค่าพารามิเตอร์เพื่อแบบจำลองดีขึ้น

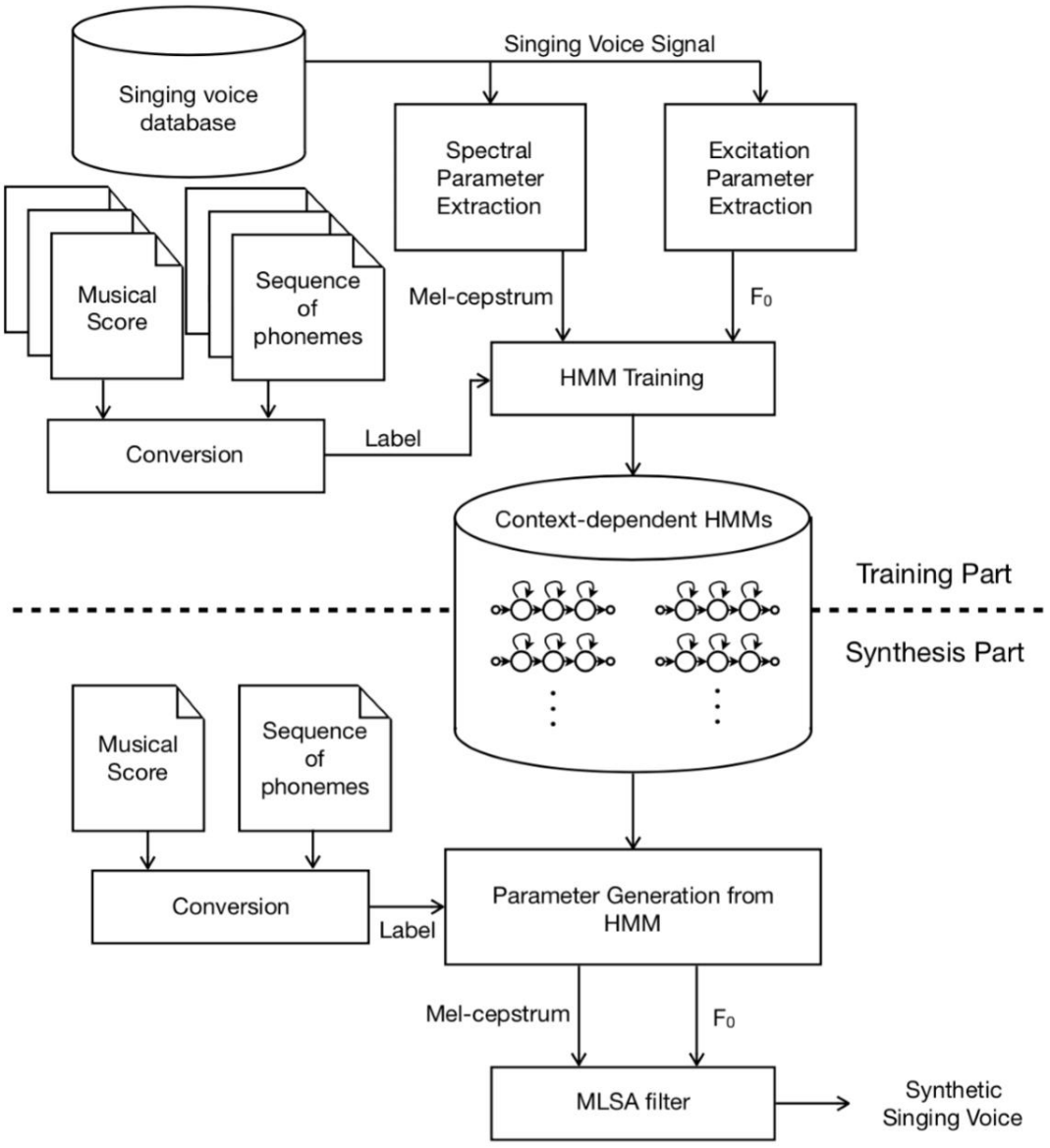
| Original | ey | ay | ow | aw | oy |
|---|---|---|---|---|---|
| Duplicated | eh, ey | aa, ay | ao, ow | aa, aw | ao, oy |




รูปที่ 11 ตัวอย่างการทำสำเนาพยางค์ภาษาอังกฤษในระบบสังเคราะห์เสียงภาษาญี่ปุ่น [4]

# บทที่ 3
# วิธีการดำเนินการวิจัย

## 3.1 ภาพรวมของระบบ (System Overview)




รูปที่ 12 แผนผังแสดงภาพรวมของระบบสังเคราะห์เสียงร้องเพลงโดยใช้แบบจำลองฮิดเดนมาร์คอฟ

3.1.1 ส่วนฝึกฝน (Training part)

ขั้นตอนแรกของส่วนฝึกฝนคือการเตรียมข้อมูลสัญญาณเสียงร้องเพลงในฐานข้อมูลที่เตรียมไว้ นำเอาไปสกัดคุณลักษณะสำคัญทางสเปคตรัมและทางการกระตุ้น พารามิเตอร์ที่ใช้คือ Mel-cepstal [27] และ log F0 ตาม ลำดับ ในขณะเดียวกันไฟล์ป้ายที่ใช้ในการฝึกฝนแบบจำลองเสียงจะสร้างขึ้น

จากโน้ตเพลงและรูปเขียนของเนื้อเพลง เมื่อเตรียมไฟล์ป้ายและสกัดคุณลักษณะสำคัญเรียบร้อยแล้ว จึงนำไปใช้ในการฝึกฝนแบบจำลองฮิดเดนมาร์คอฟที่ขึ้นอยู่กับบริบท

3.1.2 ส่วนสังเคราะห์ (Synthesis part)

ข้อมูลนำเข้าที่จำเป็นสำหรับส่วนสังเคราะห์คือโน้ตเพลงและรูปเขียนของเนื้อเพลง เมื่อรับโน้ตเพลงและรูปเขียนของเนื้อเพลงเข้ามา ทั้งสองไฟล์จะถูกแปลงเป็นไฟล์ป้ายเพื่อเป็นข้อมูลนำเข้าสำหรับการก่อกำเนิดพารามิเตอร์จากแบบจำลองเสียง นำพารามิเตอร์ที่ได้ไปผ่านตัวกรองเอ็มแอลเอสเอ (MLSA filter) สุดท้ายจะได้เสียงร้องเพลงสังเคราะห์

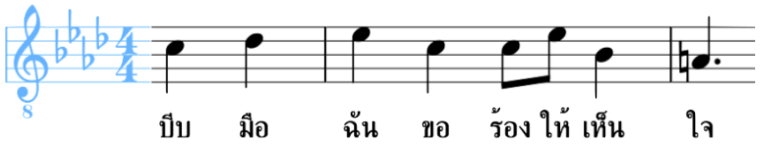
**3.2 การออกแบบปัจจัยบริบท (Contextual factors design)**

ภาษาไทยเป็นภาษาที่มีเสียงวรรณยุกต์เหมือนภาษาจีน ดังนั้นผู้วิจัยจึงเลือกปัจจัยบริบทของระบบภาษาจีนเป็นปัจจัยบริบทต้นแบบ ซึ่งปัจจัยบริบทจะมี 4 ระดับ คือ รูปเขียน (Phoneme) โน้ตดนตรี (Musical Note) ประโยค เพลง (Phrase) และเพลง (Song) ดังแสดงในตารางที่ 4

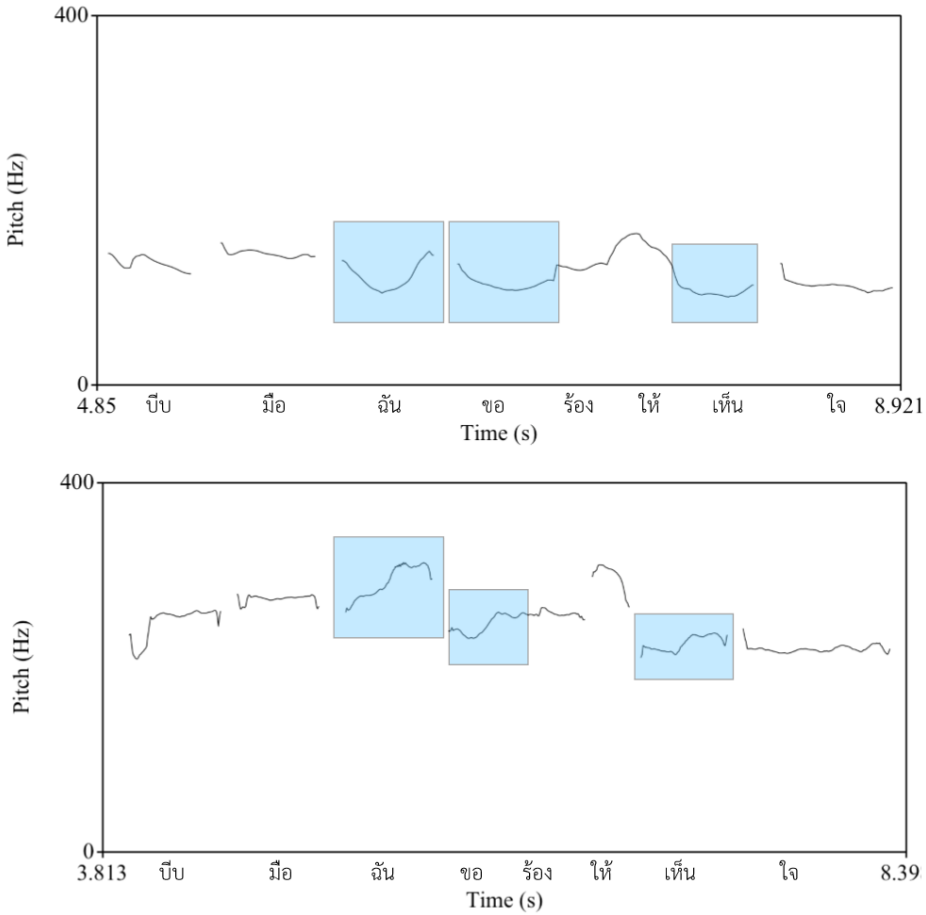
| | |
|---|---|
| **Phoneme level** | Triphone (Current phoneme with preceding and preceding phoneme) |
| | Tone of preceding, current and succeeding phoneme |
| **Musical Note level** | Absolute and relative pitch of current note |
| | Pitch difference with preceding and succeeding note |
| | Length of preceding, current and succeeding note (in 96th note) |
| | Position of current note in musical bar and phrase (in 96th note) |
| | Melisma or not |
| **Phrase level** | Number of notes in current phrase |
| | Length of current phrase (in 96th note) |
| **Song level** | Key, time signature and tempo |
| | Number of notes in current song |
| | Length of current song (in 96th note) |

ตารางที่ 4 รายละเอียดปัจจัยบริบทที่ออกแบบสำหรับระบบสังเคราะห์เสียงร้องเพลงภาษาไทย

รูปเขียนของคำอ่านภาษาไทยนั้นใช้เหมือนกับในระบบสังเคราะห์เสียงพูด [9] สำหรับรูปเขียนแทนส่วนที่เงียบ คือ sil ใช้แทนส่วนเงียบในตอนต้นและจบของเพลง และ pau ใช้แทนส่วนเงียบในตอนกลางของเพลง ในระบบสังเคราะห์เสียงร้องเพลงภาษาจีนซึ่งเป็นภาษาที่มีเสียงวรรณยุกต์นั้นไม่ได้ใช้วรรณยุกต์เป็นหนึ่งในปัจจัยบริบท เนื่องจากข้อมูลเบื้องต้นในการร้องเพลงภาษาจีนสื่อให้เห็นว่า วรรณยุกต์ไม่ส่งผลต่อเค้ารูปความถี่มูลฐานที่สังเคราะห์ได้ [7] แต่จากการสำรวจเบื้องต้นในการร้องเพลงป๊อปไทยพบว่า วรรณยุกต์มีผลต่อเค้ารูปความถี่มูลฐานของเสียงร้องเพลง




รูปที่ 13 ตัวอย่างโน้ตเพลงอ้างอิงสำหรับรูปที่ 14

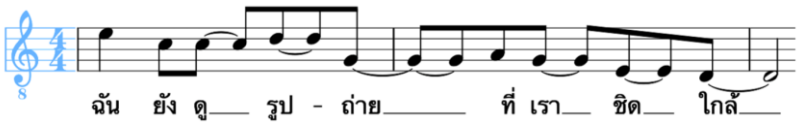




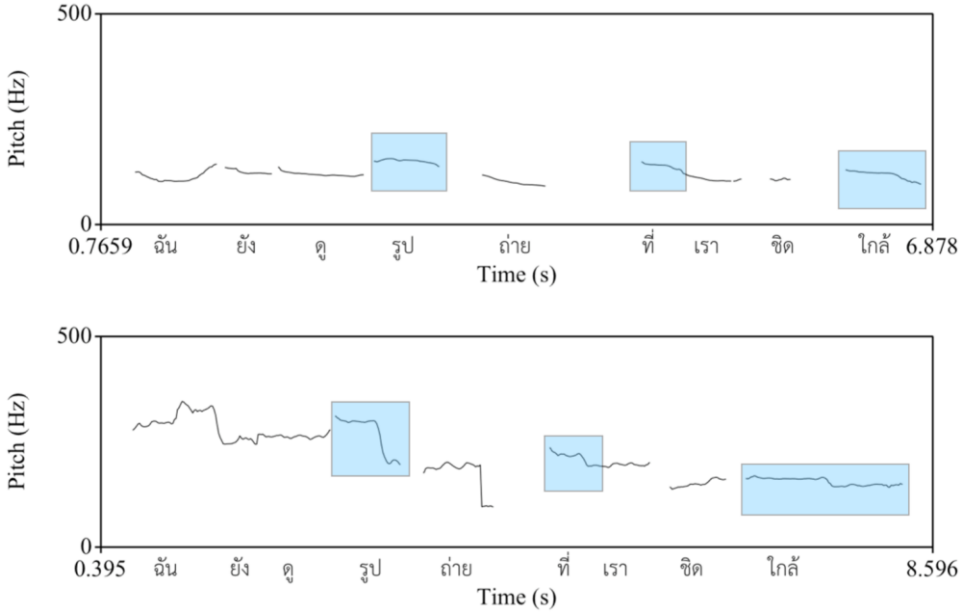

รูปที่ 14 กราฟแสดงเค้ารูปความถี่มูลฐานของเสียงพูดเนื้อเพลง (บน) และเสียงร้องเพลง (ล่าง)

จากรูปที่ 14 พิจารณาที่พยางค์คำร้องที่มีเสียงวรรณยุกต์จัตวา ได้แก่คำว่า ฉัน ขอ และเห็น ทั้งสามคำนี้มีเค้ารูปไปในทางเดียวกัน อีกทั้งในรูปที่ 16 และ 18 เมื่อพิจารณาพยางค์คำร้องที่มีเสียงวรรณยุกต์เดียวกันก็เป็นไปในทำนองเดียวกัน จากข้อสังเกตนี้ผู้วิจัยจึงเสนอปัจจัยบริบทวรรณยุกต์เพิ่มเติมในปัจจัยบริบทระดับรูปเขียน

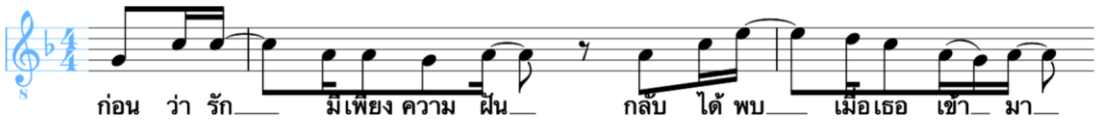
สำหรับในกรณีเมลิสมาซึ่งเป็นลำดับของตัวโน้ตหนึ่งตัวขึ้นไปในพยางค์เดียวนั้น ที่ตัวโน้ตแรกสุดของเมลิสมาจะมีเค้ารูปความถี่มูลฐานคล้ายกับการร้องพยางค์นั้นที่ไม่ใช่เมลิสมา ปัจจัยบริบทที่เกี่ยวข้องกับเรื่องเมลิสมาจึงยังคงไว้ เพื่อแยกความแตกต่างระหว่างการลากเสียงตัวโน้ตเดียวกับลากเสียงเมลิสมา ดังแสดงในรูปที่ 20



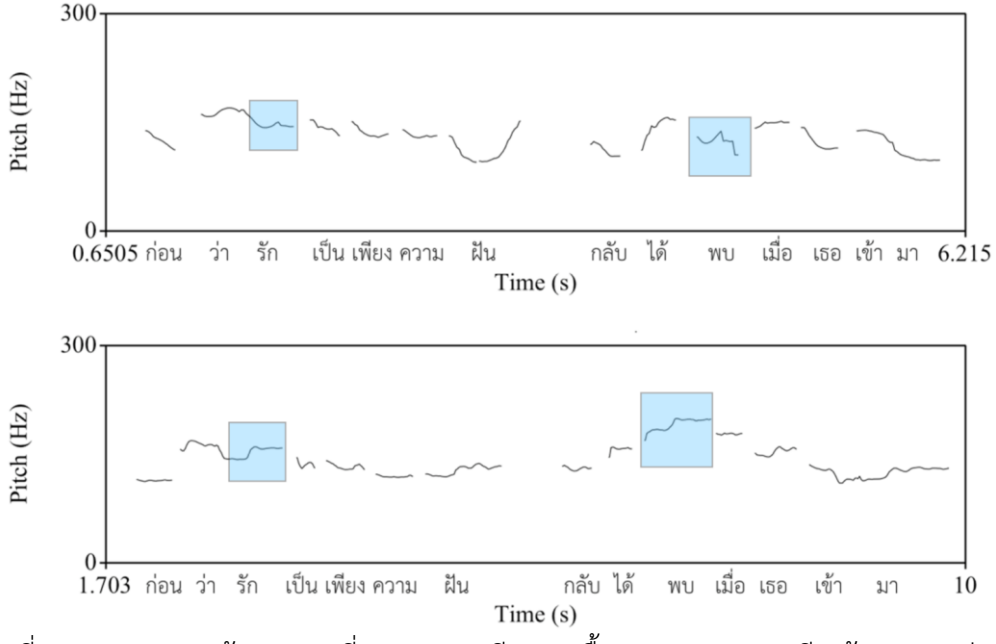

รูปที่ 15 ตัวอย่างโน้ตเพลงอ้างอิงสำหรับรูปที่ 16



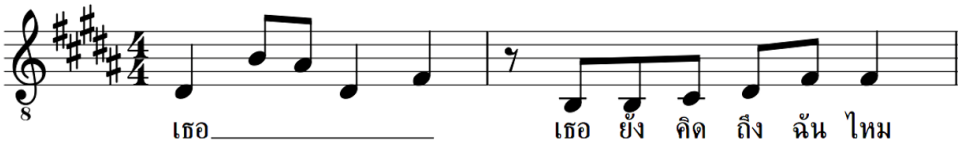

รูปที่ 16 กราฟแสดงเค้ารูปความถี่มูลฐานของเสียงพูดเนื้อเพลง (บน) และเสียงร้องเพลง (ล่าง)



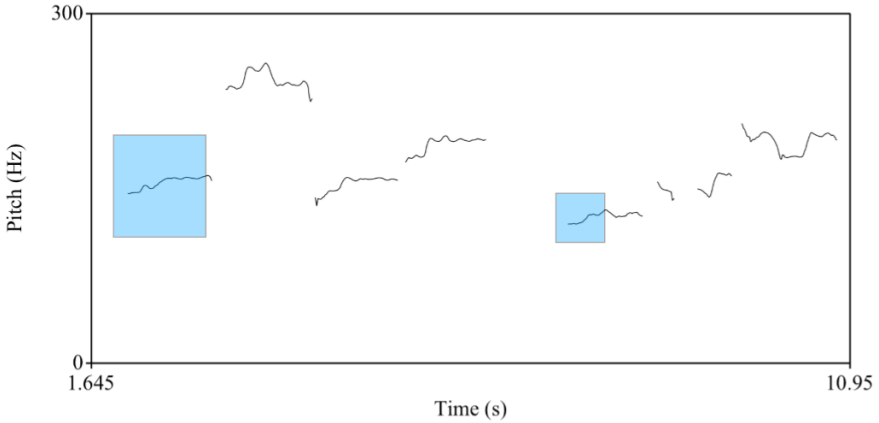

รูปที่ 17 ตัวอย่างโน้ตเพลงอ้างอิงสำหรับรูปที่ 18




รูปที่ 18 กราฟแสดงเค้ารูปความถี่มูลฐานของเสียงพูดเนื้อเพลง (บน) และเสียงร้องเพลง (ล่าง)




รูปที่ 19 ตัวอย่างโน้ตเพลงอ้างอิงสำหรับรูปที่ 20




รูปที่ 20 กราฟแสดงเค้ารูปความถี่มูลฐานเปรียบเทียบเสียงร้องเพลงธรรมดาและเมลิสมา

### 3.3 การแปลงข้อมูลนำเข้า (Input conversion)

ข้อมูลนำเข้าของระบบมี 2 อย่าง คือ โน้ตดนตรี ซึ่งเป็นข้อมูลในรูปแบบ MusicXML และลำดับของรูปเขียนของเนื้อเพลงนั้น ๆ ซึ่งเป็นข้อมูลในรูปแบบข้อความ อีกทั้งในโน้ตดนตรีนั้นจะมีสัญลักษณ์กำกับเพื่อใช้ในการกำหนดรูปเขียนจากลำดับของรูปเขียนที่นำเข้ามา

```
sil
kh-qq-j^-0
m-ii-0
khr-a-j^-0
s-a-k^-1
kh-o-n^-0
d-aa-j^-2
b-@@-k^-1
ch-a-n^-4
m-aa-0
pau
w-aa-2
w-ee-0
l-aa-0
khr-a-j^-0
th-a-m^-0
k-a-p^-1
r-a-w^-0
h-a-j^-2
c-e-p^-1
ch-a-m^-3
c-a-j^-0
…
…
…
```

```
…
…
…
<measure number="2" width="333.07">
      <note default-x="19.85" default-y="-40.00">
        <pitch>
          <step>E</step>
          <octave>3</octave>
          </pitch>
        <duration>2</duration>
        <voice>1</voice>
        <type>eighth</type>
        <stem>up</stem>
        <beam number="1">begin</beam>
        <lyric number="1" default-x="6.58" default-y="-80.00">
          <syllabic>single</syllabic>
          <text>เคย</text>
          </lyric>
        </note>
      <note default-x="54.45" default-y="-40.00">
        <pitch>
          <step>E</step>
          <octave>3</octave>
          </pitch>
        <duration>2</duration>
        <voice>1</voice>
        <type>eighth</type>
        <stem>up</stem>
        <beam number="1">end</beam>
        <lyric number="1" default-x="5.58" default-y="-80.00">
          <syllabic>single</syllabic>
          <text>มี</text>
          </lyric>
        </note>
…
…
…
```

รูปที่ 21 ตัวอย่างลำดับของรูปเขียนและโน้ตดนตรีในรูปแบบ MusicXML

เมื่อรับไฟล์นำเข้าทั้งสองแล้ว โน้ตดนตรีจะถูกดึงข้อมูลทางดนตรีออกมาพร้อมกับกำหนดรูปเขียนของพยางค์คำร้องให้โน้ตดนตรีที่สัมพันธ์กัน สุดท้ายจะเขียนข้อมูลที่ประมวลผลเสร็จแล้วออกมาเป็นไฟล์ป้ายดังแสดงในรูปที่ 22 ในแต่ละบรรทัดจะประกอบไปด้วย 3 สดมภ์ คือ เวลาเริ่มต้น เวลาจบ และสัญลักษณ์แทนแบบจำลองเสียง โดยที่หากเวลาเป็น -1 จะไม่นำมาพิจารณา

```
0 10000000 st-sil+kh/A:xx+xx@xx/B:xx-xx@xx~xx/C:xx+xx-xx@0!0/D:0@xx#xx/E:0^0/F:xx!xx@xx$0%0
10000000 -1 sil-kh+qq/A:xx+0@0/B:40-7@72~32/C:xx+12-12@0!0/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 -1 kh-qq+j^/A:xx+0@0/B:40-7@72~32/C:xx+12-12@0!0/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 14650000 qq-j^+m/A:xx+0@0/B:40-7@72~32/C:xx+12-12@0!0/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
14650000 -1 j^-m+ii/A:0+0@0/B:40-7@32~32/C:12+12-12@12!12/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 19350000 m-ii+khr/A:0+0@0/B:40-7@32~32/C:12+12-12@12!12/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
19350000 -1 ii-khr+a/A:0+0@1/B:40-7@32~35/C:12+12-6@24!24/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 -1 khr-a+j^/A:0+0@1/B:40-7@32~35/C:12+12-6@24!24/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 24050000 a-j^+s/A:0+0@1/B:40-7@32~35/C:12+12-6@24!24/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
24050000 -1 j^-s+a/A:0+1@0/B:37-4@29~29/C:12+6-24@36!36/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 -1 s-a+k^/A:0+1@0/B:37-4@29~29/C:12+6-24@36!36/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 26400000 a-k^+kh/A:0+1@0/B:37-4@29~29/C:12+6-24@36!36/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
26400000 -1 k^-kh+o/A:1+0@2/B:40-7@35~32/C:6+24-6@42!42/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 -1 kh-o+n^/A:1+0@2/B:40-7@35~32/C:6+24-6@42!42/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 35750000 o-n^+d/A:1+0@2/B:40-7@35~32/C:6+24-6@42!42/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
35750000 -1 n^-d+aa/A:0+2@1/B:40-7@32~28/C:24+6-12@66!66/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 -1 d-aa+j^/A:0+2@1/B:40-7@32~28/C:24+6-12@66!66/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 38100000 aa-j^+b/A:0+2@1/B:40-7@32~28/C:24+6-12@66!66/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
38100000 -1 j^-b+@@/A:2+1@4/B:44-11@36~31/C:6+12-6@72!72/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 -1 b-@@+k^/A:2+1@4/B:44-11@36~31/C:6+12-6@72!72/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
-1 42800000 @@-k^+ch/A:2+1@4/B:44-11@36~31/C:6+12-6@72!72/D:0@xx#xx/E:9^144/F:Amajor!4/4@64$134%1992
…
…
…
```

รูปที่ 22 ตัวอย่างไฟล์ป้ายที่อิงปัจจัยบริบท

### 3.4 การทำสำเนารูปเขียน (Phoneme duplication)

ในการสร้างไฟล์ป้ายนั้นคำร้องทุกคำจะมีตัวโน้ตกำกับ โดยคำร้องหนึ่งพยางค์อาจสัมพันธ์กับตัวโน้ตหนึ่งตัวหรือมากกว่า หากคำร้องหนึ่งพยางค์สัมพันธ์กับตัวโน้ตหนึ่งตัว รูปเขียนของพยางค์คำร้องของตัวโน้ตนั้นสามารถกำหนดได้ทันที แต่กรณีที่คำร้องหนึ่งพยางค์สัมพันธ์กับตัวโน้ตมากกว่าหนึ่งตัว ต้องมีการพิจารณารูปเขียนของพยางค์คำร้องก่อนที่จะกำหนดรูปเขียนให้แต่ละตัวโน้ต

จากงานวิจัยที่เกี่ยวข้องพบว่าวิธีที่ใช้กำหนดรูปเขียนให้ตัวโน้ตเมื่อตัวโน้ตมีมากกว่าพยางค์คำร้องคือ การทำสำเนาพยางค์คำร้อง กล่าวคือ รูปเขียนของเสียงที่สอดคล้องกับตัวโน้ตในเมลิสมาจะกำหนดจากรูปเขียนของพยางค์คำร้อง จากการสำรวจข้อมูลเบื้องต้นเกี่ยวกับเสียงร้องเพลงป๊อปไทย พบว่า ประเภทของสระและตัวสะกดมีผลต่อวิธีการลากเสียงในการร้องเพลง เพื่อกำหนดกฎในการทำสำเนารูปเขียนสำหรับภาษาไทย เสียงพยางค์สระเดี่ยวในภาษาไทยจึงแบ่งพิจารณาแยกเป็นกลุ่ม ๆ โดยแจกแจงได้ตามตารางที่ 5

| | ไม่มีตัวสะกด | กก กบ กด | กน กง กม เกย เกอว |
|---|---|---|---|
| **สระเดี่ยวเสียงสั้น** | อะ | อัก อับ อัด | อัน อัง อำ ไอ เอา |
| **สระเดี่ยวเสียงยาว** | อา | อาก อาบ อาด | อาน อาง อาม เอย อาว |

ตารางที่ 5 ตัวอย่างเสียงพยางค์สระเดี่ยวในกลุ่มต่าง ๆ

โดยปกติแล้วในการร้องเพลงเมื่อพบเมลิสมานั้น เสียงสระจะถูกลากไว้ในขณะที่เปลี่ยนระดับเสียง แต่ด้วยภาษาไทยมีสระเสียงสั้นและยาว บางกรณีเสียงที่ถูกลากไว้ในขณะที่เปลี่ยนระดับเสียงเป็นเสียงตัวสะกด กรณีที่พบนี้เกิดขึ้นเมื่อสระเป็นสระเสียงสั้นและมีตัวสะกดในมาตราแม่กน กง กม เกย และเกอว ดังแสดงในตารางที่ 6 สำหรับกรณีสระเดี่ยวที่เหลือนั้นจะลากเสียงสระ

| **มาตราตัวสะกด** | กน | กง | กม | เกย | เกอว |
|---|---|---|---|---|---|
| **การลากเสียง** | **อัน + น** | **อัง + ง** | **อัม + ม** | **อัย + อี** | **เอา + อู** |
| **รูปเขียน** | a n^ + n^ | a ng^ + ng^ | a m^ + m^ | a j^ + ii | a w^ + uu |

ตารางที่ 6 การลากเสียงสระ อะ ในมาตราตัวสะกดแม่กน กง กม เกย และเกอว

ในกรณีของสระประสมก็เป็นเช่นเดียวกับสระเดี่ยว เมื่อสระประสมเสียงสั้นสะกดด้วยแม่กน กง กม เกย และ เกอว เสียงที่ถูกลากขณะเปลี่ยนระดับเสียงคือเสียงตัวสะกด แต่เมื่อต้องลากเสียงสระขณะที่เปลี่ยนระดับเสียง เสียงสระที่ลากไว้คือเสียงสระเสียงแรกของสระประสมนั้น ๆ

| **สระประสม** | **เอียะ** | **เอีย** | **เอือะ** | **เอือ** | **อัวะ** | **อัว** |
|---|---|---|---|---|---|---|
| **เสียงสระ** | **อี + อะ** | **อี + อา** | **อือ + อะ** | **อือ + อา** | **อู + อะ** | **อู + อา** |
| **รูปเขียน** | ia | iia | va | vva | ua | uua |
| **สระที่ลาก** | **อี** | | **อือ** | | **อู** | |
| **รูปเขียน** | ii | | vv | | uu | |

ตารางที่ 7 การลากเสียงสระของสระประสมต่าง ๆ

เนื่องจากรูปเขียนพยางค์คำร้องนั้นมีเสียงวรรณยุกต์กำกับทุกรูปเขียน แต่รูปเขียนที่ทำสำเนาขึ้นมาเมื่อพบเมลิสมานั้นจะไม่กำหนดเสียงวรรณยุกต์กำกับ เนื่องจากเมลิสมานั้นไม่พบในเสียงพูดปกติ

# บทที่ 4
# การวัดประเมินผล

## 4.1 ชุดข้อมูลเสียงร้องเพลง

ชุดข้อมูลฝึกฝนประกอบขึ้นจากเพลงป๊อปไทย 23 เพลง ความยาวประมาณ 30 นาที เพลงที่เลือกมาใช้ในชุดข้อมูลนั้นเลือกโดยดูจากความครอบคลุมของรูปเขียนของเนื้อร้อง ความหลากหลายของโน้ตดนตรี คีย์เพลง และ ความเร็วของเพลง เมื่อกำหนดโน้ตเพลงที่จะใช้ทั้งหมดแล้ว จึงอัดเสียงร้องเพลงแต่ละเพลงโดยมีโน้ตเพลงกำหนดตัวโน้ตที่ต้องร้องและความเร็ว เสียงร้องเพลงอัดโดยนักร้องสมัครเล่นซึ่งเป็นสมาชิกวงขับร้องประสานเสียง สภาพแวดล้อมที่ใช้ในการอัดเสียงไม่มีเสียงรบกวนอื่น ๆ นอกจากเสียงของผู้ร้อง ขณะที่อัดเสียงนั้นมีการกำหนดความเร็วของเพลงแต่ละเพลงโดยใช้เมโทรโนมให้จังหวะผู้ร้องผ่านหูฟัง เสียงร้องเพลงสุ่มตัวอย่างที่ความถี่ 48,000 เฮิรตซ์

## 4.2 รายละเอียดในการประเมินผล

สัญญาณเสียงที่สุ่มตัวอย่างที่ความถี่ 48,000 เฮิรตซ์ นำมาประมวลผลเพื่อสกัดคุณลักษณะที่ต้องการโดยใช้ขนาดวินโดว์ 25 มิลลิวินาที และวินโดว์เลื่อนครั้งละ 5 มิลลิวินาที เวกเตอร์พารามิเตอร์ทางสเปคตรัมประกอบด้วยค่าสัมประสิทธิ์ Mel-cepstral 35 อันดับ ค่าสัมประสิทธิ์ผลต่างและผลต่างของผลต่าง ส่วนเวกเตอร์พารามิเตอร์ทางการกระตุ้นประกอบด้วย log F0 ค่าสัมประสิทธิ์ผลต่างและผลต่างของผลต่าง

แบบจำลองเสียงฝึกฝนโดยใช้ MSD-HSMM [28, 29] แบบจำลองเสียงมี 5 สถานะไม่รวมสถานะเริ่มต้นและ สถานะจบ การวางแนวของรูปเขียนในไฟล์ป้ายที่ใช้สำหรับฝึกฝนได้จากขั้นตอนวิธี DAEM [30] แบบจำลองที่อิงปัจจัยบริบทจัดกลุ่มโดยใช้เทคนิคการจัดกลุ่มด้วยต้นไม้ตัดสินใจ อีกทั้งใช้เกณฑ์ MDL [31] ในการควบคุมขนาดของต้นไม้



4.2.1 การประเมินผลเรื่องวิธีการทำสำเนารูปเขียน

รูปแบบการทำสำเนาพยางค์ที่จะใช้ในการประเมินผลแบ่งออกเป็น 2 แบบ โดยที่การทำสำเนาพยางค์ทั้ง สองแบบเหมือนกันทั้งหมดยกเว้นในกรณีสระเสียงสั้นที่สะกดด้วยแม่กน กง กม เกย และเกอว ดังแสดงในตารางที่ 8

| ชุดการทดลอง | สระเสียงสั้นสะกดด้วยแม่กน กม กง เกย และเกอว |
|---|---|
| A | ทำสำเนาเสียงสระอย่างเดียว |
| B | เพิ่มกฎการทำสำเนาเสียงตัวสะกด |

ตารางที่ 8 รายละเอียดการทำสำเนาพยางค์ในแต่ละชุดการทดลอง

การประเมินผลในหัวข้อนี้มี 3 แบบ คือ การเปรียบเทียบรูปคลื่น (Waveform comparison) แบบทดสอบความชอบ (Preference test) และคะแนนความเห็น (Mean opinion scores)

4.2.2 การประเมินผลเรื่องปัจจัยบริบทวรรณยุกต์

ปัจจัยบริบทในการประเมินผลแบ่งออกเป็น 2 แบบ คือชุดการทดลองที่ไม่มีปัจจัยบริบทเกี่ยวกับวรรณยุกต์และชุดการทดลองที่มีปัจจัยบริบทเกี่ยวกับวรรณยุกต์ ดังแสดงในตารางที่ 9

| ชุดการทดลอง | วรรณยุกต์ |
|---|---|
| C | ไม่มี |
| D | มี |

ตารางที่ 9 รายละเอียดปัจจัยบริบทแต่ละชุดการทดลอง

การประเมินผลในหัวข้อนี้มี 3 แบบ คือ การเปรียบเทียบเค้ารูปความถี่มูลฐาน (F0 contour comparison) แบบทดสอบความชอบและคะแนนความเห็น

4.2.3 การประเมินผลเรื่องจำนวนสถานะของแบบจำลองเสียง

เสียงร้องเพลงนั้นมีพลวัตในเสียงมาก ดังนั้นเพื่อให้การจำลองเสียงร้องเพลงดีขึ้น จำนวนสถานะของแบบจำลองเสียงจึงเป็นตัวแปรหนึ่งที่น่าสนใจในการศึกษา ในการประเมินผลเรื่องนี้มี 3 ชุดการทดลอง ดังแสดงในตารางที่ 10 การประเมินผลในหัวข้อนี้ใช้แบบทดสอบความชอบและคะแนนความเห็น

| ชุดการทดลอง | จำนวนสถานะ (state number) |
|---|---|
| E | 5 |
| F | 7 |
| G | 9 |

ตารางที่ 10 จำนวนสถานะของแบบจำลองเสียงของแต่ละชุดการทดลอง



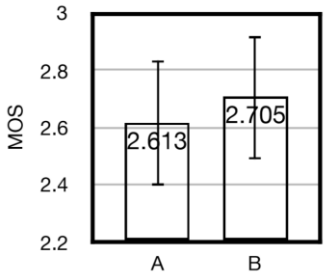
**4.3 ผลการทดลอง**

4.3.1 วิธีการทำสำเนารูปเขียน

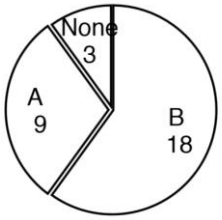
กลุ่มตัวอย่างสามสิบคนให้คะแนนความเห็นเกี่ยวกับความเป็นธรรมชาติของเสียงร้องเพลงสังเคราะห์ที่ได้จากการทำสำเนารูปเขียนแต่ละวิธี โดยที่แต่ละคนได้ฟัง 12 ท่อนเพลงที่มีเมลิสมา จาก 32 ท่อนเพลง ในแต่ละท่อนเพลงที่ฟังจะประกอบไปด้วย 2 ไฟล์เสียง คือ ไฟล์เสียงที่สังเคราะห์ด้วยการทำสำเนาเสียงสระอย่างเดียว และไฟล์เสียงที่สังเคราะห์ด้วยการเพิ่มกฎการทำสำเนาเสียงตัวสะกด

รูปที่ 23 แสดงคะแนนความเห็นเฉลี่ยที่ได้ โดยที่ A คือระบบที่ใช้การทำสำเนาเสียงสระอย่างเดียว และ B คือระบบที่เพิ่มกฎการทำสำเนาเสียงตัวสะกด

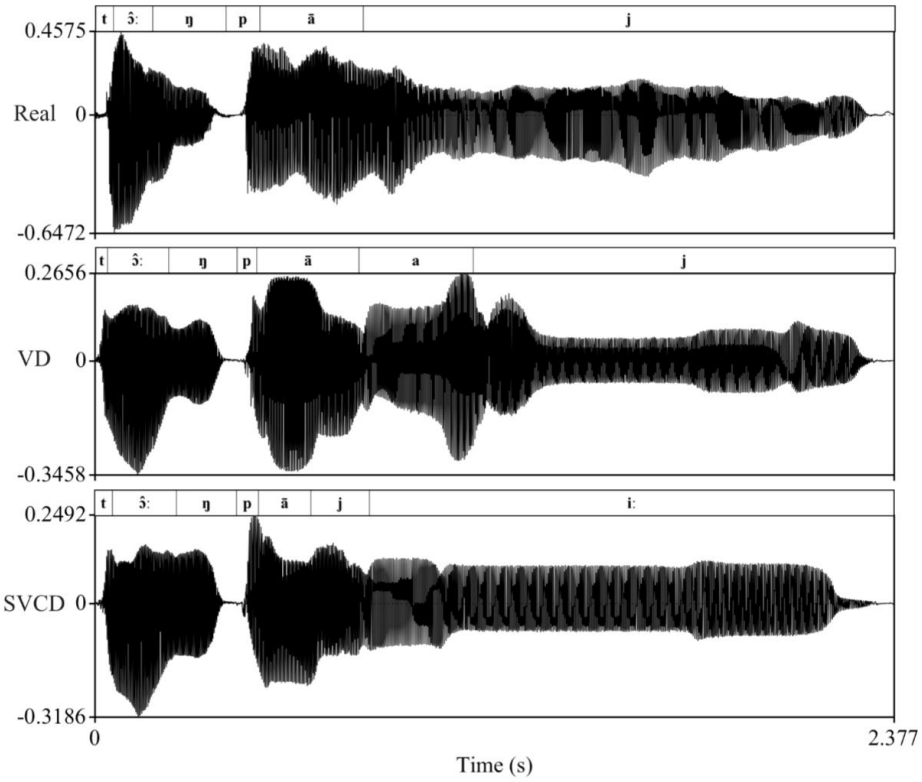
ผลของการทดสอบความชอบแสดงดังรูปที่ 24 โดยตัวเลขในแต่ละชิ้นส่วนของแผนภูมิวงกลมแสดงจำนวนผู้ทดสอบที่เลือกเสียงแต่ละระบบ และตัวอย่างความแตกต่างของรูปคลื่นเสียงร้องเพลงแสดงในรูปที่ 25




รูปที่ 23 กราฟแสดงคะแนนความเห็นวิธีการทำสำเนารูปเขียน



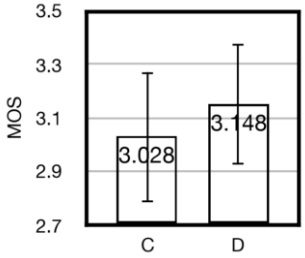

รูปที่ 24 แผนภูมิแสดงผลของการทดสอบความชอบวิธีการทำสำเนารูปเขียน



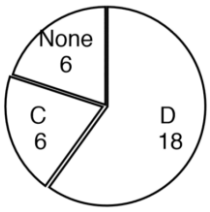

รูปที่ 25 ภาพแสดงตัวอย่างความแตกต่างของรูปคลื่นของเสียงร้องจริง (บน) เสียงร้องเพลงสังเคราะห์ที่ใช้การทำสำเนาเสียงสระอย่างเดียว (กลาง) และเสียงร้องเพลงสังเคราะห์ที่เพิ่มกฎการทำสำเนาเสียงตัวสะกด

4.3.2 ปัจจัยบริบทวรรณยุกต์

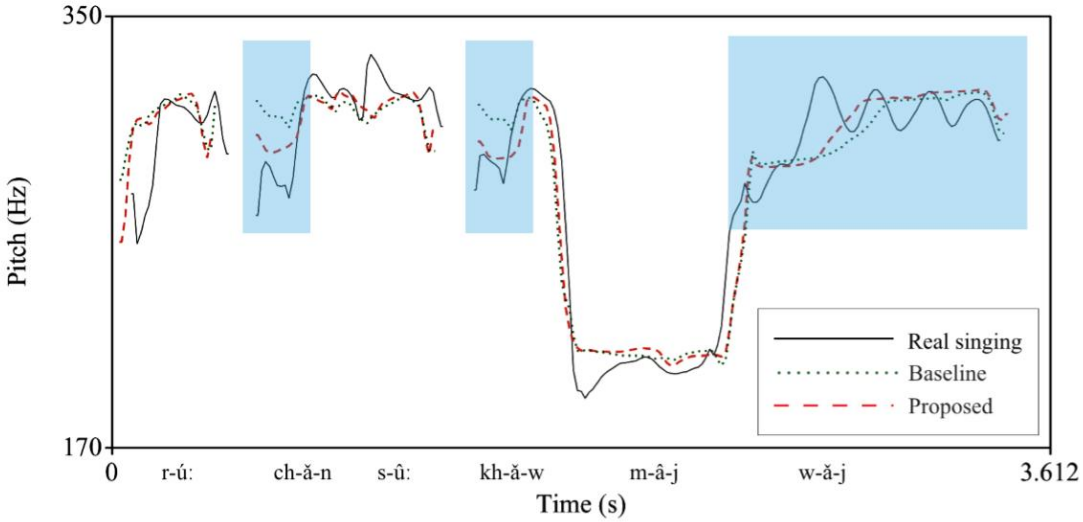
กลุ่มตัวอย่างเดิม 10 คน ให้คะแนนความเห็นเกี่ยวกับความเป็นธรรมชาติของเสียงร้องเพลงสังเคราะห์ที่ ได้จากระบบที่ไม่มีปัจจัยบริบทวรรณยุกต์และมีปัจจัยบริบทวรรณยุกต์ โดยที่แต่ละคนได้ฟัง 10 ท่อนเพลงอย่าง สุ่ม จาก 32 ท่อนเพลง ในแต่ละท่อนเพลงที่ฟังจะประกอบไปด้วย 2 ไฟล์เสียง คือ ไฟล์เสียงที่สังเคราะห์ด้วย ระบบที่ไม่มีปัจจัยบริบทวรรณยุกต์ และไฟล์เสียงที่สังเคราะห์ด้วยระบบที่มีปัจจัยวรรณยุกต์




รูปที่ 26 กราฟแสดงคะแนนความเห็นปัจจัยบริบทวรรณยุกต์




รูปที่ 27 แผนภูมิแสดงผลของการทดสอบความชอบปัจจัยบริบทวรรณยุกต์




รูปที่ 28 กราฟแสดงตัวอย่างความถี่มูลฐานระหว่างเสียงร้องเพลงจริง (เส้นทึบ) เสียงร้องเพลงสังเคราะห์ของระบบอ้างอิง (จุด) และเสียงร้องเพลงสังเคราะห์ของระบบที่มีบริบทของวรรณยุกต์ (ล่าง)

รูปที่ 26 แสดงคะแนนความเห็นเฉลี่ยที่ได้ โดยที่ C คือระบบที่ไม่มีปัจจัยบริบทวรรณยุกต์ และ D คือ ระบบที่มีปัจจัยบริบทวรรณยุกต์

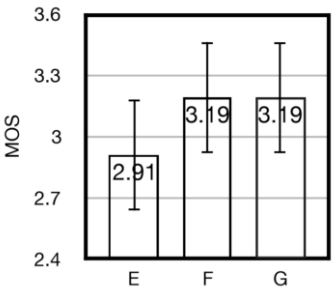
ผลของการทดสอบความชอบแสดงในรูปที่ 27 โดยตัวเลขในแต่ละชิ้นส่วนของแผนภูมิวงกลมแสดงจำนวนผู้ทดสอบที่เลือกเสียงแต่ละระบบ และตัวอย่างเค้ารูปความถี่มูลฐานระหว่างเสียงร้องเพลงจริงและเสียงร้องเพลงสังเคราะห์แสดงในรูปที่ 28

4.3.3 จำนวนสถานะของแบบจำลองเสียง

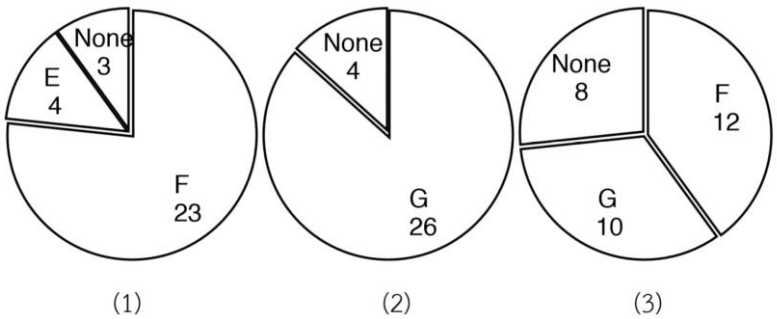
กลุ่มตัวอย่างเดิม 10 คน ให้คะแนนความเห็นเกี่ยวกับความเป็นธรรมชาติของเสียงร้องเพลงสังเคราะห์ที่ ได้จากระบบที่ใช้จำนวนสถานะของแบบจำลองเสียงแตกต่างกันสามแบบ โดยที่แต่ละคนได้ฟัง 10 ท่อนเพลงอย่างสุ่ม จาก 32 ท่อนเพลง ในแต่ละท่อนเพลงที่ฟังจะประกอบไปด้วย 3 ไฟล์เสียง คือ ไฟล์เสียงที่สังเคราะห์ ด้วยระบบที่มีจำนวนสถานะของแบบจำลองเสียง 5, 7 และ 9 สถานะ

รูปที่ 29 แสดงคะแนนความเห็นเฉลี่ยที่ได้ โดยที่ E, F และ G คือ ระบบที่มีจำนวนสถานะของแบบ จำลองเสียง 5, 7 และ 9 สถานะตามลำดับ

ผลของการทดสอบความชอบแสดงในรูปที่ 30 โดยตัวเลขในแต่ละชิ้นส่วนของแผนภูมิวงกลมแสดงจำนวนผู้ทดสอบที่เลือกเสียงแต่ละระบบ เนื่องจากการทดลองนี้มี 3 ชุดการทดลอง การแสดงผลแบบทดสอบความชอบจึงแบ่งเป็นคู่เปรียบเทียบ E-F, E-G และ G-F




รูปที่ 29 กราฟแสดงคะแนนความเห็นจำนวนสถานะของแบบจำลองเสียง




รูปที่ 30 แผนภูมิแสดงผลของการทดสอบความชอบจำนวนสถานะของแบบจำลองเสียงแต่ละคู่ E-F (1), E-G (2) และ G-F (3)

# บทที่ 5
# สรุปผลงานวิจัย

## 5.1 วิธีการทำสำเนารูปเขียน

ผู้วิจัยเสนอวิธีการทำสำเนารูปเขียนเพื่อให้ระบบสังเคราะห์เสียงที่ใช้การฝึกฝนแบบจำลองเสียงแบบปรับระดับเสียงได้รองรับกับโน้ตที่มีเมลิสมา โดยวิธีการที่นำเสนอมี 2 วิธี คือ แบบทำสำเนาเสียงสระอย่างเดียว และแบบเพิ่มกฎการทำสำเนาเสียงตัวสะกด จากการคำนวณคะแนนความเห็นเฉลี่ยที่ได้พบว่าพบว่าระบบที่มีการเพิ่มกฎการทำสำเนาเสียงตัวสะกดมีคะแนนสูงกว่าอย่างมีนัยสำคัญที่ระดับความมั่นใจ 95% และจากผลการทดสอบความชอบพบว่ามีผู้ทดสอบชอบเสียงจากระบบนี้ 18 คน มากกว่าระบบที่ทำสำเนาเสียงสระอย่างเดียวซึ่งมีเพียงผู้ทดสอบเลือก 9 คน

นอกจากนี้เมื่อสังเกตความแตกต่างของรูปคลื่นเสียงในรูปที่ 25 แล้วพบว่า ระบบที่เพิ่มกฎการทำสำเนาเสียงตัวสะกดมีรูปคลื่นเสียงใกล้เคียงกับเสียงร้องเพลงจริงมากกว่า จากการทดลองนี้สรุปได้ว่าการทำสำเนาตัวสะกดในระบบสังเคราะห์เสียงร้องเพลงภาษาไทยนั้นควรคำนึงถึงประเภทของสระและตัวสะกด

## 5.2 ปัจจัยบริบทวรรณยุกต์

ผู้วิจัยเสนอปัจจัยบริบทวรรณยุกต์เพื่อให้ระบบสังเคราะห์เสียงร้องเพลงภาษาไทยมีความเป็นธรรมชาติมากขึ้น จากการคำนวณคะแนนความเห็นเฉลี่ยที่ได้พบว่าระบบที่มีปัจจัยบริบทวรรณยุกต์มีคะแนนมากกว่าอีกระบบหนึ่งที่ไม่มีปัจจัยบริบทนี้อย่างมีนัยสำคัญที่ระดับความมั่นใจ 95% และจากผลการทดสอบความชอบพบว่ามีผู้ทดสอบชอบเสียงจากระบบนี้ 18 คน มากกว่าระบบที่ไม่มีปัจจัยบริบทวรรณยุกต์ซึ่งมีเพียงผู้ทดสอบเลือก 6 คน

จากรูปที่ 28 แสดงให้เห็นว่าเค้ารูปความถี่มูลฐานที่สังเคราะห์ได้จากระบบที่มีปัจจัยบริบทวรรณยุกต์นั้นมีความใกล้เคียงเสียงร้องจริงมากกว่าอีกระบบหนึ่งซึ่งไม่มีปัจจัยบริบทนี้ ส่วนที่ใกล้เคียงมากกว่านั้นตรงกับการเอื้อนของการร้องเพลงป๊อปไทยซึ่งมักขึ้นอยู่กับเสียงวรรณยุกต์ จากการทดลองนี้สรุปได้ว่าปัจจัยบริบทวรรณยุกต์มีความสำคัญต่อ การสังเคราะห์เค้ารูปความถี่มูลฐานของเสียงร้องเพลง



## 5.3 จำนวนสถานะของแบบจำลองเสียง

เนื่องจากเสียงร้องเพลงนั้นมีพลวัตในเสียงมากกว่าเสียงพูด เพื่อปรับปรุงให้เสียงร้องเพลงสังเคราะห์มีความเป็นธรรมชาติมากขึ้น ผู้วิจัยจึงปรับแต่งจำนวนสถานะของแบบจำลองเสียง จากคะแนนความเห็นเฉลี่ยที่ได้พบว่าการเพิ่มจำนวนสถานะของแบบจำลองเสียงจาก 5 สถานะเป็น 7 สถานะนั้นมีความแตกต่างอย่างชัดเจน ทั้งนี้เป็นเพราะจำนวนสถานะที่มากขึ้นทำให้เสียงที่สังเคราะห์ได้มีพลวัตมากขึ้น แต่เมื่อเพิ่มจำนวนสถานะจาก 7 สถานะเป็น 9 สถานะ คะแนนยังคงประมาณเท่า

เดิม จากการสังเกตพบว่าในระบบที่มีจำนวนสถานะ 9 สถานะนั้น หลายตัวอย่างเสียงมีความธรรมชาติค่อนข้างสูงเมื่ออิงจากคะแนนความเห็นเฉลี่ยในตัวอย่างเสียงนั้น ๆ แต่ก็มีหลายตัวอย่างเสียงมีความเป็นธรรมชาติต่ำมากเมื่ออิงจากคะแนนความเห็นเฉลี่ย ในตัวอย่างเสียงเหล่านั้นมีปัญหาเรื่องความต่อเนื่องของเสียงและชัดเจนมากเมื่อพบเมลิสมา ทั้งนี้อาจเกิดจากจำนวนสถานะของแบบจำลองเสียงที่มากขึ้นทำให้รอยต่อของเสียงที่เกิดจากการทำสำเนารูปเขียนมีความไม่ต่อเนื่องชัดเจนมากขึ้น

นอกจากนี้ผลการทดสอบความชอบนั้นก็เป็นไปในทางเดียวกัน คือ เสียงร้องเพลงสังเคราะห์จากระบบที่มีแบบจำลองเสียง 5 สถานะด้อยกว่าระบบที่มีจำนวนสถานะมากกว่าชัดเจน ขณะที่ในการเปรียบเทียบระบบที่มีแบบจำลองเสียง 7 และ 9 สถานะนั้นไม่มีความแตกต่างอย่างชัดเจน จากการทดลองนี้สรุปได้ว่าแบบจำลองเสียงที่มี 7 สถานะเหมาะสมที่สุดในการทำระบบสังเคราะห์เสียงร้องเพลงภาษาไทย

# บรรณานุกรม


1. Zen, H., A. Senior, and M. Schuster. *Statistical parametric speech synthesis using deep neural networks*. in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013.
2. Kenmochi, H. *Singing synthesis as a new musical instrument*. in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012.
3. Kenmochi, H. and H. Ohshita. *VOCALOID-commercial singing synthesizer based on sample concatenation*. in *8th Annual Conference of the International Speech Communication Association*. 2007.
4. Nakamura, K., et al. *HMM-Based singing voice synthesis and its application to Japanese and English*. in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference*. 2014.
5. Li, X. and Z. Wang, *A HMM-based Mandarin Chinese singing voice synthesis system.* IEEE/CAA Journal of Automatica Sinica, 2016. **3**(2): p. 192-202.
6. Wei, F.H., G.H. Min, and C. Hsuan Li, *A Research of Automatic Composition and Singing Voice Synthesis System for Taiwanese Popular Songs*, in *ICMC*. 2014. p. 1326-1331.
7. Cheng, J.Y., Y.C. Huang, and C.-h. Wu, *HMM-based Mandarin Singing Voice Synthesis Using Tailored Synthesis Units and Question Sets.* Computational Linguistics and Chinese Language Processing, 2013. **18**(4): p. 63-80.
8. Chinathimatmongkhon, N., A. Suchato, and P. Punyabukkana, *Implementing thai text-to-speech synthesis for hand-held devices.* 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2008, 2008. **1**: p. 545-548.
9. Chomphan, S. and T. Kobayashi. *Implementation and evaluation of an HMM-based Thai speech synthesis system*. in *8th Annual Conference of the International Speech Communication Association*. 2007.
10. Nectec, *VAJA - Thai text-to-speech software*. 1997.

11. Luksaneeyanawin, S. *Linguistics Research and Thai Speech Technology*. in *Proc. 5th International Conference on Thai Studies*. 1993. School of Oriental and African Studies, University of London.

12. Suchato, A., *Speech Production - Sounds in Languages*. 2014, Chulalongkorn University.

13. Hansakunbuntheung, C., V. Tesprasit, and V. Sornlertlamvanich, *Thai tagged speech corpus for speech synthesis.* The Oriental COCOSDA 2003, 2003: p. 97-104.

14. Kertkeidkachorn, N., et al. *CHULA TTS: A Modularized Text-To-Speech Framework*. in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. 2014.

15. Howard, D.M., H. Daffern, and J. Brereton, *Four-part choral synthesis system for investigating intonation in a cappella choral singing.* Logopedics, phoniatrics, vocology, 2013. **38**(3): p. 135-42.

16. Dudley, H., *Remaking speech.* The Journal of the Acoustical Society of America, 1939. **11**(2): p. 169-177.

17. Nose, T., et al., *HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling.* Computer Speech and Language, 2015. **34**(1): p. 308-322.

18. Sundberg, J., *The KTH Synthesis of Singing.* Advances in Cognitive Psychology, 2006. **2**: p. 2-3.

19. Lin, C.Y., T.Y. Lin, and J.S.R. Jang, *A corpus-based singing voice synthesis system for Mandarin Chinese*, in *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005. p. 359-362.

20. Cooper, C., et al., *Singing Synthesis With an Evolved Physical Model.* IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2006. **14**(4).

21. Saino, K., et al. *An HMM - based Singing Voice Synthesis System*. in *9th International Conference on Spoken Language Processing*. 2006.

22. Tokuda, K., et al. *Speech parameter generation algorithms for HMM-based speech synthesis*. in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2000.

23. Oura, K., et al. *Pitch adaptive training for HMM-based singing voice synthesis*. in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference*. 2012.

24. Yamagishi, J. and T. Kobayashi, *Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training.* IEICE Transactions on Information and Systems, 2007. **E90-D**(2): p. 533-543.

25. Shirota, K., et al., *Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis*, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference*. 2014. p. 2559-2563.

26. Oura, K., et al. *Recent development of the HMM-based singing voice synthesis system-Sinsy*. in *7th ISCA Workshop on Speech Synthesis*. 2012.

27. Tokuda, K., et al. *Mel-Generalized Cepstral Analysis - a Unified Approach To Speech Spectral Estimation*. in *Third International Conference on Spoken Language Processing.* 1994.

28. Tokuda, K., et al. *Hidden Markov models based on multi-space probability distribution for pitch pattern modeling*. in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference*. 1999.

29. Zen, H., et al., *A hidden semi-Markov model-based speech synthesis system.* IEICE transactions on information and systems, 2007. **90**(5): p. 825-834.

30. Ueda, N. and R. Nakano, *Deterministic annealing EM algorithm.* Neural Networks, 1998. **11**(2): p. 271-282.

31. Shinoda, K. and T. Watanabe, *MDL-based context-dependent subword modeling for speech recognition.* Acoustical Science and Technology, 2000. **21**(2): p. 79-86.

# ภาคผนวก

ข้อมูลคะแนนความเห็นเฉลี่ย

| ผู้ทดสอบ | คะแนนความเห็นเฉลี่ย | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G |
| 1 | 3.231 | 3.462 | 3.400 | 3.400 | 2.800 | 2.600 | 3.500 |
| 2 | 2.154 | 2.462 | 3.000 | 3.500 | 3.000 | 3.500 | 3.300 |
| 3 | 2.154 | 2.385 | 3.800 | 3.800 | 3.900 | 3.800 | 4.100 |
| 4 | 3.077 | 2.923 | 3.600 | 3.600 | 3.600 | 3.500 | 3.700 |
| 5 | 3.231 | 3.308 | 3.800 | 3.900 | 2.700 | 3.700 | 3.800 |
| 6 | 3.231 | 3.385 | 3.200 | 3.300 | 2.900 | 3.600 | 3.400 |
| 7 | 2.077 | 2.308 | 2.200 | 2.200 | 2.300 | 2.400 | 2.700 |
| 8 | 3.385 | 3.308 | 3.500 | 3.800 | 4.100 | 4.200 | 4.200 |
| 9 | 2.077 | 2.385 | 2.300 | 2.800 | 2.300 | 2.100 | 2.300 |
| 10 | 1.385 | 1.462 | 1.700 | 1.800 | 1.500 | 1.500 | 1.500 |
| 11 | 3.308 | 3.692 | 3.700 | 3.900 | 3.900 | 4.100 | 3.900 |
| 12 | 2.000 | 2.154 | 2.900 | 3.200 | 2.400 | 2.600 | 2.800 |
| 13 | 2.308 | 2.538 | 3.300 | 3.500 | 2.200 | 2.500 | 2.500 |
| 14 | 2.538 | 2.538 | 3.300 | 3.300 | 4.000 | 4.100 | 4.300 |
| 15 | 3.308 | 3.385 | 3.500 | 3.700 | 2.700 | 3.000 | 3.000 |
| 16 | 2.077 | 2.308 | 1.800 | 2.200 | 2.300 | 2.600 | 2.600 |
| 17 | 2.692 | 3.308 | 3.500 | 3.200 | 4.000 | 4.200 | 4.100 |
| 18 | 2.615 | 2.462 | 2.900 | 3.300 | 3.600 | 4.100 | 4.000 |
| 19 | 2.923 | 2.769 | 3.300 | 3.700 | 3.600 | 4.000 | 3.700 |
| 20 | 2.308 | 2.615 | 2.300 | 2.000 | 2.000 | 2.400 | 2.100 |
| 21 | 2.462 | 2.846 | 3.600 | 3.500 | 2.900 | 3.700 | 3.300 |
| 22 | 3.462 | 3.385 | 3.500 | 3.500 | 3.700 | 3.700 | 3.700 |
| 23 | 2.846 | 2.769 | 3.100 | 2.700 | 2.400 | 2.600 | 2.600 |
| 24 | 2.308 | 2.538 | 3.300 | 3.000 | 2.400 | 3.300 | 3.000 |
| 25 | 1.846 | 1.615 | 1.600 | 1.800 | 2.000 | 2.300 | 2.100 |
| 26 | 3.923 | 3.769 | 4.100 | 3.800 | 4.100 | 4.200 | 4.200 |
| 27 | 2.846 | 2.385 | 2.500 | 2.800 | 2.400 | 2.700 | 3.000 |
| 28 | 1.692 | 1.692 | 2.700 | 3.200 | 2.300 | 3.100 | 2.600 |
| 29 | 2.308 | 2.308 | 2.400 | 2.900 | 2.500 | 2.500 | 2.600 |
| 30 | 1.846 | 2.692 | 2.300 | 2.900 | 2.000 | 2.800 | 2.400 |
| ค่าเฉลี่ย | 2.587 | 2.705 | 3.003 | 3.140 | 2.883 | 3.180 | 3.167 |
| ค่าเบี่ยงเบนมาตรฐาน | 0.620 | 0.596 | 0.674 | 0.621 | 0.765 | 0.756 | 0.752 |

ข้อมูลคะแนนความชอบ

| ผู้ทดสอบ | คู่เปรียบเทียบแบบทดสอบความชอบ | | | | |
|---|---|---|---|---|---|
| | A-B | C-D | E-F | E-G | F-G |
| 1 | B | None | E | G | G |
| 2 | B | D | F | F | G |
| 3 | B | None | E | G | G |
| 4 | A | None | E | G | G |
| 5 | B | D | F | G | G |
| 6 | B | D | F | F | G |
| 7 | B | None | F | G | G |
| 8 | A | D | F | None | G |
| 9 | B | D | E | G | None |
| 10 | B | D | None | None | None |
| 11 | B | D | F | F | None |
| 12 | B | D | F | G | G |
| 13 | B | D | F | None | G |
| 14 | None | None | F | G | G |
| 15 | B | D | F | None | G |
| 16 | B | D | F | None | G |
| 17 | B | C | F | F | G |
| 18 | A | D | F | F | G |
| 19 | A | D | F | F | G |
| 20 | B | C | F | F | G |
| 21 | B | C | F | F | G |
| 22 | A | None | None | None | None |
| 23 | A | C | F | None | G |
| 24 | B | C | F | F | G |
| 25 | A | D | F | F | G |
| 26 | A | C | F | None | G |
| 27 | A | D | F | G | G |
| 28 | None | D | F | F | G |
| 29 | None | D | None | G | G |
| 30 | B | D | F | F | G |

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อ-สกุล** | ลัทธพล จีระประดิษฐ |
| **วัน เดือน ปี เกิด** | 25 กันยายน 2536 |
| **สถานที่เกิด** | สุราษฎร์ธานี |
| **วุฒิการศึกษา** | จุฬาลงกรณ์มหาวิทยาลัย |
| **ที่อยู่ปัจจุบัน** | 99/32 ลุมพินีเพลสพระรามเก้า-รัชดา ถนนพระรามเก้า ห้วยขวาง กรุงเทพฯ 10310 |