พจนานุกรมคำอ่านภาษาไทย

Pronunciation Dictionary for Thai words

ภัชริกา คชสำโรง

ABSTRACT - This paper reports the design and development of a machine-readable pronunciation dictionary for Thai language. A **pronunciation dictionary** is a list of words along with their associated pronunciations. This format is particularly useful for speech recognition and synthesis, as it has mappings from words to their pronunciations in the given phoneme set. The current phoneme set contains 21 phonemes of single initial consonants, 17 phonemes of cluster consonants, 24 vowels and 5 tones.

KEY WORDS - pronunciation dictionary, speech recognition, speech synthesis

บทคัดย่อ – พจนานุกรมคำอ่านสำหรับภาษาไทย ได้ออกแบบและรวบรวมคำในภาษาไทยประเภทต่างๆ ตัวอย่างเช่น ชื่อ นามสกุล สถานที่ คำภาษาไทยที่ใช้โดยทั่วไป เพื่อจัดคำอ่านของคำนั้นๆ โดยใช้สัญลักษณ์แทนเสียงที่สามารถนำไปใช้ในการประมวลผลทางคอมพิวเตอร์ได้ ซึ่งสัญลักษณ์ที่ใช้สำหรับงานนี้ได้ออกแบบและใช้ในหน่วยปฏิบัติการวิทยการมนุษยภาษามาเป็นเวลาระยะหนึ่ง โดยปรับปรุงมาจากสัญลักษณ์สากลของ IPA ระบบเสียงภาษาไทยประกอบด้วย หน่วยเสียงพยัญชนะต้นเสียงเดี่ยว 21 หน่วยเสียง เสียงพยัญชนะต้นควบกล้ำ 17 เสียง เสียงสระ 24 เสียง และเสียงวรรณยุกต์ 5 ระดับ

คำสำคัญ – พจนานุกรมคำอ่าน, ระบบรู้จำเสียงพูด, ระบบสังเคราะห์เสียงพูด, ภาษาไทย

---

1. บทนำ

ในการวิจัยและพัฒนาทางเทคโนโลยีเสียงพูดภาษาไทย การแปลงจากรูปเขียนเป็นเสียงอ่าน (Grapheme to Phoneme - G2P) นั้นมีความจำเป็นอย่างยิ่ง และเนื่องจากภาษาไทยเป็นภาษาที่มีมีข้อยกเว้นเรื่องการออกเสียงสำหรับคำบางประเภท หรือบางคำเป็นจำนวนมาก ดังนั้นการทำ G2P แบบอัตโนมัติ ก็ยังมีข้อจำกัดอยู่มาก ดังนั้นการรวบรวมคำที่มีการถอดเสียงคำอ่าน และมีการตรวจสอบอย่างถูกต้องเพื่อสร้างเป็นพจนานุกรมคำอ่าน จึงมีความจำเป็นอย่างยิ่ง เพื่อช่วยเป็นเครื่องมือสำหรับการประมวลผลทางด้านเสียงคำอ่านภาษาไทย โดยสามารถนำพจนานุกรมไปปรับใช้กับงานประมวลประเภทต่างๆ ได้ตามที่ต้องการ โดยสามารถเพิ่มคำในประเภทต่างๆ ได้เรื่อยๆ ตามความต้องการ

2. วัตถุประสงค์

เพื่อรวบรวมเสียงคำอ่านสำหรับภาษาไทยในประเภทต่างๆ ได้แก่ คำทั่วไป ชื่อคน สถานที่ ชื่อองค์กร หรือคำทับศัพท์ ที่เกิดขึ้นในภาษาไทย โดยคำจะถูกคัดเลือกจากข้อมูลที่มีการใช้จริงๆ ในภาษาไทย โดยการเพิ่มคำศัพท์ในรายการคำจะมีการตรวจสอบคำซ้ำ กับชุดข้อมูลที่ได้ทำไปแล้ว เพื่อลดโอกาสการทำงานซ้ำซ้อน และเพื่อให้ได้ข้อมูลทีมีจำนวนมากที่สุด ซึ่งข้อมูลเสียงอ่านนี้จะเป็นพจนานุกรมรายการคำ พร้อมเสียงอ่านที่มีประโยชน์สำหรับ ระบบสังเคราะห์เสียงพูดภาษาไทย และระบบรู้จำเสียงพูดภาษาไทย เป็นอย่างมากโดยจะช่วยแก้ปัญหา เรื่องการวิเคราะห์การอ่านออกเสียงที่ไม่ถูกต้องได้จากการเปรียบเทียบเสียงอ่าน จากพจนานุกรมเสียงอ่านที่เตรียมไว้แล้ว

3. รายละเอียดคลังข้อมูล

1. **คำสามัญ**
   ในรายการคำ จะประกอบได้ด้วยคำภาษาไทยสามัญทั้ง ระดับคำโดด และคำประสม โดยคำสามัญเหล่านี้ สามารถนำมาสร้างเป็นคำใหม่ได้ไม่จำกัดจำนวน ดังนั้นจึงเพิ่มคำศัพท์พื้นฐานที่เกิดขึ้นจริง มุ่งหวังที่จะสร้างเป็นชุดคำใหม่ได้ไม่จำกัด

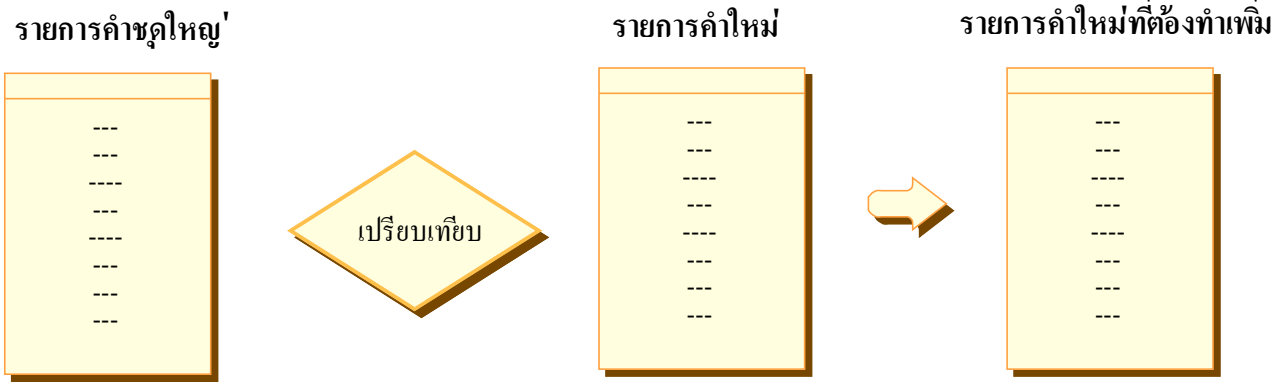
2. **ชื่อ – นามสกุล บุคคล**
   สำหรับชื่อ และ นามสกุล จะถูกแยกเป็นรายการคำที่ต่างกัน ถือเป็นรายการคำที่ไม่ใช่หน่วยเดียวกัน ชื่อและนามสกุลที่มีความกำกวม จะถอดความตามความนิยมในการออกเสียง หรืออาจถอดเสียงเป็นหลายแบบเท่าที่สามารถออกเสียงได้ เนื่องจาก เป็นการเลือกมาจากข้อความข่าว สมุดโทรศัพท์ มิได้สอบถามจากเจ้าของชื่อ ดังนั้น อาจมีความผิดพลาด สำหรับชื่อที่มีความซับซ้อนในการออกเสียง เช่น รรรรร อาจถอดความ ได้เป็น รอน-รัน (r-@@-n^-0 r-a-n^-0) หรือ รัน-รอน (r-a-n^-0 r-@@-n^-0) ได้ทั้ง 2 แบบ

3. **ชื่อสถานที่**
   ได้แก่ ชื่อถนน ชื่อสถานที่ต่างๆ ตำบล อำเภอ วัด ตรอก ซอย จะรวบรวมไว้เท่าที่สามารถค้าหามาได้ โดยเป็นคำที่เกิดขึ้นจริงในฐานข้อมูลข้อความ คลังข่าว และการรวบรวม โดยอาจมีข้อจำกัด ที่จำถอดเฉพาะชื่อของสถานที่ แต่ไม่ถอดลำดับ เช่น ตัวเลขลำดับถนน หรือซอยจะถูกตัดออก เพื่อลดความซ้ำซ้อนในการจัดทำข้อมูล

4. **ชื่อหน่วยงาน**
   ชื่อหน่วยงานจะถูกรวบรวมไว้แบบเต็ม แต่เมื่อมีคำใหม่เข้ามา จะสามารถสร้างชื่อหน่วยงานได้จาก คำสามัญหรือคำย่อย หากชื่อหน่วยงานไม่มีคำทับศัพท์ หรือเป็นชื่อหน่วยงานที่ประกอบขึ้นจากคำที่มีอยู่แล้วในรายการคำ

4. ขั้นตอนการทำงาน

เตรียมข้อมูล คัดเลือกชุดคำ - คำที่จะนำมาถอดเสียงจะถูกคัดเลือกประเภทมาเป็นชุดข้อมูล และอยู่ในรูปแบบรายการคำ เมื่อได้รายการคำแล้วก็จะนำไปเปรียบเทียบกับข้อมูลทั้งหมด ที่ได้มีการทำไว้แล้วทุกครั้ง เพื่อลดโอกาสการทำงานซ้ำซ้อน อันเนื่องมาจากรายการคำที่คัดเลือกมานั้นซ้ำกัน



ขั้นตอนการเปรียบเทียบรายการคำ เพื่อคัดเลือกคำที่จะนำไปถอดความ

ถอดความ - เมื่อได้รายการคำมาแล้ว ก็จะนำไปผ่านโปรแกรม Grapheme-to-Phoneme conversion และนำผลลัพธ์ที่ได้จากโปรแกรมมาทำการตรวจสอบต่อไป สำหรับคำที่ไม่สามารถถอดได้จากโปรแกรมก็จะถูกถอดความโดยนักภาษาศาสตร์ที่ทำการตรวจสอบต่อไป

ตรวจสอบ – เมื่อได้รายการคำ พร้อมทั้งรูปเสียงอ่านที่ได้จากโปรแกรม Grapheme-to-Phoneme conversion แล้ว ก็จะทำการตรวจสอบความถูกต้องของรูปเสียงอ่าน การแบ่งขอบเขตของพยางค์ ต้องตรงกันระหว่างรูปเสียงอ่าน กับรูปเขียน ดังนั้นจึงต้องมีการตรวจเช็คการตัดแบ่งพยางค์ที่ถูกต้องด้วย

5. สัญลักษณ์และข้อกำหนดต่างๆ

1. ใส่ **z^** แทนเสียงท้ายที่ไม่มีตัวสะกดเสมอ
2. คำที่ตัดได้สองรูปใส่ * และถอดเสียงทั้งแบบ เพื่อให้รู้ว่าสามารถอ่านได้ทั้งสองแบบ โดยถูกต้องตามหลักภาษาไทย และมีการใช้จริงในภาษาไทยปัจจุบัน เช่น ***รา|เมศร|, *รา|เม|ศร|**
3. รูปเดียวอ่านได้สองแบบใส่เครื่องหมาย **?** เช่น **?เบญจ|**
4. คำที่อ่านเสียงเชื่อมพยางค์ เช่น วิท**|**ยา**|** ให้ถอดความเสียงเชื่อมติดกับพยางค์แรก
5. พยางค์ที่ใช้พยัญชนะต้นร่วมกับ รร เช่น วรร**|**ณา**|** ถ้าสามามารถตัดแบ่งพยางค์ได้โดยที่เสียงอ่านไม่เปลี่ยน ให้ตัดแบ่งพยางค์ด้วย
6. คำที่ใช้พยัญชนะต้นและและตัวสะกดร่วมกัน ไม่ตัดพยางค์แยกจากัน เช่น อัฐิ**|** จักรา**|**
7. คำที่อ่านเสียงเชื่อมเรียงพยางค์ให้ตัดพยางค์แบบอ่านเรียงกันจนครบ เช่น วัฒน**|**ธรรม**|**
8. ตัวย่อ ใส่ **#** แล้วถอดเสียงตัวย่อเป็น **#**กทม**.|** **k@@0th@@0m@@0|** ถือเป็น **1** พยางค์
9. เสียงทับศัพท์ใช้หน่วยเสียงตามหน่วยเสียงภาษาต่างประเทศตามที่กำหนดไว้ในตารางเทียบสัญลักษณ์
10. อักษรนำ ให้ตัดตามพยางค์ ที่อ่านได้แล้วใส่เครื่องหมาย **$** หน้าคำ **$**รัศ**|**มี
11. เสียงสั้น เสียงยาว ในการถอดเสียงคำอ่านให้ถอดตามการออกเสียงจริง ไม่ขึ้นกับรูปเขียน

6. สรุป

การสร้างพจนานุกรมคำอ่านนั้นมีความจำเป็นต่องานทางด้านระบบรู้จำเสียงพูดภาษาไทย และงานสังเคราะห์เสียงพูดภาษาไทย เนื่องจากในการทำ Grapheme-to-Phoneme conversion ซึ่งเป็นส่วนงานหลักของทั้งสองระบบ ยังไม่ให้ผลความถูกต้องที่สมบูรณ์ หากต้องการความถูกต้องสมบูรณ์ของคำอ่านจำเป็นต้องมีการตรวจแก้ไขโดยนักภาษาศาสตร์ ดังนั้นการรวบรวมคำพร้อมคำอ่านที่ถูกต้องเก็บสะสมไว้จำนวนมากจึงถือเป็นคลังข้อมูลที่มีคุณค่า เพื่อใช้เป็นข้อมูลพื้นฐานสำหรับวัดความถูกต้องของระบบ Grapheme-to-Phoneme conversion ได้ต่อไป

ตารางเทียบสัญลักษณ์หน่วยเสียงภาษาไทย กับรูปสะกด

พยัญชนะต้น (C_i)			
เดี่ยว	ตัวอย่าง	ผสม	ตัวอย่าง
p	ปาก	pr	ประสาน
t	เต้น, กฎิ	phr	พราน
c	จะ	tr	เตรียม
k	ก่อน	kr	กราบ
z	อาน	khr	คร่า
ph	พบ, ภัย, ผ่าน	pl	ปลา
th	ทิ้ง, ธง, เฒ่า, ฐาน, มณโฑ	phl	พลาด
ch	ชอบ, เฌอ	thr	จันทรา
kh	คน, เขิน, ฆ่า	kl	เกลอ
b	บอก	khl	เคลื่อน
d	ด้าน, ชฎา	kw	กวาง
m	ไม่	khw	ขวา
n	นาน, เณร	เสียงทับศัพท์	
ng	เงิน	br	เบรน
l	เล่น, กีฬา	bl	บลู
r	รอ, ฤทัย	fr	ฟราย
f	ฝน, ฟัน	fl	เฟลม
s	สาย, ศิลา, รักษา, ซ่อน	dr	ดราก้อน
h	โหน, เฮฮา	17 หน่วย	
w	ว่า		
j	ย อน, หญิง		
21 หน่วย			

สระ (V)			
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 (C_f)	
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	