Pitch Target Representation of Thai Tones

Santitham Prom-on
Yi Xu

University College London, United Kingdom
King Mongkut’s University of Technology Thonburi, Thailand

Research Question / Objective

• Results of previous work were derived from the interpretation of perceptual experiments
  – Inconclusive in term of production
• There was a lack of tools to learn tone representation directly from data
• This study aim to identify pitch targets of Thai tones in continuous speech using PENTATrainer 2, a prosody modeling tool based on communicative function and target approximation

Thai Tone × Vowel Corpus

• Factors
  – 5 Tones: M (0), L (1), F (2), H (3), R (4)
  – 2 Vowel Lengths: Long, Short
  – 2 Locations: 2nd, 3rd
  – Total number of combination: 5×5×2×2 = 100
  – 5 repetitions (500 utterances per speaker)
• 5 native Thai speakers (3 males, 2 females)
  – Age: 20-25
  – Grew up and live in Greater Bangkok region

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
</tr>
</thead>
<tbody>
<tr>
<td>ka:0 “to be stuck”</td>
<td>ka:1 “galingale”</td>
<td>ka:2 “to kill”</td>
<td>ka:3 “to sell”</td>
</tr>
</tbody>
</table>

Short vowel kan0 “itch”
Long vowel ka:n0 “lever”
Implementing PENTA

Quantitative Target Approximation Model
(Prom-on et al., 2009)


Local Optimization

Parameters 1
Parameters 2
Parameters 3
Parameters 4

Best-fit parameters

Optimize

Analysis-by-synthesis

Categorical Parameters

Cat. Param. 1
Cat. Param. 2
Cat. Param. 3

Global Optimization

Raw F0 Data

Utterance 1
Utterance 2
Utterance 3
Utterance 4

Optimized Categorical Parameters

Initialize

Synthesize, Compare & Test

Raw F0 Data

Utterance 1
Utterance 2
Utterance 3
Utterance 4

Adjust


\[ f_\theta(t) = \theta_0 + (\theta_1 + \theta_2 t + \theta_3 t^2) \]
Parallel encoding – English example

You’re going to Bloom dales with Alan?

Stress: \( u \) = unstressed, \( s \# \) = stressed where \# is the number of remaining unstressed syllables
Focus: \( \text{pre} \) = pre-focus, \( \text{on} \) = on-focus
Sentence: \( q \) = question \( \{ s \) = statement \}

*Final Syllable: \( n \) = non-final, \( sf \) = word-final, non-sentence-final, \( f \) = word-final, sentence-final
*Part of Speech: \( p \) = pronoun, \( v \) = verb, \( pp \) = preposition/particle, \( na \) = noun/adjective

Functional combinations

Communicative functions
- Stress
- Focus
- Temporal Division
- Essential functional combinations

5 functional combinations = 5 parameter sets \( (m, b, \lambda) \)
= 15 parameters

Parallel encoding in PENTATrainer 2

Stress: \( u \) = unstressed, \( s\# \) = stressed where \# is the number of remaining unstressed syllables
Focus: \( \text{pre} \) = pre-focus, \( \text{on} \) = on-focus
Sentence: \( q \) = question \( \{ s \) = statement \}

*Final Syllable: \( n \) = non-final, \( sf \) = word-final, non-sentence-final, \( f \) = word-final, sentence-final
*Part of Speech: \( p \) = pronoun, \( v \) = verb, \( pp \) = preposition/particle, \( na \) = noun/adjective

Simulated Annealing Optimization

Acceptance Test
\( p_\text{acc} = \exp(\Delta E/T) \)

Total error

Accept/Reject Adjustment

QTA Model

Synthesized F0 Data

Original F0 Data

Dataset

Initialize

Synthesize

Adjust

Compare
Thai Tone vs Vowel Length

The most time consuming step

Result – Synthesis Accuracy

- 10 parameter sets per speaker (for 5 tone x 2 vowel length conditions)
- 2 types of predictions
  - Speaker dependent: parameters of each speaker are estimated and used for synthesis separately
  - Speaker independent: parameters estimated from each speakers are averaged and the averaged values are used for synthesis

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Number of Parameter</th>
<th>RMSE (st)</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker Dependent</td>
<td>50</td>
<td>0.78 ± 0.05</td>
<td>0.889 ± 0.012</td>
</tr>
<tr>
<td>Speaker Independent</td>
<td>10</td>
<td>0.90 ± 0.06</td>
<td>0.871 ± 0.014</td>
</tr>
</tbody>
</table>

Result – Visual Inspection

Speaker Independent

Just click & done?... Modeling is now easy!!!

The most computational-power consuming step
Result – Synthesis Examples

Long-Long
Original Synthesis

Short-Short
Original Synthesis

Result – Pitch Target

Result - Strength

Conclusion

• Thai tones have distinctive contrasts in the
distribution of target approximation parameters
  – These capture the unique essence of each Thai tone
  categories

• These contrasts were learned automatically from
data without constraints using PENTATrainer 2

• Tone can also be effectively generated from
  learned parameters using PENTATrainer 2
Future Work

Thai tones
- Perceptual evaluation
- Interaction with other prosodic/segmental factors
- Uncontrolled corpus
Modeling
- Anticipatory effect
- Duration
- Grouping