SYNTHESIS OF VOWELS AND TONES IN THAI LANGUAGE
BY ARTICULATORY MODELING

Keerati Inthavisas, Montri Karnjanadecha and Thanate Khaorapapong

Department of Computer Engineering
Faculty of Engineering
Prince of Songkla University
Hat Yai, Songkhla, THAILAND, 90112
E-mail: panidas@unocal.com, montri@coe.psu.ac.th, kthanate@coe.psu.ac.th

ABSTRACT

We propose a way of synthesizing Thai vowels, monophthong and diphthong which consist of nine short-long vowel pairs and three short-long vowel pairs respectively. By adjusting the fundamental frequency or pitch contour, the five tonal of these vowels can be synthesized, leading to a total of 120 phonemes. We evaluate the synthesized monophthongs by comparing the first and second formants to standard in terms of different percentages, which result in 9.89% and 7.44% respectively. For the synthesis of diphthongs, we slide from one vocal-tract vowel to another using linear, logarithmic, or exponential functions.

1. INTRODUCTION

Most Thai speech synthesis based on concatenated method [1, 2, 3] or the analysis-synthesis method by linear predictive (LP) coding [4, 5]. The concatenative method gives high quality speech, but it is usually limited to one speaker, and requires more memory than other methods. The main deficiency of LP is that it represents an all-pole model, so that phonemes containing antifomants, such as nasals and nasalized vowels are poorly modeled [6]. Our work utilizes the articulatory model due to its high-quality synthetic speech output, its small memory requirement, and its parameter-based controls. However, the method is very complicated to implement efficiently. These issues may be solved in the future.

Thai vowels and tones are briefly explained in section 2 and section 3 describes our model which has 3 main components. After that, we describe the model that separated into three parts. The excitation component uses an estimate of the local pitch period to set an impulse train. The pitch period obtained using an extraction algorithm [7] where each frame is represents an energy level by summation of the absolute value of each point. The resulting energy values are normalized in order to represent gain constant or amplitude envelope of the sound. The second component is vocal-tract model that is parameterizes in term of a cross sectional area or area function, and the last part is a radiation model. In section 4 we explain the synthesis of monophthongs which have a constant vocal-tract, and diphthongs which have a varying vocal-tract.

2. THAI VOWELS AND TONES

The Thai language consists of 24 vowels, and 5 tonal levels or tone marks. The tone marks, which are different from intonation, can be added to words changing their meaning. The tone marks can be classified into dynamic and static tones. The rising and falling tones are dynamic and exhibit a great change in fundamental frequency with respect to duration. The low, mid, and high tones are static and are quite stable.

2.1 Thai Vowels

There are 12 short-long vowel pairs in which can be classified based on their contemporaneous and non-contemporaneous properties. A phoneme is stationary or continuant if the speech sound is produced by a steady-state vocal-tract configuration. A phoneme is non-continuant if a change in the vocal-tract configuration is required during the production of the speech sound. Figure 1 shows each vowel represented in Thai and in IPA (International Pronunciation Alphabet).

2.2 Thai Tones

The five tones are:
(i) The mid tone. It is indicated by the absence of a tone marker, and is spoken in an ordinary tone of voice without any inflection.

(ii) The low tone. It is indicated by ( ], and is a level tone with no inflection but spoken at a lower pitch than the mid tone.

(iii) The falling tone. It is indicated by ( ), and is an emphatic and heavily accented tone with a falling inflection.

(iv) The high tone. It is indicated by ( ), and is a uniform tone pitched well above the level of the speaker's normal voice.

(v) The rising tone. It is indicated by ( ), and has a rising inflection pitch.

3. SPEECH SYNTHESIS MODEL

A general linear discrete-time model for speech production is shown in Figure 2.

![Figure 2: A General Discrete-time Model for Speech Production](image)

The z-domain transfer function for the speech signal $s(n)$ can be written as follow:

$$S(z) = \Theta_0 G(z) H(z) R(z)$$  

A vocal-tract model $H(z)$ and radiation model $R(z)$ are excited by a discrete-time glottal excitation signal. During periods of voiced speech activity, the excitation uses an estimate of the local pitch period to set an impulse train generator that drives a glottal pulse shaping filter $G(z)$. The gain constant or amplitude envelope, $\Theta_0$, for represents the energy of the signal to improve the sound quality.

3.1 The Excitation Model

The voiced source is taken to be a volume velocity waveform modeled by a pulse generator or asymmetrical triangular wave, repeated at each fundamental period [8]. These pulses are simulated by a convolution of equation (2) corresponding to the fundamental period that is being synthesized. First, a single glottal pulse is generated, followed by a sequence of pulses that are fed to the system.

$$g(n) = \beta^n, \beta \approx 1, \beta < 1$$  

3.2 The Vocal-tract Model

Since speech production is characterized by changing vocal tract shape, it might be expected that a realistic vocal-tract model would consist of a tube that varies as a function of time and displacement along the axis of sound propagation. The formulation of such a time varying vocal-tract shape can be quite complex. One method of simplifying this model is to represent the vocal-tract as a series of $p$ segments made up of lossless acoustic tubes. A tube has cross sectional area, $A_{k}$, at the $k$th segment, as shown in Figure 3. These parameters are cross sectional area functions.

![Figure 3: The Vocal-tract Model of p Segment from Glottis to Lips](image)

The final discrete-time transfer function for a $p$-segment vocal-tract model has the following form[9]:

$$\begin{align*}
\left( \frac{U_l(t)}{U_R} \right) &= \frac{\prod_{k=0}^{p-1} \left[ 1 - r_k z^{-1} \right]}{\prod_{k=0}^{p-1} \left[ 1 - r_k z^{-1} \right] - \prod_{k=0}^{p-1} \left[ 1 - r_k \right] z^{-1}}
\end{align*}$$  

Where $r_k = A_{k+1} - A_k / A_{k+1} + A_k$

3.3 The Radiation Model

The radiation component, $R(z)$, is a low-impedance load that terminates the vocal-tract, and converts the volume velocity at the lips to a pressure wave in the far field. A simple digital filter with these properties is a differencer [9].

$$R(z) = 1 - \alpha z^{-1}, \alpha \approx 1, \alpha < 1$$  

4. METHOD

Monophthongs were synthesized by utilizing the cross sectional area functions of six English vowels [10] whose acoustical characteristic are similar to Thai vowels. They were adapted for Thai short - long vowel pairs $\ddot{a}[i], \ddot{a}[ii], \ddot{a}[c]. \ddot{a}[e], \ddot{a}[x], \ddot{a}[z], \ddot{a}[s], \ddot{a}[\ddot{a}], \ddot{a}[\dddot{a}], \ddot{a}[\dddot{u}], \ddot{a}[\dddot{u}]$ and $\ddot{a}[\dddot{a}], \ddot{a}[\dddot{a}]$ were predicted from their physiology of articulation [11]. As indicated in Table 1.
Table 1: The cross sectional area from glottis (tube 1) to lips (last tube) of Thai monophthongs

<table>
<thead>
<tr>
<th>Vowels</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ẹ[ɨ], Ẹ[ɨ]</td>
<td>1.10</td>
<td>1.49</td>
<td>3.61</td>
<td>4.77</td>
<td>3.41</td>
<td>1.63</td>
<td>0.62</td>
<td>0.41</td>
<td>0.76</td>
<td>1.54</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ẹ[ɛ], Ẹ[ɛ]</td>
<td>1.10</td>
<td>1.49</td>
<td>3.39</td>
<td>3.50</td>
<td>2.53</td>
<td>1.68</td>
<td>1.45</td>
<td>2.40</td>
<td>4.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ẹ[ɛ], Ẹ[ɛ]</td>
<td>1.10</td>
<td>1.45</td>
<td>2.23</td>
<td>2.62</td>
<td>3.68</td>
<td>5.01</td>
<td>4.93</td>
<td>5.54</td>
<td>7.62</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Á[u], Á[u]</td>
<td>1.84</td>
<td>4.10</td>
<td>4.56</td>
<td>2.75</td>
<td>1.06</td>
<td>0.48</td>
<td>0.62</td>
<td>2.15</td>
<td>5.28</td>
<td>0.77</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Á[ɛ], Á[ɛ]</td>
<td>1.48</td>
<td>3.07</td>
<td>2.95</td>
<td>1.99</td>
<td>1.48</td>
<td>1.74</td>
<td>2.12</td>
<td>3.09</td>
<td>3.37</td>
<td>2.97</td>
<td>2.97</td>
<td>-</td>
</tr>
<tr>
<td>Á[a], Á[a]</td>
<td>1.41</td>
<td>1.32</td>
<td>0.58</td>
<td>1.39</td>
<td>4.29</td>
<td>8.30</td>
<td>8.47</td>
<td>7.67</td>
<td>10.77</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Á[u], Á[u]</td>
<td>1.41</td>
<td>1.84</td>
<td>4.10</td>
<td>4.56</td>
<td>2.75</td>
<td>1.06</td>
<td>0.48</td>
<td>0.62</td>
<td>2.15</td>
<td>5.28</td>
<td>0.77</td>
<td>0.35</td>
</tr>
<tr>
<td>Ẹ[o], Ẹ[o]</td>
<td>0.77</td>
<td>1.40</td>
<td>0.95</td>
<td>0.62</td>
<td>0.48</td>
<td>1.06</td>
<td>2.75</td>
<td>4.56</td>
<td>4.10</td>
<td>1.84</td>
<td>1.41</td>
<td>-</td>
</tr>
<tr>
<td>Á[ɛ], Á[ɛ]</td>
<td>0.77</td>
<td>1.46</td>
<td>0.48</td>
<td>0.62</td>
<td>1.04</td>
<td>5.06</td>
<td>6.75</td>
<td>8.56</td>
<td>8.10</td>
<td>5.84</td>
<td>5.41</td>
<td>-</td>
</tr>
</tbody>
</table>

Diphthong synthesis involves an intentional movement from one vowel towards another, and these are some choices about how to model the movement. We slide from one vocal-tract vowel to another using normalized linear, logarithm and exponential functions, depicted in Figure 4.

Figure 4: Linearity, logarithmic and exponential functions.

We synthesize diphthongs using dynamic area functions, which are obtained from the area functions for a vowel in the first stage. For later stages, they are plus or minus the normalized functions multiplied to the different cross sectional areas of the corresponding tube from 2 vocal-tract vowels. This continues until the end of the tube or the normalized value equals 1. Also, the number of tubes is reduced from one vowel to the next. For example both Ẹ[ɨ], Ẹ[ɨ], Ẹ[ɛ], Ẹ[ɛ] and Ẹ[ɛ], Ẹ[ɛ] are reduced from 10 tubes to 9, Ẹ[ɛ], Ẹ[ɛ] is reduced from 12 tubes to 9.

5. RESULTS AND DISCUSSION

So, we synthesize the sound using area functions from Table 1 together with fundamental frequency of 5 tones that indicated in Figure 5. The approximately duration of short and long vowels are 0.25 and 0.5 second respectively. The sound quality can be improved by adding the gain constant or amplitude envelopes of each tone.

For example, Figure 6 shows the result with the acoustic waveform for Ẹ[ɨ] in mid tone. Figure 7 gives the frequency response indicating the first and second formants of this vowel.

After that, synthesized monophthongs have been evaluated by comparing first and second formants to standard [12] in term of different percentage. The results are 9.89% and 7.44% respectively.

By listening to the synthesized sound, we found that the best result for Ẹ[ɛ], Ẹ[ɛ] is to slide the vocal-tract as an exponential function. Figure 8 shows the first and second
formant transformation of recorded sound, and Figure 9 gives
the synthesized sound. The first formant's transformational
trend of synthesized sound is similar to exponentially rising,
and the second formant appears to decrease logarithmically.
This corresponds to a vocal-tract slide using an exponential
function, while linear and logarithmic functions give reversed
results, as shown in Figure 10.

The analysis and synthesis of ʔә[ua], ʔә[uua] and ʔә[ua],
ʔә[uua] are obtained in a similar manner.

6. CONCLUSION

Thai vowels can be synthesized with articulatory modeling,
based on the characteristics of the articulated vocal-tract and
sound properties such as fundamental frequency which
separate the sound into 5 tones. Also, Short or long vowels are
classified by this parameter. Exponential functions are
satisfactory for synthesizing diphthongs such as ʔә[ai], ʔә[ii].
Logarithm function are better for synthesis of ʔә[ua],
ʔә[uua] and ʔә[ua], ʔә[uua]. General sound quality can be
improved by adding an amplitude envelope while the process
is operating.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Andrew Davison for his
review of this manuscript.

8. REFERENCES

Open Syllables using Microphonemes," Master's thesis
of Electrical Engineering, Department of Electrical
Engineering, Graduate School, Chulalongkorn
University, 1998.
[2] Luksaneeyanawin, S., "A Thai Text to Speech System,
Proceedings of The Regional Workshop on Computer
Processing of Asian Languages (CPAL), Asian Institute
to Speech Synthesis of Thai - A Tone Language,"
Journal of the Acoustic Society of Japan
, 4.2. 97-106, 1983
Master's thesis of Science, Department of Computer
Engineering, Graduate School, Chulalongkorn
University, 1990.
Linear Predictive Coding." Master's thesis of Electrical
Engineering, Department of Electrical Engineering,
Graduate School, Chulalongkorn University, 1987.
Technology," Master's Thesis of Helsinki University of
Technology, Department of Electrical and
Communications Engineering, 1999.
Extraction", IEEE Trans., Vol. Au - 6, No.2 (June 1968)
: 262-266.
[8] Sadaoki, F., Digital Speech Processing Synthesis and
Time Processing of Speech Signals, the institute of
Electrical and Electronics Engineers, Inc., New York,
2000.
/VOWELS/vowels.html
[11] Inthavisas, K., Khaorapapong, T. and
Karnjanadecha, M. "Synthesis of Thai Monophthongs by
Articulatory Modeling," Proceedings of the Information
Engineering Postgraduate Workshop 2004, pp. 34-38,
Thai: acoustical measurements and experiments.,"
Bloomington: Indiana University Research Center in