A demisyllable approach to speech synthesis of Thai —A tone language

Chatchavalit Saravari and Satoshi Imai

Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, 4259, Nagatsuta, Midori-ku, Yokohama, 227 Japan

(Received 17 May 1982)

An approach to speech synthesis of Thai which is a tone language is presented. The stored data are presently cepstral parameters of consonant-vowel- (CV-) , vowel-consonant- (VC-) and vowel (V) demisyllables. Speech is produced by concatenating stored cepstrum-coded demisyllables, providing pitch contour for each of monosyllabic sounds and synthesizing the speech through the log magnitude approximation (LMA) digital filter. Interpolation at boundaries of demisyllables is linear. For each tone, the f0 contours are simply synthesized from the information of a rudimentary contour specified by a third order polynomial in the domain of pitch period. The details on determining a rudimentary f0 contour for each tone, segmentation for demisyllable library, and rules for concatenation and pitch assignment are described, along with experimental evaluation tests and discussions.

PACS number: 43. 70. Jt, 43. 70. Gr, 43. 70. Qa, 43. 70. Ve

1. INTRODUCTION

Recently, research on speech synthesis-by-rule has become an increasingly interesting aspect.1-4) Techniques in automatic synthesis of speech, however, differ from language to language due to the differences in structural patterns of articulation and prosody. Thai, in common with many languages in Asia such as Chinese, Vietnamese, etc., is a tone language. A significant feature of tone languages is that pitch inflections are crucial in signifying distinctive speech sounds. Pike5) has stated a common feature of tone languages that "...A man and a woman may both use the same tonemes, even though they speak on different general levels of pitch. Either of them may retain the same tonemes while lowering or raising the voice in general, since it is the relative pitch of syllables within the immediate context that constitutes the essence of tonemic contrast."

In the speech of the people in central Thailand there are five tones which are called here the "level," "grave," "dropped," "acute" and "rising" tones according to So Sethaputra, New Model Thai-English Dictionary (1977). The tone names correspond to the conventionally called "mid," "low," "falling," "high" and "rising" tones, respectively. The fundamental frequency contours (f0 contours) for the tones are approximately hand-drawn in Fig. 1. For example, /ma/ inflected in the level, acute and rising tones mean "to come," "a horse" and "a dog," respectively. It can be thought that the f0 contours of the monosyllabic sounds in a message are more perceptually significant than that of the entire message. In Thai, rising and falling patterns of f0 contours are not generally used to determine whether an utterance is a question or an assertion but addition of words to the end of the utterance is used instead. A variety of perceptual experiments on Thai tones has been carried out by Abramson.6-9) In Abramson's scheme, the tones have been viewed as falling into two groups, the dynamic tones and the static tones. The dropped and rising tones have been classified as the
Fig. 1 Approximate fundamental frequency contours of Thai tones.

dynamic tones due to the sharp downward and upward movements whereas the others have been classified as the static tones. However, one of his experiments has shown that the identification of the static tones, especially the grave and acute tones, can be improved with f0 movement. Accordingly, we have viewed here all the tones except the level tone as having f0 movement.

This paper describes a demisyllable approach to speech synthesis of Thai. The stored data are presently cepstral parameters of parts of utterances of the forms consonant-vowel (CV), vowel-consonant (VC) and vowel (V). Parts of CV as well as VC are stored by keeping their mutual influence in the middle of the segment, this kind of segment has been called "dyad" by Peterson et al. For vowels, the most steady part is stored. As an initial step, the f0 contours for each tone are simply synthesized from the information of a rudimentary contour specified by a third order polynomial in the pitch-period domain. Words are produced by concatenating stored cepstrum-coded demisyllables, providing a tonal f0 contour for each of monosyllabic sounds and synthesizing the speech through the log magnitude approximation (LMA) digital filter whose log magnitude of its transfer function approximates the finite Fourier series corresponding to the log-spectral envelope passing through peaks of the fine structure. The details on determining a rudimentary f0 contour for each tone in the domain of pitch period, segmentation for obtaining the data to be stored, and rules for concatenation and pitch assignment are described, along with experimental evaluation tests and discussions. This study is the first attempt to produce Thai speech by rule and, as expected, has provided us informative data for further development in the area of automatic speech synthesis.

2. ANALYSIS-SYNTHESIS SYSTEM

The major criteria for selecting an analysis-synthesis system are as follows: 1) the model should be as natural representation of the speech signal as possible, 2) the speech information should be easily extracted from the speech signal, that is, the speech information should be automatically extractable and reliable, and 3) speech produced by interpolation of parameters representing speech must not be excessively distorted.

The speech analysis-synthesis system, in which the log-spectral envelope of the speech signal is specified by cepstral parameters and the reproduced

Fig. 2 Short-time spectral envelopes of "mIa pUan hIu kau" analyzed with a 10-ms frame period using a 25.6-ms Hamming window, in (a), and those after linear interpolation of corresponding cepstral parameters between the boundaries of diphthongs, in (b).
speech is the output of the LMA filter, has been found to be a good candidate. In this system, the vocal tract is represented as a time-varying digital filter consisting of representations of both resonances (poles) and antiresonances (zeros) which is generally a natural representations of all speech sounds. The vocal tract parameters for the LMA filter are in terms of the cepstrum corresponding to the log-spectral envelope which can be easily computed and very reliable. Since the log-spectral envelope has been further improved to pass through peaks of the fine structure, the frame-to-frame fluctuation is small. Linear interpolation of cepstral parameters at words' boundaries works as well as that of the spectral envelopes. Cepstral parameters can be directly interpolated without any transformation. As an example, Fig. 2 shows original short time spectral envelope of “mIa pUan hIu kau” analyzed with a frame period of 10 ms in (a), and those obtained after linear interpolation of cepstral parameters between boundaries of diphthongs in (b). Although some degradation of formant transition is seen, such degradation was not enough to damage its intelligibility.

3. TONE ANALYSIS

We have analyzed the tones by means of observing the \( f_0 \) contours according to a conclusion by Abramson that the \( f_0 \) contour isolated by means of acoustic analysis furnished sufficient cues by themselves for the identification of the five tones of central Thai.\(^6\)

The \( f_0 \) contours of tonally inflected /a/ in the five tones of 4 speakers, 3 males (PH, NK and CS) and a female (OS) are shown in Fig. 3. The monosyllables are consecutively inflected in the level, grave, dropped, acute and rising tones, respectively. The duration of each voiced period is about 300 ms which is about the duration of the long vowel (there are 2 kinds of vowel in Thai—long and short vowels—which are also distinctive). The speakers NK and CS were born in the capital city which is in the central part of Thailand, PH and OS were born in the northern and southern parts, respectively. However, the 4 speakers speak standard Thai. From the figure, the \( f_0 \) ranges of PH and CS who have almost the same \( f_0 \) level are both 90\(~160\) Hz. For NK whose \( f_0 \) level is about 20 Hz higher than that of PH and CS, the \( f_0 \) range is 110\(~180\) Hz. For OS whose \( f_0 \) level is about 2 times higher than that of PH and CS, the \( f_0 \) range is 180\(~320\) Hz. It is also seen in the figure that the \( f_0 \) contours for the same tones are almost in the same shapes.

Let have a closer look at the \( f_0 \) contours of

![Fig. 3 Illustration showing how 4 speakers inflected \( f_0 \) for each of the tones. The utterance is consecutive /a/'s in the five tones. The letters l, g, d, a and r denote the level, grave, dropped, acute and rising tones, respectively.](image)
the tones in the example while ignoring some gross errors at the beginning and end of each contour.

The $f_0$ contour for the level tone, $l$, is almost horizontal at the level particular to each speaker, normally in the middle of the $f_0$ range.

The $f_0$ contour for the grave tone, $g$, begins from a level below that of the level tone—about 10 Hz for the male speakers and 20 Hz for the female speaker—and performs a gradual fall to another level near the lower boundary of the $f_0$ range.

The $f_0$ contour for the dropped tone, $d$, begins from almost the same level as that for the level tone and performs a rise before falling to another level normally near the lower boundary of the $f_0$ range. The duration of rising is apparently shorter than that of falling. For the female speaker, however, the final $f_0$ level is still higher than the initial $f_0$ level; this may be because the time delay before the following voicing is short. Nevertheless, the identifiability of the tone has not been disturbed.

The $f_0$ contour for the acute tone, $a$, begins from almost the same level as that for the dropped tone and performs a gradual rise to another level about 20 Hz above the initial level for the male speakers and about 80 Hz for the female speaker. It is seen that the relative level between the initial and final levels for the female speaker is somewhat high and the contour has noticeable variation.

The $f_0$ contour for the rising tone, $r$, begins from almost the same level as that for the grave tone and performs a fall before rising to another level near the upper boundary of the $f_0$ range. The duration of falling is shorter than that of rising. It is noted that the $f_0$ contour for the rising tone is the complementary of that for the dropped tone.

Figure 4 shows the $f_0$ contours of the sentence /ag ka ra ha wan nI n3n tIa/ of the 4 speakers. The arrow mark indicates where the voicing of each monosyllabic sound begins. The figure illustrates that, for each speaker, the shape of the $f_0$ contours of the monosyllabic sounds are similar to those of the corresponding tones of /a/s in Fig. 3. The absolute $f_0$ level is apparently less perceptually significant than the shape of the contour. The second contour from left in the plot of each speaker shows the concatenation of 2 grave-tone CV-syllables in which the initial phoneme of the concatenating syllable is voiced (/r/). The shapes of the $f_0$ contours of the two syllables are similar although they begin from different levels. For all the speakers except PH, the $f_0$ contour of the concatenating syllable begins from the final $f_0$ level of the preceding syllable. As PH has stressed the concatenating syllable strongly, the initial $f_0$ level of the following syllable is somewhat lower than the final $f_0$ level of the preceding one.

![Fig. 4 Illustration showing how 4 speakers inflected $f_0$ for /ag ka ra ha wan nI n3n tIa/, the arrow mark indicates where the voicing of each of monosyllabic sounds begins.](image-url)
The third contour from left in the plot of each speaker also illustrates this. In the plot of CS shown the concatenation of 3 syllables in a breath group, /ha wan nI/, and in the others shown that of 4 syllables in a breath group, /ha wan nI n3n/. The initial phonemes of the concatenating syllables are voiced. It is seen that the shape of the f0 contour for each tone is preserved while the f0 level is varied in accordance with the final f0 level of the contour of the preceding syllable.

The f0 contour of the voicing following a pause or an unvoiced sound begins from about the level in the middle of the f0 range. In case when the syllables preceding and following the pause are in the same tone, the initial pitch value of the one following the pause is apparently somewhat higher than that of the one preceding the pause as shown in the first and second contours in the plot of each speaker.

This agrees with Abramson's hypothesis that absolute fundamental frequency heights contribute nothing to the identification of the tones while the shapes of the frequency contours carry all the information. It is apparent that the tone pattern has major effect in the intelligibility while the tone level specifies the accent and intonation of the speech. The intrinsic f0 level for each tone presumably exists in isolation but seems to disappear when the words are concatenated into a message.

3.1 Pitch Information

Abramson has found that when removing the original tonal contour of the monosyllabic sound, imposing it with the synthetic contours and resynthesizing by means of a formant synthesizer, the tones are still perfectly identifiable. This seems to be a great benefit for developing a system for speech synthesis-by-rule of Thai. The system would become simpler if the tonal contour of various durations can be synthesized from the information of the contour of a certain duration. Figure 5 shows the original 270-, 370- and 470-ms f0 contours for the utterances in the dropped tone which are plotted on the same scales. It appears that the initial as well as the final f0 level of the three contours are almost the same while the f0 values vary between the two levels with similar shapes.

Suppose that we have the information of a f0 contour of, say, 300 ms, how can we synthesize the three contours from that information? An initial simple solution proposed here is to specify the tonal pitch-period contour of a certain duration by a polynomial—the initial pitch period value is specified by its constant term and the shape by its coefficients. The reason for using the domain of pitch period is that the polynomial would be applied to another speaker easier than using the domain of frequency.

Suppose that the pitch period value \( p(v) \) of a \( L \)-frame contour is in terms of the frame number \( v \) as follow

\[
p(v) = a_0 + a_1 v + a_2 v^2 + a_3 v^3, \quad 0 \leq v \leq L - 1.
\] (1)

(The polynomial of the third order is apparently adequate for fitting the pitch-period contours for the five tones.)

We shall derive the polynomial \( \bar{p}(v) \) for a \( L_a \)-frame as

\[
\bar{p}(v) = a_0 + a_1 \left( \frac{L-1}{L_a-1} v \right) + a_2 \left( \frac{L-1}{L_a-1} v \right)^2 + a_3 \left( \frac{L-1}{L_a-1} v \right)^3, \quad 0 \leq v \leq L_a - 1.
\] (2)

It is seen that \( \bar{p}(0) = p(0) \) and \( \bar{p}(L_a-1) = p(L-1) \). If \( L_a \) is not so much smaller than \( L \), the variation of the values of \( \bar{p}(v) \) should be almost in the same shape as that of \( p(v) \). Figure 6 shows the 270-, 370- and 470-ms f0 contours synthesized from the information of a 300-ms (30-frame) contour. An experiment has shown that the identifiability of the tone was not disturbed when replacing the original contours in Fig. 5 with the synthetic contours in Fig. 6 and resynthesizing the speech by means of the analysis-synthesis system.

Another experiment has been made by using
Fig. 6 Synthetic 270-, 370- and 470-ms \( f_0 \) contours for the dropped tone produced from the information of a 300-ms \( f_0 \) contour.

Fig. 7 The rudimentary 30-frame (300-ms) \( f_0 \) contours for speech synthesis of Thai.

The 30-frame \( f_0 \) contours shown in Fig. 7 as the rudimentary contours. The original \( f_0 \) contours of the monosyllabic sounds in five 3-s continuous messages were replaced with the synthetic contours and the messages were resynthesized. The \( f_0 \) level at the beginning of a voiced breath group was adjusted to be the same as that of the original contour, the \( f_0 \) level at the beginning of each tonal monosyllabic sound in the breath group was adjusted to be the same as the final \( f_0 \) level of the preceding monosyllabic sound. The speech was presented to 12 Thai listeners who wrote down the identification of the tones in Thai script. The result has shown that the tones are perfectly identifiable by the listeners. We have also found that the identifiability of the dropped, grave and acute tones was not damaged when \( L_d \geq 4 \) frames (40 ms) while that of the level and rising tones is when \( L_d \geq 15 \) frames (150 ms). In Thai, the voiced duration shorter than 200 ms is not normally inflected in the level and rising tones. These tones are almost always for the long vowels whose durations vary from 280~400 ms for normal speed. For the short vowels which can be inflected in all the five tones, the durations vary from 50~150 ms. It can therefore be thought that synthetic \( f_0 \) contours may be efficiently used in the speech synthesis of Thai.

The 30-frame rudimentary pitch-period contours whose \( f_0 \) contours are shown in Fig. 7 are specified by the following polynomials.

For the level tone:

\[
p(v) = 70 + 9.20 \times 10^{-1} v - 4.205 \times 10^{-2} v^2 + 6.497 \times 10^{-4} v^3.
\]

Table 1  Confusions of consonantal parts of 1,170 synthetic tonal nonsense CV- and VC-words presented to 12 Thai listeners.

<table>
<thead>
<tr>
<th>STIMULI</th>
<th>WRONG ANSWER (NUMBER OF OCCURRENCES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/g/-</td>
<td>/k/- (37) , /t/- (21)</td>
</tr>
<tr>
<td>/b/-</td>
<td>/w/- (103) , /b/- (24)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/a/- (35) , /g/- (20)</td>
</tr>
<tr>
<td>/p/-</td>
<td>/t/- (26) , /p/- (13)</td>
</tr>
<tr>
<td>/t/-</td>
<td>/g/- (61) , /a/- (38) , /k/- (22) , /g/- (18)</td>
</tr>
<tr>
<td>/f/-</td>
<td>/t/- (61) , /f/- (61) , /a/- (23)</td>
</tr>
<tr>
<td>/y/-</td>
<td>/b/- (62) , /t/- (55) , /a/- (45) , /k/- (17)</td>
</tr>
<tr>
<td>/j/-</td>
<td>/b/- (111) , /a/- (67)</td>
</tr>
<tr>
<td>/m/-</td>
<td>/b/- (42)</td>
</tr>
<tr>
<td>/n/-</td>
<td>/m/- (17)</td>
</tr>
<tr>
<td>/l/-</td>
<td>/n/- (83) , /m/- (20)</td>
</tr>
<tr>
<td>/r/-</td>
<td>/t/- (36)</td>
</tr>
<tr>
<td>/l/-</td>
<td>/l/- (103) , /k/- (61)</td>
</tr>
<tr>
<td>/l/-</td>
<td>/l/- (39)</td>
</tr>
<tr>
<td>/l/-</td>
<td>/b/- (35) , /w/- (25)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/o/- (177) , /b/- (62)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/g/- (115) , /b/- (80)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/g/- (119) , /b/- (65)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/y/- (36) , /m/- (20)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/a/- (22)</td>
</tr>
<tr>
<td>/a/-</td>
<td>/a/- (45)</td>
</tr>
</tbody>
</table>

error rate: \[
\frac{2026}{1020} \times 100 = 14.4 \%
\]
For the grave tone:
\[ p(v) = 80 + 2.720v - 1.726 \times 10^{-1}v^2 + 3.767 \times 10^{-5}v^3. \]

For the dropped tone:
\[ p(v) = 68 - 6.413 \times 10^{-1}v - 1.155 \times 10^{-2}v^2 + 2.018 \times 10^{-5}v^3. \]

For the acute tone:
\[ p(v) = 68 + 5.441 \times 10^{-2}v - 1.045 \times 10^{-2}v^2 + 4.718 \times 10^{-6}v^3. \]

For the rising tone:
\[ p(v) = 80 + 1.254v - 3.758 \times 10^{-2}v^2 - 9.926 \times 10^{-4}v^3. \]

The constant terms are chosen for isolated monosyllabic words for a medium-pitch male speaker.

4. DEMISYLLABLE LIBRARY

Speech materials consisted of the level-tone utterances of Thai CV- and VC-syllables and vowels spoken by the speaker NK. The materials involved 243 isolated utterances. The utterances were recorded with a high-quality microphone in a quiet room. The acceptability of each utterance was evaluated by a Thai observer. The utterance that had not been agreed upon by the observer was repeated. The durations of the utterances were about 150–250 ms. Before analyses were made, the recorded utterances had been evaluated again by the observer.

Thai has 9 vowels, /i/, /e/, /ae/, /a/, /3/, / /, /u/, /U/ and /o/; 7 diphthongs, /ai/, /au/, /ia/, /ia/, /ua/, /ua/ and /ul/; 4 semivowels, /w/, /l/, /r/ and /y/; 3 nasals, /m/, /n/ and /n/; 3 voiced stops, /g/, /b/ and /d/; 3 voiced stops, /k/, /t/, /t/ and /p/; 3 voiced fricatives, /f/, /s/ and /h/; a voiced affricate, /d3h/; and an unvoiced affricate, /t/.

Voiced fricative does not exist in Thai speech. The major difference between the stops /t/ and /t/ as well as /p/ and /p/ is that for the formers, more pressure is built up behind the occlusion in the vocal tract and the air is released with much shorter period of aspiration than that for the laters.

The consonants and semivowels all exist in the initial phonemes for CV-utterances. Only the voiced stops and nasals exist in the final consonants for the VC utterances. Only the vowels are used as the vocalic sounds since the diphthongs can be intelligibly produced by rules.

The segmentation was performed by a segmenter. First, the speech cepsrum was computed and stored. The symbolic representation of the isolated utterance to be segmented was inputted by the user and the demisyllable was outputted to library automatically. The segmentation was performed in the quefrency domain. The major cues for the segmentation were the cepstral gain, the first and second coefficients which are the major cues specifying the transitions between the fricative and vowel, and the semivowel and vowel, respectively, cepstral distance measures, and the average powers of the spectral envelope components in various frequency regions. We shall state here some criteria for the segmentation.

4.1 Initial Segmental Boundary (ISB)

The initial plosives—including the voiced stops, the unvoiced stops, /d3h/ and /tç/—are dynamic sounds involving the explosion followed by a period of aspiration or a period of friction. Accordingly, the ISB must be the initial boundary of the phoneme unless the degradation in intelligibility is observed. For the initial unvoiced stops, /d3h/ and /tç/, the ISB was the beginning of the explosion. For the initial voiced stops /b/ and /d/, however, the vocal cords vibrate before the explosion generating a buzz that was found to have much effect on the intelligibility. The ISB was therefore the initial boundary of the buzz. In case of /g/, it was found that the buzz has negligible effect and was excluded, the ISB was the beginning of the explosion.

The initial nasals are continuant sounds involving a steady state period. The ISB was the earliest frame entering the steady state. The condition that the ISB must be in the steady state period is necessary since it will be used as an interpolating boundary when concatenated to some voiced phonemes.

The initial fricatives and the whisper are continuant sounds. The ISB can be the earliest frame in which a fraction of the consonant exists or the earliest steady state frame. There is no restriction like in case of the initial nasals. However, we have considered the earliest steady state frame as the ISB.

The initial semivowels are continuant sounds, the condition for the ISB is similar to that of the initial nasals.
4.2 Final Segmental Boundary (FSB) and Vowel Onset

The FSB was one of the steady-state vowel frame. First, the vowel onset (the earliest frame behind the transitional region consisting of only a vowel part) was determined, the FSB was the seventh frame from the vowel onset. The vowel onset was also considered as the initial timing point for the duration assignment of the syllable nucleus.

The vowel itself was stored by including 5 most steady state speech frames which were determined by means of cepstral distance measures.

4.3 Voiced-Unvoiced Sign

After the ISB and FSB had been determined, the voiced-unvoiced sign was assigned to each frame. For initial voicing, all speech frames were assigned voiced sign. For initial voiceless, the voiced sign was assigned to each of speech frames beginning from the earliest frame in the transitional region where the average power of the spectral envelope components in the f0 region is greater than a threshold.

5. CONCATENATION RULES

The block diagram for producing speech from the stored cepstrum-coded demisyllables is shown in Fig. 8. The input sequences consist of symbols denoting demisyllables, tone, vowel duration and stress level for each of monosyllabic sounds. Generally, the major factors determining vowel durations are the position of the vowel in the word and of the word in the breath group, tones and stress levels. For normal speech of the speaker NK, the durations of the long vowel vary from 280 to 400 ms and those of the short vowel vary from 50 to 150 ms. It appears that the duration is the longest for rising-tone utterances and gradually becomes shorter for the dropped-tone, acute-tone, grave-tone and level-tone utterances, respectively.

In this study vowel durations were assigned according to those of the corresponding natural speech. Consonant durations were unchanged from those of the stored demisyllables.

Patterns of Thai monosyllabic words are mostly a vowel or diphthong preceded or followed, or preceded and followed by consonants. Words of CCV pattern also exist but they were excluded in this study. The words produced were of the forms CV, VC, CVC and V, where the V in each form can be both the vowel or diphthong and Cs in CVC can be both the same and different phonemes. The vowel duration was adjusted by repeatedly transmitting the cepstral parameters of the final segmental boundary in case of CV and V, and the initial segmental boundary in case of VC to the LMA filter. Words of CVC pattern in which V is a vowel were produced by concatenation of CV- and VC demisyllables of the same vowel with 20-ms (2-frame) linear interpolation of cepstral parameters between the boundaries, the vowel durations of both demisyllables were adjusted to be the same. When V is a diphthong, the concatenation was between those of different vowels with 80- and 40 ms linear interpolation for the long and short vowel, respectively, the duration of the concatenating vowel was adjusted to be two-third of that of the preceding one.

The parameter specifying the intensity of the synthetic speech was the cepstral gain. Without stress, the cepstral gain was adjusted to gradually rise from an initial state level to a steady state level, stay in that level for some times and gradually fall to a final state level. The intensity contour was modified by stress assignments. Two stress levels, unstress and primary stress, were used in our system. For primary stressed vowel, there is an increase in the value of the cepstral gain on the vowel. Besides, the initial f0 level of the word involving the primary stressed vowel was adjusted to be higher than that when the vowel was unstressed (0.5 ms lower in the domain of pitch period).

To synthesize polysyllabic words, four cases were considered when concatenating syllables into a word: (Let FPP denotes the final phoneme of the preceding syllable and IPC denotes the initial phoneme of the concatenating syllable.)

1. When IPC was an unvoiced stop: An interval of silence was allowed for 80 ms between FPP and IPC. The f0 contour of the voiced region
C. SARAVARI and S. IMAI: SPEECH SYNTHESIS OF THAI: A DEMISYLLABLE APPROACH

of the concatenating syllable began from similar \( f_0 \) level as that of the preceding syllable did. For primary stress vowels, the initial \( f_0 \) level was adjusted to be higher than that when it was unstressed (0.5 ms lower in the domain of the pitch period).

2. When IPC was an unvoiced fricative or a whisper: IPC was directly concatenated to FPP. The \( f_0 \) information was the same as that case 1.

3. When FPP was a vowel or a nasal and IPC was voiced except a stop: Linear interpolation was performed between FPP and IPC. When IPC was a stop, it was directly concatenated to FPP. The \( f_0 \) contour of the concatenating syllable with unstressed vowel began from the final \( f_0 \) position of the preceding syllable. The \( f_0 \) contour of the concatenating syllable with primary stressed vowel began at higher level (0.5 ms lower in the domain of pitch period) than the final \( f_0 \) position of the preceding syllable.

4. When FPP was a stop and IPC was voiced: An interval of silence was allowed after the stop for 100 ms. The \( f_0 \) information was the same as in case 1.

To produce a message, words were produced and concatenated. Last word in the breath group was automatically assigned primary stress, other primary stressed word was marked in the input sequences unless it was unstressed word.

6. EVALUATION TESTS

In the first test we conducted intelligibility tests of nonsense CV, VC and V words. Monosyllabic sounds of 300-ms voicing duration were produced from the stored data, each in the five tones. There were therefore 5 tokens in each monosyllabic sound resulting in 1,215 synthetic words. The words were randomized and presented to 12 Thai listeners, most of whom were unfamiliar with synthetic speech. Each word was played 2 times in a row with about 2-s interval. A sufficient interval of time was allowed for the listeners to write down what they heard before playing the next word. The listeners were told to guess whenever in doubt. The answers were written in Thai script. The intelligibility scores averaged about 86%; the tones were perfectly identifiable. No wrong answer was observed in the vowel part of the nonsense words. The confusions of the consonantal part of the words is shown in Table 1. It is seen that the major confusions involved the stops and the semivowels /t/ and /l/. For CV and VC words, the confusions when the vowel part was /u/ occurred about 80% of all the confusions, the others are /U/ and /o/.

In the second test, 150 meaningful disyllabic words were produced by rule and presented to the same group of listeners. The words were those in common use. Again, each word was played 2 times in a row with about 2-s interval. The intelligibility scores averaged about 97% while the tones were perfectly identifiable. No misidentification occurred for both monosyllables at a time. The major confusions involved final stops.

Finally, 10 simple sentences were experimentally produced and presented to the listeners. The sentence intelligibility was perfect. No formal test was conducted to measure the naturalness of the synthetic speech. The informal comments of the listeners indicate that the improvements in intonation, consonant duration and rhythm were necessary to improve the naturalness in the synthetic speech.

7. CONCLUSION

This study has shown that Thai speech can be arbitrarily produced by concatenation of cepstrum-coded demisyllables while an appropriate pitch contour must be provided for each of monosyllabic sounds in a message. Further investigation is necessary to find out the optimum number of demisyllables required for synthesizing natural-sounding speech. The cepstrum analysis and synthesis method has been found to be one of the most useful methods because 1) the parameter estimation is reliable and very accurate, 2) the synthesis filter consists of representations of both resonances (poles) and antiresonances (zeros) which is the natural representation of all speech sounds, and 3) the linear interpolation of cepstral parameters at words' boundaries do not damage the intelligibility of the speech.

Most of the words synthesized were very intelligible in isolation but did not always fit smoothly into sentences, usually because of lack of proper consonant duration assignment, since the consonant duration was unchanged from that of the stored demisyllables. Although words in the sentences were tonally intelligible, they were sometimes unnatural due to improper overall intonation. To improve the naturalness in the synthetic speech, the following major topics will be considered for
further works:

1) Instead of using only one rudimentary polynomial for specifying each of the tonal \( f_0 \) contours, it may be more appropriate to specify the tonal \( f_0 \) contours of various durations by corresponding polynomials while the \( f_0 \) values of the contour of desired length can be extrapolated from some of those polynomials.

2) The effects of stress on levels and shapes of the \( f_0 \) contours of the tones.

3) The intrinsic duration for each of the vowel and also the effects of the position of the vowel in the word or of the word in the breath group on the duration.

4) The implementation of an appropriate model for vowel duration assignment which involves the effects of the pitch pattern, stress, position and also consonant environment.

5) The interpolation method other than linear that may produce less distortion in the transitional region between demisyllables' boundaries and also the appropriate duration of the transitional region particular to each of interpolating pairs.

REFERENCES


5) K. Pike, Tone Languages (University of Michigan, Ann Arbor, Mich., 1948).


