ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Psychological and Physiological Acoustics
Session 2pPPb: Speech. Attention, and Impairment (Poster Session)

2pPPb16. The roles of temporal envelope and temporal fine structure in speech synthesis for cochlear implants for tonal language speakers
Nantaporn Saimai*, Charturong Tantibundhit, Chutamanee Onsuwan and Chai Wutiwiwatchai

*Corresponding author's address: Electrical and Computer Engineering, Thammasat University, 99 Moo 18 Phaholyothin Rd., Amper Khlonghuang, 12120, Pathumthani, Thailand, 5310030076@student.tu.ac.th

Since most cochlear implants (CIs) have been developed for non-tonal languages, their level of hearing improvement is significantly decreased when used by speakers of tonal language, e.g., Thai. Temporal envelope (TE) and temporal fine structure (TFS) are important acoustic cues for languages with lexical tones. Specifically, TE is shown to carry manner and voicing cues for consonants, while TFS correlates with vowel formant transitions. Therefore, TFS and TE are expected to enhance intelligibility of lexical tones for CI patients. We proposed the use of six-channel bandpass filters to extract spectral information. Then, TE is extracted by half-wave rectification and smoothed by lowpass filler at 500Hz cutoff frequency. TFS is extracted by the Hilbert transform to construct carrier signals. TE from each channel is modulated with its corresponding carrier signal and then combined to generate synthesized speech. Synthesized speech tokens from this study and two others (Fu et al. (1998) and Chen and Zhang (2008)) are evaluated by sixteen Thais with normal hearing. The results showed that the intelligibility scores from the proposed algorithm are significantly higher than the other two for initials (by 32.2%) and final consonants (by 16.7%) and significantly higher for tones (by 48.8%) than Fu et al.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Because most of the cochlear implant (CI) devices have been developed for English language, a stress-timed language. However, when it is used by a tonal language speaker, the recognition rate could be dramatically dropped. Recently, several researchers have been interested in developing CI specifically for tonal languages, e.g., Chinese. This device is composed of two parts: an external audio processor sometimes called sound processor and a local device to be implanted in the skull, called the implant (Loizou, 1999).

Moreover, the CI devices can be divided into two categories, i.e., single-channel and multi-channel. A continuous interleaved sampling (CIS) strategy is one of the multi-channel approaches that has been developed for both English (Geurts and Wouters, 2001) (Wilson et al., 1991) and Chinese (Fu et al., 1998) (Chen and Zhang, 2008) (Luo and Fu, 2004) (Lan et al., 2004) (Nie et al., 2005) languages. For Chinese, the standard CIS extension includes temporal envelope (TE) and fundamental frequency (F0), which are important for tone recognition (Fu et al., 1998) (Luo and Fu, 2004).

Fu et al. used CIS strategy and amplitude envelope modulation with white Gaussian noise bands (Fu et al., 1998). They compared speech recognition results from different lowpass cutoff frequencies to extract TE (50Hz and 500Hz) and different number of channels (1–4 channels). Chen and Zhang emphasized on TE and temporal fine structure (TFS), which are important acoustic cues of speech intelligibility (Chen and Zhang, 2008). The Hilbert transform (HT) was employed in finding the TE and TFS of the analytic signal and then used to generate a carrier signal. They showed that the resulting synthesized speech provided more correct responses on tone identification and higher scores on speech intelligibility. Xu and Pfingst developed CI and emphasized on TFS and TE cues on speech recognition like Chen and Zhang, but they varied the number of channels and lowpass cutoff frequencies (Xu and Pfingst, 2008). The experimental results from normal hearing listeners showed varying degrees of trade-off between spectral and temporal information in the recognition of consonants, vowels, and tones (Xu and Pfingst, 2008).

Saimai et al. developed a speech synthesis algorithm for tonal language (Saimai et al., 2012). They used envelope extraction using CIS strategy and applied TFS to capture formant transitions extraction by Hilbert transform (Fogerty and Humes, 2012). The developed algorithm was evaluated and compared with Fu et al.’s algorithm (Fu et al., 1998) and Chen and Zhang’s algorithm (Chen and Zhang, 2008) by using intelligibility tests on meaningful rhyming words, which differ only in their initials. The experimental results showed that the proposed algorithm achieved highest percent intelligibility scores and percent correct responses. They suggested that both TE and TFS are important acoustic cues for languages with lexical tones and could provide higher speech recognition rate of synthesized speech.

Moreover, Saimai et al. (Saimai et al., 2013) designed the intelligibility test for final consonants and tones to evaluate the synthesized speech of their previously developed algorithm (Saimai et al., 2012) with the other two algorithms, i.e., Fu et al. and Chen and Zhang. The experimental results showed that percent intelligibility scores from the proposed algorithm are significantly higher than those of the other two algorithms for initials and final consonants and significantly higher for tones than that of (Fu et al., 1998). The percent recognition rate improvement of (Saimai et al., 2012) when compared with the other two algorithms and divided by type of phonemes shows that the initial nasals are significantly higher than (Fu et al., 1998), voiceless initial plosives are significantly higher than (Chen and Zhang, 2008), and tones are significantly higher than (Fu et al., 1998).

In this present study, we further investigate our developed algorithm (Saimai et al., 2012) with Fu et al.’s algorithm and Chen and Zhang’s algorithm by constructing tone-confusion matrix generated from the results of the intelligibility test to analyze the confusions (Saimai et al., 2013) in tones.

BACKGROUND

Thai Phonology

Thai is a tonal language composed of 21 initial phonemes, i.e., /p/, /pʰ/, /b/, /bʰ/, /tʰ/, /d/, /tɕ/, /tɕʰ/, /k/, /kʰ/, /Ɂ/, /f/, /l/, /m/, /n/, /ŋ/, /w/, and /j/ and 9 final phonemes, i.e., /p/, /t/, /k/, /Ɂ/, /m/, /n/, /ŋ/, /w/, and /j/. Thai have 9 short vowels, i.e., /i/, /ɨ/, /e/, /ɛ/, /æ/, /a/, /u/, /uː/, and 9 long vowels /ii/, /ɨː/, /ee/.
FIGURE 1. Fundamental frequency (F0) contours of five Thai tones (Wutiwiwatchai and Furui, 2007).

The Proposed Speech Synthesis Algorithm

The proposed algorithm illustrated in FIGURE 2 is composed of two paths, i.e., the TE extraction and the TFS extraction. For the first path of the envelope extraction, we follow the CIS approach (Fu et al., 1998). Specifically, speech signal was down-sampled from 44.1kHz to 22.05kHz and put through a pre-emphasis filter at 1,200Hz (first-order Butterworth high-pass filter). The cutoff frequencies of 6-channel bandpass filters (fourth-order elliptic with 50dB attenuation in the stop band and a 0.01dB ripple in the passband) are 25, 260, 600, 1,240, 2,420, 4,650, and 8,820Hz based on (Chen and Zhang, 2008). The temporal envelope is extracted by half-wave rectification and then smoothed by a low-pass filter (fourth-order Butterworth low-pass filter at 500 Hz).

For the second path, the temporal fine structure extraction is used to construct the carrier signal based on the Hilbert transform (HT) (Chen and Zhang, 2008).

The Hilbert transform $f(t)$ of the real signal $f(t)$ is defined for all $t$ by

$$f(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{t-\tau} d\tau.$$  \hspace{1cm} (1)

Temporal FS is $\cos(\Phi(t))$, where $\Phi(t)$ is a phase of the analytic signal expressed as

$$\Phi(t) = \arctan \left( \frac{f(t)}{f'(t)} \right).$$  \hspace{1cm} (2)

A carrier signal is constructed from the sinusoidal pulse with a period of 0.9 msec (Chen and Zhang, 2008) based on the peak of the sine wave corresponding to the peak of the temporal fine structure. The using criterion is to select the peak which is greater than or equal to 0.9. Applying the carrier signal to amplitude modulated with the temporal envelope in each band, which is obtained from the CIS strategy will give the results as

$$f_i(t) = a_i(t) C_i(t).$$  \hspace{1cm} (3)

The synthesized speech is a summation of amplitude of all of the modulated bands.
EXPERIMENTAL SETUP

The psychoacoustic test is composed of three test sets. Test 1 is the test of initial phonemes which is composed of 210 initial phonemes pairs (half of 2 times of combination of 21 initial phonemes, (i.e., $1/2 \times 2^{21}$) and 10 pairs of filler words. To bring out a balanced confusion matrix, the rhyming word in each pair is presented once as a stimulus in a trial, resulting in a total of 420 trials for initial consonants and 20 pairs of filler words.

Test 2 is a test set of final phonemes composed of 84 rhyming pairs across 8 final phonemes and 7 pairs of filler words. To be in line with the initial consonant test, there are the 175 trials ($84 \times 2 + 7$). Test 3 is a test set of tones composed of 80 pairs (40×2) and 5 pairs of filler words.

Each test set is composed of 3 sets of synthesized speech from 3 different algorithms: Fu et al. (Fu et al., 1998), Chen and Zhang (Chen and Zhang, 2008) and the proposed algorithm, respectively. Before the test session, a short training (with 4 trials) is given. In each trial, listeners hear one stimulus word, then choose the corresponding response from two words (forced choice) that appear on a computer screen by pressing a key on the keyboard. After each response was chosen, the listener will hear the next trial. After each set, there is a break of 5 minutes.

EXPERIMENTAL RESULTS

Table 1 shows percent correct responses, percent intelligibility scores and 95% confidence intervals (CI) of the results from the intelligibility test in each test set. For initial consonants, the percent correct responses and the percent intelligibility scores based on Fu et al. are 78.5% and 57.0%, Chen and Zhang 76.6% and 53.1%, while those of the proposed algorithm are 94.6% and 89.2%, respectively. The 95% CI of the proposed algorithm is the narrowest. For final consonants, the percent correct responses and the percent intelligibility scores based on Fu et al. are 80.65% and 61.31%, Chen and Zhang 74.55% and 49.11%, while those of the proposed algorithm are 88.99% and 77.98%, respectively. However, 95% CI of Fu et al. is the narrowest. For tones, the percent intelligibility scores of Chen and Zhang are higher than those of Fu et al., while the proposed algorithm achieves the highest percent intelligibility scores and has the narrowest 95% CI.

Table 2 shows $p$-values of tone confusion matrices calculated from paired t-test between percent correct responses from the proposed algorithm and those of the other two algorithms. The experimental results show that percent correct responses of the proposed algorithm are significantly higher in mid, high, and rising tones than those of Fu et al.

Table 3 shows confusion matrix for Thai tones from the proposed algorithm. Overall, stimuli with mid, falling, and rising tones (92.19%) are more accurately perceived than those with low and high tones. It can be observed that high tone is the most confusable.
TABLE 1. Percent correct responses and percent intelligibility scores in 3 test sets from the 3 different algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Initials correct</th>
<th>intelligibility</th>
<th>95% CI</th>
<th>Finals correct</th>
<th>intelligibility</th>
<th>95% CI</th>
<th>Tones correct</th>
<th>intelligibility</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fu et al.</td>
<td>78.51</td>
<td>57.02</td>
<td>48.92 to 65.13</td>
<td>80.65</td>
<td>61.31</td>
<td>53.68 to 66.94</td>
<td>65.94</td>
<td>31.88</td>
<td>19.85 to 43.90</td>
</tr>
<tr>
<td>Chen and Zhang</td>
<td>76.55</td>
<td>53.10</td>
<td>44.36 to 61.83</td>
<td>74.55</td>
<td>49.11</td>
<td>38.60 to 59.62</td>
<td>88.44</td>
<td>76.88</td>
<td>64.77 to 88.98</td>
</tr>
<tr>
<td>Proposed</td>
<td>94.58</td>
<td>89.17</td>
<td>84.13 to 94.20</td>
<td>88.99</td>
<td>77.98</td>
<td>68.81 to 87.14</td>
<td>90.31</td>
<td>80.63</td>
<td>72.97 to 88.28</td>
</tr>
</tbody>
</table>

TABLE 2. P-values of tone confusion matrix.

<table>
<thead>
<tr>
<th>Paired t-test between</th>
<th>Mid</th>
<th>Low</th>
<th>Falling</th>
<th>High</th>
<th>Rising</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fu et al. vs. Proposed</td>
<td>&lt; 0.01*</td>
<td>0.15</td>
<td>0.08</td>
<td>0.03*</td>
<td>&lt; 0.01**</td>
</tr>
<tr>
<td>Chen and Zhang vs. Proposed</td>
<td>0.27</td>
<td>0.08</td>
<td>0.42</td>
<td>1</td>
<td>0.33</td>
</tr>
</tbody>
</table>

TABLE 3. Confusion matrix for Thai tones from the proposed algorithm.

<table>
<thead>
<tr>
<th></th>
<th>Mid</th>
<th>Low</th>
<th>Falling</th>
<th>High</th>
<th>Rising</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid</td>
<td>92.19</td>
<td>4.69</td>
<td>1.56</td>
<td>0</td>
<td>1.56</td>
</tr>
<tr>
<td>Low</td>
<td>3.13</td>
<td>89.06</td>
<td>3.13</td>
<td>1.56</td>
<td>3.13</td>
</tr>
<tr>
<td>Falling</td>
<td>1.56</td>
<td>0</td>
<td>92.19</td>
<td>4.69</td>
<td>1.56</td>
</tr>
<tr>
<td>High</td>
<td>6.25</td>
<td>0</td>
<td>4.69</td>
<td>85.94</td>
<td>3.13</td>
</tr>
<tr>
<td>Rising</td>
<td>1.56</td>
<td>3.13</td>
<td>3.13</td>
<td>0</td>
<td>92.19</td>
</tr>
</tbody>
</table>

DISCUSSION AND FUTURE WORK

Confusion matrix of five Thai tones shows that high tone is the most confusable tone probably due to the fact that its contour looks very similar to that of the mid tone. Mid, falling and rising tones are the least confusable tones probably because their contours are clearly different from those of other tones; hence could easily be distinguished.

For tones, the intelligibility of the synthesized speech from (Chen and Zhang, 2008) is higher than that of (Fu et al. 1998). This is probably because the algorithm in (Chen and Zhang, 2008) was specifically developed for Chinese, which is a tonal language. The proposed algorithm provides the best speech intelligibility for almost all tones except for the low tone, which was perceived lower than (Chen and Zhang, 2008). We will investigate further to improve the intelligibility of the low tone. Finally, we plan to come up with a prototype device that can be used with Thai cochlear implant patients.

ACKNOWLEDGMENTS

The first author was supported in part by Thailand Graduate Institute of Science and Technology (TGIST), NSTDA (TGIST 01-55-016) and Faculty of Engineering, Thammasat University.

REFERENCES


