An Experimental Analysis of Pitch Patterns in Japanese Speakers of English with Verification by Speech Re-synthesis

Tomoko Nariai, Kazuyo Tanaka

Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
nariai@slis.tsukuba.ac.jp, ktanaka@slis.tsukuba.ac.jp

Abstract

Certain irregularities in utterances of a word or phrase occur in English as spoken by Japanese native subjects (Japanese English, henceforth). This study considers such pitch patterns as one of the most common causes of deficiencies in Japanese English, and that Japanese English would have better pitch patterns if its peculiarities are modified. Firstly, pitch patterns of Japanese English are statistically analyzed. The analytical results provide a rule for modifying the pitch patterns of Japanese English, in order to improve naturalness. To check the appropriateness of the rule, the pitch patterns of several samples of Japanese English are acoustically modified and re-synthesized. The modified speeches are evaluated in a listening experiment taken by native English speakers. Averagely, over threefold subjects support the proposed modification against original speeches. Therefore, the results indicate practical verification of modifying ways of Japanese English.

Index Terms: speech synthesis, second language learning, Japanese learners of English

1. Introduction

Previous researches on Japanese English have listed a number of characteristics [1]. However, none of these characteristics have yet been confirmed in terms of actual speech modification.

For example, previous study [2] analyzed the range of maximal and minimal pitch in Japanese English sentences. It was revealed that the dynamic range of pitch in Japanese English sentences was smaller than that in English speakers sentences. On the basis of this finding, we tried to improve the characteristics of Japanese English by means of a speech synthesizer. This type of modification, however, cannot cover the gap in pitch between Japanese English sentences and English speakers sentences. As a method of research, knowledge extraction by analysis with verification by re-synthesis is both informative and reliable in investigating second speech prosody.

In this study, the pitch patterns of Japanese English sentences are analyzed, and then the extracted characteristics are confirmed by speech re-synthesis. We analyze pitch of Japanese English on the assumption that the prosodic difference between the English and Japanese languages appears in focus or prominence [3].

We comparatively analyze the pitch peak and pitch range for words in Japanese English sentences and English speakers sentences with regard to the word class. The analytical results give rules of modifying ways of the pitch patterns in Japanese English. Then, the pitch patterns of Japanese English samples are modified on the basis of the rules, and then re-synthesized. Finally, the synthesized speech is evaluated by a listening experiment to confirm the approach.

2. Analysis of Pitch Patterns of Japanese English

2.1. Sample Speech

The group of English speakers consisted of 10 subjects, 5 males and 5 females, aged between 20 and 40. Most of the subjects were English teachers living in Japan.

The group of Japanese English consists of 17 subjects, 9 males and 8 females, aged between 20 and 30. Most of the subjects are undergraduate students. A native English speaker, who is an English teacher in Japan, judges that they are not proficient in English.

One hundred sentences are chosen from the MOCHA-TIMIT data set [4], the sentence numbers of which are timit001-030, 211-260, and 441-460. The 10 English speakers utters 100 sentences each. A group of 9 Japanese speakers utters 50 sentences each. A second group of 8 Japanese speakers utters the remaining 50 sentences.

The subjects were given sufficient time to practice reading the speech materials before recording. They were asked to enunciate clearly and to utter a sentence repeatedly until the speech sample was recorded properly. No other specific instruction for utterances of English was given to subjects.

2.2. Statistical Measure Used in Analysis

Each sentence utterance is sampled at the rate of 16 kHz and quantized into 16 bits. The acoustic feature extraction is conducted by WaveSurfer, which is provided from KTH. Extracted pitch patterns of individual sentences are segmented into word sequences, where word boundary is determined by observing the waveform and the spectrogram patterns. The word with cases such as vowel epenthesis or reduction, or a phonological clitic of monosyllabic word is not deleted as long as it does not interfere in word boundary detection.

The values of \( \text{peak}(i) \) and \( \text{range}(i) \) of individual words \( i \) are estimated as characterizing its prosodic patterns. The values are defined as:

\[
\text{peak}(i) = \text{maximal fundamental frequency of word } i \\
\text{range}(i) = \text{maximal minus minimal fundamental frequency of word } i
\]

The value of each word is normalized by the average of words in a sentence, as follows.

\[
x_j(i)' = x_j(i)/x\text{ave} \\
y_j(i)' = y_j(i)/y\text{ave}
\] (1)

where

\( x\text{ave} \) and \( y\text{ave} \) are the averages of the pitch peak and range in a sentence. The words are basically divided into three types as follows:

\[
\begin{align*}
x_j(i)' & = \frac{x_j(i)}{x\text{ave}} \\
y_j(i)' & = \frac{y_j(i)}{y\text{ave}}
\end{align*}
\] (1)
\[x_{ave} = \frac{\sum_{i=1}^{L} x_j(i)}{L}\]
\[y_{ave} = \frac{\sum_{i=1}^{L} y_j(i)}{L}\]

\(x_j(i): \text{peak}(i)\) or \(range(i)\) for a English speaker \(j\)
\(y_j(i): \text{peak}(i)\) or \(range(i)\) for a Japanese speaker \(j\)

\(L: \text{number of words in a sentence}\)

Statistical significance of the difference in sample distributions between the two groups can be evaluated by criterion used in statistical pattern recognition, that is, a ratio of the between-group variance to the within-group variance, known as Fisher’s ratio in linear discriminant analysis. This ratio is denoted by \(R\). If \(R\) is large, it indicates that considerable difference exists in sample distributions of the two groups. A procedure to calculate \(R\) is as follows:

\[R = \frac{(\tau - \mu)^2}{\sigma_x + \sigma_y}\]  \hspace{1cm} (2)

where:
\[\tau = \frac{1}{N} \sum_{j=1}^{N} x_j(i) '\]
\[\mu = \frac{1}{M} \sum_{j=1}^{M} y_j(i) '\]
\[\sigma_x = \frac{1}{N} \sum_{j=1}^{N} (x_j(i) ' - \tau)^2\]
\[\sigma_y = \frac{1}{M} \sum_{j=1}^{M} (y_j(i) ' - \mu)^2\]

\(N: \text{number of English speakers}\)
\(M: \text{number of Japanese speakers}\)

In the analysis, \(R > 0.1\) is used for detecting the difference between two distributions. In the analysis, “ntv>jpe” indicates that \(\text{peak}(i)\) or \(\text{range}(i)\) of word for English subjects “ntv” is higher or larger than those for Japanese subjects “jpe,” and “ntv<jpe” indicates the reverse. We count the words of “ntv>jpe” or “ntv<jpe” that satisfy \(R > 0.1\). Parameters \(\text{peak}(i)\) and \(\text{range}(i)\) are redefined using normalized prosodic patterns, \(\frac{x_j(i) '}{\tau}\) and \(\frac{y_j(i) '}{\mu}\), as

\(\text{peak}(i) ': \text{maximal value of } x_j(i) '\) (or \(y_j(i) '\))
\(\text{range}(i) ': \text{maximum minus minimum of } x_j(i) '\) (or \(y_j(i) '\))

2.3. Results

2.3.1. Result for Pitch Peak

Table 1 shows the results of \(\text{peak}(i)\) of pitch for content words. The columns in the table indicate noun, represented by ‘noun’, adjective, ‘adj’, verb, ‘verb’, and adverb, ‘adv’, respectively.

From the table, we can see that nouns of males and females amounts to 380, adjectives to 270, verbs to 210 and adverbs 44. For each word class, the number of words satisfying “ntv>jpe” and “ntv<jpe” is counted.

Out of 380 nouns, 244 satisfy \(R > 0.1, 61\%\) of which satisfy “ntv>jpe.” Out of 270 adjectives, 160 satisfy \(R > 0.1, 53\%\) of which satisfy “ntv>jpe.” In contrast, out of 210 verbs, 144 satisfy \(R > 0.1, 53\%\) of which satisfy “ntv<jpe.” Out of 44 adverbs, 26 satisfy \(R > 0.1, 58\%\) of which satisfy “ntv<jpe.” These suggest that nouns and adjectives for Japanese English are lower pitch than those for English speakers.

Table 1: Results of \(\text{peak}(i)\) of pitch for content words.

<table>
<thead>
<tr>
<th>number of word</th>
<th>noun</th>
<th>adj</th>
<th>verb</th>
<th>adv</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R &gt; 0.1)</td>
<td>244</td>
<td>160</td>
<td>144</td>
<td>26</td>
</tr>
<tr>
<td>ntv&gt;jpe</td>
<td>150</td>
<td>84</td>
<td>67</td>
<td>11</td>
</tr>
<tr>
<td>ntv&lt;jpe</td>
<td>94</td>
<td>77</td>
<td>77</td>
<td>15</td>
</tr>
</tbody>
</table>

2.3.2. Result for Pitch Range

Table 3 shows the results of \(\text{range}(i)\) of pitch for content words. The results suggest that more than half of nouns and

Figure 1: Pitch peak of \(\text{peak}(i)\) distribution of words in timit004 for English subjects (left-side thin bar) and Japanese subjects (right-side bold bar). Each bar indicates the range from (mean - SD) to (mean + SD).

Table 2 shows the results for function words. Function words of males and females amount to 512 words: 24 interrogative/negative, ‘int’, ‘ng’; 168 conjunction/preposition, ‘cnj’, ‘prp’; 74 be/auxiliary/verb/do, ‘be’; 148 articles, ‘art’; and 98 pronouns, ‘pron’.

For 22 interrogative/negative, 15 satisfy \(R > 0.1, 93\%\) of which satisfy “ntv<jpe.” On the contrary, over 80% of the conjunction/preposition, be/auxiliary/verb/do, article, and pronoun satisfy “ntv>jpe.” These suggest that most function words for Japanese English are higher pitch than those for English speakers. Also, interrogative/negative for Japanese English is lower pitch than that for English speakers.

In Figure 1, the means and standard deviations of \(\text{peak}(i)\) of words in timit004, “\(J\)ane may earn more money by working hard,” are plotted. As a result, the words that satisfy “ntv>jpe” are “\(J\)ane\(=0.53\)(\text{noun}) more\(=0.62\)(\text{adjective}) more\(=0.21\)(\text{noun})\), where the \(R\) value for each word is given in first parenthesis and the word class is given in the second parenthesis. The words for “ntv<jpe” are “\(J\)ane\(=0.82\)(\text{auxilialyverb}) earn\(=0.81\)(\text{verb})\(=1.93\)(\text{verb})\) hard\(=0.31\)(\text{adverb})\).
adjectives satisfy “ntv>jpe.” In contrast, for verbs and adverbs, more than half of which satisfy “ntv<jpe.” These suggest that nouns and adjectives for Japanese English are smaller pitch range than those for English speakers.

Table 4 shows the results of \( \text{range}(i) \) for function words. The results suggest that the majority of function words satisfy “ntv>jpe.” However, only for interrogative/negative, the majority satisfies “ntv<jpe.” These suggest that most function words for Japanese English are larger pitch range than those for English speakers.

Figure 2 shows the result of words in timit241, “Clear pronunciation is appreciated” and “pronunciation\([0.10:]\text{noun}\) satisfy “ntv>jpe,” however, “\( \text{ntv}<\text{adj} \)” and “\( \text{ntv}<\text{verb} \)” satisfy “ntv<jpe.”

3. Methods to Improve Pitch Patterns of Japanese English

Statistical analysis in section 2.3 suggests a knowledge for modifying way of Japanese English. Then the modification of pitch patterns of Japanese English can be stated as follows.

Japanese English will have improved pitch patterns if:

1. peak\((i)\) and range\((i)\) of each word in a sentence are ordered as follows:
   \[ \text{[rule]} \text{ function word } < \text{ (verb, adverb) } < \text{ (noun, adjective, interrogative, negative) } \]

2. English phonetics reveals that a sentence is phrased by a decline in pitch \([5, 6]\). Therefore, this study defines the word with the lowest pitch is the one at the end of sentence: peak\(i)\) of end word has the lowest pitch in a sentence, \([\text{rule2} \text{ the end word } < \text{ the word within a sentence} ]\)

Japanese English samples are modified to adjust to the rules, if there includes an erroneous order. Japanese English samples are analyzed to list what needs to be modified. The analysis process has the following five steps:

(i) A speech sample of Japanese English is analyzed to extract the pitch patterns. The pitch patterns are manually aligned with the word boundaries.

(ii) peak\(i) and \( \text{range}(i) \) for words in a sentence are measured.

(iii) peak\(i) and \( \text{range}(i) \) of words in the sentence are ranked according to its values. The ranking is compared with rule 1. Then, the defects are detected.

(v) The pitch height is checked by rule 2, and then, the defects are detected.

4. Modification by Speech Synthesis

4.1. Sample Speeches

Six Japanese speakers (four males, two females), aged between 20 and 30, were chosen. Most were Japanese university students.

Six sample sentences were chosen at random from the MOCHA-TIMIT data set, the sentence numbers of which are timit 009, 021, 022, 216, 246 and 452.

Each subject was allocated a different sentence, which they uttered once. Subjects were assigned number sample-1 to sample-6.

4.2. Analysis Result

Sample speeches are analyzed to list which rules needs to be fixed. The list of defects of the subjects of sample-1 to sample-6 are indicated in Table 5.

4.3. Modification Method

The pitch patterns are modified as accurate as possible to adapt the rules. The pitch patterns of a word were modified according to the following equation.

\[
f_0(t) = f_{\text{mean}} + \left(f_0(t) - f_{\text{mean}}\right) \times a + b \tag{3}
\]

\(f_0(t)\): pitch frequency pattern
\(a\): dynamic range modification factor
\(b\): peak adjustment factor

where \(f_{\text{mean}}\) denotes the mean value of the pitch patterns of a corresponding word. If \(a > 1\), then the pitch range is
and additionally with those pitch patterns are re-synthesized by STRAIGHT [7].

The results of the listening experiment are shown in Table 7, where S indicates an answer that supports the modified speeches, N indicates one that does not support the modified speeches, and f indicates that the subject could not distinguish between the contrasting speech samples.

Table 7 shows that the modified speeches sounded more natural to the native English speakers than the original versions. This was true for all six samples, and averaging, over threefold subjects support the proposed modification against the original speeches. Therefore, our approach is considered to be practically verified.

5.4. Discussion

The results supported the following generalization. Our results confirm the knowledge indicated in previous studies that Japanese English tends to utter an important word (i.e., content word) non-emphatically [1]. In addition, our results confirm that the irregularities occur on a noun, adjective, interrogative and negative in Japanese English.

Some issues toward practical use of this approach need further investigations. One of them is to develop an automatic algorithm for determining optimal values of a and b in equation in section 4.3, which are used for modifying the pitch patterns in Japanese English.

6. Conclusion

This paper has described the prosodic difference between Japanese English and English speakers. As mentioned in the introduction, previous general studies on related themes adopted a bottom-up approach for extracting the difference. However, the results found in those studies was often found to be inaccurate when confirmed by re-synthesis of speech. In this study, the analysis was conducted by focusing on prominence. In addition, the effect of extracted prosodic features was verified by a listening test using speech re-synthesis. We thought this framework made it possible to discover meaningful prosodic features of Japanese English.

7. References


