interspeech 2008 incorporating SST 08

A Study of Pitch Patterns of Japanese English Analyzed Via Comparative Linguistic Features of English and Japanese

Tomoko Nariai [1], *Kazuyo Tanaka* [1]

[1] Graduate school of Library, Information and Media Studies, University of Tsukuba, Japan

{nariai, ktanaka}@slis.tsukuba.ac.jp

Abstract

Certain defects in utterances of a word or phrase occur in English as spoken by Japanese native subjects, referred to in this article as Japanese English. This study considers such prosodic feature patterns as one of the most common causes of deficiencies in Japanese English. Japanese English is linguistically supposed to have the phonetic characteristics of the native language, Japanese. This supposition leads to the hypothesis that pitch patterns of Japanese English can be interpreted from the point of view of comparative linguistic features of English and Japanese, and that Japanese English would have better (i.e. closer to English) prosodic patterns if its particular characteristics were removed. In this study, the hypothesis was acoustically examined by means of a synthesis-by-analysis system, STRAIGHT, and then tested by listening experiments. Results of the experiments indicate practical verification of the hypothesis.

Index Terms: speaking foreign language, prosodic features, native language, defectiveness, synthesis-by-analysis system

10.21437/Interspeech.2008-239

1. Introduction

English is studied as a second language in junior high school or high school in Japan, and we Japanese study hard to master English during our schooldays. However, we often have difficulty in making ourselves understood in English when we actually talk to a native English speaker. There are certain differences in utterances between native speakers of English and ourselves. Thus we have the phenomenon known as *Japanese English*.

This has been a matter of some concern to people involved in studying ways to improve Japanese English. The number of computer-based studies of Japanese English has increased markedly over the last decade [1]. Recent research on Japanese English has produced some findings about its characteristics via statistical analysis, but few studies so far have come up with concrete proposals for improving it. Also many studies [2] have shown that a major cause of defects in Japanese English is the occurrence of phonemes in English utterances that are unfamiliar to Japanese.

On the other hand, few studies have considered prosodic patterns unfamiliar to Japanese, which we consider also to be one of the most common causes of defects in Japanese English. This is because delicate handling of the relation between pitch and stress is required when considering prosodic patterns of English, linguistically classified as a stressed language. We consider a stressed syllable in a word or a stressed word in a sentence in English to be realized via a dynamic pitch range. Statistical analysis of thousands of samples of Japanese English has revealed that the dynamic range of pitch patterns in Japanese English is narrow, compared with the range of native speakers [1]. According to this finding, we tried to widen the dynamic range of pitch patterns in several samples of Japanese English by means of a speech synthesizer, but this type of modification cannot cover the gap in pitch between Japanese English and the English of native speakers. There are definite differences in pitch patterns between Japanese English and the English of native speakers.

Japanese English is linguistically supposed to have the phonetic peculiarities of Japanese, particularly when the contents of foreign language sentences are unfamiliar to the speaker. Our speaking English with faint or strong Japanese-based pitch patterns is supposed to cause the defects in Japanese English. Although several characteristics of Japanese English have been identified over many years [1][2], none have been confirmed as yet in terms of actual speech modification. Therefore this study focuses on the pitch patterns of Japanese English, analyzing their concrete acoustic features, with the expectation that this analysis will contribute to developments in educational systems of language learning in future studies.

In this paper, we first summarize the particular phonetic characteristics of English and Japanese respectively, inferred from comparison of the linguistic features of the two languages. Then the Japanese-based pitch patterns in Japanese English are predicted. Focus and accented phrases are defined to supplement the phonetic characteristics of English. Next, we create several rules for modifying pitch patterns for evaluation tests, based on the hypothesis that the pitch patterns of Japanese English will become more natural and fluent if the predicted peculiarities due to Japanese are removed. Some samples of Japanese English are analyzed and pitch patterns of these samples are modified on the basis of the above rules, and then synthesized utilizing a synthesis-by-analysis system STRAIGHT [3]. Evaluation tests for these attempts to improve the pitch patterns are carried out by listening experiments.

Concrete procedures, experimental conditions and evaluation results are given in the following sections.

2. Phonetic characteristics of English and Japanese

There are marked differences in pronunciation between English and Japanese. One major difference is that a successive sequence of consonants is usual in English, but in Japanese a consonant basically followed by a vowel. Also, English is a stressed language and Japanese is a tonal language. Here we consider that stress is realized in terms of width of the pitch range and tone is realized in terms of high or low pitch, thus resulting in different pitch patterns for each language.

Phonetic characteristics of the two languages are defined in terms of the phonetics of English and Japanese as follows.

(1) *In English*, there are four or less levels of stress in a word. The syllable with primary stress in a word is indicated by a wider range in pitch [4].
In Japanese, there are two or less levels of tones in a word. In a word or a word with a postpositional particle of Japanese, there are four main types of sequence of tones: all-low, low-high, low-high-low or high-low sequences of tones [5].

(2) *In English*, a diphthong in a word is uttered with one syllable, the first vowel of which is indicated by a wider range of pitch than that of the second vowel [4].
In Japanese, a vowel, excluding a semi-vowel in the contracted sound, is uttered with one mora [5].

(3) *In English*, there are differences in pitch ranges between content words and function words in a sentence: the former have a wide range and the latter have a narrow range [6].
In Japanese, tones of content words in a sentence are determined by phonetics, but those of function words are determined by the tone preceding them [7].

(4) *In English*, prominence is given to some words in a sentence in order to express which word contains the most important meaning of the sentence, such prominences are indicated by a wide range in pitch and a high pitch [8, 9].
In Japanese, prominence is often given to some words in a sentence, and those prominences are indicated by a slightly wider range in pitch, but the tonal patterns of individual words in a sentence also strongly affect sentence prominences [7].

(5) *In English*, a sentence is uttered with some phrases, which correspond to the lexical or phonetic units, of which the sentence is composed. Each end of phrase is indicated by a decline in pitch, as is stated to many as a phenomenon in English utterance [9, 10, 11].
In Japanese, the same is true, but the tonal patterns of each word in a sentence also strongly affect sentence phrasing in Japanese [7].

3. Pitch ranges and pitch heights in real speech samples of Japanese English and native English

In order to preliminarily investigate acoustic phenomena described in (3), (4) and (5) in section 2, the pitch ranges and the peak pitch values in individual words of Japanese English and the English of native speakers were analyzed for comparison. Here, the pitch range is defined as:

(pitch range) = (maximum pitch value in the word)/ (minimum pitch value in the word)

Twenty declarative English sentences were chosen from the MOCHA-TIMIT sentence set. Speakers for the comparison consisted of 16 Japanese (8 males and 8 females) and 10 English natives (5 males and 5 females). The national origins of native speakers were UK, USA, Canada, Australia and New Zealand. 8 Japanese uttered 10 sentences and other 8 Japanese uttered another 10 sentences. Utterance samples were manually divided into words by checking the waveforms and listening. Then the pitch ranges and the peak pitch values in individual words were obtained.

Results are summarized as follows:

(a) In the English of native speaker, there are differences in pitch range between content words and function words, but in Japanese English such differences are ambiguous.

(b) In the English of native speaker, almost all subjects give prominence to the same word, but in Japanese English, prominence are realized irrelevant word for sentence structure or not realized at all.

(c) In the English of native speaker, each sentence is clearly phrased by pitch height and the pitch peaks in latter phrases are lower than those of the foregoing phrases, but in Japanese English, such phrase are not realized.

Fig. 1 shows the relative values of the pitch ranges for words of 8 Japanese English and those of the 10 English of native speakers, for the utterance "*I gave them several choices and let them set the priorities*", and Fig. 2 shows the relative values of the peak pitch in each word for the same sample set.

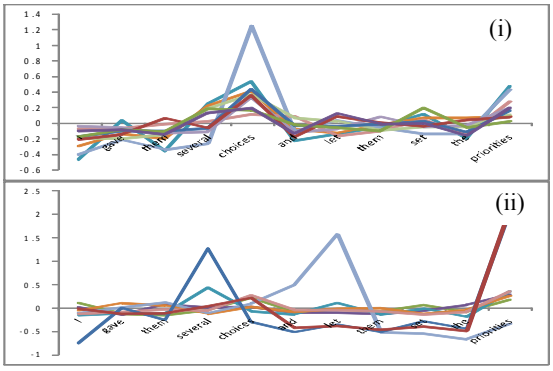

Fig. 1 *Variation patterns of the pitch range for Native English* (i) *and Japanese English* (ii). *Each value for a word is subtracted by the average value of all words contained in a phrase.*

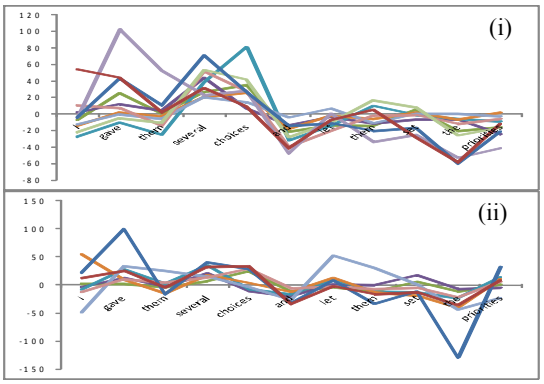

Fig. 2 *Variation patterns of the peak pitch value in each word section for Native English* (i) *and Japanese English* (ii). *Each value for a word is subtracted by the average value of all words contained in a sentence.*

4. Predictable pitch patterns of Japanese English

4.1. Predictable Japanese-based pitch patterns of Japanese English

Japanese-based pitch patterns in Japanese English are presumed, based on the phonetic characteristics defined in section 2 to satisfy the following 5 conditions, corresponding to the enumeration of section 2.

(1) In Japanese English, each syllable of a word is uttered in a high or low tone, without considering the stress level of each syllable or the falling pitch of the stressed syllable.

(2) In Japanese English, diphthongs, each with two syllables, are uttered in a high-low or low-high tone sequence.

(3) In Japanese English, there is a tendency for the pitch range of content words and function words to be same.

(4) In Japanese English, the prominence is inappropriately performed.

(5) In Japanese English, there is a tendency for the pitch peaks of content words and function words are to be almost same and the phrasing of utterances performed by the pitch pattern is inappropriate.

4.2. Focus and accented phrase

In dealing with the rules of sentence prominence and sentence phrasing of English, described in (4) and (5) above, different approaches have been taken in linguistics [8, 12]. So a definitive characterization has not yet been provided.

This study defines sentence prominence as a focus and sentence phrasing in terms of an accented phrase. Sentence prominences in English are defined by two foci: the first and second focus. Accented phrases are also defined in terms of the pitch of focus, as follows.

[*First Focus*]

A study of English phonetics [4] reveals that a English sentence is arranged in order of the importance of information of each word. The most informative word in a sentence is put at the end of a sentence on the basis of the End-Focus Principle of English structure. So this study defines the first focus as the ends of phrases, clauses or sentences. The prominence corresponding to the first focus is indicated by the wider range in pitch than that of non-focus.

[*Second Focus*]

A study of English pragmatics reveals that prominence is given to the word with the outstanding role in the information structure of a sentence. The information structure measures how much additional information the word provides to the listener. From the concept of activation of a discourse reference in the study [6, 8], this study defines the second focus as a noun, an adjective, an adverb or a quantifier. The prominence corresponding to the second focus is indicated by the wider range in pitch than that of non-focus.

[*Accent Phrase*]

A sentence is phrased by the lexical unit, e.g., the ends of phrases, clauses or sentences. The ends of the phrases are each realized by a decline in pitch. So this study defines the first focus as the end of an accent phrase and that the first focus is changed into the sharpest fall. If a word is defined as the second focus and also defined as the first focus, then the sentence is phrased on that word. And if there are two first foci in a sentence, the pitch height of the focus in latter phrase is changed to be lower than that of the foregoing phrase.

5. Hypotheses about methods to improve pitch patterns of Japanese English

The hypotheses that Japanese English has improved pitch patterns if the Japanese-based pitch patterns, defined in section 4, are removed can be stated as follows.

Japanese English will have improved pitch patterns if:

(1) The pitch pattern of the syllable with the primary stress in a word is changed to wider pitch range.

(2) The pitch pattern of the first vowel of a diphthong is changed to a wider range than that of the second vowel.

(3) The pitch patterns of function words in a sentence are made narrower, compared to the pitch patterns of content words.

(4) The pitch patterns of the first focus and second focus, defined in 4.2, are extended to a wider range than other words.

(5) The pitch peaks of function words are controlled to be lower than content words. The pitch patterns of words in a sentence are modified to form an accented phrase, defined in 4.2, at the end of which is the first focus. The first focus is changed to the sharp fall and the pitch peak of it is controlled to be higher than others.

6. Realization of hypotheses by speech synthesis

The hypotheses defined in section 5, were investigated utilizing the speech analysis-synthesis system STRAIGHT [3]. Test samples of Japanese English consisted of two sentences and each sentence was uttered once by two Japanese speakers.

The test speech samples were analyzed by STRAIGHT to extract power spectral envelopes and fundamental frequency patterns. The fundamental frequency patterns of the original speech samples were manually aligned with the word boundaries. (Henceforth, the fundamental frequency pattern is considered as equivalent to the pitch pattern.) They were modified according to hypotheses (1) to (5) described in section 5, and modified speech waves were synthesized.

Concrete procedures of acoustical realization for hypotheses (1) to (5) are as follows: For hypothesis (1), the accent syllable of each word is designated and widened its pitch range according to the following equation:

$$\tilde{f}_0(t) = f_{mean} + sign[f_0(t) - f_{mean}]\,|f_0(t) - f_{mean}|\cdot a \qquad \text{(i)}$$

where f_{mean} denotes a mean value of the pitch frequency pattern in each corresponding syllable (or word), and a is a parameter for amplification (if a >1, then pitch range is amplified). For change of the diphthong in hypothesis (2), pitch expansion is carried out about the first half of each diphthong. For the pitch range of the function words in hypothesis (3), a = 0.1 to 0.6 were used. For the focus of the hypothesis (4), a =1.3 to 2.2 was used. For the hypothesis (5), to form the accent phrase, fixed value 30 Hz was substituted for $|f_0(t)-f_{mean}|$ and a down step of (-20) Hz was added for sharp fall in pitch.

Figs. 3 and 4 show the contrasting pitch patterns of test-sample: "*I gave them several choices and let them set the priorities*". Fig. 3 illustrates the pitch pattern of the original, in which the size of the pitch range does not suit an English rule, as presumed in section 4. Fig. 4 illustrates the pitch pattern of the modified synthesized speech of the same utterance, in which the pitch patterns of the first focus [*choices*] [*priorities*] were changed to a wider and sharp fall in pitch. The second focus [*several*] was assigned a wider pitch range than others. Function words [*I*] [*them*] [*and*] [*them*] [*the*] were assigned a narrower pitch range than others. Two accent phrases were identified, and at the end of each, the first focus was indicated by a sharpest fall in pitch.

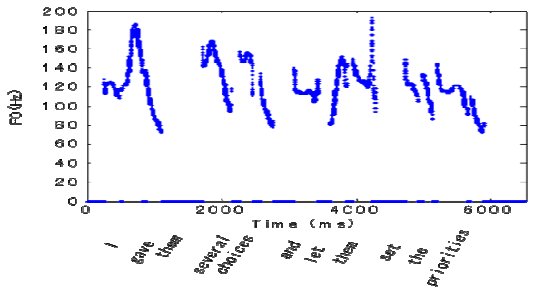

Fig. 3 *Pitch pattern of the original speech of a test sample.*

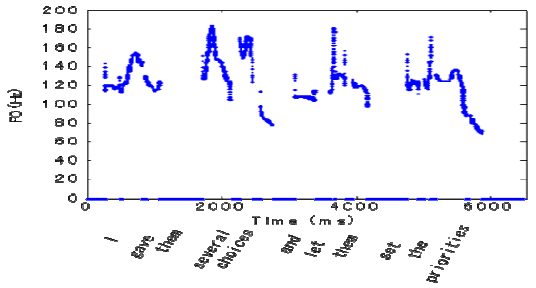

Fig. 4 *Pitch pattern of synthesized speech modified from the sample in Fig.3.*

7. Experiment for verification of the hypotheses

7.1 Experimental conditions of the listening test

The hypotheses were verified by an evaluation test, in which a pair of contrasting speech samples, the original one and its modification, i.e., those shown in Fig.3 and Fig.4, were presented randomly to subjects. The subjects were basically requested to answer which one sounded better as an English utterance.

The listening test was carried out in a quiet room. The subjects were requested to listen at least twice to each speech sample, and to answer the following two questions: whether the subject could catch the difference in pitch patterns of the two samples, and which sample had more natural pitch patterns in English.

The subjects were 7 persons from countries where English is usually spoken, e.g. America, New Zealand, Canada and Philippines. Most were foreign students attending the University of Tsukuba.

7.2 Result

The results of the evaluation test are shown in Table 1, where *S* indicates an answer that supported the hypotheses, *R* indicates one that did not support the hypotheses, and *I* means that the subject was not able to differentiate between the contrasting speech samples. As indicated by the results, 8 out of 14 answers supported the hypothesis, one answer did not support it, and five answers out of 14 were inconclusive.

Results of the experiments show that this study is an appropriate way to investigate the pitch patterns of Japanese English.

Table 1. *The results of the experiments*
(a-g: subjects; S:support; R:reject; I:inconclusive answer)

Subject name	a	b	c	d	e	f	g
Test-sample1	S	S	I	S	R	S	I
Test-sample2	S	S	S	S	I	I	I

8. Conclusions

The pitch patterns of Japanese English were characterized by considering comparative linguistic features of English and Japanese. The hypothesis that Japanese English would have improved pitch patterns if presumed peculiarities were removed was tested by listening experiments using synthesized speech. The result of the experiments almost confirmed the hypothesis. The experiments need more samples and subjects, and also need to be discussed in more detail in a future study.

In the last we thank Prof. Hideki Kawahara, Univ. Wakayama, for kindly licensing STRAIGHT.

9. References

[1] H. Obari, R. Tomiyama, M. Yamamoto, S. Itahashi, " Differentiation of English utterances of Japanese and native speakers by several prosodic parameters," Proc. Oriental COCOSDA-2005, pp.143-147. Jakarta Indonesia, 2005.

[2] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, "Development of English speech database read by Japanese to support CALL research," Proc. Int. Cong. Acoustics (ICA'2004), pp.557-560, 2004.

[3] H. Kawahara, I. Masuda-Katsuse, A. Cheveigné, "Re-structuring speech representations using pitch adaptive pitch frequency smoothing and instantaneous-frequency-based F0 extraction," Inter. J. of Speech Communication, Vol.27, No.3-4, pp.187-207, 1999.

[4] I. Yasui, " Phonetics," (in Japanese), Kitakusha, 1995.

[5] M. Sugito, "Accent Intonation, rhythm and pause," (in Japanese), *Rhisym and Pause*, Sanshodo, 1997.

[6] S. Takebayashi, "English Phonetics," (in Japanese), Kenkyusha, 1996.

[7] M. Sugito, "English Spoken by Japanese," (in Japanese), Izumishoin, 1996.

[8] K. Lambrecht, "Information Structure and Sentence Form; topic, focus and mental representation of discourse referents," Cambridge Studies in Linguistics 71, Cambridge University Press, 1994.

[9] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," JASA, 89(4), pp.1768-1776, 1991.

[10] D.R Ladd, "Declination reset and the hierarchical organization of utterance," JASA, Vol. 84(2), pp.530-544, 1988.

[11] N. Erteschik-Shir, "The dynamics of focus structure," Cambridge Studies in Linguistics 84, Cambridge University Press, 1997.

[12] J. Pierrehumbert, "Synthesizing Intonation," JASA. Vol.102(1), pp. 985-955, 1981.