筑波学院大学紀要第14集　77 ～ 85 ページ　　2019年

Perceptual differences of the fundamental frequency at the sentence level between native English speakers and native Japanese speakers

Tomoko Nariai *

Abstract

Certain differences are often recognized in utterances between native speakers of English and Japanese speakers of English. In this paper, perceptual differences in the change in a fundamental frequency (F0, henceforth) between native speakers and Japanese speakers of English will be investigated. A hypothesis was created to improve the F0 of Japanese speakers of English so that they will be more natural for native speakers of English. It was based on predictions induced by linguistic differences in the F0 between English and Japanese. The hypothesis was acoustically examined by means of synthesis-by-analysis software. Synthesized speech was evaluated in a listening test taken by native English speakers and native Japanese speakers. The results from both the native English speakers and native Japanese speakers almost always provided positive support of the hypothesis. These results indicated practical verification of the hypothesis. In addition, the results also showed that it sounded more natural for native speakers of English to decline in the F0 only at the end of accent phrases, whereas for native Japanese speakers to decline in the F0 not only at the end of accent phrases but also on some words in sentences in English.

Key words: second language learning, speech analysis, speech synthesis, speech modification, voice conversion

1. Introduction

Many Japanese have a good command of written English after having studied for the university entrance examination, but are often confused when trying to in make themselves understood when they actually talk to native English speakers. There are certain differences in utterances between native speakers of English and Japanese speakers of English.

The number of computer-based studies of Japanese speakers of English has increased markedly over the last decade. Recent research on Japanese speakers of English has produced some findings about its characteristics via statistical analysis[1–3], but few studies so far have come up with concrete proposals for improving those findings; although several characteristics of Japanese speakers of English have been identified over many years, none have been confirmed as yet in terms of actual speech modification. Also many studies have shown that a major cause of

＊　筑波学院大学経営情報学部、Tsukuba Gakuin University

pronunciation problems in Japanese speakers of English is the prevalence of phonemes in English utterances that are unfamiliar to Japanese, but few studies have considered prosodic patterns unfamiliar to Japanese, which we consider also to be one of the most common causes of speech problems in Japanese speakers of English.

One reason for this might be the delicate handling of the relation between F0 and stress in English, which is linguistically classified as a stressed language. A stressed syllable in a word or a stressed word in a sentence in English is thought to be realized via a dynamic F0 range. As regards this, several statistical analyses of thousands of samples of Japanese speakers of English[1, 4] have revealed that the dynamic F0 range in Japanese speakers of English is narrow compared with the range of native speakers. According to this finding, attempts were made to widen the dynamic F0 range in several samples of Japanese speakers of English by speech synthesizer, but this type of modification cannot cover the gap in F0 between Japanese speakers of English and the English of native speakers. There appear to be definite differences in F0 patterns between Japanese speakers of English and the English of native speakers.

Previous research in linguistic phonetics on this issue[5] has revealed that Japanese speakers of English are linguistically supposed to have phonetic characteristics peculiar to Japanese. According to this view, we assume that the Japanese-based F0 patterns can be estimated by comparison of particular phonetic features of English and Japanese, and the F0 patterns of Japanese speaking English will become more natural and fluent if the Japanese-based F0 patterns are removed from "Japanese English."

To verify this assumption, the particular phonetic characteristics of English and Japanese respectively were first summarized, inferred from a comparison of the linguistic features of the two languages. Then the Japanese-based F0 patterns in Japanese English were predicted using an analysis of sentence samples of English spoken by Japanese and native speakers. Thereafter several rules for modifying F0 patterns were created as hypotheses that improve Japanese speakers of English. The Japanese-based F0 patterns of several sentence speech samples were modified on the basis of the hypotheses and re-synthesized utilizing the synthesis-by-analysis software STRAIGHT[6].

In our preliminary experiment, several speech samples had become predictably distorted after the F0 patterns had been modified on the basis of the hypotheses. This distortion prevented proper evaluation of re-synthesized speech quality, so the actual effect of F0 pattern modification was unclear. To resolve this ambiguity, the experiment employed two ways of modifying the F0 patterns, that is, about half of the samples were modified based on the hypotheses and the others were modified based on some pseudo-hypotheses. Finally, these synthesized speech samples were tested using a listening experiment. Experimental results are discussed referring to differences between the two ways of modification.

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

2. Phonetic characteristics

In this section, linguistic differences between English and Japanese are examined. 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[7].

In Japanese, there are two or less levels of tone in a word. In a word, or a word with a postpositional particle of Japanese, there are four main types of sequences of tone: all-low, low-high, low-high-low or high-low[8].

(2) *In English*, a diphthong in a word is uttered as one syllable, the first vowel of which is indicated by a wider range in pitch than occurs for the second vowel[7].

In Japanese, a vowel, excluding a semi-vowel in the contracted sound, is uttered as one mora[8].

(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[9].

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[10].

(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 prominence is indicated by a wide range in pitch and a high pitch[11, 12].

In Japanese, prominence is often given to some words in a sentence, and such prominence is indicated by a slightly wider range in pitch. The tonal patterns of individual words in a sentence also strongly affect sentence prominence[10].

(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 a phrase is indicated by a decline in pitch. This is considered to be generally true for spoken English[12–14].

In Japanese, the same is true, but the tonal patterns of each word in a sentence also strongly affect sentence phrasing in Japanese[10].

3. Speech samples of native and Japanese speakers of 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 for the individual words of Japanese speakers of English and the English of native speakers were analyzed for comparison. Details were included in our previous study[15]. Results are summarized in relation to (3), (4) and (5), as follows:

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

(b) In the English of native speakers, almost all subjects give prominence to the same word, but in Japanese speakers of English, prominence is given for irrelevant words in certain sentences or not realized at all.

(c) In the English of native speakers, 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 speakers of English, such phrases are not utilized.

4. Presupposition

In this section, Japanese-based characteristics is predicted by linguistic differences in the F0 between English and Japanese.

4.1. Predictable Japanese-based Pitch Patterns of Japanese Speakers of English

Japanese-based characteristics for Japanese speakers of 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) With Japanese speakers of 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) With Japanese speakers of English, diphthongs, each with two syllables, are uttered in a high-low or low-high tone sequence.
(3) With Japanese speakers of English, there is a tendency for the pitch range of content words and function words to be same.
(4) With Japanese speakers of English, the words for prominence are often inappropriately selected.
(5) With Japanese speakers of English, there is a tendency for the pitch peaks of content words and function words to be almost the same and therefore the phrasing of utterances conforming to this pitch pattern is inappropriate.

4.2. Focus and Accent Phrases

In dealing with the rules of sentence prominence and sentence phrasing in English, described in (4) and (5) above, different approaches have been taken in linguistics[11, 17], so a definitive characterization has not yet been provided. Therefore this paper defines sentence prominence as focus and sentence phrasing in terms of an accented phrase. Sentence prominence in English is defined by two foci: the first and second focus. Accented phrases are also defined in terms of the pitch of the focus, as follows.

[*First Focus*]

A study of English phonetics[16] reveals that an English sentence is arranged in order of the End-Focus Principle, which put the most informative word in a sentence at the end of a sentence. Therefore, this paper defines the first focus as the ends of clauses or sentences. Prominence corresponding to the first focus is indicated by a wider range in pitch than that of a non-focus.

[*Second Focus*]

A study of English pragmatics reveals that prominence is given to a word with an outstanding role in the information structure of a sentence. The information structure measures how much additional information the word provides to the listener. Based on this study[11], this paper

defines the second focus as a noun, an adjective, or an interrogative. An instance of prominence corresponding to the second focus is indicated by a wider range in pitch than that of a non-focus.

[*Accent Phrase*]

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

5. Hypothesis

In this section, a hypothesis is created to improve the F0 of Japanese speakers of English so that they will be more natural for native speakers of English. It was based on predictions induced by linguistic differences in the F0 between English and Japanese. F0 pattern is considered to be the equivalent of the pitch pattern.

Then, Japanese speakers of English will have improved F0 patterns if:

(1) The F0 pattern of the syllable with the primary stress in a word is changed to a wider F0 range.

(2) The F0 pattern of the first vowel of a diphthong is changed to a wider range than that of the second vowel.

(3) The F0 patterns of function words in a sentence are made narrower, compared to the F0 patterns of content words.

(4) The F0 patterns of the first focus and second focus, defined in subsection **4.2**, are made wider than those of other words. Hence, the F0 patterns of verbs and adverbs are made wider or narrower according to the F0 patterns of function words and foci in the sentence.

(5) The F0 peaks of function words are modified to be lower than those of content words. The F0 patterns of words in a sentence are modified to form an accented phrase, defined in subsection **4.2**, at the end of which is the first focus. The first focus is changed to the sharp fall and its F0 peak is modified to be higher than others.

6. Realization of the Hypothesis by Speech Synthesis

In this section, the hypothesis was acoustically examined by means of synthesis-by-analysis software.

6. 1. Materials

Eleven subjects (5 male, 6 female), aged between 20 and 30, were chosen; each was a native Japanese speaker. Most were Japanese university students. Eleven sample sentences were chosen at random from the MOCHA-TIMIT sentence text set[18]. Those were timit010, 021, 022, 026, 027, 216, 246, 249, 259 and 452. Each subject was allocated a different sentence, which he or her uttered

once.

Subjects were required to utter a sentence repeatedly until the speech sample was recorded properly. There is one sample per one sentence for each subject.

6. 2. Synthesized Speech

As describe in the introduction, in our previous experiment[15], several speech samples had become distorted after the F0 patterns had been modified on the basis of the hypotheses. Subjects in the listening experiment are affected by these types of distortions and the actual effect of F0 pattern modification was unclear. To resolve this ambiguity, the present experiment employs two ways of modifying the F0 patterns of speech, that is, the six samples, timit009, 021, 022, 216, 246 and 452 were modified based on the hypotheses and the five samples, timit010, 026, 027, 249 and 259 were modified according to the pseudo-hypotheses. By employing this method, both samples of re-synthesized speech involve almost the same distortions, and the effect of the modifications can be discussed by observing the differences between the evaluation results for the two types of re-synthesized speech samples.

6. 3. Synthesized Speech Based on the Hypotheses

The test speech samples were analyzed by STRAIGHT to extract F0 patterns. The F0 patterns of the original speech samples were manually aligned with the word boundaries. They were modified according to hypotheses (1) to (5) described in section **5**, and modified speech waves were synthesized.

Although the F0 range of each word depends on its number of syllables, and some restrictions have to be added to the amount of modification allowed for the F0 pattern of the speech samples, the way of modifying speech is to change the F0 range of words, depicted as follows:

$(Function\ word) < (Verb,\ adverb) < (2^{nd}\ Focus) < (1^{st}\ Focus)$

Concrete procedures to realize hypotheses (1) to (5) acoustically were as follows:
For hypothesis (1), the accent syllable of each word was designated and its F0 range widened, according to the following equation:

$$\tilde{f0}(t) = f_{mean} + (f0(t) - f_{mean}) \cdot a$$

where f_{mean} denotes the mean value of the F0 frequency pattern in each corresponding syllable (or word), and a is a parameter for amplification (if $a > 1$, then the F0 range is amplified).

For changes in the diphthong according to hypothesis (2), F0 expansion was carried out for the first half of each diphthong. For the F0 range of the function words in hypothesis (3), values for a between 0.1 and 0.5 were used. For the focus of hypothesis (4), values for a between 1.1 and 1.5 were used. For the non-focus words, values for a between 0.5 and 1.1 were used. If the F0 for focused words went on rising in declarative sentences, values for a between 1.0 and 1.2 were used in order to comply with the accent phrase definition in section **4.2**. For hypothesis (5), to form the accent phrase, a F0 of 10 to 30 Hz was substituted for $|f_0(t) - f_{mean}|$ and a down step of (-20) to (-40) Hz was added to create the sharp fall in F0.

6. 4. Pseudo- synthesized Speech

The procedure for modifying speech was the same as explained in 6.3; however, the pseudo-hypotheses was for the F0 range to be ordered as follows:

(*Function*)<(*Verb, Adverb*)<(*2 nd Focus*)<(*1 st Focus, Pseudo_Focus*)

7. Experiments for Testing the Hypotheses

The synthesized speech is evaluated in a listening test.

7. 1. Subjects

There were two groups of subjects. The first group consisted of 18 (3 male, 15 female) native English speakers, aged between 19 and 50. Most were undergraduate or graduate students in Michigan. The second group consisted of 20 (11 male, 9 female) native Japanese speakers, aged between 20 and 40. All were undergraduate and graduate students, not majoring in linguistics, phonetics or acoustics.

7. 2. Procedure

The hypotheses were examined using an evaluation test, in which a pair of contrasting speech samples, the original one and its modification were presented randomly to subjects.

The listening test was carried out in a quiet room. The subjects were requested to listen to each speech sample, and to answer the following question "Which sample of the two had more natural F0 patterns in English". The subjects were instructed to answer "*N*" if they could not catch the difference in F0 patterns of the two samples or could not decide which should be chosen.

8. Results and Discussions

8.1. Results

The results of the evaluation test are shown in Table 1, where "○" indicates the sum of the answer that supports the hypotheses, "×" indicates the one that does not support the hypotheses, and "*N*" indicates that the subject could not distinguish between the contrasting speech samples.

Table 1 for native speakers of English suggests that 54 out of 86 support the hypothesis, whereas 44 out of 64 of them reject the pseudo-hypotheses. In contrast, Table 1 for Japanese speakers suggests that 61 out of 103 support the hypothesis, and 51 out of 75 of them support the pseudo-hypotheses.

The results for the hypothesis in Table 1 show the sentences modified according to the hypotheses sounded more natural to both the native speakers of English and the Japanese than the original versions.

The results for the pseudo- hypotheses in Table 1 show the sentences modified according to the pseudo-hypotheses sounded less natural to the native speakers of English than the original versions. On the contrary, this was not the case for the Japanese; the modified versions sounded more natural.

Table 1 Results of the listening experiment for the modification based on the hypotheses and pseudo-hypotheses. native speakers of English (ntv) and native speakers of Japanese (jpe), where (○: support, × : reject, N: inconclusive answer).

	ntv			jpe		
	○	×	(N)	○	×	(N)
Hypothesis	**65 (75.6%)**	21	(22)	**61 (59.2%)**	42	(17)
Pseudo-hypothesis	20	**44 (68.8%)**	(26)	**51 (68%)**	24	(25)

8. 2. Discussions

This study hypothesizes that Japanese spoken English would have improved F0 patterns if Japanese-based prosodic features of Japanese speakers of English were removed from it. The hypotheses were examined by creating two types of synthesized speech based on the hypotheses and pseudo-hypotheses. Pseudo-hypotheses that did not reflect the hypothesis was also created. The main difference between the hypothesis and pseudo-hypotheses lie in the F0 at the ends of accent phrases: if only one of them declines, then it conforms to the hypothesis; if not, then it comes under the pseudo-hypotheses.

The results in Table 1 showed that for native English speakers, it sounded more natural to decline in F0 only at the ends of accent phrases. Therefore, the hypotheses were practically verified.

On the contrary, for native Japanese speakers, it sounded relatively natural to decline in F0 not only at the ends of accent phrases but also on some words in a sentence. Those words were nouns in this experiment, so that this can be read to experimentally prove an indication described in the study by Sugito et al.[10] ; namely, that Japanese speakers of English tend to utter each word clearly. Also, it sounds relatively natural for native Japanese speakers that the ranges of F0 patterns were widened for words other than focus.

Through the experiment, distinctions between the two contrasted speeches were not definite in several samples, so that there were a lot of answers indicating that the subject could not decide which to choose. However, employing two types of hypotheses and observing differences between the results for those hypotheses, it could be clarified which hypotheses were appropriate.

One reason for the ambiguity in contrasting re-synthesized speech samples is that the modification was made only for F0 patterns in this experiment. The effect of integration of prosodic and articulatory features on this issue needs to be discussed in a future study.

9. Conclusion

This study hypothesizes that Japanese spoken English would have improved F0 patterns if Japanese-based prosodic features of Japanese speakers of English were removed from it. The hypotheses were examined by creating two types of synthesized speech based on the hypotheses and pseudo-hypotheses. These synthesized speech samples were then tested in

listening experiments with native speakers of English and native Japanese speakers. The listening experiments show that results for the native speakers of English support the hypotheses but do not support the pseudo version. Results of the experiments indicated practical verification of the hypotheses. On the other hand, results for the native Japanese speakers were inconclusive for the experiment testing the hypotheses and pseudo-hypothesis. It suggested that it sounded more natural for native Japanese speakers to decline in F0 not only at the ends of accent phrases but also some word in a sentence in English.

10. References

1) H. Obari, R. Tomiyama, M. Yamamoto and S. Itahashi, "Differentiation of English utterances of Japanese and native speakers by several prosodic parameters," *Proceedings of Oriental COCOSDA,* pp.143–147. 2005.
2) N. Minematsu, Y. Tomiyama, K. Yoshimoto and K. Shimizu, S. Nakagawa, M. Dantsuji, "Development of English speech database read by Japanese to support CALL research," *Proceedings of International Congress of Acoustics*, pp.557–560, 2004.
3) Y. Mori, "The initial high pitch in English sentence produced by Japanese speakers," *English Linguistics*, 22, 23–55 (2005).
4) T. Nariai, Y. Nanbu and K. Tanaka, "A study of the relationship between sentence structure and pitch pattern of Japanese speaking English," *Proceedings of Spring Meeting of Acoustical Society of Japan,* pp.413–414, 2009.
5) Y. Takehuta, *Science of Japanese speaking English*, (Kenkyusha, Tokyo, 1982).
6) H. Kawahara, I. Masuda-Katsuse and A. Cheveigné, "Re-structuring speech representations using pitch adaptive pitch frequency smoothing and instantaneous-frequency-based F0 extraction," *Speech Commun.,* 27, 187–207 (1999).
7) I. Yasui, *Phonetics*, (Kaitakusha, Tokyo, 1995).
8) M. Sugito, *Accent, Intonation, Rhythm and Pause*, (Sanshodo, Tokyo, 1997).
9) S. Takebayashi, *English Phonetics*, (Kenkyusha, Tokyo, 1996).
10) M. Sugito, *English Spoken by Japanese*, (Izumishoin, Tokyo, 1996).
11) K. Lambrecht, *Information Structure and Sentence Form; Topic, Focus and Mental Representation of Discourse Referents*, (Cambridge University Press, 1994).
12) J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *Journal of Acoustical Society of America.* 89 (4), 1768–1776 (1991).
13) J.B. Pierrehumbert and M.E. Beckman, *Japanese Tone structure*, (MA: MIT Press, 1988).
14) D.R. Ladd, "Declination reset and the hierarchical organization of utterance," *Journal of the Acoustical Society of America,* 84 (2), 530–544 (1988).
15) T. Nariai and K. Tanaka, "A study of pitch patterns of Japanese English analyzed via comparative linguistic features of English and Japanese," *Proceedings of InterSpeech*, pp.776–779 (2008).
16) S. Greenbaum and R. Quirk, *A Student's Grammar of the English Language*, (Longman, London and New York, 1990).
17) N.E. Shir, *The dynamics of focus structure*, (Cambridge University Press, 1998).
18) http://data.cstr.ed.ac.uk/mocha/mocha-timit.txt (May 1, 2009)