Japanese Repetition of Normal and Prosodically-Modified English Words

Anthony O. Okobi † *& Keikichi Hirose* ‡

† Harvard-MIT Division of Health Sciences and Technology, MIT, USA
‡ Graduate School of Frontier Sciences, University of Tokyo, Japan
† tokobi@mit.edu ‡ hirose@gavo.t.u-tokyo.ac.jp

Abstract

Native speakers of Japanese were asked to repeat 2 and 3-syllable English words, with varying lexical stress placements, spoken by a native speaker of American English. One group of words consisted of unaltered sonorant and semi-sonorant nouns, while the second group contained the same words, but with artificially flattened fundamental frequency (F0) and intensity contours. Speakers' accuracy in producing the prosodic characteristics of the English words was determined acoustically by analyzing the F0 contour, intensity contour, and syllable-duration of their utterances. Production of the correct prosodic characteristics was affected by the number of syllables and place of the lexical accent, as well as by the individual subject's level of familiarity and understanding of the word. Furthermore, the results show that subjects' accuracy was influenced by the prosodic modification of the words.

1. Introduction

Both Japanese and English have tonal properties that are used in the production of lexical accent. According to Beckman and Pierrehumbert (1986), the precise nature of accent is not identical in the two languages, but both languages share an underlying characteristic of association between some well defined prominence and the syllable of a word that, by virtue of the association, is 'accent'. Suprasegmental characteristics of a word, such as fundamental frequency (F0) and intensity contours, as well as syllable durations, give rise to the lexical prosodic prominence.

The accented syllable of a word generally displays a F0 prominence, more intensity and longer duration than a similar unaccented syllable of another or of the same word (Fry, 1955). In examining homographs, Lieberman (1960) found that in no case did the stressed syllable have both a lower intensity and a lower F0 than the unstressed syllable. In fact, there seemed to exist a "trading" relationship between F0 and intensity.

2. Experimental Paradigm

The purpose of this study was to compare the accuracy of native speakers of Japanese in repeating the prosodic characteristics of normal English words with their accuracy in producing the normal prosodic characteristics given that the words are prosodically-modified. A 28 year old male native speaker of American English was chosen as the target speaker. Stimuli were delivered to subjects using headphones and all recordings were made in a sound-proof room using a dynamic microphone. Initial sampling was done at 44.1kHz, with resampling being done at 11kHz for speech analysis.

2.1. Subjects

A total of 14 subjects were involved in this study. All were male native speakers of Japanese from the University of Tokyo. The average age of the subjects was 24 years and all had about 11 years of experience studying English.

2.2. Hearing Evaluation

An auditory discrimination task, Wepman test (1958), was used to determine how well each subject was able to distinguish between different English phonemes. The subjects were asked to distinguish between words that differed by one phoneme by listening to the sound of the words and indicating if the pair sounded the same or different. Performance was scored based on a level expected for a 6-year old native speaker of American English.

Subjects were also given a lexical stress discrimination test to determine their ability to distinguish between different lexical stress placements. Two and 3-syllable English homographs with similar vowel qualities, but with differing stress patterns such as *sn* and *ns*, for 2-syllable words and *snn*, *nsn*, and *n***ns*, for 3-syllable words (where *s* indicates the stressed syllable carrying the primary lexical accent, *n* indicates a unstressed syllable carrying no lexical accent, and *n** indicates a weak stressed syllable possibly carrying the secondary lexical accent) were low-pass filtered at 350Hz and presented to the subjects. Although the subjects were unable to hear the higher formant frequencies, they had access to information about the syllable duration, the intensity contour and the fundamental frequency.

2.3. Word Repetition

Thirty sonorant and semi-sonorant English concrete nouns were chosen for this study. The words were 2 and 3-syllables in length and of various stress patterns, *sn* and *ns* for 2-syllable words and *snn*, *nsn*, and *n***ns* for 3-syllable words. For semi-sonorant words, only the first phoneme was allowed to be non-sonorant due to the difficulty in finding all sonorant English words with varying lexical accent positions and of varying familiarity to native Japanese speakers. Table 1 lists the words according to their stress patterns.

Table 1: *Stress pattern of words used in Word Repetition.*

sn	***ns***	***snn***	***nsn***	***n*ns***
Army	Alarm	Animal	Arena	Alienee
Cherry	Balloon	Cinema	Aroma	Cannoneer
Dinner	Canoe	Cinnamon	Banana	Millionaire
Pollen	Cologne	Dowery	Hyena	Nominee
Sirloin	Marine	Mineral	Tiara	Pioneer
Vermin	Saloon	Oriole	Vanilla	Violin

2.3.1. Normal Words

Subjects were asked to repeat 3 words, 7 times, from each stress pattern group, for a total of 105 repetitions. The target speaker was also asked to listen to his own utterances and repeat each word 7 times. Information obtained by analyzing his productions was used to construct criteria utilized subsequently in determining the accuracy of the native Japanese speakers.

2.3.2. Prosodically-Modified Words

The same words uttered by the target speaker were manipulated using the acoustics and phonetics software, Praat, to artificially flatten F0 and intensity contours, while keeping higher frequency and duration information intact. Thus the words could be understood, but sounded monotone. The subjects were then asked to follow the procedure outlined in section 2.3.1. However, the 15 words were a different set from the words they repeated in the normal word repetition experiment. Figure 1 shows subject 2's attempt to correctly produce prosodic characteristics of the modified word, Nominee.

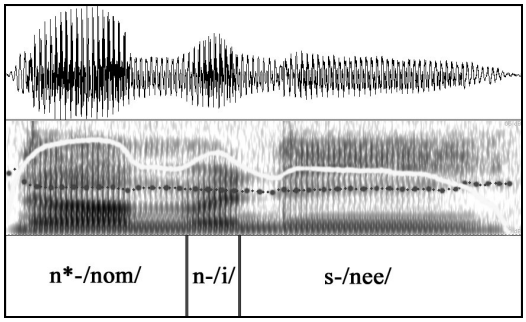

Figure 1: *Third repetition of prosodically-modified Nominee by subject 4. The F0 contour (dotted black line) and the intensity contour (solid white line) overlay the spectrogram of the utterance. The wave form is above, while the syllable boundaries are represented on the bottom.*

2.4. Familiarity and Understanding Survey

At the conclusion of the word repetitions, the subjects' prior experiences with the words were determined by means of a survey. The first part of the survey required the subjects to indicate, from 0 to 2, how familiar they were with the sound of the words. The second part of the survey required the subject to read the 30 words and indicate, from 0 to 2, their level of understanding of the meaning of the words (Okobi and Hirose, 2003).

3. Data Extraction

3.1. Fundamental Frequency Contour

An autocorrelation method developed by Paul Boersma (1993) was used to estimate a signal's short-term autocorrelation function on the basis of a windowed signal. Using a Hanning window, the function is given by equations 1 and 2.

$$r_x(\tau) \approx \frac{\mathrm{r}_{\mathrm{xw}}(\tau)}{\mathrm{r}_{\mathrm{w}}(\tau)} \tag{1}$$

where normalized autocorrelation of a Hanning window is

$$r_W(\tau) = \left(1 - \frac{|\tau|}{T}\right)\left(\frac{2}{3} + \frac{1}{3}\cos\left(\frac{2\pi\tau}{T}\right)\right) + \frac{1}{2\pi}\sin\left(\frac{2\pi|\tau|}{T}\right) \tag{2}$$

and

$$T = \frac{3}{(\min F0)} \tag{3}$$

The windowed signal is $\mathrm{r}_{\mathrm{xw}}(\tau)$, where τ represents the lag and T is the length of the Hanning window.

3.2. Intensity Contour

The extraction of the intensity values in the acoustic signal involved convolving the signal with a Kaiser-20 window (sidelobes below -190dB). The effective length of the Kaiser-20 window was 3.2 divided by the minimum F0 (75Hz). Since the intensity of sound in air relative to the human auditory threshold, dB-SPL, is defined as

$$\text{dB-SPL} = 10\log_{10}\left(\frac{1}{TP_0^{\,2}}\int x^2(t)dt\right) \tag{4}$$

where x(t) is the sound pressure in units of Pascal (Pa) at time t, T represents the duration of the acoustic signal, and P_0 at 0.00002 Pa is the human auditory threshold. In order to represent the intensity relative to the human auditory threshold, dB-SPL, the intensity was computed as

$$\text{dB-SPL} \approx 10\log_{10}\left(\frac{1}{nP_0^{\,2}}\sum_{i=1..n} x_i^{\,2}\right) \tag{5}$$

where n is the number of samples.

4. Results

The subjects tested were all able to distinguish between English phonemes used in the auditory discrimination task, as well as between the different lexical stress placements. Analysis and quantitative measurements were conducted on

the F0 contour, as well as the intensity contour of the words uttered by each subject. Comparative measurements were also made on duration differences between the syllables of each word. Criteria used to determine accuracy were mostly derived from analysis of the target speaker's utterances and English linguistic rules, as well as from other related studies (Fujisaki *et al.*, 1986).

4.1. Fundamental Frequency Prominence Placement

Results obtained by using the error detection method proposed by Okobi and Hirose (2003) indicated that subject accuracy in reproducing the F0 contour of normal English words was, in general, greater for 2-syllable words then for 3-syllable words (Figure 2). Furthermore, the percentage of accurately produced F0 contours for 3-syllable words, that the subjects were familiar with and understood (Fam-Und rating of 22), was higher than for unfamiliar and unknown words (Fam-Und ratings of 00). This effect, however, was not observed for 2-syllable (Figure 2).

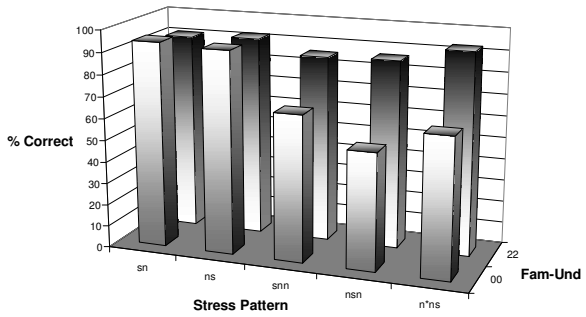

Figure 2: *Percentage of correctly reproduced F0 contours of normal English words.*

In the case of prosodically-modified words, subjects' accuracy in producing the correct F0 contours was greater for words with Fam-Und rating of 22 then for words rated as 00. Furthermore, 3-syllable prosodically-modified words with *nsn* and *n*ns* lexical stress patterns had the lowest percentage in accuracy. Within both the 2-syllable and the 3-syllable prosodically-modified groups, the last syllable carrying the lexical stress had the lowest percent correct for words with Fam-Und ratings of 22, as well as for words with ratings of 00 (Figure 3).

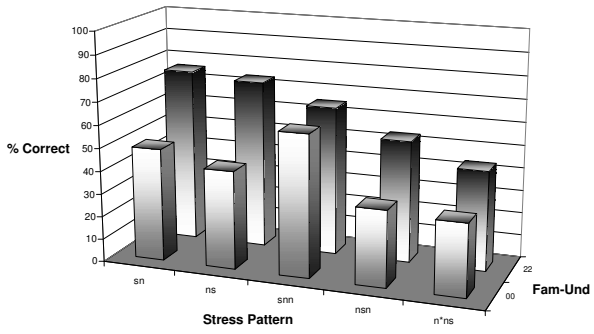

Figure 3: *Percentage of correctly reproduced F0 contours of prosodically-modified English words.*

4.2. Intensity Prominence Placement

An overall higher percentage of correct stress placement was observed for the production of the intensity contour of normal 2-syllable words, when compared to the 3-syllable words. Although subject accuracy in producing the correct intensity contour of normal words was very high, for all the stress patterns and for both Fam-Und ratings of 22 and 00, words with stress patterns *nsn* and *n*ns* had slightly lower percentage of correctly produced intensity contours (Figure 4).

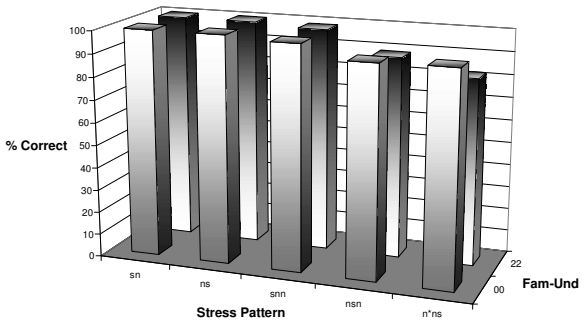

Figure 4: *Percentage of correctly reproduced intensity contours of normal English words.*

Results from the prosodically-modified word repetition experiment show the same effect of location of the lexical accent on accuracy of production of the correct intensity contour as was observed for the F0 contour production of modified words. The percent correct decreased as the lexical accent placement moved from the first syllable to the last syllable, with 3-syllable words with stress pattern *n*ns* having the lowest percentage correct (Figure 5). Prior experiences of a subject with a word also influenced the subject's ability to accurately produce that word's intensity contour. In general, words with Fam-Und rating of 22 had a higher percent correct than words with a rating of 00.

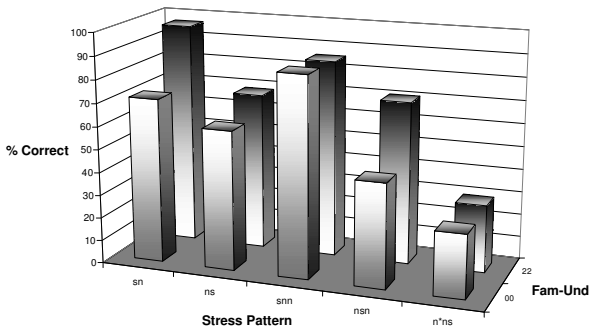

Figure 5: *Percentage of correctly reproduced intensity contours of prosodically-modified English words.*

4.3. Stressed Syllable Duration

In contrast to the results obtained for the production of the correct F0 and intensity contours of normal words, the percentage of accurately produced stressed syllable durations

increased as the lexically stressed syllable moved from the first syllable location to the last syllable location (Figure 6). The effect of subject experience and exposure to a word on the accuracy of syllable duration reproduction could not be clearly discerned from the obtained results.

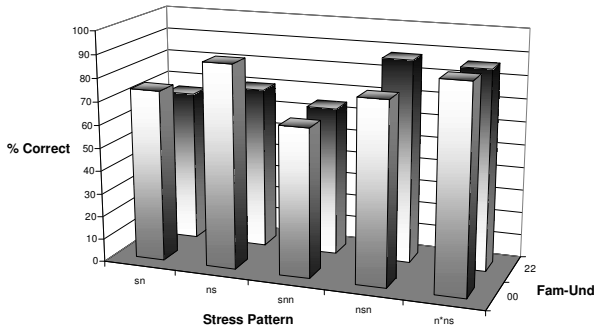

Figure 6: *Percentage of correctly reproduced accented-syllable durations of normal English words.*

Similar correct syllable duration production results were obtained for both prosodically-modified and normal words. Although the overall percent correct was lower for modified words, the pattern of increasing percentage of accurately produced stressed syllable durations was still observed as the lexically accented syllable moved from the first syllable location to the last syllable location (Figure 7). Likewise, the effect of subject experience and exposure to a word on the accuracy of syllable duration production could not be clearly discerned from the results obtained for modified words.

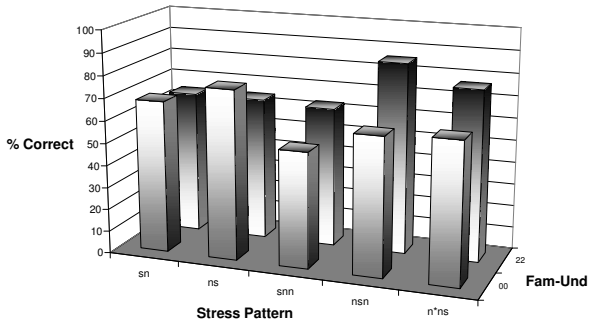

Figure 7: *Percentage of correctly reproduced accented-syllable durations of prosodically-modified English words.*

4.4. Normal and Modified Word Comparisons

Results for the normal word repetition and for prosodically-modified word repetition differed in the case of the intensity contour. Prosodic modification reduced the percent correct for intensity contours of words with Fam-Und ratings of 22 and 00, with 00 rated words being slightly lower. A similar pattern of accuracy was observed for both F0 and intensity stress placements, in the case of modified words. Words with first syllable stress were the least influenced by the prosodic modifications.

In the case of F0 contour, in general, the percent correct was lower for modified words. Comparison between the normal and modified words for the production of stressed syllable duration revealed little difference in the pattern of accuracy. However, percent correct for modified words with Fam-Und ratings of 00 were slightly lower.

5. Conclusion

The lexical-prosody production errors indicate that the placement of the lexical stress had considerable effect on prosodic accuracy for both normal and modified words. This effect was to decrease accuracy as the lexical accent placement moved from the first syllable to the last syllable, for F0 contour of normal and modified words, as well as for the intensity contour of modified words. An opposite effect of increasing accuracy, for the same stress placements, was observed for production of stressed syllable duration.

Subject familiarity and understanding of words increased the accuracy of production of the correct F0 contour, for both normal and modified words. Greater familiarity and understanding of a word also led to increased accuracy in the production of the correct intensity contour of that word, if the word was prosodically-modified. Studies by Erickson (2000) and others indicate that better accuracy in later syllable duration stress placement can be attributed to middle and final vowel lengthening. Quantitative differences between Japanese and English speakers in the location of the prosodic prominences within the accented syllable are currently being determined.

6. References

[1] Beckman, M.E.; Pierrehumbert J.B., 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3, 255-309.

[2] Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Inst. Phonetic Sci. Univ. Amsterdam* 17, 97-110.

[3] Erickson, M.L., 2000. Simultaneous effects on vowel duration in American English: A covariance structure modeling approach. *J. Acoust. Soc. Am.* 108, 2980-2995.

[4] Fry, D.B., 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* 27, 765-768.

[5] Fujisaki, H.; Hirose, K.; Sugito, M., 1986. Comparison of acoustic features of word accent in English and Japanese. *J. Acoust. Soc. Jpn.* 7, 57-63.

[6] Lieberman, P., 1960. Some acoustic correlates of word stress in American English. *J. Acoust. Soc. Am.* 32, 451-454.

[7] Okobi, A.O.; Hirose, K., 2003. Acoustic Analysis of English Lexical-Prosody Reproduction by Japanese Speakers. *IEICE Tech. Report* 103, 37-42.