

# Perception of English lexical stress by English and Japanese speakers: effect of duration and “realistic” intensity change

*Shinichi Tokuma*

Chuo University  
Tokyo, 192-0393, Japan  
tokuma@tamacc.chuo-u.ac.jp

## Abstract

This study investigated the effect of duration and intensity on the perception of English lexical stress by native and non-native speakers of English. The spectral balance of intensity was manipulated in a “realistic” way suggested by Sluijter et al. [1], which is to increase intensity level in the higher frequency bands (above 500Hz) as shown in the realisation of vocal effort. A non-sense English word /nə:nə:/ embedded in a frame sentence was used as the stimuli of the perceptual experiment, where English speakers and two levels of Japanese learners of English (advanced and pre-intermediate) were asked to determine lexical stress locations. The result showed: (1) “realistically” manipulated intensity serves as a strong cue for lexical stress perception of English for all subject groups; (2) advanced Japanese learners of English are, like English speakers, sensitive to duration in lexical stress perception, whereas pre-intermediate Japanese learners are, to a very limited extent, duration-sensitive; and (3) intensity, if altered in a proper way, could be as significant a cue as duration in perceiving English lexical stress.

## 1. Introduction

The phonetic correlates of lexical stress in English have been a central issue in English prosody research, with four acoustic parameters contributing to the perception of English lexical stress: F0, segmental duration, intensity and vowel quality [2]. Of these four parameters, pitch and duration are the most important cues, while intensity is less significant and vowel quality contributes little to the perception. (See [3] and the papers reviewed in [1]) However, Sluijter et al. [1] maintain that intensity, if its spectral balance is altered in a more “realistic” way, can serve as a strong lexical stress cue in Dutch; they further suggest that this will also be the case in English.

It is also generally agreed that Japanese has a lexical stress system which uses high/low pitch difference to distinguish words [2]. This observation has some phonetic evidence: in production, Japanese speakers tend to use only F0 and intensity to realise linguistic stress [4], while in stress perception, Japanese listeners heavily rely on F0 in the stress perception of disyllabic words [2]. Nevertheless, in [2], Beckman altered the overall intensity spectrum, unlike Sluijter et al. [1], who manipulated the spectral balance of intensity, and therefore the true role played by intensity in English lexical stress perception remains moot.

This discrepancy in the cues for lexical stress perception between Japanese and English listeners raises an

interesting issue of L2 perception: the production and perception of English lexical stress by Japanese listeners. A number of studies have been conducted on this topic (e.g. [5], [6]). Notably, Mochizuki-Sudo & Kiritani [6] claim that Japanese listeners with poor English command do not perceive inter-stress intervals, and that they are “not less sensitive than Americans in discriminating durations of stressed vowels,” (p.247). This is ascribed to the phonological distinction in vowel length in Japanese. However, Mochizuki-Sudo & Kiritani [6] did not manipulate the intensity, partly because they used natural words as perceptual stimuli.

In this paper, the following issues were investigated: (1) how the lexical stress perception of English and Japanese listeners is affected by the durational change or the balance change of the intensity spectrum as manipulated in Sluijter et al. [1]; (2) how differently Japanese listeners perceive English lexical stress according to the level of their English command; and (3) how the perceptual interaction of duration and intensity changes influences Japanese and English listeners. To investigate these issues, synthesised non-sense words were used to facilitate the manipulation of two perceptual parameters: duration and intensity. Furthermore, to study issue (2) above, Japanese listeners at pre-intermediate or advanced levels of English participated in the test.

## 2. Experiment

### 2.1. Subjects

Three groups of subjects participated in the perceptual experiment.

- (A) Native speakers of English (henceforth called NE) : Six native speakers of British English, three of whom are postgraduate students of University College London, the rest being university lecturers in Japan. None of them speaks with a noticeable regional accent.
- (B) Advanced Japanese learners of English (henceforth called AJ): Six Japanese postgraduate students, who were studying for an MA or PhD in English phonetics/linguistics at Sophia University in Tokyo. All of them had studied English for more than 12 years, and spoke very fluent English. Four of them had lived in an English-speaking country for at least one year.
- (C) Pre-intermediate Japanese learners of English (henceforth called PJ): Thirty-five Japanese first-year undergraduate students of Saint Margaret Junior College in Tokyo. Eighteen of them had taken TOEIC tests, with an average

score of 464 points. This, and informal evaluation by a professor at the College put their English abilities around the pre-intermediate level.

## 2.2. Materials

Since existent English word pairs with a lexical stress contrast, such as “object” (noun) - “object” (verb), are not symmetrical in their syllable and segmental structure, and this could possibly affect subjects’ performance, we used a nonsense word /nə:nə:/ for the experiment, as in Sluijter et al. [1]. The choice of /n/ facilitated the manipulation of intensity and duration, and the vowel /ə:/ is assumed to be the closest in vowel quality to an English weak vowel /ə/. The duration of the syllable /nə:/ in the word was varied in 6 steps of 20ms from 160ms to 260ms although the total duration of the word was kept constant. This range of duration variance was determined by the normal duration figures proposed by Umeda [7] [8], as well as the acoustic data obtained from one native speaker of South-East British English. This process produced the stimuli of 6 durational types shown in the Table 1 below:

	1st syllable duration	2nd syllable duration
Stimulus 1	160ms	260ms
Stimulus 2	180ms	240ms
Stimulus 3	200ms	220ms
Stimulus 4	220ms	200ms
Stimulus 5	240ms	180ms
Stimulus 6	260ms	160ms

Table 1: Durational structure of stimuli

These stimulus words /nə:nə:/ were embedded in a frame sentence “I can’t say .... now” to avoid the perceptual intervention of sentence-final lengthening. In the synthesis, MBROLA synthesiser devised by Dutoit et al. [9] was used and its British male voice database was implemented. F0 was set to reach the peak of 180Hz in the initial part of /a:/ in “can’t”, down to 100 Hz at the end of /eɪ/ in “say”, and linearly down to 95 Hz at the end of the sentence, so that auditorily the frame sentence would have a nucleus on “can’t” and no pitch change in the tail, thus eliminating the perceptual effect of pitch on lexical stress perception. The synthesised sentences were checked by one native speaker of South-East British English, who ensured that they had an acceptable quality of synthesised speech.

The intensity of the synthesised stimulus words was next manipulated by the Speech Filing System Windows Version (henceforth SFS) devised by Mark Huckvale of University College London.

Sluijter et al. [1] maintain that in order to reflect the reality of role played by loudness in lexical stress perception, varying the spectral balance of loudness, and specifically increasing intensity level in the higher frequency bands as shown in the realisation of vocal effort, serves as a more effective cue for lexical stress. They proved this by increasing the levels of the frequency components of a syllable above 500Hz by 3, 6 or 9 dB, based on their production data [1]. The present study followed their procedure.

The actual intensity manipulation procedure is shown in Figure 1. First, the target syllable, either the first or

the second syllable of /nə:nə:/, was annotated in SFS. Then the whole sentence, i.e., “I can’t say /nə:nə:/ now.” was divided into two frequency components: above 500Hz and below 500Hz, by linear phase filtering using a non-recursive high-pass/low-pass filter, with a cut-off frequency of 500Hz. This filter, implemented in SFS, was used to minimise phase distortions when combining two components as the final output. The output speech wave of the filter was carefully inspected to identify any considerable phase distortions.

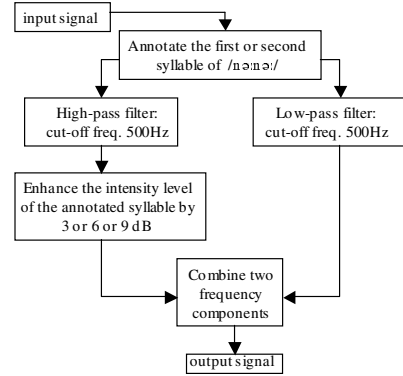


Figure 1: Diagram of intensity manipulation procedure

Next, the intensity of the annotated target syllable in the component above 500Hz was increased by either 3, 6 or 9 dB, while the rest of the component and the component below 500Hz remained unchanged. Finally, the two components were combined, again with careful attention to the phase. Thus, in these stimuli, no intensity was added in the base band of 0-500Hz, and only the intensity of the annotated syllables was increased above the 500Hz band.

The whole intensity manipulation process produced six intensity level patterns. Thus, by also including the stimulus with the original intensity level, seven intensity steps were created and presented to the subjects, as shown in Table 2. Note that some of the stimuli have conflicting lexical stress cues, such as Stimulus 6 with the intensity pattern of Step 7, where the duration of the first syllable (260ms), although its intensity is lower than that of the second syllable by 9dB, is much longer than that of the second syllable (160ms). This enabled the investigation of how the perceptual interaction of durational and intensity changes influenced the Japanese and English subjects.

	1st syllable intensity	2nd syllable intensity
Step 1	+9dB	0dB
Step 2	+6dB	0dB
Step 3	+3dB	0dB
Step 4	0dB	0dB
Step 5	0dB	+3dB
Step 6	0dB	+6dB
Step 7	0dB	+9dB

Table 2: Intensity manipulation steps of stimuli

## 2.3. Experimental procedure

Each stimulus sentence was presented to the subjects twice, producing a total of 84 presentations (6 durational patterns x 7 intensity manipulation types x 2 repetitions) per subject, and they were preceded by five trial presentations designed to

make listeners familiar with the experimental setting and the nature of the stimuli. The interval between each presentation was 3 seconds, and a longer pause of 5 seconds with a beep was inserted after every 10 presentations. These stimulus sentences were recorded in a random order onto a DAT and an audiocassette tape, the latter of which was used to test the PJ subjects.

The task of the subjects was to listen to the stimulus words embedded in a sentence and to judge which of the syllables in the stimulus word /nə:nə:/ was stressed. They were asked to circle or tick the syllable of the word (written as “Nur-Nur” in an answer sheet) which they thought was stressed. For NE and AJ groups, the test was carried out in a sound-proof room or a quiet study room, and the subjects listened to the stimuli through covered-ear headphones. For the PJ group, the Language Laboratory Room in Saint Margaret Junior College was used to test the subjects, where the stimuli were played through covered-ear headphones. The NE and AJ subjects were tested one by one in a different time-slot, while PJ group was tested altogether in the Language Laboratory Room. None of the PJ subjects, however, reported that their attention had been diverted by the presence of other students.

## 2.4. Results

After the experiment, it was found that four subjects of PJ group had not followed the instructions and had left some answers blank. These four subjects, therefore, were excluded from the analysis in order to validate the reliability of the data. This reduced the PJ group to thirty-one members.

In the analysis, the responses were accumulated and the numbers of the first or second syllable choices were counted for each intensity and duration type across all the subjects within the group, before the percentages of the first/second syllable choices were calculated. Figures 2 to 4 show the percentages of the first syllable choices for each intensity and durational pattern and for each subject group. Figure 2 is for the NE group, Figure 3 for the AJ group, and Figure 4 for the PJ group. In these figures, intensity manipulation steps are plotted on the X axis; for example, ‘Step 1’ means that the first syllable was increased by +9dB above 500Hz (See Table 2). On the other hand, durational patterns of the first/second syllables are plotted as a separate category. For example, S1 stands for Stimulus 1 in Table 1, which means that the duration of the first and second syllable was 160ms and 260ms respectively.

Figures 2 to 4 indicate that the whole subject groups were sensitive to intensity changes, which is shown in the rising trend from left to right. Furthermore, Figures 2 and 3 show that the durational change significantly affected the performance of NE and AJ groups. In Figure 4, the influence of the durational change on the performance of PJ group can be scarcely discerned, but Figure 5, where the mean percentages of the first/second syllable choices were obtained across all intensity patterns for each duration type, demonstrates that the subjects of PJ group are, to a very limited extent, sensitive to durational change.

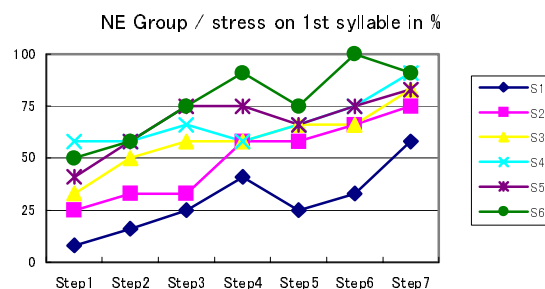


Figure 2: Results of NE group; ‘S’ stands for Stimulus (See Table 1).

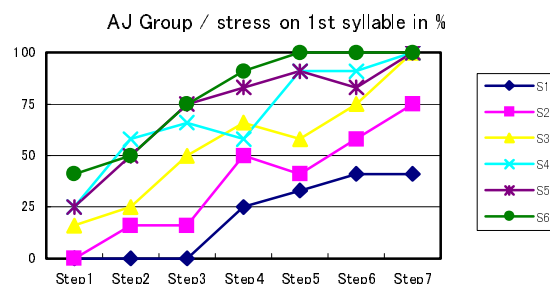


Figure 3: Results of AJ group; ‘S’ stands for Stimulus (See Table 1).

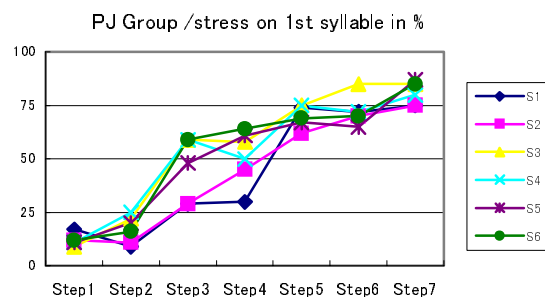


Figure 4: Results of PJ group; ‘S’ stands for Stimulus (See Table 1).

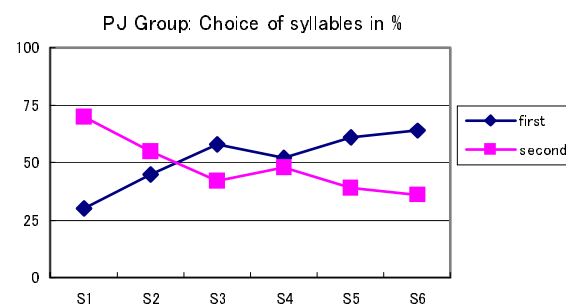


Figure 5: Mean responses of PJ group across all intensity patterns.

All the above observations suggest that, although the subjects of PJ group hardly use duration as a cue for English lexical stress, they are very sensitive to intensity, whereas AJ and NE groups rely both on the durational and intensity cues for lexical stress perception of English.

To address the issue of the perceptual interaction of durational and intensity changes, the mean perceptual

results of the ends of the stimulus continuum, together with the stimuli with no intensity manipulation, are shown in Table 3. In the table, Step 1 means that the first syllable was increased by 9dB, Step 7 the second syllable by 9dB. Step 4 means that no intensity manipulation was made. The shaded cells in Table 3 display the results of the stimuli with conflicting lexical stress cues. Note that for S1, displayed figures are percentages of the second syllable choices.

S1 (160ms-260ms): Stress on second syllable in %				S6 (260ms-160ms): Stress on first syllable in %			
	Step1	Step4	Step7		Step1	Step4	Step7
NE	41%	59%	92%	NE	91%	91%	50%
AJ	59%	75%	100%	AJ	100%	91%	41%
PJ	25%	70%	83%	PJ	85%	64%	12%

Table 3: The results of the stimulus continuum ends. Results of the stimuli with the conflicting stress cues are in shaded cells. See Table 2 for actual intensity-manipulation figures of Steps.

In Table 3, the mean perceptual scores of NE and AJ groups in shaded cells are around 50%, which suggests that manipulated intensity could be as influential as duration in the perception of English lexical stress. However, it remains to be seen which acts as a more dominant cue in lexical stress perception. The shaded cells in Table 3 also show that PJ group responded almost solely to an intensity cue, suggested by the much lower scores (25% for S1 and 12% for S6) than predicted by the durational patterns.

There is one issue to be mentioned here: Table 3 demonstrates that the NE group gave strong preference to stress on the first syllable, as shown in the results of the Step 4 (non-manipulated) stimuli: 91% first syllable choice for S6 (260ms, 160ms), while 100-59=61% first syllable choice for S1 (160ms, 260ms). van Heuven & Menert [10] reported on this strong preference in English and Dutch and called it “stress bias”, although they claim that it disappears when words are embedded in a sentence-final position in Dutch. Further research is required to establish whether stress bias truly affects the words in a sentential context in English.

### 3. Discussion

The results of the experiment in this study support the claim by Mochizuki-Sudo & Kiritani [6] that Japanese listeners with poor English command did not perceive foot-level shortening of unstressed syllables. This was demonstrated in this study by the PJ group’s poor sensitivity to duration. Moreover, the results confirm the claim by Sluijter et al. [1] that intensity, if properly altered, can be an effective cue in English stress perception.

There is one apparent discrepancy between the results of this study and the predictions made by Sluijter et al. [1]. In [1], they hypothesised that since Japanese uses pitch accent, “Japanese listeners will be insensitive to those more realistic loudness manipulation...” (p. 511) while the present results show that intensity serves as a significant stress cue for PJ group. This is perhaps due to English teaching methods in Japanese schools, which put less emphasis on oral communication and where lexical stress is taught as saying a vowel loudly, not as highlighting a syllable by loudness, pitch and most importantly, length. (The fact that durational change caused by lexical stress is not taught in

Japanese schools also accounts for the fact that the subjects of PJ group paid little attention to the durational change of the stimuli.) The subjects of PJ group may not respond to manipulated intensity if the stimuli are synthesised Japanese words, which may encourage them to adopt a different perceptual strategy, but this speculation is beyond the scope of this study.

### 4. Conclusion

Overall, the results of the experiment demonstrate that the “realistic” balance change of the intensity serves as a strong cue to lexical stress perception of English for English speakers and advanced/pre-intermediate learners of English. The findings also indicate that advanced Japanese learners of English are, like English speakers, sensitive to duration in lexical stress perception, while pre-intermediate learners hardly rely on it. Moreover, it is also suggested that manipulated intensity could be as influential as duration in perceiving English lexical stress.

### 5. Acknowledgements

The author cordially appreciates the helpful suggestions made by Mark Huckvale of University College London and the comments from Professor Naoki Ogawa of Saint Margaret Junior College. This research was funded by a Chuo University Grant for Special Purposes.

### 6. References

- [1] Sluijter, A.M.C., van Heuven, V.J. and Pacilly, J.J.A. (1998) “Spectral balance as a cue in the perception of linguistic stress.” *J. Acoust. Soc. Amer.*, vol. 101, 503-513.
- [2] Beckman, M. (1986) *Stress and Non-Stress Accent*. Foris Publications.
- [3] Wouters, J. and Macon, M.M. (2002) “Effects of prosodic factors on spectral dynamics: II. Synthesis.” *J. Acoust. Soc. Amer.*, vol. 111, 428-438.
- [4] Fujisaki, H., Hirose, K. and Sugito, M. (1986) “Comparison of acoustic features of word accent in English and Japanese.” *J. of the Acoust. Soc. of Japan*, vol.7, 57-63.
- [5] Ueyama, M. (2000) *Prosodic Transfer: An Acoustic Study of L2 English vs. L2 Japanese*. Ph.D. Diss. UCLA.
- [6] Mochizuki-Sudo, M. and Kiritani, S. (1991) “Production and perception of stress-related durational patterns in Japanese learners of English.” *J. of Phonetics*, vol. 19, 231-248.
- [7] Umeda, N. (1975) “Vowel duration in American English.” *J. Acoust. Soc. Amer.*, vol. 58, 434-445.
- [8] Umeda, N. (1977) “Consonant duration in American English.” *J. Acoust. Soc. Amer.*, vol. 61, 846-858.
- [9] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Vreken, O. (1996) “The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes” *Proc. ICSLP’96*, vol. 3, 1393-1396.
- [10] van Heuven, V.J. and Menert, L. (1996) “Why stress position bias?” *J. Acoust. Soc. Amer.*, vol. 100, 2439-2451.