# Unfilled pauses in Japanese sentences read aloud by non-native learners

*Hiroko Hirano[1], Goh Kawai[2], Keikichi Hirose[3], Nobuaki Minematsu[4]*

[1,4] School of Frontier Sciences, University of Tokyo,
[2] Institute of Language and Culture Studies, Hokkaido University,
[3] School of Information Science and Technology, University of Tokyo
[1,3,4]{hiran,hirose,mine}@gavo.t.u-tokyo.ac.jp, [2]goh@kawai.com

## Abstract

Perception experiments suggest that natives judge non-native unfilled pauses as indiscriminate and indecisive. Multiple regression analyses of unfilled pauses indicate a connection between syntactic structure and pause location and duration. Native speakers uniformly pause at large syntactic breaks with marked duration, whereas non-natives' unfilled pauses are spread over various locations, possibly reflecting limited syntactic planning. Our method might be used to synthesize appropriate unfilled pauses in text-to-speech systems, and to train pausing behavior in automated pronunciation learning systems for non-native learners.

**Index Terms**: unfilled pauses, second language learning, syntactic structure, perception experiments, multiple regression analyses

## 1. Introduction

Appropriate unfilled pauses in fluent, natural speech help listeners understand the message [1]. The speaker may choose, to a certain extent, the location and duration of unfilled pauses. Text-to-speech systems determine where to place pauses based on factors such as the degree of syntactic boundaries, the connection between adjacent phrases, and the number of contiguous phrases or syllables [2][3][4][5].

Seeking to develop an automated pronunciation learning system for non-native speakers of Japanese, we have compared native and non-native prosodic patterns [6]. Regarding pauses, compared with native speakers, non-native speakers are known to pause frequently (particularly at left-branching syntactic boundaries and within phrases), and shorten or omit pauses at the end of sentences [7]. This paper reports on 1) duration analysis of unfilled pauses considering branching structure levels, 2) perception experiments of unfilled pauses in spoken sentences, and 3) multiple regression analyses that predict what factors affect naturalness.

## 2. Speech corpus

Chinese learners comprise over two-thirds of learners of Japanese as a second language. We used 7 sentences read aloud by 10 Tokyo-dialect native speakers and 10 Mandarin-dialect speakers (intermediate to advanced learners of Japanese as a second language who had been living in Japan between 8 and 18 months at the time of recording). Subjects read a script with *furigana* (i.e., written pronunciation guide for readers unfamiliar with kanji), but without punctuation. Subjects read a fair amount of material besides the 7 sentences we analyzed; the 7 sentences we used were spoken after the subjects became sufficiently comfortable with the recording task (after warming up, so to speak), thus we believe these utterances represent each subjects' speech styles. Recordings were made in a soundproof booth (2 meters square), using a Sony TCD-D10PRO digital audio tape recorder (sampling rate 48 kHz, 16 bits) and a Sony C-38B microphone (desktop, monaural, cardioid directivity, FET condenser, frequency response 80–18000 Hz).

In native speech analyses, unfilled pauses are often defined as fairly long silence intervals, common minimum durations being anywhere between 100 and 300 ms [8]. In non-native speech, short but clearly audible gaps in speech also need to be analyzed. We tagged unfilled pauses as silence intervals equal or greater than 40 ms, which is one-third of the average mora duration of native speakers. A native Japanese language instructor trained in phonetic labeling labeled utterances at mora boundaries.

Table 1. *Descriptive statistics for mora and pause durations in the corpus. 10 native and non-native speakers each read aloud 7 sentences.*

| | Non-native (L1=Chinese) | Native (L1=Japanese) |
|---|---|---|
| number of mora | 25 | 25 |
| mora duration [sec] | 0.17 | 0.12 |
| SD | 0.06 | 0.04 |
| mean duration[sec] | 5.01 | 3.59 |
| SD | 1.26 | 1.17 |
| pause included | Non-native (L1=Chinese) | Native (L1=Japanese) |
| max duration/median | 1.52 | 1.28 |
| min duration/median | 0.72 | 0.77 |
| SD | 0.16 | 0.10 |
| pause excluded | Non-native (L1=Chinese) | Native (L1=Japanese) |
| max duration/median | 1.30 | 1.30 |
| min duration/median | 0.76 | 0.57 |
| SD | 0.12 | 0.14 |

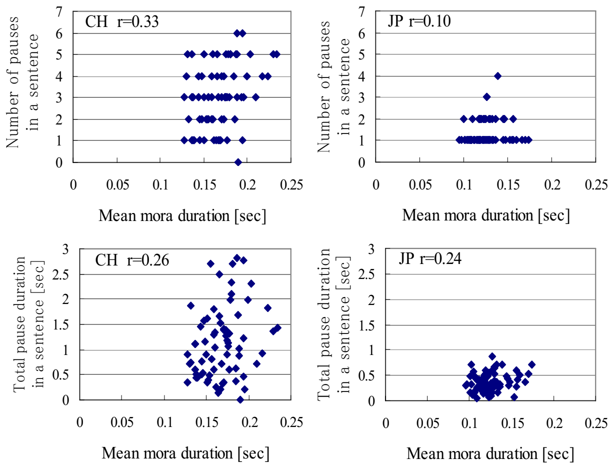

Figure 1. *Scatter plots of mora durations and number of pauses in utterance (top) and mora durations and total pause duration in utterance (bottom).*

## 3. Duration analysis

Unfilled pause duration studies report that instead of absolute physical measures (such as the number of milliseconds in an unfilled pause) phonetic-phonological relative measures (such as the lengths of unfilled pauses expressed as a ratio of average syllable duration) depict phenomena more accurately [3][8]. Our results show that both absolute and relative measures depict duration phenomena (Table 1, Figure 1), although relatively represented data tend to be normally distributed (Figure 2).

Duration of utterances (Table 1) including pauses is the duration of the utterance less leading and trailing silence. Duration of utterances excluding pauses is the duration of the utterance less leading and trailing silence, and utterance-internal pauses. Durations were normalized per sentence by taking the ratio of each utterance's duration to the median utterance duration for that sentence.

Compared with utterance durations of native speakers, non-native speakers have larger variance and significantly slower speech rate ($p<0.01$). There is no apparent connection between speech rate and either the number or total duration of unfilled pauses. Non-native speakers pause frequently (up to 6 times per utterance) while most native speakers pause only once or twice. Figure 1 shows how non-native pausing behavior is diffuse while natives are more uniform.

Table 2 shows the number of pauses and their breakdown based on absolute length (longer or shorter than 200 ms) and relative length (longer or shorter than 2 average mora; average mora durations being calculated per speaker per utterance as the total duration of filled segments divided by the number of mora in the utterance). Non-native speakers have numerous long pauses; hence their aggregate pause duration is larger than natives.

At first sight, little difference exists between pauses longer or shorter than 2 average mora. Native speakers and non-native speakers seem to use long and short pauses with similar frequency. A different picture emerges when syntactic structure is considered.

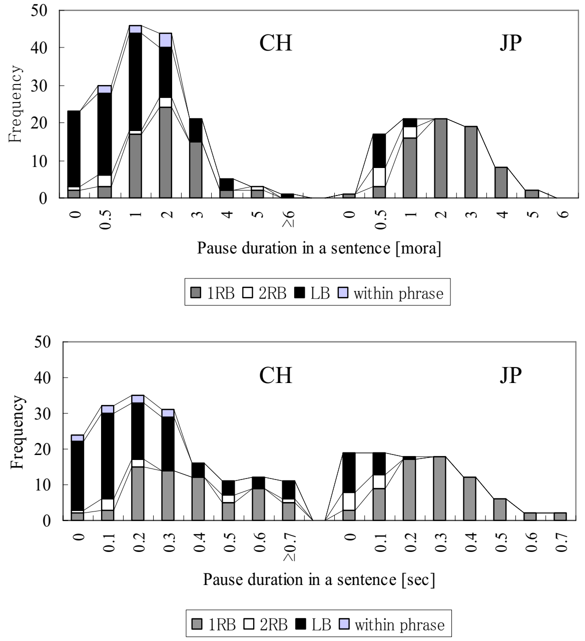

Figure 2. *Histogram of relative pause durations (top) and absolute pause durations (bottom). The relative measure shows a cleaner bimodal distribution. 1 RB means the first-level (deepest) right-branching boundary, 2 RB means the second-level (less deep) right-branching boundary and LB means left-branching boundary.*

Table 2. *Breakdown of number of pauses based on absolute and relative measurements. For relative measurements, mora were divided into two groups depending on whether the mora was longer or shorter than twice the average duration of mora in the utterance.*

| | number of pauses | ≥ 0.20 sec | <0.20 sec | ≥ 2 mora | <2 mora |
|---|---|---|---|---|---|
| CH | 172 | 116 (67%) | 56 (33%) | 74 (43%) | 98 (57%) |
| JP | 96 | 58 (60%) | 38 (40%) | 50 (52%) | 46 (48%) |

Figure 3 shows a syntactic tree structure of an utterance. In this study, first-level right-branching boundaries (where adjacent phrases do not modify each other) are typically near the center of the utterance, and second-level right-branching boundaries contain only left-branching structures (where adjacent phrases modify each other). Due to their syntactic significance and balanced phonetic mass on either side, first-

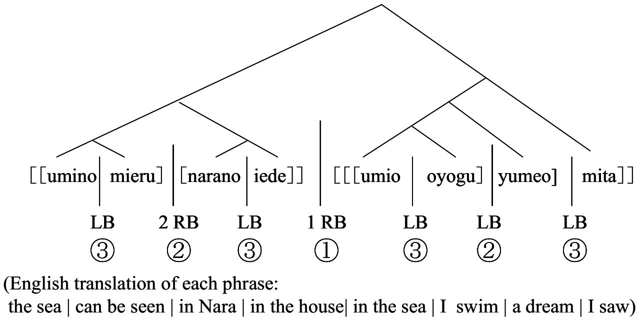
[[umino | mieru] | [narano | iede]] | [[[umio | oyogu] | yumeo] | mita]]
LB
2 RB
LB
1 RB
LB
LB
LB
③
②
③
①
③
②
③
(English translation of each phrase:
the sea | can be seen | in Nara | in the house| in the sea | I swim | a dream | I saw)

Figure 3. *Syntactic tree of one of the sentences. The first-level (deepest) right-branching boundary is located near the center of the utterance in the sense that on either side there are roughly equal amounts of phonetic material and syntactic complexity.*

level boundaries are obvious choices for pause insertion. Native speakers use long pauses exclusively at first-level boundaries (see Figure 2). Non-native speakers are not as consistent in reserving long pauses for prime candidates. This suggests that a straightforward measure of correct pause usage is counting the ratio of long pauses used in phonologic-syntactic significant locations vs. other locations.

It is tempting to imagine that native speakers rank pause usage in the order of first-level right-branching boundaries, second-level right-branching boundaries, and left-branching boundaries, but our data size is too small to be conclusive at this level of analysis (but see the following section). Non-native speakers show a similar, albeit weaker, tendency (Table 3).

Another easily discerned difference between native speakers and non-native speakers is that non-native speakers insert pauses within phrases, while native speakers never do in read speech tasks.

To summarize, based on duration analysis of unfilled pauses, the key rules are:

- Pause once in the middle.
- Pauses should not be too short or too long (200 to 500 ms).
- Do not pause too often.
- Rank-order your pauses. Associate long pauses with important locations.

## 4. Perception experiments and multiple regression analysis

14 trained native judges scored non-native speech (10 speakers, 7 sentences) by listening to each utterance 5 times to score (a) overall impression, (b) pitch accent, (c) rhythm, (d) pauses, and (e) intonation. As judges often disagree, we compared each judges scores against mean scores of the other 13. We chose judges (there were 4) whose correlation coefficients were no less than 0.7 for all 5 score categories.

We ran multiple regression analyses using the factors mentioned in the previous section as explanatory variables and the judges' scores as the dependent variables (Table 4).

Table 3. *Number of pauses broken down by location. Both absolute and relative measures indicate roughly the same tendency.*

| | ≥ 0.20 sec | | ≥ 2 mora | |
|---|---|---|---|---|
| | CH | JP | CH | JP |
| 1st-level right-branch | 60<br>(52%) | 58<br>(100%) | 43<br>(58%) | 50<br>(100%) |
| 2nd-level right-branch | 5<br>(4%) | 0<br>(0%) | 4<br>(5%) | 0<br>(0%) |
| left-branch | 47<br>(41%) | 0<br>(0%) | 23<br>(31%) | 0<br>(0%) |
| within phase | 4<br>(3%) | 0<br>(0%) | 4<br>(5%) | 0<br>(0%) |
| | < 0.20 sec | | < 2 mora | |
| | CH | JP | CH | JP |
| 1st-level right-branch | 5<br>(9%) | 11<br>(29%) | 22<br>(22%) | 19<br>(41%) |
| 2nd-level right-branch | 4<br>(7%) | 9<br>(24%) | 5<br>(5%) | 9<br>(20%) |
| left-branch | 43<br>(77%) | 18<br>(47%) | 67<br>(68%) | 18<br>(39%) |
| within phase | 4<br>(7%) | 0<br>(0%) | 4<br>(4%) | 0<br>(0%) |

The combination of 5 explanatory variables showed multiple correlation coefficient R=0.74. Further improvement might be obtained by adding the following factors to consideration: (a) phrase-final lengthening as a perceived pause, (b) pauses shorter than 100 ms, (c) F0 and power.

We repeated multiple regression analysis by interpreting phrase-final lengthening (mora with duration greater than 1.5 times the mean mora duration in the utterance) as pauses. Table 5 shows the result by the combination of explanatory variables yielding the highest accuracy. We obtained R=0.80, which suggests the 5 explanatory variables we chose have considerable predictive power. Each variable seems to contribute at roughly equal amounts. Multiple regression analysis results indicate that well-formed unfilled pauses have the following features:

- Insert a pause (200 to 600 ms) at the first-level right-branching boundary.
- Never insert a long pause (over 400 ms) at locations other than the first-level right-branching boundary.
- Insert and lengthen pauses in the descending order of priority of first-level right-branching boundaries, second-level right-branching boundaries, and left-branching boundaries.
- Do not pause within phrases.
- The shorter the total pause duration in the utterance, the better.
- Unless you want to pause, do not lengthen word-final phones, or abruptly drop F0 or power, because doing so sounds like you inserted a pause.

Non-native learners should find the above guidelines useful. Our explanatory variables used in multiple regression analysis can be implemented on automatic pronunciation learning systems. As far as automated scoring of pauses is concerned, there seems to be no advantage in normalizing pause durations.

Table 4. *Explanatory variables used in multiple regression analysis.*

| variable | description | value range |
|---|---|---|
| x1 | one pause inserted at first-level right-branching boundary (1RB) | 0 or 1 |
| x2 | 200-600 ms pause at 1RB | 0 or 1 |
| x'2 | ≥ 2 mora pause at 1RB | 0 or 1 |
| x3 | ≥ 400ms pause at other than 1RB | 0 or 1 |
| x'3 | ≥ 3 mora pause at other than 1RB | 0 or 1 |
| x4 | durations are 1RB > 2RB > LB | 0 to 1 |
| x5 | normalized number of pauses | 0 or 1 |
| x6 | phrase-internal pause | 0 or 1 |
| x7 | total duration of pause in utterance | measured value (ms) |
| x'7 | | relative value (number of mora) |
| x8 | duration of 1RB pause | measured value (ms) |
| x'8 | | relative value (number of mora) |

## 5. Discussion

Multiple regression analysis predicting human perception scores using (a) pausing at the first-level right-branching boundary, (b) long pauses at locations other than the first-level right-branching boundary, (c) whether (duration of the first-level right-branching boundary) > (duration of second-level right-branching boundary) > (duration of left-branching boundary), (d) pausing within phrases, and (e) total pause duration in the utterance, suggest that, at least for our data, the following guidelines apply when using unfilled pauses:

- Pause sparingly and briefly.
- Pause at the first-level right-branching boundary. This pause can be long or short. If it is long, it is the only long pause allowed. If it is short, shorten or eliminate your other pauses.
- Optionally, pause at second-level right-branching boundaries, and as a further option, at left-branching boundaries.

Our method might be used to synthesize appropriate unfilled pauses in text-to-speech systems, and to train pausing behavior in automated pronunciation learning systems for non-native learners.

Table 5. *Combination of explanatory variables that yielded best correlation with judges' scores.*

| variable | standardized regression coefficient | correlation |
|---|---|---|
| x2 | 0.16 | 0.30 |
| x3 | -0.19 | -0.60 |
| x4 | 0.30 | 0.61 |
| x6 | 0.20 | 0.38 |
| x7 | -0.32 | -0.67 |
| R | 0.80 | |
| $R^2$ | 0.63 | |
| adjusted R | 0.78 | |
| adjusted $R^2$ | 0.61 | |

## References

[1] A. Kurematsu "Technical applications of prosodic information", in "Japanese Language Phonetics series, volume 2, accent, intonation, rhythm and pauses", Tokyo, Sanseido, 303-318, 1997

[2] H. Kawai, K. Hirose and H. Fujisaki, "Rules for generating prosodic features for text-to-speech synthesis of Japanese", The Journal of Acoustic Society of Japan, 50-6, 443-442, 1994

[3] N. Kaiki and Y. Sagisaka, "Study of Pause Insertion Rules Based on Local Phrase Dependency Structure", IEICE Trans. Inf. & Syst. (Japanese Edition), vol. J79-D-II, no.9, 1455-1463, Sept. 1996

[4] H. Tsukada, "A left-to-right processing model of pausing in Japanese based on limited syntactic information", ICSLP-1996, Philadelphia, PA, USA, 1353-1356, Oct. 1996

[5] K. Hakoda and H. Sato, "Prosodic Rules in Connected Speech Synthesis", IEICE Trans. Inf. & Syst. (Japanese Edition), vol. J63-D, no.9, 715-722, Sept. 1980

[6] H. Hirano, G. Kawai, "Pitch patterns of intonational phrases and intonational phrase groups in native and non-native speech", Proceedings of Eurospeech 2005, Lisbon, Portugal, 761-764, Sept. 2005

[7] A. Ishizaki, "How do learners leave a pause when reading Japanese aloud?: A comparison of English, French, Chinese and Korean learners of Japanese and native Japanese speakers", Japanese-Language Education around the Globe vol.15, 75-89, June 2005

[8] A. Ishizaki, "Survey of research on pausing: Basic structure of pause research in learner speech", Japanese Language Education, Nov. 2003 Special Issue, 128-146, July 2003