Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 113-116.

Evaluation of vowel hiatus in prosodic boundaries of Japanese

Shigeyoshi Kitazawa

Shizuoka University, Hamamatsu, Japan

Abstract

We investigated V-V hiatus through J-ToBI labeling and listening to whole phrases to estimate degree of discontinuity and, if possible, to determine the exact boundary between two phrases. Appropriate boundaries were found in most cases as the maximum perceptual score. Using electroglottography (EGG) of the open quotients OQ, pitch mark and spectrogram, the acoustic phonological feature of these V-V hiatus was found as phrase-initial glottalization and phrase-final nasalization observable in EGG and spectrogram, as well as phrase-final lengthening and phrase-initial shortening of the morae. A small dip was observable at the boundary of V-V hiatus showing glottalization. The test materials are taken from the "Japanese MULTEXT", consisting of a particle - vowel (36), adjective - vowel (5), and word - word (4).

1. Introduction

In normal fluent speech, phrase as well as word boundaries become obscure because of fluency and become difficult to segment. This is the salient problem of speech recognition and speech synthesis. Marks such as juncture, punctuation, focus, and prominence in a stream of speech sound are crucial for effective use of prosodic corpus. Resolution of such hiatus plays an important role in listening comprehension. In Japanese, there are very few studies about hiatus, but Kawahara states that preceding vowels spread into following syllables [1].

This paper presents results of a study concerning the boundary between morphological units, i.e., words and phrases in a Japanese sentence. Here we investigate the phrasal boundary in an utterance comprising a transition between a final mora of a preceding accentual phrase and an initial mora of the succeeding accentual phrase consisting of the same two vowels, i.e., vowel-vowel hiatus.

J-ToBI, a prosody annotation scheme, defines the phrase structure vaguely as BI label with 5 different degrees as perceived disjuncture [2]. We tried to measure this ambiguous disjuncture quantitatively through a series of perceptual experiments. Results were also investigated using EGG analyzed data (open quotient), F0, speech waveform, and spectrogram. These observed disjunctures matched with discontinuities of articulatory measurements.

2. Vowel-Vowel Hiatus in Japanese

Since Japanese almost exclusively consists of open syllables ending with a vowel (with the exception of some syllables ending in ～ん), if the following phrase begins with an initial vowel, a vowel-vowel (V-V) hiatus arises, the same vowel continues without pause. This vowel sequence is very common in Japanese:

body of a phrase	vowel	|	vowel	body of a phrase

2.1. Morphonology of vowel-vowel hiatus

Possible Japanese vowel-vowel hiatus consists of the following structure:

front phrase		rear phrase
noun + particle	→	predicate
morpheme + adverb	→	adjective
a part of compound word	→	a part of compound word

Example hiatus was taken from our corpus (3.1). The most frequent occurrence is with particles, and the next most frequent occurrence is with adjectives.

The most common phrasal unit is a morpheme (e.g. a noun) + a particle which bears an accent to compose a particle (joshi) | vowel initial phrase.

a morpheme (e.g. a noun) + a particle (joshi) | a vowel initial phrase
ga | *aru*, *wa* | *ame*, *sika* | *arimaseN*, *ni* | *iQte*, *te* | *ekizo*, *to* | *omou*, *wo* | *osiete*, *no* | *otaku*

The second type is an adjective (fukushi) | vowel initial phrase.

an adjectives (fukushi) | vowel initial phrase
mada | *atarasii*, *iQtai* | *itu*, *mosi* | *ikite*, *seQkaku* | *utouto*, *kitiNto* | *okonau*

The third less frequent type is a compound word (word | word), such as, *komugi* | *iro*, *takusii* | *ichidai*.

compound word
komugi | *iro*, *takusii* | *ichidai*

Similar phenomena are observable in the TIMIT.

examples in TIMIT
She | *is thinner than I am* (sx5: /iy | ih/). *Combine all the* | *ingredients in a large bowl* (sx118: /iy | ix/). *Where were you while we were* | *away?* (sx9: /axr | ax/)

2.2. Phonological realization of Japanese hiatus

There are a number of possible factors that help perception of Japanese V-V hiatus.

2.2.1. Phrase-initial glottalization

Glottalization of word-initial vowel is a common phenomenon of world languages [3]. It is more strongly pronounced if the word has a stress or accent at the beginning of the word.

2.2.2. Phrase-final nasalization

Voiced velar consonant is nasalized at the non-word-initial position in Tokyo Japanese. This nasalization contrasts with the following word-initial vowel that should not be nasalized. This sort of hiatus resolution occurs very often since noun phrases consisting of a noun + a particle *ga* are very common in Japanese, and such phrases can be followed by a predicate *aru* for example, composing a *a*|*a* hiatus.

2.2.3. Lengthening and shortening

Phrase-initial syllable or mora is shortened, while phrase-final syllable or mora is lengthened. This mora timing is a built in rhythm of Japanese as well as other languages. Duration of the concatenated vowel might be segmented with a built in

timer of the perception mechanism. The mechanism will help the human hearing to resolve the hiatus.

2.2.4. Morphonological constraints

Part of speech plays some role in realization of the hiatus. Vowel sequence at the phrase boundary often occurs in the environments stated in 2.1. Such constraints help to resolve the hiatus.

3. Prosody data base

Phonetic prosodic labeling is performed on voice data collected for Japanese prosody database.

3.1. Japanese MULTEXT prosody corpus [4]

The Japanese version of MULTEXT (multi-language prosody corpus) is created by the specification of EUROM1 [5]. It aims at recording same-content of speech consisting of 40 small paragraphs, then the extraction of prosody parameter, and the prosody notation of five languages.

Speakers are native speakers of the Tokyo dialect. A text is given for a reading and to evoke a simulated spontaneous utterance. Speech was recorded with apparatus based on the specifications of EUROM1, in an anechoic chamber, using a B&K 1/2 capacitor microphone, a DAT recorder (SONY PCM2300). In addition, electroglottograph is recorded with an EGG (KAY (Co.) 4338) from which F0 and open quotient are extracted.

3.2. Phonetic and prosodic labeling

Phoneme segmentation by hand-eye is good, but still is difficult to segment when the same two vowels connect. Those cases were conventionally marked at the mid point to achieve equality of morae duration [6].

J-ToBI labeling is applied for prosodic annotation according to the manual [2]. Although, the X-JToBI [7] extended the J-ToBI in spontaneities of speech, e.g. descriptions of fillers and disfluencies, it does not describe V-V hiatus. J-ToBI is sufficient for our prepared speech.

4. Method of hiatus analysis

The prosodic boundary of phrases was segmented with reference to the waveform (speech and EGG) and the spectrogram of wide-band and narrow-band, and then evaluated by listening to the separated accentual phrases.

4.1. Perceptual analysis of phrase

The hiatus we treat is a V-V boundary between adjacent accentual phrases in Japanese. Samples were taken from the Japanese MULTEXT prosodic corpus spoken by a female speaker fhk. The examined phrases consist of 45 phrases producing hiatus of /a/|/a/, /i/|/i/, /u/|/u/, /e/|/e/, /o/|/o/. There is no gap or transition between these two vowels.

4.1.1. Preparation of speech materials

In order to investigate deviations of V-V segment boundary, the following short speech waveforms are prepared. Referring to the hand labeled boundary as a fixed point, a front phrase and a rear phrase are separated and excised for speech materials in a perceptual experiment. The excising points are moved forward and backward from the fixed point with a step width of one vocal cord vibration period up to 5 periods (vertical lines in Figure 1 like pitch marks). As a result, it amounted to 11 speech sounds for each side, to a total of 22 speech sounds per hiatus.

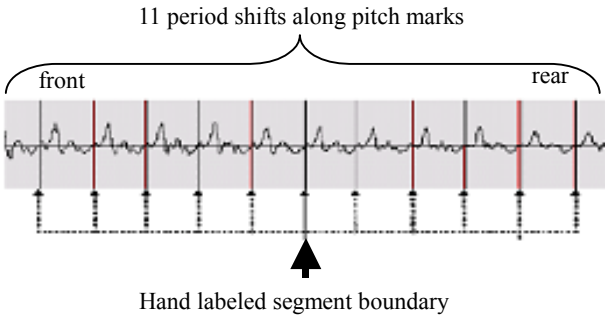

Figure 1: Pitch marks at zero-crossings with trimming for evaluation of perceptual sharpness of either cut. Trimming is done from the right side for front parts of the shift, and from the left side for rear parts

4.1.2. Phrase listening [6]

Speech sounds are presented in random order for each subject. Subjects were asked to judge the naturalness or the sharpness of each phrase sound, paying special attention to the ending and beginning. Responses were scored on a scale from 5 to 0, with 5 points awarded for natural clear-cut speech, and 0 for utterances appearing completely unnatural or contaminated with the adjacent component. Each answer is scored from +2, +1, 0, -1, -2 accordingly. Subjects' answers are summed and averaged for individual speech materials. The listeners participating in the perceptual experiments were 6 male students and 2 female students.

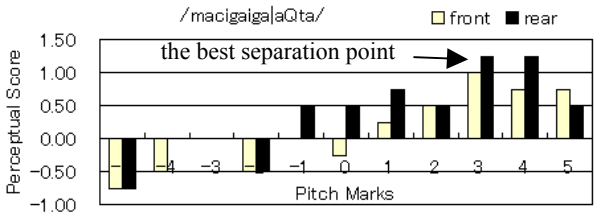

Figure 2: Phrase listening result for */macigaiga\aQta/* "There was some mistake." Perceptual score shows sharpness of either cut and separation of the front (white bar) and the back (black bar) phrases consisting with an */a|a/* hiatus.

4.2. Electroglottography waveform analysis

Electroglottography waveforms were analyzed for the open quotient (as shown in the bottom tier of the Figure 4) and the fundamental frequency (the middle tier in Figure 4) was computed from each glottal cycle using the KAY CSL tool [8]. The open quotient is related to voice quality, i.e., over 50% is harsh voice, 50% is modal voice, and 20-30% is breathy voice. The quotient changes smoothly along time, but abrupt change can be an evidence of glottalization. The fundamental frequency extracted from EGG is an instantaneous F0, i.e., an inverse of the pitch period, drops simultaneously with glottalization as well. This F0 differ from the conventional F0's that are smoothed by a filter.

5. Analysis results of vowel-vowel hiatus

Hiatus resolution is possible based on glottalization in most cases and nasalization in some cases. And the resultant lengthening of phrase-final vowel and shortening phrase-initial vowel is common.

5.1. Hiatus observed in phrasal tone

An intermediate phrase is composed as a chunking of several (only rarely more than three) accentual phrases in Japanese. An intermediate phrase boundary is often marked by a pause or *pseudo-pause*. Also, F0 declining characteristic of the intermediate phrase, however, is known as *catathesis*. We sometimes call this a *pitch-reset*.

Figure 3 shows an accent phrase *iQtai* "in the world" is emphasized, while *icunji* "when?" is focused and emphasized. A pitch-reset is observable in the succeeding phrase, and the pitch-range is extended. Perceptual results show that there is a clear hiatus at +2 period backward with scores 1.0 for the front phrase and 1.25 for the rear phrase. The F0 has began to increase already at the end of the former phrase /i/. A dip in open quotient indicates glottalization. Lengthening at the phrase ending as well as shortening at the phrase start up was not observed because the rear phrase was emphasized.

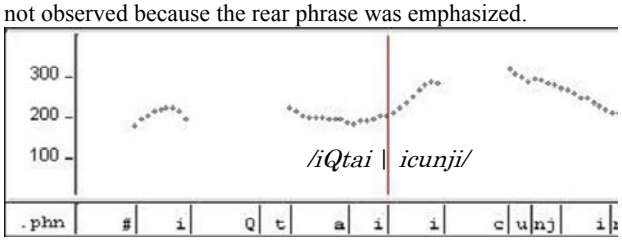

Figure 3: An example "Pitch Reset" from */iQtai | icunji/* "When in the world…?".

5.2. Phrase-initial glottalization

In most cases, the phrase-initial vowel is stressed on its phonation by glottalization. This is also true in cases where the preceding phrase-ending vowel is the same as the following phrase-initial vowel. An example is shown in Figure 4 below, showing that the EGG open quotient goes up once to 65% then decreases down to 56% and then goes up again to 61%. The most appropriate point to separate the phrases is before this bottom (8.907s.). Simultaneously, the F0, which is computed from the EGG period, showed lowering: 249 Hz to 195 Hz then 239 Hz movements.

Similar phenomena were observed in other cases in different degrees of prominence. Among 45 hiatuses analyzed, 17 items showed clear glottalization with dip in F0 and open quotient, and 23 items showed weak glottalization accompanied with other features, with the remaining 5 items showing phrase-final nasalization. If the phrase-initial word is emphasized, the boundary showed prominent glottalization, while a particle *ga* | *a*-initial phrase and a particle *wo* | *o*-initial phrase depict rather vague features of glottalization if the following phrase is not emphasized. In those cases, the acoustic features are too small in magnitude to detect visually. Although syntactic condition may help to resolve hiatus, probably human perception of glottalization must be far more sensitive than present signal processing.

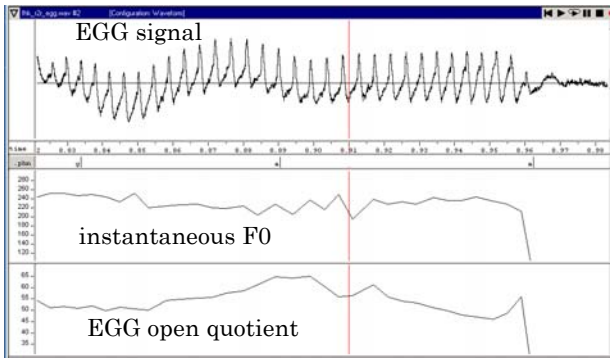

Figure 4: A boundary of a */a|a/* hiatus in */macigaiga | aQta/* "There was some mistake." Perceptual scores were shown in Figure 2.

5.3. Phrase-final nasalization

Switching from nasalization to non-nasalization may sign a perceptual cue, and indicates a segmental boundary by spectrogram texture as relatively lower high frequency energy for nasalized speech. The nasalization contrast is observable even in a continuation of the same vowel. A phrase-final particle *ga* has to be nasalized in Tokyo Japanese. Around the central part of Figure 5., a long vowel /*a*/ in the context of /*ga | aru*/ is shown with a vertical bar that is the best separation point between two phrases. The left part is nasalized (albeit weakly) and the right part is not nasalized. Although the speaker fhk, as a younger generation in the drift of phonological change, tends to pronounce these *ga* with non-nasals or at least weakly nasalized, spectrographic contrast is not clear.

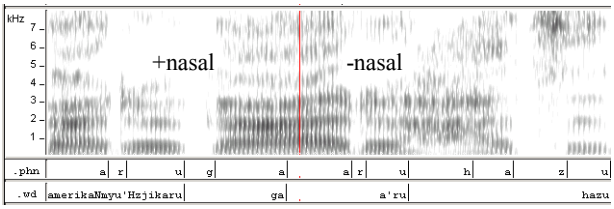

Figure 5: A contrast of *nasalized+/-* at the phrase boundary of */myuHzjkaruga-aruhazu/* "There must be a musical program".

Nasalization is helpful to detect hiatus since the preceding phrase-final nasalization ends and the following phrase starts without nasalization. This +/-nasal contrast as well as glottalization resolves many hiatus (in our example, +/-nasal contrast alone was 4 cases and 12 cases accompanied with glottalization as well as nasal contrast).

5.4. Segmental duration analysis

A phrase ending mora is lengthened to indicate the end of a phrase, while a phrase initiating mora is shortened in order to catch up with the isosyllabic mora timing. A V-V sequence of the same vowels has duration of about two morae, however, the boundary is usually searched in the right half region.

Statistics of our 45 hiatus, the ratio of duration of vowel segment in the preceding phrase-final position to the following phrase-initial position was 1.7 in average. In cases the following word is emphasized, the preceding vowel is not lengthened, but the following vowel keeps normal duration, then the ratio becomes as low as 0.76 for example.

Figure 7 plots 45 hiatuses examined along their measured durations of individual vowels in pairs of the frontal and the following. In reference to the diagonal line, which shows the equal duration, the figure suggests that the former (phrase-ending) vowel is usually longer than the following (phrase-initial) vowel. In a few cases, the following vowel is a little longer than the former vowel, and the former vowel is usually shortened.

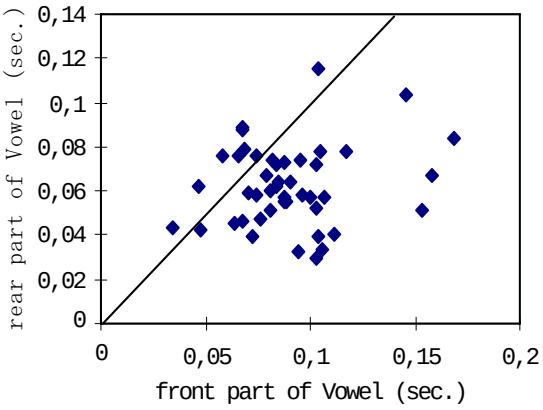

Figure 6: Durations of vowels in hiatus. Diagonal line, being the equal duration of connected vowels, suggests phrase-final lengthening and phrase-initial shortening.

6. Conclusion

J-ToBI labeled phrase boundaries are examined through perceptual evaluation of disjuncture, i.e. tidiness or flawless perfection. We investigated V-V hiatus by listening to whole phrases. The best perceptual score was obtained in most cases as the maximum perceptual score of a single peak.

A phrase-final particle | a vowel, the most common pattern of V-V hiatus, was found to have the following acoustic features: (1) *ga* | *a, nji* | *i, no* | *o*: these show +/-nasal contrast in the spectrographic pattern, since *ga* is normally nasalized while the following *a* is not nasalized. (2) *wa* | *a, sjika* | *a, te* | *e, to* | *o*: phrase initial vowel is glottalized. This glottalization is observable in F0 drop and a dip in EGG open quotient. (3) In *wo* | *o*:, another frequent pattern, glottalization is not so distinct since EGG open quotient is not stable, but spectral change is also useful.

A phrase-final adjective | a vowel and a word ending a vowel | a word beginning a vowel are cases characterized with stronger glottalization than the above-mentioned cases.

Phrase-initial glottalization observable in EGG open quotient, F0 or period of each cycle, and phrase ending nasalization are all important in resolving the hiatus phenomena.

Duration of vowels, constitutes hiatus, depend on mutual emphasis of the phrases adjacent, however, usually the former is longer than the follower.

The findings in this paper indicate that some small abrupt discontinuity in vocal source generator is sharply sensed by our auditory system to effectively segment phrases, words, and phonemes. Accordingly, speech synthesizers may need to take much more care in their smoothness and discontinuity of the artificial vocal source generator so as to cause effective phonological prosodic signs as well as to prevent unnecessary signs to confuse listeners.

7. Acknowledgements

This research is based on the domain research specific (B) subject number 12132204.

8. References

[1] Kawahara, Shigeto, 2003. On certain type of hiatus resolution in Japanese, *Phonological Studies*, 6, 11-20, ed. Phonological Society of Japan, Tokyo: Kaitakusha.

[2] Venditti, Jennifer J., 2002. The J-ToBI model of Japanese intonation. In S. - A. Jun (ed.) *Prosodic Typology and Transcription: A Unified Approach*. Oxford: Oxford University Press.

[3] Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M., 1996. Glottalization of word-initial vowels as a function of prosodic structure, Journal of Phonetics, 24, 423-444.

[4] Kitazawa Shigeyoshi, Kitamura Tatsuya, Mochiduki Kazuya, and Itoh Toshihiko, 2001. Preliminary Study of Japanese MULTEXT: a Prosodic Corpus. International Conference on Speech Processing, Taejon, Korea, 825-828.

[5] Campione, E., & Veronis, J., 1998. A multilingual prosodic database. 5th International Conference on Spoken Language Processing (ICSLP'98), Sidney, 3163-3166.

[6] Kitazawa Shigeyoshi, Kiriyama Shinya, Itoh Toshihiko, and Yukinori Toyama, 2004. Perceptual Inspection of V-V Juncture in Japanese, SP2004, 349-352.

[7] Maekawa, K., Kikuchi, H., and Igarashi, Y., 2001. "X-JToBI: An Intonation Labeling Scheme for Spontaneous Japanese", Technical Report of IEICE, SP 2001-106, 25-30. (in Japanese)

[8] Instruction Manual Electroglottograph (EGG) Model 4338, Kay Elemetrics Corp., Lincoln Park, NJ 07035-1488 USA (April 1995).