Tonal alignment of spontaneous Tokyo Japanese

石原　健
Takeshi ISHIHARA

Keywords：tonal alignment, segmental anchoring, prosodic structure, spontaneous speech corpus
キーワード：イントネーション、基本周波数パターン、韻律構造、話し言葉コーパス

1　Introduction

In the autosegmental-metrical theory of tone and intonation (Bruce, 1977; Pierrehumbert, 1980; Ladd, 1996, among others), an intonation contour is seen as a sequence of tones which occurs at well-defined locations in the structure. Some of the tones in an utterance are seen to be associated with specific elements of the segmental string such as mora or syllable (and others with the edges of larger domains such as phrase). This relationship is known as *association*, and tones and tone-bearing units on the *tiers* are linked via *association lines*. Association represents temporal relationship between tones and tone-bearing units, so it can be interpreted that tone phonetically occurs during the temporal interval of tone bearing unit. However, it has been reported that the phonetic realization of such association does not simply follow this relationship, and that the temporal synchronization (*alignment*) of F0 events with segmental events may vary in complicated ways across languages. It is pointed out that there are at least two types of complications to the relationship between association and alignment (Ladd, 2003). For one thing, in some languages there appear to be linguistically significant contrasts of alignment. To deal with these alignment differences, bitonal pitch accents are proposed in the autosegmental-metrical theory. A bitonal pitch accent consists of a starred tone which associates with the tone-bearing unit, and an unstarred tone which leads or trails the starred tone. For example, Pierrehumbert (1980) applies L*+H, L+H*, H*+L, H+L* and H*+H (as well as monotonal pitch accents, H* and L*) for the description of American English intonation. An unstarred tone then is considered to be phonetically realized outside the tone bearing unit in languages which have bitonal pitch accents. For another, there is a phenomenon called *peak delay* in which the F0 peak of a pitch accent may be aligned outside (usually after) the stressed syllable with which it is intuitively associated. Peak delay has been widely reported across languages, and how it happens seems different from language to language.

いしはらたけし：目白大学外国語学部英米語学科准教授

Among the alignment studies, earlier studies focused more on factors affecting the temporal alignment of the F0 targets, and some of them attempted to model the extent to which those factors affect the alignment, using a multiple regression model (Silverman and Pierrehumbert, 1990; Caspers and van Heuven, 1993; Prieto et al., 1995). These earlier studies demonstrated that the alignment of F0 targets is influenced by contextual factors—word boundary, phrase boundary, stress clash, tonal crowding and so on. In other words, the F0 peak location is a consequence of the complex interaction between various factors. Although the extent of the effects of these factors differs between the languages, this can be basically regarded as avoiding a clashing situation: shifting the pitch accent earlier to put distance, or lengthening the associated syllable to provide more time (or both).

More recent studies provided various evidence in different languages that the F0 targets of a pitch accent are consistently aligned with a specific point in the segmental string, particularly by Arvaniti, Ladd and their colleagues (Arvaniti et al., 1998; Ladd et al., 1999, 2000; Atterer and Ladd, 2004; Schepman et al., 2006, among others). They revealed that, when the factors affecting pitch accent alignment are properly controlled, both the F0 maxima and minima for pitch accents are consistently aligned with specific segmental landmarks. Moreover, some of the work by Arvaniti, Ladd, and their colleagues share a view that both the beginning and end of a pitch accent are independently anchored at a specific point in the segmental string, which they call ‘segmental anchoring’, and that these anchored F0 turning points (F0 maxima and minima) mainly contribute to shaping an F0 movement of an utterance.

In Japanese, there is a well-known phenomenon called *ososagari* (‘late fall’) in which the beginning of the F0 fall (i.e. the F0 peak) for a pitch accent occurs after the end of the associated mora (Neustupný, 1966; Sugito, 1982). Ishihara (2006) investigated tonal alignment in Tokyo Japanese (including *ososagari*), and demonstrated consistent alignment of the F0 targets with specific places in the prosodic structure in a language-specific way, which were rather resistant to changes caused by differences of speaking mode. The overall results of Ishihara (2006) indicated that both the F0 valley and peak were consistently aligned with specific segmental landmarks, and the alignment of the F0 peak depended on the syllable/mora structure of the accented syllable. Thus, the accentual F0 peak was aligned with as follows:

1. The beginning of the vowel of the syllable following the accented light syllable
2. The end of the first mora of the accented syllable with a coda nasal
3. About 70 percent of the two-mora vowel of the accented syllable with a long vowel or a diphthong.

For all the structures, the beginning of the second mora is the segmental landmark for the F0 peak.

Moreover, in terms of the organization of the mora and syllable in Tokyo Japanese, it seems feasible to interpret these alignment patterns as a consequence of segmental anchoring. According to one of the proposals of moraic phonology (Hayes, 1989), onset consonants are attached directly to a syllable node, rather than to a mora, in the organization of the prosodic structure (see Figure 1). As suggested in Ishihara (2006), it can be claimed that the accentual F0 peak is anchored to the beginning of the second mora across all the syllable/mora sequences. This may lead to a more precise description of *ososagari* (i.e. peak delay in Japanese). That is, it is one of the consequences of the anchoring of a pitch accent to the beginning of the second mora. These alignment patterns were seen across the speakers and regardless of the durational variation of the target segments.

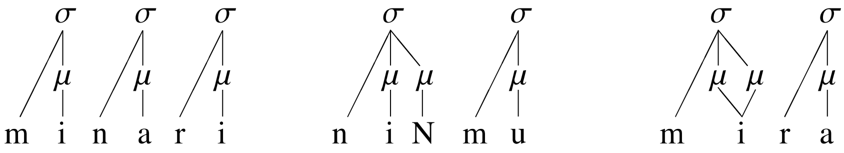

Figure 1: Examples of words with different syllable/mora structures. ' σ ' stands for a syllable, and 'μ' for a mora.

The purpose of this study is to investigate how the F0 peak of pitch accent is aligned with segmental string in spontaneous speech. The alignment patterns reported in Ishihara (2006) were based on lab speech data, which were obtained in controlled and less natural experimental settings. On the other hand, the patterns were found in the data of different speaking styles, which supported the alignment consistencies. In the present study, with spontaneous speech data, which were collected in more natural settings, we can examine if there are similar alignment patterns of pitch accent.

2 Methods

The analysis was conducted with Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003) which contains about 662 hours (about 7.5 million words) of various speech recordings by Japanese native speakers. Part of CSJ (called ‘core data’) was taken from Tokyo Japanese speakers. It was first automatically segmented and labeled, and then manually corrected by experienced labelers trained by phoneticians. This subset of CSJ was analyzed for this study.

Annotation and acoustic measurements were performed using Praat. The segmentation was carried out manually on the basis of the visual display of the oscillogram and the spectrogram with a controlled cursor. If there were two candidates for a segmented point, the earlier was always taken. The segmentation points were basically marked at zero in amplitude (zero crossing), except for cases like the release of voiced stops. When it was hard to place a boundary in the waveform, I relied on the wide-band spectrogram. Target F0 minima and maxima were located via parabolic interpolations between selected portions around them.

```
(1)      |   C   |   V   |   C   |   V              MI.na.ri ’dress’
         C0      V0      C1      V1

(2)      |   C   |   V   |   N    C   |   V.        NIN.mu ’duty’
         C0      V0      C1           V1

(3)      |   C   |   V    V   |   C   |   V.        MII.ra ’mummy’
         C0      V0           C1      V1
```

Figure 2: Labelling scheme of target sequences. A vertical line (‘|’) is a segmented point. The letters below the vertical line are the labels. The first segment of a target sequence starts from zero (e.g. ‘C0’). The first sequence is for CV.CV; the second for CVN; the third for CVV (a long vowel or a diphthong).

Figure 2 shows the labelling scheme of target sequences. Target sequences were labelled at the beginning—i.e. point, not section—of each segment, and the first segment of a target sequence was labelled starting from zero. So the beginnings of the first consonant and the first vowel of a target sequence were labelled ‘C0’ and ‘V0’, respectively. After the labelling, semi-automatic data extraction was performed on the duration of portions of interest and the annotated F0 maxima and minima, with the aid of Praat scripts. Twenty-five data points for each group were selected randomly from the annotated data and used for statistical analysis. The analysis of the alignment patterns were performed based on those found in Ishihara (2006). Thus, the alignment of accentual F0 maxima relative to C1, V1 and the onset of the second mora of the target word, was examined across different segmental sequences (CV.CV, CVN and CVV).

3 Results and discussion

The following analyses aim to investigate where the accentual F0 peak was aligned across different segmental sequences. According to the findings of Ishihara (2006), the accentual F0 peak is expected to be closely aligned with V1 in CV.CV, while, in CVN, it is expected to be aligned with C1. In CVV, it is expected to occur in the middle of the diphthong or the long vowel.

Figure 3 shows the alignment of H to C1. The F0 peak was on average aligned 48 ms after C1 in CV.CV; 13 ms after C1 in the CVN cases; 81ms before C1 in CVV. The data were analyzed with a one-way ANOVA, with items as the random factor and structure (structures of the target syllable) as a single within-items factor. The ANOVA showed that there was a statistically significant difference between target sequences beyond the 1% level: $F(2, 72) = 39.72$; $p < 0.001$.

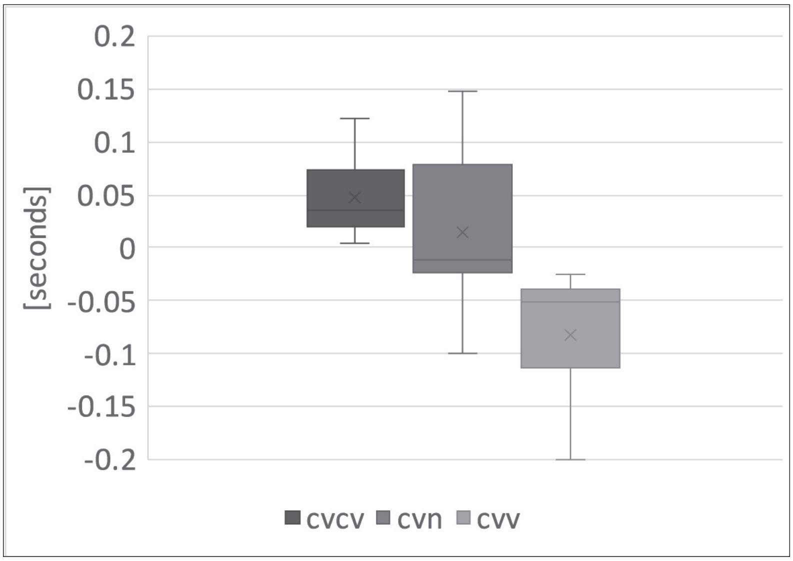

Figure 3: Mean duration from C1 to H (in seconds). 'C1' is the beginning of the second consonant of the target sequence. The value zero in the graph amounts to C1.

Figure 4 shows the alignment of H to V1. The F0 peak was on average aligned 0.02 ms after V1 in CV.CV; 117 ms before V1 in the CVN cases; 81 ms before V1 in CVV. The data were analyzed with a one-way ANOVA, with items as the random factor and structure (structures of the target syllable) as a single within-items factor. The ANOVA showed that there was a statistically significant difference between target sequences beyond the 1% level: $F\ (2,\ 72) = 33.18$; $p < 0.001$.

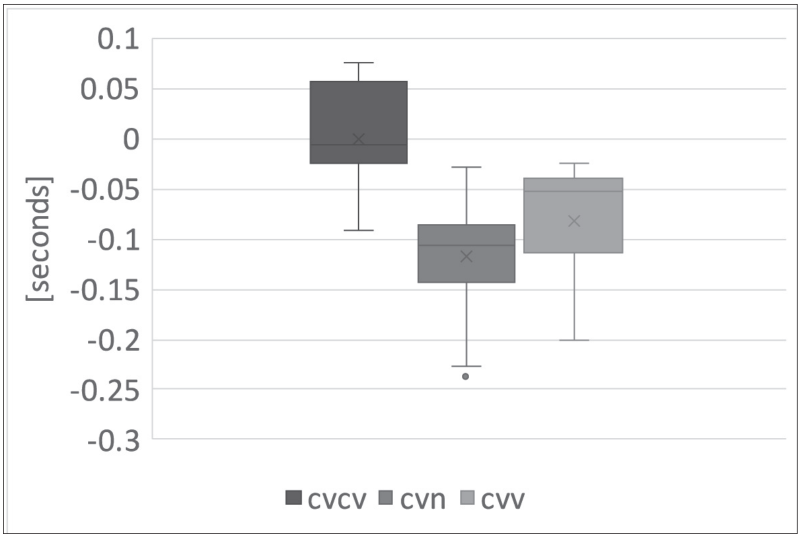

Figure 4: Mean duration from V1 to H (in seconds). 'V1' is the beginning of the vowel of the syllable following the accented syllable. The value zero in the graph amounts to V1.

The data of the accentual F0 peak alignment relative to the two segmental landmarks (C1 and V1) demonstrated similar regularities reported in Ishihara (2006). Since it was proposed by Ishihara (2006) that the accentual F0 peak is anchored to the beginning of the mora following the accented syllable, another analysis of the data was conducted in order to examine how the peak is aligned with proposed anchoring points in different segmental structures. Figure 5 shows the alignment of H to the onset of the second mora of the target word: V1 for CV.CV; C1 for CVN; the middle of the diphthong or the long vowel for CVV. The F0 peak was on average aligned 35 ms after the reference point in CV.CV; 13 ms after in CVN; 34 ms after in CVN. The data were analyzed with a one-way ANOVA, with items as the random factor and Structure as a single within-items factor. There was no significant difference between the structures.

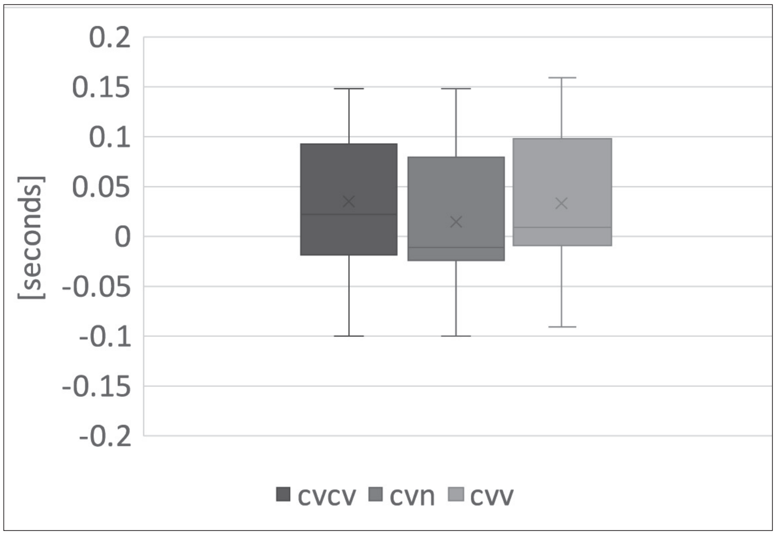

Figure 5: Mean duration from the onset of the second mora to H (in seconds). Refer to the text for detailed explanations.

The three kinds of data described above revealed where the accentual F0 peak was aligned with depending on the different segmental structures. It was aligned with the beginning of the vowel of the syllable following the accented syllable for CV.CV; with the end of the first mora of the accented syllable for CVN; and with the middle of the two-mora vowel of the accented syllable for CVV. The overall results thus indicated that the accentual F0 peak was consistently aligned with a specific segmental landmark based on the prosodic structure of the target words.

Although the current study with a spontaneous speech corpus showed similar alignment patterns to those found in Ishihara (2006), since the data used was a subset of the large corpus, it is necessary to explore it further in order to gain a better picture of F0 alignment regularity.

【References】

Amalia Arvaniti, D. Robert Ladd, and Ineke Mennen. Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, 26: 3–25, 1998.

Michaela Atterer and D. Robert Ladd. On the phonetics and phonology of "segmental anchoring" of *F0*: evidence from German. *Journal of Phonetics*, 32: 177–197, 2004.

Gösta Bruce. *Swedish Word Accents in Sentence Perspective*. Gleerup, Lund, 1977.

Johanneke Caspers and Vincent J. van Heuven. Effects of time pressure on the phonetic realisation of the accent-lending pitch rise and fall. *Phonetica*, 50: 161–71, 1993.

Bruce Hayes. Compensatory lengthening in moraic phonology. *Linguistic Inquiry*, 20: 253–306, 1989.

Takeshi Ishihara. *Tonal Aligment in Tokyo Japanese*. PhD thesis, University of Edinburgh, 2006.

D. Robert Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, England, 1996.

D. Robert Ladd. Phonological conditioning of F0 target alignment. In Maria-Josep Solé, Daniel Recasens, and Joaquín Romero, editors, *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 249–252, Barcelona, Spain, August 2003. Causal Productions.

D. Robert Ladd, D. Faulkner, H. Faulkner, and Astrid Schepman. Constant segmental anchoring of *F*0 movements under changes in speech rate. *Journal of the Acoustical Society of America*, 106: 1543–54, 1999.

D. Robert Ladd, Ineke Mennen, and Astrid Schepman. Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America*, 107: 2685–96, 2000.

K. Maekawa. Corpus of Spontaneous Japanese: its design and evaluation. *Proceedings of The ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pages 7–12, 2003.

Jiri Vaclav Neustupný. Is the Japanese accent a pitch accent? *Onsei Gakkai Kaihoo*, 121: 1–7, 1966. (In Japanese).

Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, 1980.

Pilar Prieto, Jan P. H. van Santen, and Julia Hirschberg. Tonal alignment patterns in Spanish. *Journal of Phonetics*, 23: 429–51, 1995.

Astrid Schepman, Robin Lickley, and D. Robert Ladd. Effects of vowel length and "right context" on the alignment of Dutch nuclear accents. *Journal of Phonetics*, 34(1): 1–28, 2006.

Kim Silverman and Janet B. Pierrehumbert. The timing of prenuclear high accents in English. In John Kingston and Mary E. Beckman, editors, *Between the Grammar and Physics of Speech*, Papers in Laboratory Phonology I, pages 71–106. Cambridge University Press, Cambridge, 1990.

Miyoko Sugito. *Nihongo Akusento no Kenkyuu [Studies on Japanese Accent]*. Sanseido, Tokyo, 1982.

（平成31年1月25日受理）