X-JTOBI: AN EXTENDED J-ToBI FOR SPONTANEOUS SPEECH

Kikuo Maekawa*, Hideaki Kikuchi*†, Yosuke Igarashi**, and Jennifer Venditti***

* National Institute for Japanese Language, †Waseda University
** Graduate School, Tokyo University of Foreign Studies
*** Institute for Research in Cognitive Science, Univ. of Pennsylvania
{kikuo,kikuchi}@kokken.go.jp

ABSTRACT

We report on the new X-JTOBI prosodic labeling scheme, the eXtended version of J_ToBI which has grown out of our work on annotating prosodic features of spontaneous speech. Among the new characteristics of X-JTOBI are 1) Exact match between the time-stamp of tone labels and the timing of physical events, 2) Enlargement of the inventory of boundary pitch movements, 3) Extension and ramification of the usage of break indices, and 4) Newly defined labels for filled-pause and non-lexical prominence.

1. INTRODUCTION

The study of spontaneous speech requires large database because spontaneous speech (henceforth 'SS') is inherently more variable than read or laboratory speech. Such a database must be annotated with segmental and intonational labels.

In an attempt to compile the *Corpus of Spontaneous Japanese* (henceforth 'CSJ'), a large-scale database of spontaneous Japanese of about 7 million words [1], we plan to provide phonetic annotations for a subset of the database containing about 500,000 words or 45 hours of speech. The aim of the present paper is to report the new intonation-labeling scheme that we devised for the work. We call it X-JTOBI, because we regard this to be a natural extension of the J_ToBI scheme [2].

2. MOTIVATIONS

There are three motivations for X-JTOBI. First, it provides full access to the time information associated with all the tone labels. This issue will be discussed in the next section. Another motivation is that there are SS-specific phenomena that we cannot treat in a satisfactory manner in the current J_ToBI scheme; filled pauses (fillers), word fragments, repairs, and false starts are some examples. In addition, there are also phenomena that are clearly beyond the coverage of the underlying theory of the J_ToBI system (See 3.6).

3. NEW FEATURES

3.1 Match between the tone label and physical event

In the J_ToBI convention, the time location of a tone label is not necessarily simultaneous with the physical event, i.e. a peak, valley, or other F0 contour characteristics. The J_ToBI convention is to mark accent and phrasal tone labels within the vowel to which the tones are aligned phonologically. When the corresponding physical events do not occur in the vowel segments, the timing of physical event are shown by the labels '<' and '>'; early and late peak respectively. All boundary tone labels are placed in phrase final position, regardless of the complexity of the tones. As a consequence, a single time-point is given to a complex tone like L%H%.

In X-JTOBI, on the other hand, maximum match between the label location and physical event is pursued. In principle, all tone labels (phrase tone, accent, and boundary tones) are located at the place where the corresponding F0 events occur. Concomitantly, early and late peak labels are repealed; we use the '>' label for a different purpose, instead (See 3.2.2). The repeal is possible because segment labels are an integral part of the X-JTOBI labeling. The phonological location of phrase tones and accents can be obtained from a combination of lexical-accent information in the word tier and segment labels.

Needless to say, providing segment labels is not an easy task, but they are indispensable for detailed phonetic study. Also, the segmentation task can be automated, fully or partially, by use of the Hidden Markov Model-based speech alignment algorithm commonly used in today's speech recognition systems.

3.2 Pointers and extenders

In X-JTOBI, two new labels are introduced: *pointers* and *extenders*.

3.2.1 Pointers

Table 1 is the inventory of phrase-final boundary tones and their corresponding X-JTOBI labels. 'pH' and 'pL' in Table 1 are newly introduced labels called *pointers*. Pointers are used to mark the second tonal element of the tri-tonal and the second and third elements of the quadri-tonal boundary tones. This convention is introduced for the ease of database query; if we were to time-decompose all three elements, a complicated query would be needed to distinguish between the simple L% and the last element of L%HL%, for example.

Table 1. List of the phrase-final boundary tones and the corresponding X-JTOBI labels.

Type	X-JTOBI Labels
L%	L%
L%H%	L%, H%
L%HL%	L%, pH, HL%
L%LH%	L%, pL, LH%
L%HLH%	L%, pH, pL, HLH%

Figure 1 shows how decomposed labels are located on the time axis. We can tell that the Boundary Pitch Movements (henceforth 'BPMs') used in these utterances are tri-tonal just by looking at the last tone labels; also the exact time location of F0 change can be determined by looking the time stamps associated with the two preceding tone labels in the tone tier (e.g. L% and pH in panel A).
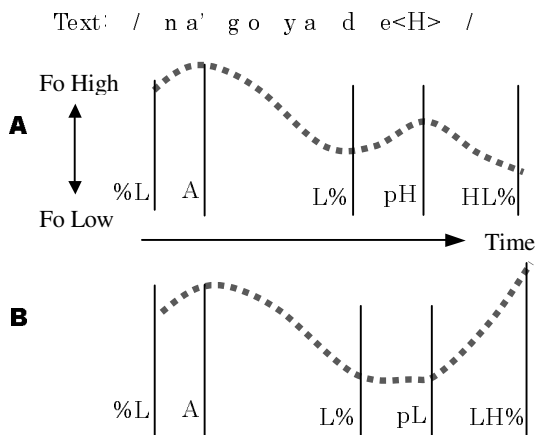

Figure 1. Schematic examples of the tone labeling of phrase-final complex boundary tones. Pointers (pH and pL) are used to denote the time-points of H and L of HL% and LH% respectively. Tone label 'A' stands for the lexical accent, which is represented by an apostrophe in the text. <H> stands for the prolongation of vowel. The text means '*at the city of Nagoya*'.

3.2.2. Extender

Extender ('>') is a label used to express variants of boundary tones in terms of prolongation at both edges of a phrase. Figure 2 shows examples. In panel A, extender '>' is used to show the prolongation of phrase initial %L before the rise up to the accentual peak. In panel B, extender is used to show the prolongation of the phrase final H%.

Extender is used always with a boundary tone. In Figure 2, %L and H% are located at the beginning of the prolonged tones, and extenders mark the ends of the prolongation.

Prolongations shown in Figure 2 are not mere phonetic variation. They convey expressive, or paralinguistic, meaning ('suspicion/disbelief', in the case of panel A) or discourse function (in the case of panel B). See [3-5] for paralinguistic meaning of intonation in Japanese.

3.2.3 Repeal of strong vs. weak %L

The phonological distinction between strong vs. weak %L at the beginning of an accentual phrase (henceforth 'AP') has been repealed for two reasons. First, the distinction is completely predictable from the syllable weight and/or accent location of the initial syllable of an AP. Second, in spontaneous or expressive speech, the strong %L (i.e. an ordinary initial L tone) appears in the context where the weak %L is supposed to be. Panel A of Figure 2 is an example. In this panel, the F0 at the beginning is quite low (i.e. 'strong'), even though the AP is initially accented and predicted to have a weak %L. This is because the utterance conveys the speaker's attitude of 'suspicion/disbelief' [3,5].
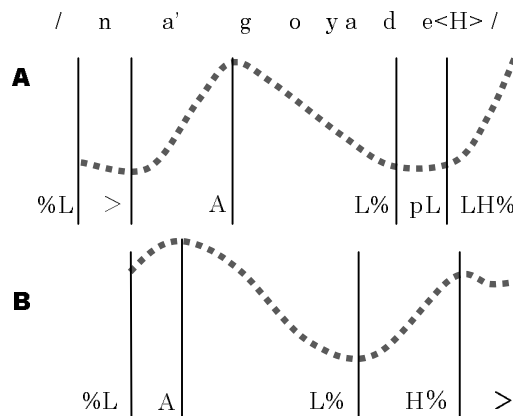

Figure 2. Examples of the usage of extender at the utterance initial (A) and utterance final positions (B). The rendition A is a typical contour of suspicion/disbelief.

3.3 Break Indices

The selection of break indices, or BI, is the most difficult task in SS labeling. Inter-labeler agreement decreases because the canonical prosodic structure is obscured by factors like fragmental and/or agrammatical sentence structure, frequent use of BPMs at the end of nearly every AP, AP-internal filled pauses, and so on.

To cope with these issues, X-JToBI introduces two devices in the BI labeling. First, intermediate values between two successive integers are allowed. The label '*n*+' (where *n* is either 1,2, or 3) stands for the intermediate strength between BI=*n* and BI=*n*+1. This alone is not good enough, however, because labelers can feel an intermediate boundary strength for various reasons. One way to deal with this problem is to use additional diacritics to indicate the reason why an intermediate strength was chosen. Table 2 shows the inventory of all X-JToBI break indices. The letters coupled with the '*n*+' labels show the reason to choose the intermediate BI value.

Figure 3 shows the typical cases of '1+p' and '2+b'. The prosodic boundary marked with '1+p' is stronger than the usual word boundary (BI=1), because there is an AP-internal pause between the penultimate mora (/ri/) and the final mora (/to/) of the AP (This pause insertion happens frequently in Japanese monologue especially when the last mora of the AP is a particle and bears prominence).

The '2+b' label is used when there is a BPM at the end of an AP, but the AP is not followed by a pause and there is no pitch resetting between the two consecutive APs. This boundary is stronger than the usual AP boundary (BI=2) because it ends with a BPM, but weaker than the usual intonation phrase boundary (BI=3) because the resetting of pitch range does not accompany it.

BI labels marking intermediate values without a 'reason' (such as plain '1+' or '2+') are not recommended, but are allowed nonetheless, because the 'reasons' listed in Table 2 are by no means exhaustive. When using these labels for reasons not listed, the labeler should note his/her reason in the miscellaneous tier.

It is impossible to explain here all the label combinations listed in Table 2, though some of them will be explained in the following sections (See our forthcoming X-JToBI

Reference Manual for details of each label).

Figure 3 shows also the example of 'x' diacritic. When the F0 value corresponding to a tone is missing for reasons like vowel devoicing, the tone label is followed by this diacritic. There is also another tone diacritic '?' which is applied when the corresponding F0 is existing but unreliable.

Table 2. Inventory of the X-JToBI BI labels. OP and NR in the remarks column stand for Optional and Not Recommended respectively

| Label	Usage	Remark
0	Same as the J_ToBI usage	
1	Same as the J_ToBI usage	
1+	Between 1 and 2	NR
1+w	Compound word internal prosodic boundary	OP
1+p	Accentual phrase internal pause	
2	Same as the J_ToBI usage	
2+	Between 2 and 3	NR
2+p	No BPM at the end & Followed by a pause & Without pitch range resetting	
2+b	BPM at the AP end & No following pause & Without pitch range resetting	
2+bp	BPM at the AP end & Followed by a pause & Without pitch range resetting	
3	Same as the J_ToBI usage	
3+	Identical to the "finality" of the J_ToBI	OP
D	Word fragment without pitch range resetting	
D+	Word fragment with pitch range resetting	OP
P	Word internal pause without pitch range resetting	
P+	Word internal pause with pitch range resetting	OP
PB	Consecutive BPMs	
<F	Beginning of a prosodic filler	
F	End of a prosodic filler	
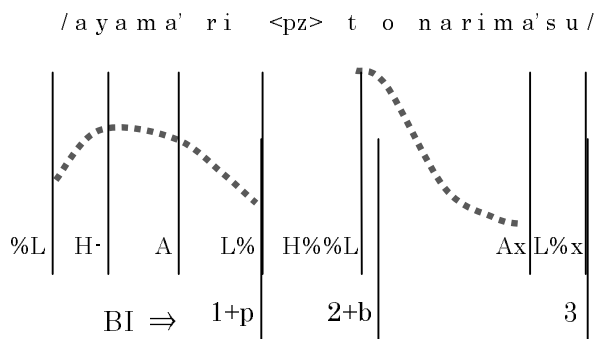
Figure 3. Example of the usage of '1+p' and '2+b'. <pz> is a perceptible pause inserted before the final mora of the first AP (/ayama'rito/) that ends with L%H% boundary tone. Fo values corresponding to "A" and "L%" at the phrase end are lacking due to vowel devoicing. The text means '*It causes an error*'; /ayama'ri/ 'error', /to/ particle, /narima'su/ 'result in'.

3.4 Disfluency
Disfluency is the most remarkable characteristic of SS. In the X-JToBI BI system, capital letter symbols 'P', 'D', and 'F' are used to mark various disfluency phenomena.

3.4.1 Word internal pause and word fragment
'P' is used to mark the disfluency resulting in a word internal pause. If a word is separated into two or more portions by an internal pause, the end of each portion is marked with 'P' in the BI tier. And if there is a resetting of pitch range between the two consecutive portions, 'P+' is used instead.

'D' is used to mark word fragments, i.e. cases where speaker stops speaking in the middle of a word and does not completed the word. 'D+' is for the cases where pitch resetting is observed after the fragment.

3.4.2 Prosodic filler
In the X-JToBI, filled pauses, or 'fillers', are sometimes treated as normal APs, and sometimes treated separately. Japanese fillers can be classified into two classes from a prosodic point of view. Some fillers have the textual properties of fillers (e.g. /eH/, /aH/, /maH/, etc. Note /H/ stands for a long vowel), but show the full characteristics of an ordinary AP, i.e. the phrase initial pitch rise and/or accent-like local pitch fall. These fillers are labeled as ordinary APs. On the other hand, there are also fillers that are quite monotone in pitch. If a filler does not show neither a pitch rise at its beginning nor a local pitch fall anywhere, we consider it a 'prosodic filler'.

The beginning and end of a prosodic filler are demarcated in the BI tier by '<F' and 'F' respectively. Also, the overall pitch height is judged as either high or low and labeled as 'FH' (filler- high) or 'FL' (filler-low) in the tone tier.

3.5 Non-lexical prominence
By non-lexical prominence we mean a local pitch peak that is not a realization of lexical accent. For example the F0 contour of the text /de'Hta ga/ (*data* followed by the nominative particle) can have two peaks: one on /de'H/ and one on /ta/. The first peak is the realization of the lexically-specified accent, but the second peak is not specified by the lexicon. In addition, this is not a case of an ordinary BPM in that the peak does not occur on the final mora.

In the X-JToBI, this kind of prominence is treated as the special case of the L%HL% BPM, where the pitch peak location (i.e. the location of 'pH') is shifted from the last mora to the penultimate mora. No special tone label is used in the tone tier, but the label 'PNLP' (*Penult Non-Lexical Prominence*) is added in the miscellaneous tier to show the difference from the ordinary BPM. Whether non-lexical prominence can occur in positions earlier than the penultimate mora is an interesting issue, but we haven't observed any such case so far.

3.6 Parasitic boundary
In SS, there are cases where it seems that two BPMs occur consecutively at the end of an AP. One typical case involves quotation: a quotation phrase ending with a BPM (L%H%, for example) can be followed by a particle having its own

BPM.

This type of intonation is interpreted in X-JToBI as having a "*parasitic*" boundary before the main boundary of the same AP. Figure 4 shows an example of a WH question /da'rega/ (Who?) is followed by the quotation particle /to/ and the main verb /iQta/ ('(he) said'). The end of the quoted portion has a L%H% BPM, and the quotation particle carries a H%. The H% of the quote belongs to the 'PB' boundary while the L% in the quote and the last H% make an authentic L%H% that belongs to the '2+b' boundary.
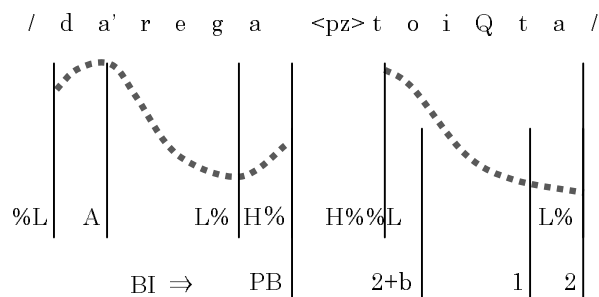

Figure 4. Example of 'parasitic boundary'. The 'PB' label is used for succession of two BPMs and demarcates the end of the first one.

4. SAMPLE LABELING OF CSJ

Table 3 shows the frequency of phrase-final boundary tones in two types of monologue: academic presentation speech (APS) and simulated public speech (SPS), both recorded for *CSJ* [1]. The total amount of speech (excluding pauses) in the APS and SPS styles is 1.02 and 2.75 hours respectively. An expert labeler labeled all samples.

It is interesting to see that speakers use L%H% more frequently in APS than in SPS, presumably to signal the so-called 'continuation rise'. It is also noteworthy that L%LH%, a newly introduced BPM, appears across speech types, albeit in much smaller numbers.

Table 3. Frequency of phrase-final boundary tones in simulated public speech (SPS) and academic presentation speech (APS). Percentage is shown in parentheses.

	L%	L%H%	L%HL%	L%LH%	Total
SPS	3141 (81.7)	437 (11.4)	232 (6.0)	33 (0.9)	3843 (100)
APS	3402 (66.8)	1331 (26.1)	346 (6.8)	12 (0.2)	5091 (100)

Table 4 compares the frequency of the X-JToBI BI labels in SPS and APS. The frequency of '0', '1+w', and '3+' are not shown because we do not currently use these labels in the labeling of CSJ.

The most salient difference between the two speech styles is the higher relative frequency of '2+b' in APS. This label marks cases in which downstep is continued across one or more prosodic boundary marked with BPM(s). This particular intonation is found in the presentation of debutant researchers who seem to memorize his/her entire talk.

Another interesting difference is the higher frequency of

'F' and '<F' in APS. This means that APS contains relatively more prosodic fillers than SPS. In addition the ratio of '<F' to 'F' is higher in APS than SPS. This suggests that in the APS relatively more fillers occur either utterance-internally or finally, because the '<F' label is used to mark the beginning of prosodic fillers that occur after (or in between) accentual phrases, and is not used for post-pausal fillers.

Finally, the newly introduced BI labels occurred 4299 (8.26%) and 2502 (17.17%) times in SPS and APS respectively. This suggests that the new X-JToBI scheme is able to capture and distinguish certain SS phenomena that the old J_ToBI scheme could not handle.

At the same time, however, there still remain some unrecommended labels (i.e., '1+' and '2+'). The analysis of the reasons for these cases may require further extension of the current labeling scheme.

Table 4. Frequency of the X-JToBI BI labels. Numbers in parentheses indicate percentages to the total number of BIs.

Break Indices	SPS N (%)	APS N (%)
1	36196 (69.53)	8121 (55.73)
1+	42 (0.12)	9 (0.06)
1+p	254 (0.70)	164 (1.13)
2	4071 (7.82)	1333 (9.15)
2+	15 (0.03)	9 (0.06)
2+p	214 (0.41)	67 (0.46)
2+b	466 (0.90)	526 (3.61)
2+bp	48 (0.09)	8 (0.05)
3	7478 (14.37)	2617 (17.96)
D	280 (0.54)	97 (0.67)
D+	31 (0.07)	26 (0.18)
P	19 (0.04)	6 (0.04)
P+	12 (0.02)	5 (0.03)
PB	61 (0.06)	10 (0.07)
<F	431 (0.83)	464 (3.18)
F	2426 (4.66)	1111 (7.62)
Total	52047 (100.0)	14573 (100.0)

REFERENCES

[1] Maekawa, K. et al. (2000) "Spontaneous speech corpus of Japanese." *Proc. LREC2000*, Athens, 2, 947-952.

[2] Venditti, J. (1997). "Japanese ToBI Labeling Guidelines." *OSU Working Papers in Linguistics*, 50, 127-162. (http://www.ling. ohio-state.edu/ phonetics/J_ToBI/).

[3] Kawakami, S. (1963) "Bunmatsunadono jooshoochooni tsuite." *Kokugo Kenkyu*, 16, 25-46.

[4] Maeda, K. & J. Venditti (1998) "Phonetic investigation of boundary pitch movements in Japanese." In *Proc. ICSLP1998*, Sydney, Australia.

[5] Maekawa & Kitagawa (2002) "How does speech transmit paralinguistic information?" *Cognitive Studies*, 9-1, 46-66.