Evaluation of the effectiveness of "X-JToBI": A new prosodic labeling scheme for spontaneous Japanese speech

KIKUCHI Hideaki†‡ and **MAEKAWA Kikuo**†

† National Institute for Japanese Language
‡ Waseda University
{kikuchi,kikuo}@kokken.go.jp

ABSTRACT

Formerly we proposed a new intonation labeling scheme called X-JToBI. There were two reasons that we needed a new scheme: a) to describe prosodic features more accurately and b) to improve inter-labeler reliability. In this paper, we mainly discuss the verification of improvement of inter-labeler reliability by comparing the results of X-JToBI labeling with that of J_ToBI. An analysis of inter-labeler reliability using Cohen's kappa showed that kappa was higher for X-JToBI in all tone labels including boundary pitch movement, accent, and phrasal tone. Exact matching between the time-stamp of tone labels and the timing of physical events helps improve inter-labeler agreement. We also found that kappa for X-JToBI was higher than for J_ToBI in all of break index labels. However, there was no significant difference when the labels for disfluencies and filled-pauses were excluded. This suggests the effectiveness of newly introduced tone and BI labels for filled-pauses and various disfluencies. The observed rate of overall agreement rose from 75% to 88% overall.

1 INTRODUCTION

Study of spontaneous speech requires large database because spontaneous speech is inherently more variable than read or laboratory speech. Such database must be annotated with segmental and prosodic labels. Since 1999, the authors have been involved in the compilation of a large-scale corpus of spontaneous speech known as the *CSJ*, or Corpus of Spontaneous Japanese [1]. This corpus involves the digitized speech, transcribed speech, POS annotation of about 650 hour spontaneous speech corresponding to about 7 million words. In addition, we will provide segmental labels and prosodic labels for a true subset of the *CSJ*, called the Core, containing about 45 hour speech, or 500 thousand words.

It is generally believed that labeling spontaneous speech is more difficult than labeling read speech, because of wider variety of acoustic and linguistic features. The aim of this paper consists in making report of problems in labeling spontaneous speech and verification of effectiveness of our solutions to the problems.

According to our pilot experiment, inter-labeler reliability of J_ToBI[2] lowered considerably when the scheme was applied to spontaneous speech. To solve this problem, we proposed a new prosodic labeling scheme named X-JToBI[3], the extended version of J_ToBI.

Generally speaking, in designing a new labeling scheme, it is important to assess the scheme objectively from various points of view. Having applied the X-JToBI labels to spontaneous speech of more than 20 hours, we now believe that the new scheme has higher reliability compared to the traditional one. This paper reports verification of effectiveness of X-JToBI in terms of its correctness, stability, and reproducibility. After presenting the new features of X-JToBI, frequency of newly introduced labels in the *CSJ* will be reported. Then, accuracy and inter-labeler reliability will be reported based upon the results of experiments. Finally, reproducibility of F0 from prosodic labels will be discussed.

2 X-JToBI

There are three motivations for X-JToBI. First, in J_ToBI, it is not always possible to get exact time information of tones especially in the analysis of complex boundary pitch movements (BPM, hereafter). The second motivation is that there are spontaneous speech specific phenomena that we cannot treat in a satisfactory manner in the J_ToBI; filled-pauses, word fragments, repairs, and false starts are some examples. Lastly, there are also some phenomena that are clearly beyond the coverage of the underlying theory of the J_ToBI system, which presupposes a strictly hierarchical structure of intonation.

Among the new characteristics of X-JToBI are 1) Exact match between the time-stamp of tone labels and the timing of corresponding F0 events, 2) Enlargement of the inventory of BPMs, 3) Extension and ramifica-

ISBN 1-876346-48-5 © 2003 UAB

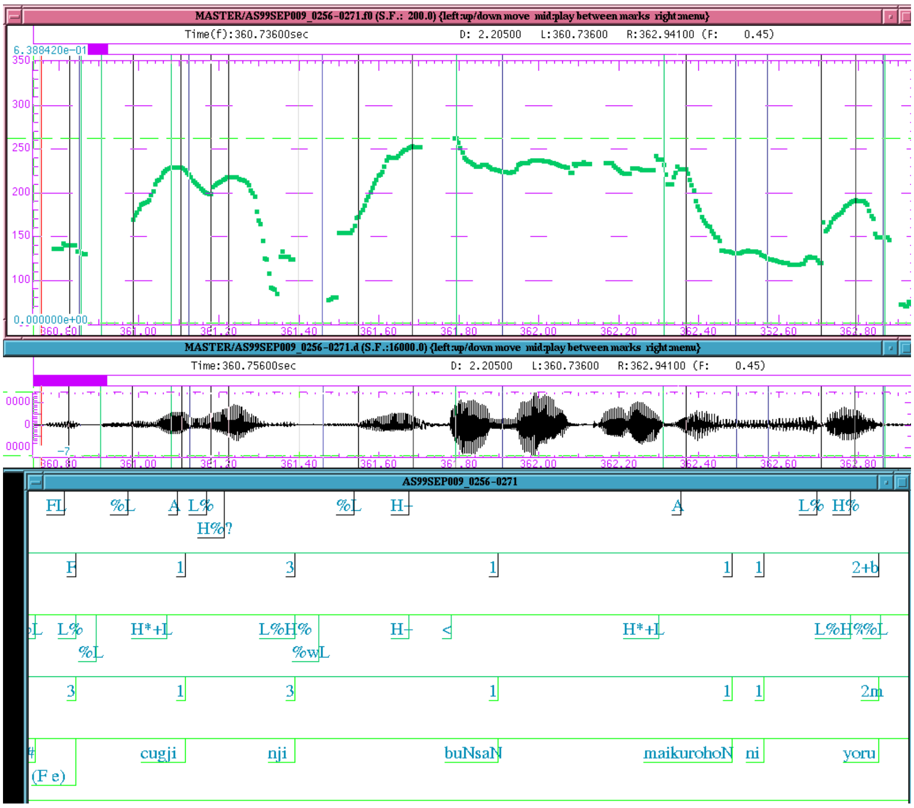

Figure 1: Labeling result by J_ToBI and X-JToBI for the same speech sample. (from above, tone tier and BI tier of X-JToBI, tone tier and BI tier of J_ToBI.)

tion of the usage of break indices, and 4) Newly defined labels for filled-pauses and non-lexical prominence.

Figure 1 compares the results of J_ToBI and X-JToBI labeling applied for the same speech sample. In J_ToBI, “e” in the initial part, which is a filled-pause, is dealt as an accentual phrase. On the other hand, in X-JToBI, it is dealt as a filled-pause, and the newly introduced label “FL”(“filled-pause low”) is used in the tone tier. On the last half of this figure, tonal decomposition of a BPM can be observed. This decomposition is based on the above-mentioned characteristics 1). We can tell that the BPM used in this utterance is bi-tonal; also the exact time location of F0 change can be determined by looking at the time stamps associated with the two constituent tone labels of the BPM, *i.e.* L% and H%.

3 Frequency of labels

In this section, frequency of newly introduced labels and the performance of the X-JToBI labeling are described.

Table 1 shows the frequency of tone labels of BPM in two types of monologue. ‘APS’ and ‘SPS’ are the two main speech types recorded in *CSJ*, namely, academic presentation speech (APS) and layman’s public speech (SPS) [1]. The total amount of speech (excluding pauses) in the APS and SPS were 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, appeared across speech types.

Table 2 compares the frequency of the X-JToBI break index(BI) labels. The most salient difference between the two speech types was 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.

Newly introduced BI labels for disfluency(D,D+,P and P+) and filled-pause(<F and F) appeared frequently. From these results, we can say that newly introduced labels of tones and BIs are useful in the prosodic labeling of *CSJ*.

Table 1: Frequency of tone labels of BPM
(numbers in () stands ratio in all BPMs.[%])

tone label	SPS		APS	
L%+H%	1683	(58.05)	1331	(78.80)
L%+HL%	1121	(38.67)	346	(20.49)
L%+LH%	95	(3.28)	12	(0.71)

Table 2: Frequency of BI labels
(numbers in () stands ratio in all BI labels.[%])

BI label	SPS		APS	
1	31696	(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+pb	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)
<F	431	(0.83)	464	(3.18)
F	2426	(4.66)	1111	(7.62)
PB	61	(0.06)	10	(0.07)

4 Accuracy and Inter-labeler reliability

In this part, the accuracy and inter-labeler agreement will be analyzed. Many labelers labeled *CSJ* data of about 3 minutes long with J_ToBI or X-JToBI. By calculation of accuracy of each labelers, the results by the best three labelers were selected in both of J_ToBI and X-JToBI for analysis. Cohen's κ(kappa)[4] was used as the index of label agreement.

4.1 BI tier

Table 3 compares the accuracy of J_ToBI and X-JToBI labeling. Here, accuracy is defined as the ratio of labels that agreed with the labeling result of an expert labeler. Accuracy of X-JToBI labels were higher than that of J_ToBI labels almost always.

But, the accuracy of BI=2 and BI=3 were not high enough. Figure 2 shows the accuracy of BI=2 judgement as a function of the accentedness of the preceding and following phrases. 'A' and 'N' stand respectively for the presence and absence of lexical accent. 'N,N' means that both precedent and following accentual phrases of the target break have no accent. 'A,N' means that the precedent accentual phrase has an accent but the following accentual phrase does not. In this figure, the accuracy of judgments of BI=2 in the case of 'A,A' is high, but the accuracy of other cases are worse. This figure suggests the importance of establishing clear criteria for the cases involving unaccented accentual phrase.

Table 3: Accuracy of BI labels.
(numbers in () stands frequency of labels.)

J_ToBI			X-JToBI		
label	accuracy		label	accuracy	
1	91.3	(593)	1	94.9	(531)
2	74.0	(123)	2	74.4	(112)
3	70.5	(182)	3	75.1	(181)
2-	33.3	(1)	1+	0.0	(1)
1p	47.9	(16)	1+p	81.5	(9)
3-	–	(0)	2+	–	(0)
2m	80.5	(29)	2+b	66.7	(30)
2p	27.3	(11)	2+p	33.3	(8)
3m	0.0	(1)	2+pb	33.3	(4)
—	—	—	D	83.3	(4)
			D+	44.4	(3)
			<F	76.2	(7)
			F	91.0	(63)
			PB	33.3	(2)
Total	83.2	(956)	Total	86.1	(956)

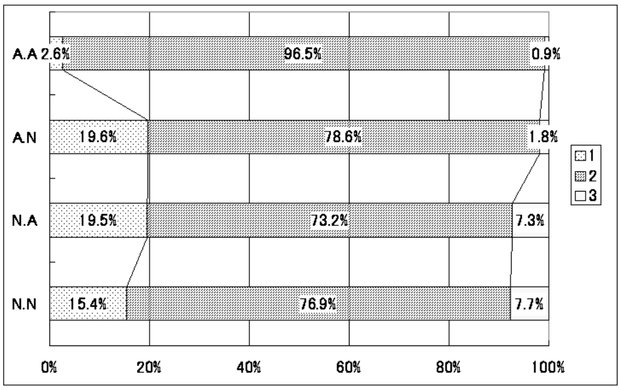

Figure 2: Ratio of labeling result where the correct labels were marked as BI=2.

As for inter-labeler agreement, κ is 0.64 and 0.73 in J_ToBI and X-JToBI respectively, but if we exclude the labels of disfluency and filled-pauses from the calculation, there is no significant difference(κ is 0.69 and 0.71 in J_ToBI and X-JToBI). This suggests that newly proposed BI labels for disfluency and filled-pauses has improved the inter-labeler reliability of prosodic labeling of spontaneous speech.

4.2 Tone tier

Table 4 shows accuracy of tone labels in BPMs. The accuracy of L% and L%H% was high, but that of L%HL% was noticeably low in the J_ToBI. On the other hand, the accuracy of L%HL% in X-JToBI was not very low.

Also, κ was 0.41 and 0.61 in J_ToBI and X-JToBI respectively as far as BPMs are concerned. Moreover,

ISBN 1-876346-48-5 © 2003 UAB

same tendencies were observed in "H-" and "H*+L".

Presumably, these improvements of the X-JToBI labeling were the byproducts of the very fact that it was a demanding labeling scheme. Because X-JToBI had more tone inventories than J_ToBI, and it required labels to be located at the exact locations of the corresponding physical (*i.e.* F0) events, labelers had to pay more attention for speech than in J_ToBI labeling which had fewer inventories and no strict requirement on the label location.

Table 4: Accuracy of tone labels in BPMs.

(a)J_ToBI

	Correct label		
Labeling result	L%	L%H%	L%HL%
L%	144	15	11
L%+H%	7	57	6
L%+HL%	1	0	10
Accuracy[%]	94.7	79.2	37.0

(b)X-JToBI

	Correct label			
Labeling result	L%	L%H%	L%HL%	L%LH%
L%	360	34	17	0
L%+H%	12	157	3	0
L%+HL%	5	7	40	0
L%+LH%	0	0	0	0
Accuracy[%]	95.5	79.3	66.7	-

5 REPRODUCIBILITY

Prosodic labels are also useful for the study of F0 contour synthesis [5][6]. In X-JToBI, maximum match between the label location and physical event is pursued. All tone labels are located to the places where the corresponding F0 events occur. It is expected that this principle showed result in better reproducibility of F0 contour than in J_ToBI. Figure 3 shows the F0 contour synthesized from X-JToBI labels. The method of synthesis is linear interpolation of F0 points corresponding to tone labels. Average difference between the original and synthesized F0 contours of all samples used in Section 4 are 10.5[Hz]. This results are worse than the results of RFC model[5] (3.6-7.3[Hz], for read speech and dialogue speech). In Figure3, serious deviation from the original contour is observed after the location of accent, which is realized as a local steep F0 fall in Japanese. The analysis of accentual F0 fall will certainly help us in developing better interpolation method.

6 CONCLUSION

Performance of the prosodic labeling of spontaneous speech was evaluated using the sample data of CSJ. The new X-JToBI scheme showed higher accuracy and inter-labeler agreement compared to the J_ToBI scheme. Introduction of new BI labels for filled-pauses and disfluency and enlargement of BPM inventory seemed to be the main factors of the improvement. Particulary, analysis of inter-labeler reliability using Cohen's kappa showed that inter-labeler reliability of X-JToBI was higher than that of J_ToBI with respect to both tone and BI labels. Lastly, for all the good characteristics of X-JToBI, its agreement rate remained at around 80%. Further clarification of the labeling criteria will increase the rate.

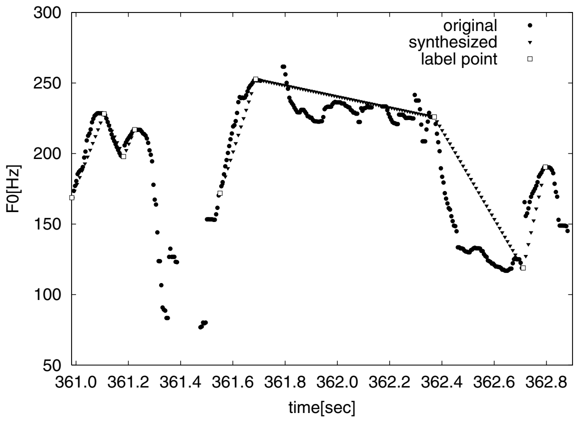

Figure 3: Comparison of synthesized and original F0 contour. (This sample is the same as the one used in Figure 1.)

REFERENCES

[1] K.Maekawa, H.Koiso, S.Furui, H.Isahara, "Spontaneous speech corpus of Japanese," Proc. 2nd International Conference on Language Resources and Evaluation, Athens, Greece, pp.947-952 (2000).

[2] Venditti, J., "Japanese ToBI Labelling Guidelines," Manuscript. Ohio State University, USA., 1995.

[3] K.Maekawa, H.Kikuchi, Y.Igarashi, J.Venditti, "X-JToBI: An extended J-ToBI for spontaneous speech," Proc. International Conference on Spoken Language Processing, Denver, U.S., vol.3, pp.1545-1548 (2002).

[4] Cohen, J., "A Coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, Vol.20, No.1, pp.37-46 (1960).

[5] Taylor, P., "The rise/fall/connection model of intonation," Speech Communication, Vol.14, pp.169-186 (1994).

[6] D.Hirst, A.D.Crist, R.Espesser, "Levels of representation and levels of analysis for description of intonation systems," Merle Hoenw (ed.) Prosody: Theory and Experiment, Academic Press, pp.51-87 (1998).