

Japanese Prosodic Labeling Support System Utilizing Linguistic Information

Shinya Kiriyaama, Yoshifumi Mitsuta, Yuta Hosokawa,
Yoshikazu Hashimoto, Toshihiko Ito, and Shigeyoshi Kitazawa

Faculty of Information, Shizuoka University, JAPAN

{kiriyaama, cs8085, cs9077, cs9069, t_ito, kitazawa}@cs.inf.shizuoka.ac.jp

Abstract

A prosodic labeling support system has been developed. Large-scale prosodic databases are strongly desired for years, however, the construction of databases depend on hand labeling, because of the variety of prosody. We aim at not automating the whole labeling process, but making the hand labeling work more efficient by providing the labelers with the appropriate support information. The methods of auto-generating initial phoneme and prosodic labels utilizing linguistic information are proposed and evaluated. The experimental results showed that more than 70% of prosodic labels were correctly generated, and proved the efficiency of the proposed methods. The results also yielded the useful knowledge to support the labelers.

1. Introduction

In order to realize more intelligent speech understanding system, or to generate synthesized speech with higher quality, large-scale prosodic databases are indispensable. The progress of the speech recognition technology based on HMM and the increase of computer power enabled us to construct large-scale speech databases with phoneme information. Automatic labeling of prosodic information is very difficult, however, because of the variety of prosody. So the labeling of prosodic information must be conducted manually by the experienced labelers. Because the present technology level does not allow us automating the prosodic labeling, we aim at making the hand labeling work more efficient by providing the labelers with the appropriate support information. Our final purpose is development of the support system for the prosodic labeling.

Not a few prosodic features can be estimated from linguistic information. We propose the automatic prosodic labeling methods positively utilizing linguistic information. The Japanese-ToBI[1] system was adopted as our prosodic labeling scheme. In order to generate J-ToBI labels automatically, the method based on decision tree learning[2], or the method using the generation process of F_0 contour in a text-to-speech system[3] are studied. Linguistic information is used preliminarily in these studies, and there are few researches aiming at constructing a labeling support system.

In Section 2, the J-ToBI system is described in detail. The methods of automatic prosodic labeling using linguistic information are proposed and evaluated in Section 3. The experimental results design a labeling support system in Section 4.

2. Prosodic labeling scheme

Japanese ToBI (J-ToBI)[1] was proposed as an extension of the ToBI (Tones and Break Indices) prosodic labeling system originally designed for English[4]. It provides a systematic phonological transcription for recording the F_0 events and prosodic

boundaries in Tokyo Japanese speech. Like its predecessor, it consists of four tiers:

2.1. Word tier

The word tier contains the romanised transcription of the words of utterance. A minimal dictionary entry is used as the working definition of a “word”, and as such we mark postpositions and particles as separate words. Accented words are marked with an apostrophe (') after the vowel of the relevant mora.

2.2. Tone tier

The tone tier in J-ToBI marks the distinctive pitch events in the F_0 contour, and is consistent with the work on Japanese intonation by Beckman and Pierrehumbert[5]. The following is a list of the core labels in this tier.

H*+L bitonal pitch accent marking the lexically specified accent of accented phrases.

H- phrasal tone marking the high F_0 of unaccented phrases, also used in some accented phrases in which the phrasal H is higher than the accentual H. It is one of the two tones that delimit the accentual phrase (most commonly *bunsetsu*) in Japanese (see break index 2 below). In Tokyo Japanese, this tone usually occurs on the second mora of the phrase.

L% Along with the phrasal H-, this final low boundary tone characterizes the accentual phrase in Japanese. Together, these two tones produce the familiar rise-fall pattern of the accentual phrase. There is also a “weak” variant of this tone (**wL%**) used in cases where the next phrase begins with a long syllable, or is initially accented.

%L initial low boundary tone marked at the beginning of post-pausal phrases. It provides an anchor from which the F_0 rises at the beginning of utterances and after pauses. As with the final low boundary tone, this initial tone also has a “weak” variant (**%wL**), used in the same contexts.

In addition to these core tones, the tone tier also includes labels for marking the boundary tones conveying various nuances at the end of intonation phrase (**L%H%**, **L%HL%**), the late or early F_0 events (<, >), and uncertainty about the de-phrasing of accented words (*?).

2.3. Break Index tier

Break indices are a measure of the degree of association between two sequential units. They indicate the prosodic grouping of words at various levels. These are measures of perceived juncture that have observable physical correlates, such as tonal

markings and pre-boundary lengthening. J-ToBI distinguishes 5 degrees of disjuncture in the prosodic structure of Japanese.

- 0 break index marking junctures common in fast speech phenomenon, e.g., /kore+wa/ → [korya].
- 1 marks the juncture between sequential words.
- 2 marks the juncture between prosodic units corresponding to the accentual phrase[5]. This unit is delimited by the rise-fall of the phrasal **H-** and **L%** boundary tones. It often consists of a noun plus following postposition (*bunsetsu*). However, it is also common to find two or more content words grouped together into a single accentual phrase, delimited by these two tones.
- 3 marks the boundary between successive intermediate (major) phrases. This is the domain in which the high-tone line is specified, and therefore at a break index **3** boundary, a new pitch range is chosen for the following phrase. The prosodic juncture marked by a **3** is stronger than that marked by a **2** (accentual phrase), but lacks the percept of “finality” which accompanies the stronger break index **4**.
- 4 marks the boundary of an intonation phrase. It is a strong juncture marked by a sense of “finality.”

In addition to the 5 labels described above, the J-ToBI break index tier also contains labels for marking labeler uncertainty (e.g. **1-**, **2-**, **3-**, **4-**) about the strength of the boundary, and also labels for marking hesitations or other disfluencies (e.g. **1p**, **2p**, **3p**), or mismatches between tones and perceived juncture(e.g. **2m**, **3m**).

2.4. Miscellaneous tier

This tier is used for other phenomena present in the speech signal which cannot be properly described by the phonological events marked in the tone and break index tiers. Such phenomena include repairs, disfluencies, laughing, etc.

3. Automatic prosodic labeling

Not a little label information of J-ToBI can be estimated by linguistic information:

- **Word tier:** The labels can be determined from a morphological analysis.
- **Tone tier:** Peaks of the F_0 contours can be presumed by the information of accent types. The foot boundaries of the F_0 contours correspond to the positions of the pause label.
- **Break Index tier:** Syntactic information is usable to estimate the labels on the BI tier. **BI 1**, **BI 2**, and **BI 3** correspond to the boundaries of word, *bunsetsu*, and sentence, respectively. The treatment of **BI 3** needs to be considered.

We propose the methods of automatic labeling on the tone and BI tiers. Additionally, we inspect the accuracy of automatic phoneme alignment for each phoneme. Finally, the results of studies are utilized to develop the labeling support system with high usability.

The experiments was conducted using the Japanese-MULTEXT prosodic corpus[6]. This corpus is the Japanese version of MULTEXT (a multi-language prosody corpus), which was created in March 2001 according to the specifications of EUROM1[7]. It aims at recording spoken Japanese

Table 1: Results of automatic labeling on the tone tier. The numbers of all labels(A), correct labels(C), substitutions(S), deletions(D), and insertions(I) are shown with their rates to the number of all labels.

	A	C	S	D	I
Number	45172	36260	3655	5257	5383
Rate(%)	100.0	80.3	8.1	11.6	11.9

with the same contents, consisting of 40 short paragraphs, and accompanied by a phonemic and the phonetic notation according to the above standard. The speakers were aged between 20 and 40 years, a total of 6 persons, three male and three female speakers of the Tokyo Japanese. A text was given for each reading and a simulated spontaneous utterance was also recorded. The corpus has the speech data of totally 480 paragraphs, with the information of hand labeled phoneme boundaries, and the J-ToBI labels.

3.1. Tone tier labeling

The linguistic information, phoneme labels and accent types for each word in the reading text, was utilized to the automatic labeling on the tone tier. The accent type information was generated manually based on the indices of an accent type dictionary, and the rules of accent sandhi. In this study, the 6 symbols($\%L$, $\%wL$, $L\%$, $wL\%$, $H-$, and H^+L) were selected as the targets of automatic labeling. The generation rules for each symbol were following:

$\%L$ is put at the end of short pause label in the phoneme label tier. If the current phrase begins with a ‘heavy syllable,’ which means a long syllable, or that the current accentual phrase is initially accented, $\%wL$ is used instead of $\%L$.

$L\%$ is put at the end of accentual phrase. If the next phrase begins with a ‘heavy syllable,’ $wL\%$ is used instead of $L\%$.

$H-$ is placed at the center of the vowel of the second mora.

H^+L is placed at the end of the mora where the accent nucleus is located.

Accuracy of automatic labeling was inspected using the answer labels annotated manually by the labelers. Not considering lags on the time axis, and only taking correctness of the label’s symbol into account, the following numbers were counted out; number of all labels(A), number of correct labels(C), number of substitutions(S), number of deletions(D), and number of insertions(I). The results were shown in Table 1. It was proved that as for about 80% of labels, their symbols were correctly auto-generated.

Figure 1 indicates distributions of lags on the time axis for the labels whose symbols are correct. Most labels for H^+L had the lags around 20ms to 40ms, however, the lags of the labels for the other symbols were less than 20ms.

The rates of the numbers of correct/incorrect labels to the number of all labels were shown in Table 2 for each symbol. ‘Correct labels(C)’ mean their time lags were not more than 50ms, and N is (the rate of) the number of labels which have the correct symbols, but their time lags exceed 50ms. The results proved that 71.6% of all labels were placed on the correct positions with the correct symbols, and that the proposed method generated the labels for $\%L$ and $\%wL$ almost perfectly. The rates of insertions for $L\%$, $wL\%$, and $H-$ were not low. This

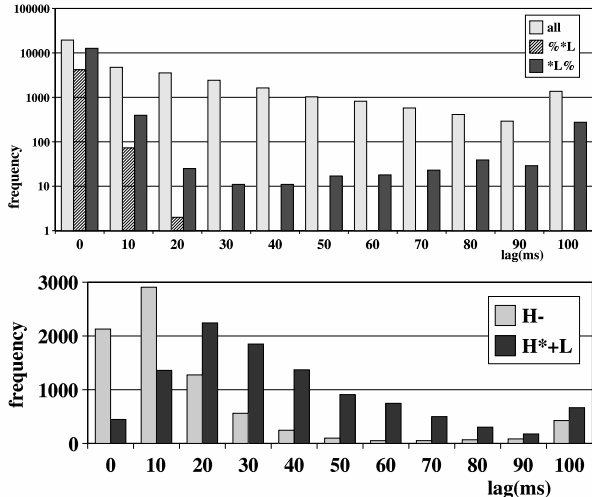


Figure 1: Distribution of timing errors in automatic labeling on the tone tier.

Table 2: The rates of the numbers of correct/incorrect labels for each tone symbol and all labels.

Symbol	<i>C</i> (%)	<i>N</i> (%)	<i>S</i> (%)	<i>D</i> (%)	<i>I</i> (%)
%L	91.2	0.0	8.4	0.3	0.4
%wL	96.4	0.0	3.2	0.5	0.4
L%	75.2	2.4	16.2	6.3	10.8
wL%	81.8	2.1	6.4	9.6	13.3
H-	83.8	8.5	0.0	7.7	29.0
H*+L	67.2	24.2	0.0	8.6	7.9
*?	0.0	0.0	75.4	24.6	0.0
all	71.6	8.7	8.1	11.6	11.9

fact indicated that many accent sandhi events had occurred more than we had expected.

3.2. Break Index tier labeling

The labels on the Break Index tier were estimated automatically using the results of syntactic analysis. The reading text of 40 paragraphs in Japanese-MULTEXT corpus was analyzed by the Kurohashi-Nagao Parser(KNP)[10], a Japanese syntactic structure analysis system. The BI labels of the core 4 symbols(‘1’, ‘2’, ‘3’, and ‘4’) were generated by the following rules:

BI 1 is put on the word boundaries in the current *bunsetsu*.

BI 2 is put at the end of *bunsetsu*.

BI 3 is substituted for **BI 2** with comma.

BI 4 is put at the end of sentence.

The estimation accuracy was investigated by comparing the auto-generated labels with the answer labels annotated manually. Table 3 shows the rates of the labels whose symbols were correctly given, for each symbol, and for the whole labels. The results proved that 73.7% of all labels were generated correctly. The results also indicated that the estimation of the labels for **BI 1** and **BI 4** is pretty accurate, and that the rules for **BI 2** and **BI 3** should be improved. A confusion matrix of Table 4 showed that the **BI 3** labels tended to be substituted for **BI 2** by mistake. This revealed that the current rule for **BI 3** is insufficient.

Table 3: Results of automatic labeling on the BI tier. *NC* and *NA* indicate number of correct labels and number of all labels for each index, respectively. *CR* is the rate of *NC* to *NA*. * The total number of answer labels includes the number of labels for the optional BI symbols(2-, 2m, 2p, 3-, and 3p).

Index	<i>NC</i>	<i>NA</i>	<i>CR</i> (%)
1	13543	15014	90.2
2	6767	9322	72.6
3	1700	2834	60.0
4	1860	2220	83.8
all	23870	32377*	73.7

Table 4: A confusion matrix indicating the distribution of substitution errors for BI labels.

	1	2	3	4
1	13543	1334	60	56
2	2421	6767	88	32
2-	715	901	8	1
2m	130	66	0	0
2p	0	2	0	0
3	170	876	1700	79
3-	161	941	31	1
3m	2	18	6	0
4	150	115	28	1860

3.3. Phoneme boundary labeling

The phoneme labels were generated by the HMM-based automatic phoneme alignment, using the reading text as linguistic information.

The speech recognition engine was Julius ver. 3.2[8]. A monophone model set of gender independent(number of states, 3,000, number of mixtures for each state, 16) included in the ‘‘Continuous Speech Recognition Consortium Software, 2001 version[9]’’ was used as the acoustic model. The boundary positions of phonemes generated by automatic phoneme were compared with those of hand labeled phoneme boundaries as the answer. The average and standard deviation of time lags were calculated for each phoneme and shown in Figure 2.

Accuracy of the automatic segmentation was very high; the longest time lag was around 20ms. Especially, voiced consonants, such as semivowels(/r/, /y/), nasals(/n/, /m/), and explosives(/b/, /d/, /g/), were labeled very exactly.

The directions of lags on the time axis were inspected for each phoneme. For most phonemes, the boundaries were shifted to the ‘late’ direction on the time axis. The boundaries of phoneme /w/, however, were shifted to the ‘early’ direction. It is expected that the information of ‘lag direction’ is efficient for the labelers to modify the lags more easily.

4. Discussion

Using linguistic information only, more than 70% of all labels on the tone and BI tiers had been auto-generated correctly. The time lags for most of the phoneme labels did not exceed 20ms. Concerning the phoneme labels, it is reported that the working time for hand labeling was shortened by the half when the initial labels were available[11]. It might be true that the labels generated by the proposed method are sufficiently useful as the initial labels for hand labeling.

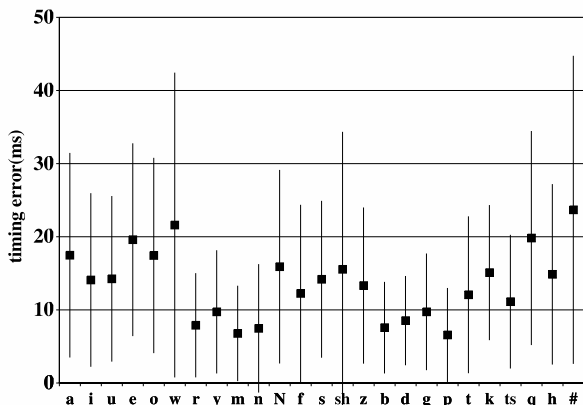


Figure 2: The average and standard deviation of time lags for each phoneme.

An efficient prosodic labeling support system shown in Figure 3 will be realized by considering the followings:

- **Improvement of the quality of the initial prosodic labels:** The estimation ability of accentual phrase boundaries using linguistic information would be improved by introducing more detailed accent sandhi rules.
- **Effective use of duration information:** When a symbol sequence on the tone tier represents an accentual phrase with a small number of mora, the labelers should be given the instruction that the sequence is likely to include insertion errors caused by accent sandhi. The advice for the BI tier that the initial **BI 2** label may be a substitution error for **BI 1** should be also provided. As for the labels on the tone and BI tiers corresponding to the accentual phrase with a large number of mora, the support system should indicate the possibilities of deletion errors on the tone tier and substitution errors on the BI tier, because such a long phrase tends to be divided into plural accentual phrases.
- **Providing the information of the lag direction on the time axis:** Concerning the modification of phoneme boundaries, the label for the phoneme which boundaries tend to be located on the previous/following position, the system give the labelers the instruction that the label should be moved to the following/previous position on the time axis.

5. Conclusions

The methods of automatic prosodic labeling utilizing linguistic information were proposed. The prosodic labels on the tone and BI tiers in the J-ToBI labeling scheme were auto-generated using the information of phoneme boundaries, accent types, and syntactic structure. Ability of the HMM-based automatic phoneme alignment was also investigated. The experimental results proved that 71.6% of all labels on the tone tier, and 73.7% of them on the BI tier were correctly placed only applying linguistic information. The accuracy of auto-generated labels seems to be sufficient as an initial label set for hand labeling. The results also provided the design of a prosodic labeling support system. The system should navigate the labelers appropriately, using the information such as the trends of timing errors for each phoneme, and the duration of each accentual

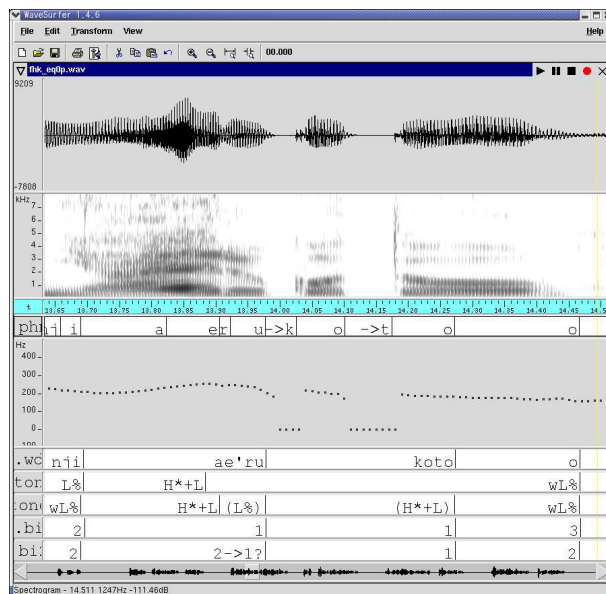


Figure 3: A prosodic labeling support system. The utterance is /ae'ru koto o/ (*to be able to meet (you)*) '→k' and '→t' on the phoneme label tier mean that these labels should be modified to the directions of the arrow. The symbol sequence "**L% H*+L wL%**" on the upper tone tier is the hand labeled answer. '**(L%)**' and '**(H*+L)**' on the lower tone tier indicate that they are the candidates of insertion error, because they correspond to the rather short accentual phrases (/ae'ru/ and /koto o/). The sequence "**2 1 1 3**" on the upper BI tier is correct. '**2→1?**' on the lower BI tier should be a substitution error for **BI 1**.

phrase. From now on, the efficiency of the support system needs to be verified through the experiments using the system.

6. References

- [1] J. J. Venditti, "Japanese ToBI Labelling Guidelines," Technical Report, Ohio-State University, 1995.
- [2] H. Noguchi, et al, "Automatic labeling of Japanese prosody using J-ToBI style description," Proc. Eurospeech99, pp.2259-2262, 1999.
- [3] N. Campbell, "Autolabelling Japanese ToBI," Proc. ICSLP-96, pp.2399-2402, 1996.
- [4] M. Beckman, and G. Ayers, "The ToBI Handbook," Technical Report, Ohio-State University, 1993.
- [5] J. Pierrehumbert, and M. Beckman, "Japanese Tone Structure," Cambridge, MA: MIT Press. 1988.
- [6] S. Kitazawa, et al, "Preliminary Study of Japanese MULTEXT: a Prosodic Corpus," Proc. ICSP2001, pp.825-828, 2001.
- [7] E. Campione, and J. Veronis, "A multilingual prosodic database," Proc ICSLP98, pp.3163-3166, 1998.
- [8] <http://julius.sourceforge.jp/>
- [9] <http://www.lang.astem.or.jp/CSRC/>
- [10] S. Kurohashi and M. Nagao, "Building a Japanese Parsed Corpus while Improving the Parsing System," Proc. ICLRE98, pp.719-724, 1998.
- [11] H. Kikuchi, and K. Maekawa, "Accuracy of Automatic Phoneme Labeling on Spontaneous Speech," Proc. 2002 Spring Meeting of ASJ, pp.97-98, 2002 (in Japanese).