Japanese MULTEXT: a Prosodic Corpus

Kitazawa Shigeyoshi, Kiriyama Shinya, Itoh Toshihiko, Nick Campbell*

Department of Computer Science, Faculty of Information, Shizuoka University
3-5-1 Johoku,Hamamatsu, 432-8011, Japan
{kitazawa, kiriyama, t-itoh}@cs.inf.shizuoka.ac.jp
*ATR Human Information Science Research Labs.
Seika-cho, Hikari-dai, Kyoto 619-02, Japan
nick@atr.co.jp

Abstract

A prosodic corpus of Japanese was developed as a scheduled project by the university researchers in Japan. This paper describes the contents of the corpus: speakers, speaking style, recording conditions, prosodic annotations. The corpus is a Japanese version of the MULTEXT prosodic database of EUROM1. We adopted a J-ToBI prosodic labeling scheme as well as additional labels such as pitich range, prominence, devoicing, and nasalization. We developed an automatic generation of J-ToBI labels. It was proved that 71.6% of tone labels were placed on the correct positions with the correct symbols, and that 73.7% of BI labels were generated correctly. Automatic prosodic label generator was evaluated by expert labeler team and beginner team and found to be helpful for both of them.

1. Introduction

There are strong needs of prosodic corpus for speech synthesis and recognition technologies. A prosodic corpus of Japanese was developed as a scheduled project by the university researchers in Japan. A new project focusing on "Realization of advanced spoken language information processing from prosodic features" headed by Professor Keikichi Hirose (Professor, Department of Frontier Informatics, School of Frontier Sciences, The University of Tokyo) has started from last October 2000. This is a project sponsored by the government, the Ministry of Education, Culture, Sports, Science and Technology, with the fund called the “Grants-in-Aid for Scientific Research”. The term of the project is from 2000-2003. Because our database will be used as the foundation for future prosody research, it is important to identify the prosodic events by labeling by Tones and Break Indices (as in ToBI). We therefore evaluated whether the extent to which J-ToBI labels (Venditti, 1995) provide useful information.

2. Prosodic Corpus

MULTEXT (a multi-language prosody corpus) (Campione Veronis, 1998) is a prosody corpus with annotation of prosodic parameters, and the prosody notation about five languages of EUROM1 (Chan, Fourcin et al., 1995). The Japanese version of MULTEXT was created in March 2001 according to the specifications of EUROM1. It aims at recording spoken Japanese with the same contents, consisting of 40 short paragraphs, and accompanied by a phonemic and the phonetic notation according to the above standard. 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), and calibration signal generating equipment (94dB). In addition, electroglottograph is recorded with an EGG (KAY (Co.) 4338).

2.1 Speakers for the database

Being able to act out a role for each of the various utterance styles was a basic selection condition for our speakers, as well as being a speaker of the Tokyo dialect. We selected those speakers who had not lived outside Tokyo at some time in their lives or less influenced in their speech accentuation by the experience of living in the provinces.

The speakers were aged between 20 and 40 years, a total of 6 persons, three male and three female speakers of the Tokyo dialect. A text was given for each reading and a simulated spontaneous utterance was also recorded. The speakers practiced before reading and took natural pauses at will. In the case of an incorrect utterance or an accentual error, the speaker was instructed to record the whole paragraph again.

2.2 Speaking style

Each speaker produced each sentence in two different speaking styles: One was a reading style, where the speakers were instructed to read each text aloud. The other was in one of the acting styles; the speakers pretending that they were taking part as a participant in each of the situations. A classification into emotional speech and emotionless speech would not be appropriate. In these passages, no speaker expressed his or her emotions, because they were only performed according to the text of each situation.

Our analysis has shown that frequency of emphasis marked the clearest difference between the two speaking styles. More accentual phrases were emphasized in the acted speech than in the reading style. This tendency was observed not only in the F_0 contours, but also from subjective impressions when listening to the speech. In some passages, although there might be no difference in emphasis between the two styles, however, there were paralinguistic differences when feelings such as sarcasm, dissatisfaction, and a sincere apology were expressed.

3. Prosodic Annotation with J-ToBI

Prosodic labeling was performed on the voice data collected for the prosody database. The following phenomena are considered to be of special interest.

3.1 Phoneme labels

Segmentation of phonemes and labeling is important to study prosodic parameters because we need precise timing of prosodic events and the events are aligned to the phoneme boundaries. After an automatic segmentation based on an HMM, phoneme labels are corrected by hand.

3.2 Accent types

Two experts of accentuation in the Japanese language marked the morae of each perceived accent kernel. This was performed from listening alone, without reference to acoustic data such as the F_0 contour. Some accents were considered difficult to perceive. Some may have been placed by prejudice, or missed by chance. The two experts processed the data independently and then crosschecked their results for reconsideration afterward

The corpus includes words that could be ascribed multiple possible accent types in the Japanese accent dictionary. The accent types, which had been actually realised, were annotated. Most of the cases converged on one of accent type, perhaps according to context. The number of examples where two or more accent types were actually realised was very few.

Indecision about whether verbs at the end of a sentence were accented or not, was common; however, such types of accent (so-called semi-accents), are rarely perceived. It sounds unnatural to stand out a semi-accent. In J-ToBI, these accents are noted with a question mark. For example, in "*reNraku wo to'Qte moraemase'Nka (Would you like to make a contact?),*" *to'* is a main accent and *se'* is a semi-accent that is not clearly marked in the speech. There were a lot of semi-accents in our corpus.

3.3 Chunking by Break Index

Accentual phrases can be connected into new accentual phrases. There are many examples of such compound accentual phrases: consisting of lists of numbers or items etc., found in reporting speech. The emphasized phrases are marked by a short pause and a pitch-reset. This type of emphasis is an important phenomenon for Japanese prosody.

It may be possible to mark such emphasis using the BI labels in J-ToBI. The BI=2- indicates that a compound accentual phrase is composed after the accentual phrase weakens independence and is connected. The BI=3- label can indicate that reset of the F_0 contour occurred for emphasis etc. Labelers are asked to judge this by noting 2m and BI=2-, 3-, 3m from the F_0 contour, and to determine where the speech is emphasized by listening.

3.4 Boundary tone

This tone represents various nuances in the speech, such as doubt, request, and continuity, at the boundary of an intonation phrase. J-ToBI includes only L%, H%, HL%, and LH%, etc. according to the raising and lowering etc. of F_0, but it is likely that these nuances are also indicated by extending the syllable length, though J-ToBI describes nothing about the lengthening. Maekawa, Kikuchi, and Igarashi (2001) proposed extensions to describe the varieties observed in spontaneous speech. There is an experiment on F_0 and phoneme durations.

3.5 Miscellaneous tier of the J-ToBI

Many interesting phenomena were observed which are inexpressible within the frame of J-ToBI. The following have been selected as the target items: phoneme label, position of accent nucleus, F_0 peak value, and whether the speech is emphasized or not.

3.5.1 Paralinguistic description

There were, for instance, expressions of nuance from the difference of the intonation in addressing someone, such as anger, sarcasm, dissatisfaction, irritation, etc., though these might be of a character best described in the miscellaneous tier. These are problems of para-language rather than prosody. Naturally, these are thought to be outside the framework of J-ToBI labeling.

Various nuances in the speech are communicated in the realized utterance by use of a special tone. There is a sense of incompatibility when the spoken tone is inappropriate to the content. J-ToBI does not readily allow us any way to mark this, but we can freely describe such nuances in the miscellaneous tier. We noted here the types of tone located at the end of each sentence or phrase, whether for continuation, inquiry, request, determination, confirmation, sarcasm, dissatisfaction, doubt, compulsion, desire, sincere apology, and excuse, etc.

We propose some new prosodic labels that are not yet in the J-ToBI standard, but which we feel may be useful for an understanding of the prosody of Japanese speech. Trained labelers annotated these labels from inspection of the F_0 contours, spectrograms, speech waveforms and by listening to the speech.

3.5.2 Emphasis, Prominence, Focus

The label that specifies the emphasized part is not in J-ToBI. Emphasis is important since it is related with prosodic phenomena. We are labeling every emphasized part, that is accentual phrase, by hearing, observing F_0 contour for pitch reset, short pause etc. The maximum value of F_0 in each accent phrase was also recorded.

3.6 Fluency and emphasis

A voice trainer examined the whole utterances from the point of view of accentuation, phrasing, and prominence. From the prosodic point of view, smoothness of speech flow with a jaunty rhythm is very important. This smoothness is closely related with Japanese specific prosodic phenomena: devoicing of vowels in surrounding voiceless consonants environments and nasalization of voiced velar stop consonants and continuation of connected vowels. These three phonological rules should be achieved in order to realize a "beautiful" Japanese. That is in order the emphasized part to be clearly transmitted to the receiver, the other parts should be deemphasized to achieve a clear contrast.

3.6.1 Devoicing

An unaccented vowel /i/ or /u/ sandwiched between voiceless consonants /k, s, h, t, p, m, n/ or put at the end of a word is devoiced, e.g. *kik(u)* (chrysanthemum), *kuk(i)* (stalk), *s(u)mai* (house), *k(i)Qp(u)* (ticket). Where vowels in ( ) are devoiced. Occasionally inserted devoiced vowels produce a feeling of modern and a light pleasant sensation to a Japanese speech sound. This part is paid no attention therefore it become deemphasized. On the other

hand, if these vowels are voiced, a heavy rugged feeling of wrongness is produced, which disturbs smooth perception of content of speech because a bogus focus interferes the real focus.

3.6.2 Nasalisation

A voiced velar stop consonant becomes nazalized except in the phrase initial position, e.g., *kaigi* (meeting), *usugoori* (thin ice). Where *g*'s with a underline are nasalized. The effect of this is similar to the above devoicing, this phenomena deemphasize and smooth the flow of speech sound of the phrase. This style is classic Tokyo Japanese, and those young peoples do not follow this manner.

3.6.3 Consecutive vowels

Consecutive vowels, appearing in a word or at the connection of words, are expressed various ways: *ei* in a word changes into a long vowel *e*- such as *e-ga* instead of *eiga* (cinema), *toke*- instead of *tokei* (clock), however, if *ei* is word boundary *tameiki* (sigh) is correct and *tame-ki* is incorrect. In case the same vowel follows, the vowel is lengthened: *oka-san* (mother), *ku-ko-* (airport). Although the following examples are hiatus of vowels that is there is a pause between two vowel sounds: *baai* (case), *kiiro* (yellow color). We have made perceptual experiments where is the most appropriate boundary between two vowel sounds (Kitazawa, 2004). These two cases have to be distinguished whether marked or unmarked to convey the real meanings.

4. Automatic prosodic labeling

It is considered that much of the J-ToBI system of prosodic labeling can be automatically generated from knowledge of the word and phoneme labels, the accent nucleus, the F_0 peak value, and the emphasis. In order to test this hypothesis, we implemented a system that can generate J-ToBI labels automatically from a text, and it was evaluated, e.g., using the front-end of a speech synthesis system (Kiriyama, Mitsuta, Hosokawa, Hashimoto, Ito, & Kitazawa, 2003) in the following way:

_ **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.

The experiments were conducted using the Japanese-MULTEXT prosodic corpus (Kitazawa, 2001).

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

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.

4.1. 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 numbers were shown in Table 1. It was proved that as for about 80% of labels, their symbols were correctly auto-generated. Amongst 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 fact indicated that many accent sandhi events had occurred more than we had expected.

Symbol	*C*(%)	*N*(%)	*S*(%)	*D*(%)	*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

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

4.2. 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) (Kuriashi, & Nagao 1998), 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.

Index	*NA*	*NC*	*CR(%)*
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 3: Results of automatic labeling on the BI tier. And 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**).

BI lab.	put 1	put 2	put 3	put 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

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

4.3 Evaluation of the auto-labeling

Experiments of J-ToBI labeling performance conducted using the auto-generated tone and BI labels by both experts and beginners. The results are shown in Table 5. The control case is without any initial labels. The second case is provided an auto-generated J-ToBI label as an initial value. The third case is provided with an initial label accompanied with information of possible alternative labels. Total time spent for labeling and total operations to modify/add/delete labels are counted.

Beginners take much time than experienced labelers. Auto-generated initial labels are helpful resulting saving time and actions. Suggestions are useful for beginners but not so much for experienced labelers.

Condition	Experienced		beginners	
	Time(s)	actions	Time(s)	actions
no label	1468	287.0	2337	236.7
initial label	1116	164.1	1397	119.5
suggestions	1141	166.9	1333	114.3

Table 5: Average time and operations to annotate a passage for experienced labellers and beginners on conditions without initial labels, with initial labels and with initial labels as well as suggestions of possible alternatives

5. Conclusions

We have developed a prosodic corpus of Japanese in parallel to the MULTEXT with annotations of phonemic and prosodic including J-ToBI and some additional comments. Automatic generation of tone and BI labels of the J-ToBI was useful in saving works of labelers.

Acknowledgements

We are grateful to the Ministry of Education, Culture, Sports, Science and Technology for supporting this research through the Grants-in-Aid for Scientific Research, project number 12132204, during 2000-2001.

References

Chan, D., Fourcin, A. et al., (1995) EUROM - A Spoken Language Resource for the EU (pp.867-870) Eurospeech95, vol.1,.

Campione, E. and Veronis, J. (1998) A multilingual prosodic database (pp.3163-3166) ICSLP98.

Kitazawa, S. et al, (2001) "Preliminary Study of Japanese MULTEXT: a Prosodic Corpus," ICSP2001, pp.825-828.

Kitazawa, S., Kiriyama, S., Itoh, T., Toyama, Y. (2004) Perceptual Inspection of V-V Juncture in Japanese (to appear in SP2004 in Nara).

Kiriyama, S., Mitsuta, Y., Hosokawa, Y., Hashimoto, Y., Ito, T., and Kitazawa, S. (2003). Japanese Prosodic Labeling Support System Utilizing Linguistic Information (pp. 181-184) EUROSPEECH 2003 – GENEVA.

Kurohashi, S. and Nagao, M. (1998) Building a Japanese Parsed Corpus while Improving the Parsing System (pp.719-724) Proc. ICLRE98,.

Maekawa, K., Kikuchi, H., and Igarashi, Y. (2001) "X-JToBI: An Intonation Labeling Scheme for Spontaneous Japanese", Technical Report of IEICE, SP 2001-106, pp25-30. (in Japanese)

Venditti, J. J., (1995) "Japanese ToBI Labelling Guidelines," Technical Report, Ohio-State University.