AUTOMATIC LABELING OF JAPANESE PROSODY USING J-TOBI STYLE DESCRIPTION

Hiroaki Noguchi*,†, Kazuhisa Kiriyama†, Hiroshi Matsuda†, Miki Taniguchi‡, Yasuharu Den†, and Yasuhiro Katagiri*

*ATR Media Integration & Communications Research Laboratories
†Graduate School of Information Science, Nara Institute of Science and Technology
‡Graduate School of Language and Culture, Osaka University

ABSTRACT

Speech corpora with prosodic labels are getting more and more important not only for speech synthesis but also for discourse modeling. A widely used labeling system for Japanese prosody, J-ToBI, however, is insufficient for applications like discourse modeling and it even lacks an accurate method for automatic labeling. In this paper, we propose an automatic labeling method for J-ToBI style description of tonal events in Japanese speech, aiming at applying it to a general-purpose labeling of Japanese prosody. The proposed method takes into account the linguistic constraints on the tone structure, which improves the accuracy of automatic labeling. We achieve a fairly good performance in a preliminary experiment using a read speech corpus.

1. INTRODUCTION

In recent years, speech corpora with rich annotations for prosodic information are getting more and more important not only for speech synthesis but also for discourse modeling such as the prediction of speech act types and of discourse boundaries. The development of speech corpora with prosodic labels, however, requires huge human labor and time, and thus it is indispensable to provide an automatic labeling method for any kind of prosodic labeling system.

In Japanese, J-ToBI [10] has been used as a standard labeling system for some purposes including speech synthesis. It, however, is not sufficient for applications such as speech act type prediction and discourse structure modeling, which require precise information other than locations of accents and boundary tones. Nor does it provide accurate ways of automatic labeling; the only attempt at auto-labeling J-ToBI [1] resulted in poor accuracy.

The aim of our study is to propose a general-purpose labeling system of Japanese prosody and an automatic method for labeling Japanese speech based on it. The proposed model is an enhancement of the tilt model [8, 9] by incorporating J-ToBI style description of tonal events. The tilt model precisely models linguistically meaningful events with continuous parameters, which, we believe, is suitable for applications like discourse modeling. One drawback of the model, however, is that it does not take into account the tone structure of the language to be modeled; it distinguishes only two types of tonal events, i.e., accents and boundary tones, which might be sufficient for English but is obviously insufficient for Japanese. The incorporation of linguistically adequate constraints on the tone structure, such as the ones utilized in J-ToBI, would enhance the model and improve the accuracy of automatic labeling when applied to Japanese.

In our model, like the original tilt model, the labeling process consists of two parts: (i) detecting tonal events and (ii) modeling each event with continuous parameters. In this paper, we focus on the first part, that is, the automatic detection of tonal events. Since tonal events in our model are described using a variant of the J-ToBI system, the proposed method can also be applied to auto-labeling of J-ToBI. Section 2 introduces our model of Japanese prosody. Section 3 describes the method for automatic detection of tonal events. Section 4 reports the result of a preliminary experiment using a read speech corpus.

2. MODEL

Since the main focus of this paper is to describe the method for automatic detection of tonal events, we present here only the relevant part of our model. Please refer to the original tilt model [8, 9] for the modeling of tonal events with continuous parameters.

In Tokyo Japanese, the accentual phrase can be seen as a basic unit of phonetic descriptions. Accental phrases are classified into (at least) three types according to their surface pitch patterns (Table 1). An accented word like yamaza’kura (wild cherry) begins with low pitch, rises to high on the second mora, and
Table 1: Pitch patterns for Tokyo Japanese

<table>
<thead>
<tr>
<th>Pitch pattern</th>
<th>J-ToBI label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ya ma za ku ra</td>
<td>L% H- H*+L L%</td>
</tr>
<tr>
<td>ka ge bo o si</td>
<td>L% H*+L L%</td>
</tr>
<tr>
<td>mu ra sa ki i ro</td>
<td>L% H- L%</td>
</tr>
</tbody>
</table>

Figure 1: Constraints on the tone structure of Tokyo Japanese

These patterns can be represented by different sequences of J-ToBI labels. In all cases, a boundary tone L% is placed at (the beginning and) the end of each phrase. In the case of an accented word, a lexical accent H*+L is placed within the accented mora. Furthermore, in all cases other than initially accented words, a phrasal tone H- is placed within the second mora. This yields three different sequences for the patterns in Table 1.

The sequences in Table 1 are the only possible sequences for accentual phrases in Tokyo Japanese. Any other sequences, such as ‘L% H*+L H*+L L%’, are illegal. Therefore, these patterns can be represented by different sequences of J-ToBI labels. In all cases, a boundary tone L% is placed at (the beginning and) the end of each phrase. In the case of an accented word, a lexical accent H*+L is placed within the accented mora. Furthermore, in all cases other than initially accented words, a phrasal tone H- is placed within the second mora. This yields three different sequences for the patterns in Table 1.

The overall process of tonal event detection is illustrated in Figure 2. The input to the system is both the speech waveform and the transcription; the output is the tonal events detected by the system. The detection process consists of the following five parts:

1. F0 extraction, which outputs the F0 value at each sampling point.
2. Morphological analysis, which outputs a sequence of words and their parts of speech.
3. Word alignment, which outputs the beginning and the ending times of each word.
4. J-ToBI labeling by hand, which produces the training data for the decision tree learner. This part (enclosed region in Figure 2) is required only in training phase.
5. Decision tree learning, which, in training, produces rules for tonal event detection from the training data and, in execution, predicts a sequence of tonal events from the input.

We utilize various kinds of information, prosodic, syntactic, and segmental, which can be obtained (semi-)automatically by the current speech/NL technologies: F0 values, words and their parts of speech, and the time at word boundaries. We also employ a statistical learning method to automatically obtain a set of rules for tonal event detection from the hand-labeled J-ToBI training data. The learner inspects a set of features computed from the information described above, and produces a set of queries about these features in the form of a decision tree which best explains the distribution of the tonal events in the training data.

In the rest of this section, we describe each of the five processes in turn.

3.1. F0 Extraction

F0 values are extracted at every 10 msec point by using ESPS/waves+ [2]. A median filter with the window size of 5 points is used for smoothing.

3.2. Morphological Analysis

We produce a morphological analysis of the transcription by using the Japanese morphological analyzer ChaSen [4]. The lexicons, as well as word bigrams for selecting the best analyses, were automatically obtained from a corpus of Japanese newspaper articles. It is suitable for the read speech corpus used in our preliminary experiment, showing near perfect accuracy. However, since our ultimate goal is to ap-
ply our method to natural conversational data, we are also working on the development of a dictionary for conversational speech.

### 3.3. Word Alignment

We produce a word alignment of the input speech by using the forced alignment function of the speech recognizer HTK [6]. It outputs a speech recognition result which matches a possible pronunciation for the input when the transcription is given a priori. By this, the beginning and the ending times of each word in a morphological analysis are obtained.

This part also works reasonably well for the read speech corpus, but we might need phoneme models for conversational speech in the future.

### 3.4. J-ToBI Labeling by Hand

We produce J-ToBI style descriptions of tonal events by hand for each sentence in the training data. One of the co-authors was engaged in this process. Tonal events were labeled with H*+L, H-, and L%, and initial boundary tones, %L, were mechanically placed at the beginning of words immediately following a pause. Boundary tones other than L% were not used since our read speech corpus contains only declarative sentences.

Unlike the original J-ToBI, we put H*+L and H- on the basis of an actual F0 contour; when the beginning of the fall in F0 appears later than the accented mora or the end of the rise appears later than the second mora, we place H*+L or H- at the position where the limit of the fall or rise is observable in the F0 contour, rather than placing it within a phonologically specified mora. This is for two reasons. First, we do not use phonological information such as accent patterns of words, which vary according to dialect. Although our current target is Tokyo Japanese, we would like to make our method also applicable to other dialects. Second, the late F0 event itself poses interesting research problems; for instance, a delayed ending of initial rise is said to convey some pragmatic/social meaning [3]. Thus, it is important to describe a tonal event in its actual appearance.

### 3.5. Decision Tree Learning

The task of tonal event detection consists of two sub-tasks: (i) identifying the locations of linguistically meaningful events and (ii) specifying the type of each event. We simplify this task in the following way.

We introduce dummy labels, r, c, and f, which should be placed at every sampling point (with 10 msec interval) between L% and H-, between H- and H*+L, and between H*+L and L%, respectively. Thus, the tone structure in Tokyo Japanese, now, is modeled as in Figure 3. By this, the tonal event detection can be seen as assigning to every sampling point one of the J-ToBI labels or the dummy labels, analogous to part-of-speech tagging.

The labeling process is implemented by using the decision tree learning software C4.5 [7]. The features at each sampling point x, which are used for learning, are as follows:

**Prosodic features** F0 values at x and at 10, 20, 50, and 100 msec prior or posterior to x; difference of F0 values between x and 10, 20, 50, or 100 msec prior or posterior to x.

**Syntactic features** Parts of speech of the word containing x, of the preceding word, and of the succeeding word.

**Segmental features** Duration between x and the beginning or the end of the word containing x; duration between x and the preceding or the following pause.

**Contextual features** Tonal event label assigned to the preceding sampling point.

We designed the learning process so that it can automatically obtain the constraints on the tone structure. In Tokyo Japanese, the possible sequences of J-ToBI labels are restricted as shown in Figure 1. With the dummy labels, the constraints depicted in Figure 1 are now restated as in Figure 4. (%L is used for a segment initial boundary tone.) We implement these constraints in our decision tree learning through the contextual feature, i.e., the tonal event label of the preceding point. If the training data is labeled so as to
Table 2: Experimental results (average of 8 trials)

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>H*+L</td>
<td>28.2%</td>
<td>53.1%</td>
</tr>
<tr>
<td>H-</td>
<td>70.4%</td>
<td>68.6%</td>
</tr>
<tr>
<td>L%</td>
<td>92.0%</td>
<td>91.4%</td>
</tr>
</tbody>
</table>

Table 3: Categories of errors (total = 1950)

<table>
<thead>
<tr>
<th>Deletion</th>
<th>H*+L</th>
<th>34.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-</td>
<td></td>
<td>12.8%</td>
</tr>
<tr>
<td>L%</td>
<td></td>
<td>8.3%</td>
</tr>
<tr>
<td>Insertion</td>
<td>H*+L</td>
<td>9.8%</td>
</tr>
<tr>
<td>H-</td>
<td></td>
<td>9.3%</td>
</tr>
<tr>
<td>L%</td>
<td></td>
<td>8.5%</td>
</tr>
<tr>
<td>Substitution</td>
<td>H*+L ⇒ H-</td>
<td>10.9%</td>
</tr>
<tr>
<td>H- ⇒ H*+L</td>
<td></td>
<td>5.7%</td>
</tr>
<tr>
<td>others</td>
<td></td>
<td>0.8%</td>
</tr>
</tbody>
</table>

obey the constraints, the rules obtained by the learner should contain conditions like the one that H- is assigned only when the preceding label is r, etc. By this mechanism, our method is expected to surpass the auto-labeling method in the original tilt model [9], which does not use this kind of detailed constraints.

4. PRELIMINARY EXPERIMENT

We evaluated the performance of our method by using 8-fold cross-validation. The data used for training and test are 400 utterances of a professional male speaker taken from a corpus of phoneme-balanced sentences [5]. It is read speech rather than dialogue speech, but is a good starting point to provide the baseline of labeling by our method. Since the complete identification of event locations is unrealistic and is not necessary for application, we use approximate matching, which allows a label within a margin of 50 msec from the right location to be correct.

The results are shown in Table 2. This is fairly good compared to earlier works [1]. However, a low recall of H*+L indicates that we missed a great number of lexical accents which had been identified by the human-labeler. Table 3 shows the categories of errors. Missing H*+L amounts to one third of the total errors. Other major types of errors are missing H- and confusion between H*+L and H-, both of which amount to over 12% of the errors.

An investigation on the error data showed that most of the major errors come from either of two sources: (i) gentle slope or short duration in a rise/fall region and (ii) an unvoiced region which obscures potential rise/fall. The former is difficult also for human labelers, and it is unclear whether those cases should really bear tone labels.\(^3\) The latter is more critical:

\(^3\)Note that, in labeling these cases, our labeler might refer to phonological information such as accent patterns of words.

5. CONCLUSION

We proposed a new method for automatically labeling Japanese prosody. We utilized J-ToBI style description of tonal events, and applied a decision tree learning method to obtain the rules for tonal event detection, which reflect the linguistic constraints on the tone structure of Japanese. We achieved a fairly good performance in a preliminary experiment using a read speech corpus.

6. ACKNOWLEDGMENTS

We would like to thank Yoko Ohta for her lecture on the J-ToBI labeling and John Fry for proofreading the paper.

7. REFERENCES

2. Entropic Research Laboratory, Inc. ESPS/waves+ 5.1.1 Reference Guide, 1996.