A Study On Japanese Prosodic Pattern And Its Modeling In Restricted Speech

Tomoki HAMAGAMI †, *Ken-ichi MAGATA and Mitsuo KOMURA*
SECOM Intelligent Systems Laboratory, SECOM CO.,LTD.
8-10-16 Shimorenjaku, Mitaka, Tokyo, 181 JAPAN.

ABSTRACT

This report proposes a simple and practical model for generating relatively monotonous, but sufficiently natural, prosodic features by analyzing restricted natural speech. The basic assumption of this model is that the natural F0 pattern can be obtained without complicated linguistic analysis. To achieve this prosodic control, we have analyzed and modeled this speech subject that is recoded so that it will appear in the following. First we composed the hypothesis that a Japanese Major Phrase (MP) could be modeled with the combination of a minor phrase (mp) sequence limited to fewer than three. The number of the combination is decided by the accentual type of minor phrase and intrasentence position.

The combination types have 28 patterns. To confirm the hypothesis, the restricted speech (RSP) subjects were collected and analyzed by having the speaker utter the subject sentence without emotional effect or attention to prosodic features. Furthermore, to evaluate the performance of the model, a pattern-matching process (two-level DP) was used between the synthesized F0 pattern and the restricted real F0 pattern. We thus confirmed that our model would create a synthesized F0 pattern sufficiently similar the restricted-speech patterns.

The synthesized speech using this model sounds relatively monotonous, but is sufficiently natural as compared with general spontaneous speech.

1.INTRODUCTION

Prosodic features, especially the F0 pattern, generally play an important role in spontaneous (or unrestricted) natural speech. It has been pointed out in many studies that the details of linguistic information need to be analyzed to construct prosodic rules for synthesizing high-quality speech [1]. This approach, however, depends on the accuracy of linguistic analysis. Thus, a large amount of computation is necessary to perform accurate linguistic analysis. Moreover, a poor linguistic analysis can often lead to serious errors in prosodic features. Consequently, until a perfect system of language analysis is formulated, no high-quality synthesized speech is possible. Conversely, Japanese native speakers read out texts without deep understanding of meaning and structure. This is because to speak naturally it must be possible to employ the "basic rule" obtained by long-term recitation experience. If this "basic rule" is used for prosodic control for the TTS (Text To Speech) system, the burden of language treatmentcan be reduced. With these considerations, restricted speech subjects were collected by making the speaker enunciate a subject sentence without emotional effect or attention to prosodic features. Such utterance is completely natural although comparatively dull. After the (restricted) prosodic pattern was analyzed, classified, and registered in the database, it is applied to the prosodic control of the TTS system.

2. PROSODY IN RESTRICTED SPEECH

In this section our way of thinking is outlined concerning the Japanese prosodic model for this research. First, it was decided that the prosodic model would be divided into a Major Phrase (MP) and a minor phrase (mp). Then it was examined regarding mismatches between the MP model and syntax information. Finally, this research states a hypothesis regarding the targeted prosodic model.

2.1 Minor Phrase Model

A Japanese prosody can generate because of the continuance of a high-low accented word. All the accented words are represented by the accented mora position (the last high-pitched mora) in the word and are categorized according to an accented nucleus (type A) or without (type N). An accented nucleus is a characteristic of an accented word that accompanies a mora, and causes the rapid-descending phenomenon of a pitch.

The minor phrase (mp) is corresponded with the word (type-A, type-N, or these coupled items), and is also represented by the tonal pattern of high-low pitch marks that are assigned in every mora. This pattern can be approximately represented by the piecewise-linear interpolation of the F0 at vowel centers, which is called the "point-pitch pattern [2]." The mp's shape is maintained by having the point-pitch pattern in the unit of mp.

In the case of natural utterances, we can observe the phenomenon that spreads the dynamic range of the accentual phrase with the word that it wishes to emphasize. It works effectively when the intentions of such utterances and deep meaning are transmitted.

It is difficult, however, to analyze a word's emphasis and sense from the surface structure of the language. Conversely, Japanese native speakers read out a text without deep understanding of its meaning and structure. The prosody of such utterances is the most fundamental accentual pattern that does not depend on the meaning of the word or the overall structure. In other words, it is the F0 pattern that is decided by the number of mora and the accentual type (type A or N).

Furthermore, the point pitch pattern of the F0 that is

†Graduate School of Science and Technology, CHIBA Univ.

normalized with its dynamic range is called the "normalized point pitch pattern[3]." Variations are defined by its definition and limited mora number. For example, when the mora character is limited to ten mora/word, the number of normalized point pitch patterns become 55 types. In this study the prosody is controlled by adjusting the dynamic range of this "normalized point pitch pattern" and the "base pitch." The normalized point pitch pattern is considered the model for minor phrases, and it is thus called a model Nmp.

2.2 Major Phrase Model

The restriction of the branch parting based on syntax structure effects the intonation rule that is the connection rule of the accentual phrase. The most typical example of this theory is as follows. In this example the unified meaning unit is uttered in one prosodic unit.

a [[私と 彼の] 友達] → a'((私と 彼の) 友達)
My and his friend

b [私と [彼の 友達]] → b' (私と (彼の 友達))
Myself and his friend

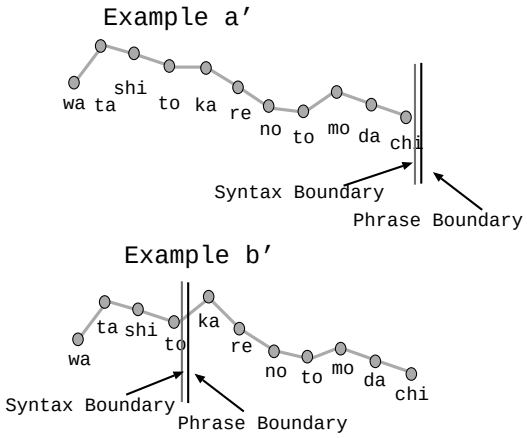

Figure 1: Example 1

Where, syntax structure is shown with [ ], and prosodic structure is shown with ( ). The structure of the surface of these two sentences is the same. The syntax structure, however, is distinguished by the difference in the prosodic structure. Generally, the boost of F0 occurs at the moment when the right-hand branch parting occurs. The unit of the prosody is then punctuated on the syntax boundary.

The downtrend phenomenon of a pitch occurs in the unity regarding this single prosodic unit, and this unit is equivalent to one Major Phrase (MP). When this downtrend is observed in detail, a pitch downtrend pattern is found that depends on a fixed phonological structure.

Generally the magnitude of an mp's dynamic range depends on the type of the mp (A or N) that precedes it. If the preceding mp type is "A," the following mp's dynamic range is smaller. Conversely, if the preceding type is "N," the subsequent range is comparatively large. This phenomenon is caused in conjunction with "catathesis" or "downstep." [4]

This process changes the dynamic range of the sequence of mp's one after another, and plays a part that lends naturally to the overall phrase pattern. To achieve the process artificially, the Major Phrase (MP) Model is needed. The purpose of using MP is to tie mp naturally within the unit concerning the prosodic unit. As mentioned above, close relationships are found between the MP boundary and the right-hand branch-parting syntax structure.

As for the following examples, however, the prosodic pattern of the utterance is not always reflected in the syntax structure.

c [[[私と 彼の] 友達の] 親達] →
c' ((私と 彼の) (友達の 親達))
The parents of myself and his friend.

d [私と [[彼の 友達の] 親達]] →
d' (私と ((彼の 友達の) 親達))
I and the parents of his friend.

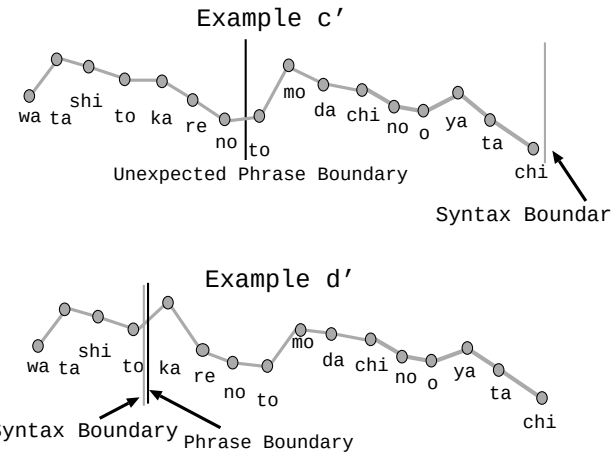

Figure 2: Example 2

Example d is read in a way that reflects syntax structure d'. In example c, however, four phrases are not necessarily always made the unit of one MP. In many cases it becomes an example such as c', and a division is observed that differs from the syntax structure. A "mismatch" often and sometimes occurs between the unit of syntax structure and unit of the MP, and finds it easily formed as the number of mp to combine becomes large or as the structure of the sentence becomes complex. This phenomenon was estimated from previous research[5].

On the other hand, Japanese native speakers can read out text without deep understanding of its meaning and whole structure. Furthermore, because human's short term memory may not be so large, a length of MP is expected to consist of a few mps at simple reading process. This prospect is observed from the result of our pre-experiment that analyzed the restricted-utterance data. According to the pre-experiment, MP of that utterance data was expected to be made from less than about three mps.

Moreover, there is another our study that supports this estimation[6]. In that study, there exists a possibility that we would utilize restricted speech by using a fixed prosodic pattern. This prosodic pattern could be characterized by a surface key which is classified into three types.

We paid attention to the estimation of mention above and formulated the following hypothesis.

When it is uttered very simply (or is restricted), the prosodic-formation process is controlled into a "window" of fewer than three words("*bunsetsus*"). A natural prosody can be given in the window by repeating a basic prosodic pattern. Accordingly, one MP can be observed with an mp sequence that is fewer than three.

To collect the utterance subject that is a match for this hypothesis, and to build an MP model, the restricted speech (RSP) subjects were collected. The RSP is speech that arranged the prosody by making the speaker enunciate a sentence without emotional effect or attention to prosodic features. The prosodic pattern in this speech is assumed to be one that could be generated only by surface information. An easy and breakdown-resistant prosodic model will be obtained by analyzing and modeling this prosodic pattern.

3. DMP PATTERN AND ITS MODELING

Modeled was the MP pattern based on the above hypothesis. Moreover, it was proposed that the prosodic-formation technique be combined with the mp-Model (normalized point pitch pattern). The validity of the model is then verified by using a pattern-matching technique (two-level DP) between the restricted F0 pattern and the synthesized ones.

3.1 Discrete Major Phrase (DMP) Model

The mp's dynamic range pattern is handled with this model as an MP shape. In other words, the MP shape is decided to be shown by a given value. This model was decided to be called "DMP (Discrete Major Phrase." Figure 3 shows an example of the "DMP" pattern. The types of DMP were divided into 28 patterns that depend on (a) the mp type (A or N) and (b) the MP's position within the sentence. These patterns, however, could be added in the process of improvement. We nevertheless believe that the patterns are sufficient for representing the prosodic features of restricted speech. Each of the 28 patterns–"the type of DMP"–is as follows.

1. $[mp-type]/[location]$
2. $[mp-type][mp-type]/[location]$
3. $[mp-type][mp-type][mp-type]/[location]$

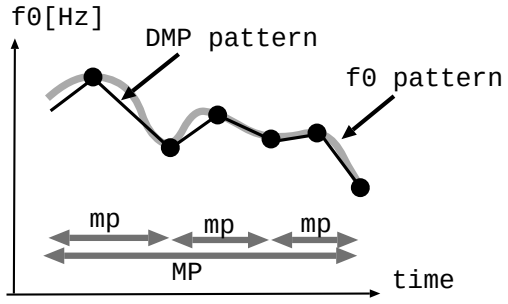

Figure 3: Discrete Major Phrase (DMP) Pattern

One mp-type can take two types of "A" (with accented nucleus) or "N" (no accented nucleus). The location can take two types of "C" (within the sentence) or "F" (end of the sentence). The items in the number of the chain phrase are shown in Table 1. The examples of DMP pattern are shown in Figure 4. Figure 5 explains the manner of prosodic control that uses the DMP and Nmp models. With this system the prosodic pattern is created in the following order.

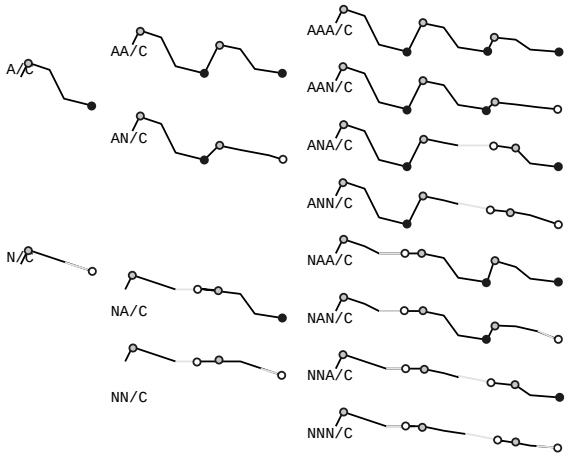

Figure 4: Examples of DMP pattern

$mp-number$	number of pattern	DMP type
1	4	A/C,A/F,N/C,N/F
2	8	AA/C,AA/F,AN/C ···
3	16	AAA/C,AAA/F,AAN/C ···
total	28	

Table 1: The types of DMP

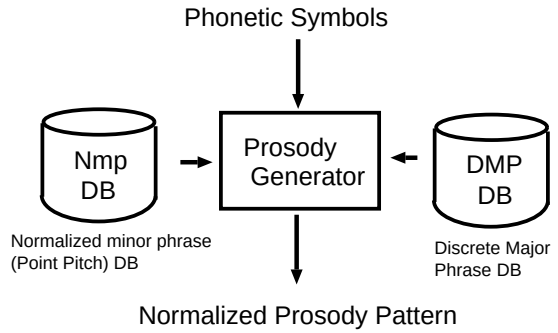

Figure 5: Prosodic pattern generation system

1. The DMP type and mp pattern that comprise the MP are input.
2. The normalized point pitch pattern corresponding to the mp pattern is read from Nmp DB.
3. The DMP corresponding to the MP pattern is read from DMP DB.
4. The mp amplitude is made elastic in accordance with the value of DMP, and the pitch pattern is created.

3.2 Discussion

To evaluate the performance of the system described above, the following experiment was attempted. Restricted speech was recorded by a professional female narrator. This utterance was checked strictly, and 300 sentences fully elucidated the nature of the item. The average length of each sentence was 4.23 s, the average number of mora was 31.2 mora, and the average speech utterance was 7.4 mora/s. The contents included an editorial, a novel, and an essay. The F0 patterns of this restricted speech were analyzed after the following manner, and the DMP pattern was obtained.

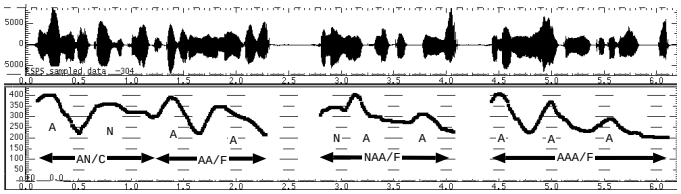

Figure 6: Example of analysis

1. The continuance F0 pattern was analyzed and taken for smoothing.
2. Labeling was executed for the maximum and minimum values of each mp.
3. Labeling was performed for the boundary of each MP.
4. The peak and valley of each mp were extracted.
5. Normalized dynamic range was performed for each MP.
6. A parameter was kept for every normalized pattern, and an DMP database was created.

Next the synthesized F0 pattern comprised of DMP and Nmp was examined so that the F0 pattern of the RSP could be shown in accordance with the following process.

1. The RSP was newly recorded and an F0 was analyzed. The average of the standard pitch height (base pitch) and dynamic range were analyzed in advance.
2. Reference F0 patterns were made by the parameter-set made from the combination of 28 DMPs and a typical Nmp.
3. Synthesized F0 patterns were formulated by using the interpolated point pitch pattern (Nmp). The interval on point pitch was adjusted to the utterance average (7mora/s=143ms/mora).The base pitch and dynamic range were fitted to the RSP.
4. Two-level DP matching was done between the RSP F0 pattern and the synthesized F0 pattern.
5. The sequence of the synthesized F0 pattern closest to the REP F0 pattern was selected. This was done because the validity of the synthesized F0 pattern could be evaluated from the error.

Accordingly, the average of the mp's dynamic range error was about 5 percent. It was thus ascertained with RSP's F0 that it could be expressed further and fully with the combination of a synthesized F0. Figure 7 shows the results of attempted two-level DP.

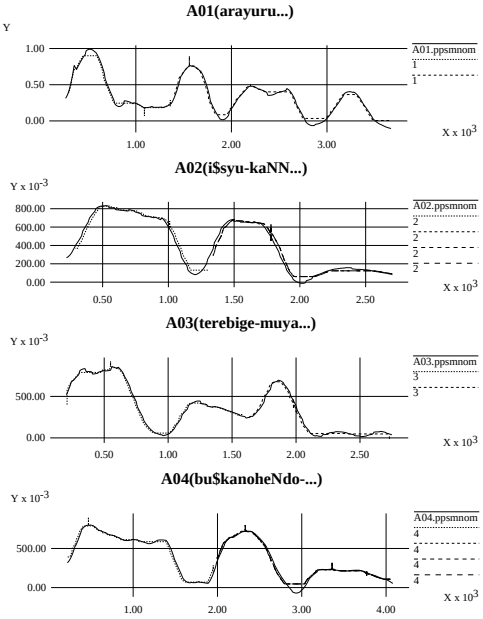

Figure 7: The results of 2-level DP

4.SUMMARY

The Discrete Major Phrase (DMP) Model has been proposed for the Japanese TTS system. It is useful for creating synthesized speech having natural prosodic features. This model is based on restricted speech. It restricts the number of minor phrases to three and prescribes its dynamic ranges. The pattern type is classified in 28 variations according to being nuclear or nuclear-free, by the accentual-phrase sequence, and by the intrasentence position.

It was found by using these DMPs that fairly satisfactory precision could be achieved as a result of the two-level DP matching between the restricted F0 pattern and synthesized ones. Natural Japanese utterances could be modeled by combining this model with a language-processing portion that treats the sentence's surface information.

References

[1] Fujisaki,H., Hirose, K. and Takahashi, N., "Manifestation of linguistic and para-linguistic information in the voice fundamental frequency contours of spoken Japanese", Proc. ICSLP-90 (1990)

[2] Hashimoto,S., " Several Features of Japanese Word Accent",Trans. IECE,vol.56-D No.11 (in Japanese)(1973)

[3] Hamagami,T. and Komura, M., " The Extended Point-Pitch Model for Prosody Contral",Proc. ASJ Fall Meeting (in Japanese)(1994)

[4] Pierrehumbert and Beckman.,"Japanese Tone Structure" ,The MIT Press (1988)

[5] Kubozono,H.,"The Organization of Japanese Prosody", Kuroshio Publishers (1993)

[6] Magata,K., Hamagami,T. and Komura, M., " A Method For Estimating Prosodic Symbol From Text For Japanese Text-To-Speech Synthesis", Proc. ICSLP-96 (1996)