A Method For Estimating Prosodic Symbol From Text For Japanese Text-To-Speech Synthesis

Ken-ichi MAGATA, Tomoki HAMAGAMI and Mitsuo KOMURA

SECOM Intelligent Systems Laboratory, SECOM CO.,LTD.
8-10-16, Shimorenjaku, Mitaka, Tokyo, 181 JAPAN.

Abstract

This report describes a method for estimating the separation degree at the bunsetsu boundary (SD) for Japanese text-to-speech synthesis. Our method gives us the prosodic symbol without using complicated linguistic analysis.

First we classify bunsetsus according to the final morpheme. Each classified bunsetsu has a temporary separation degree in advance. We call this "the estimated separation degree" (ESD). ESD is derived from the SD's statistical tendency regarding each bunsetsu. The SD is decided by rules that correct the ESD as an initial degree. Correction rules are constructed by comparing the ESD, and the SD is observed from natural speech to cancel the frequently occurring mismatches.

An absolute evaluation test of five grades was performed upon 300 sentences with prosodic symbols given by our method. As a result, the ratio of "Natural" and "Somewhat unnatural but tolerable" exceeded 2/3. The proportion of "Serious error" was less than 10%, thus giving us satisfactory results.

1. Introduction

In the text-to-speech synthesis system it is very important to decide proper prosodic features. Prosodic features, especially intonation and pause length, strongly depend on linguistic information. Several studies have pointed out that "the separation degree at the *bunsetsu* boundary" (SD) provides many prosodic features of spoken Japanese[1][2]. *Bunsetsu* is the smallest meaning block consisting of more than one Japanese morpheme. An SD shows the strength of the connection between *bunsetsus*. In the text-to-speech synthesis system, the general form of the prosodic unit, pause length, etc. can be generated by SD. In general, SD has been decided by modifier/modifyee relationships, meanings, etc. as obtained by complicated linguistic analysis[3][4]. It is difficult correctly to obtain all SD because natural language is complex. A poor linguistic analysis often leads to serious errors in prosodic features. Furthermore, we have already known cases where linguistic information cannot explain actual speech phenomena. For instance, we have limits of words for utterance when we read out a sentence without taking a breath. Such cases only become clear by analyzing natural speech. Thus the previous method could not apply these cases to the rule giving SD. (For the decision of pause insertion points, a method using a statistical analysis of utterance has been reported[5].) We observe that Japanese native speakers read out a text without deep understanding of its meaning and structure. We therefore set up the hypothesis that Japanese speakers estimate SD based on the last morpheme of *bunsetsu* while reading out a Japanese sentence. In our method the database has the advance statistical tendency of SD. This is called the "estimated separation degree" (ESD). This database gives the ESD as an accurate initial degree without complicated linguistic analysis.

Next the ESD is modified based on the relationships with the adjacent *bunsetsus*. Correction rules are constructed by referring to natural speech. Accordingly, these rules can accurately estimate SD. The process of estimating SD is shown in Figure 1.

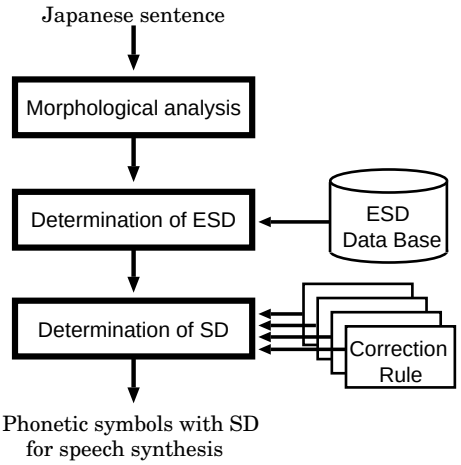

Figure 1: System overview

In this report we will first present how to make the rules. Next will be an example of applying the rules. Finally we will report the evaluation of SD as a prosody symbol and its result.

2. Estimated Separation Degree and Correction Rules

In this chapter we explain the "*bunsetsu*" process unit and its type. Next discussed is how to make the ESD and correction rules.

2.1 Bunsetsu Type

A Japanese sentence is comprised of a row of *bunsetsus*. *Bunsetsu* consists of a content word with or without the succession of a string of dependent words. Japanese text has several kinds of writing methods, such as *kanji* (Chinese characters), *kana* (Japanese syllabaries), Roman alphabet etc. Inasmuch, however, as a dependent word is only written in *hiragana* (one of the Japanese syllabaries), it often becomes a surface key for understanding sentence meaning. We assumed that the reason why Japanese native speakers can read out a text without fully understanding its meaning and structure is that Japanese have empirically known the SD tendency of dependent words. We thus classified *bunsetsu* under 48 types according to the last *bunsetsu* morpheme. We call these types "*bunsetsu type*" (BT). Specifically, if the final *bunsetsu* morpheme is a particle, its BT is denoted by that morpheme (ex., "*ga*" and "*wo*"). If the final *bunsetsu* morpheme is inflectional, its BT is denoted by this inflection (ex., *renyoh-form* and *rentai-form*). Otherwise BT is denoted by the content word's part of speech (POS).

2.2 Estimated Separation Degree

We assumed that Japanese speakers estimate SD based on the final *bunsetsu* morpheme. To validate this proposition we investigated the statistical tendency of the relationship between SD and BT. We studied the modification distance to obtain objectively a large number of data. The "modification distance" (MD) is the number of *bunsetsus* between the *bunsetsu* modifier and the modifyee. It has already been reported that there is a strong correlation between SD and MD[1]. According to this report, the longer the MD, the weaker the relationship with the next *bunsetsu*. Thus the SD of the *bunsetsu* is high. In other words, the longer the MD of *bunsetsu*, the higher its SD. Conversely, SD becomes the lowest degree if a *bunsetsu* modifies the succeeding *bunsetsu*. Because MD is given an upper limit according to the number of *bunsetsus* before the end of the sentence, the sentence position of *bunsetsu* must be investigated. To express the position within a sentence, we use the number of *bunsetsus* from the end of the sentence[6]. When the total number of *bunsetsus* is n, the position of the i-th *bunsetsu* equals $n - i$. An example of our investigation is shown in Figure 2.

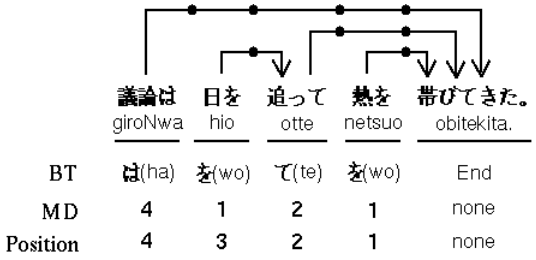

Figure 2: Example of BT, MD and position in a sentence. Input sentence means that "As a day passes, discussion has gotten heated".

We analyzed all MD using 1,645 sentences taken from newspaper articles. (The average number of *bunsetsus* in a sentence is 8.0.) From our analysis, it became clear that each BT has own modification tendency. This shows that it is possible to estimate MD from BT. As for the relationship between the position of a *bunsetsu* in a sentence and its MD, we obtained that MD falls within three types as in Figure 3.

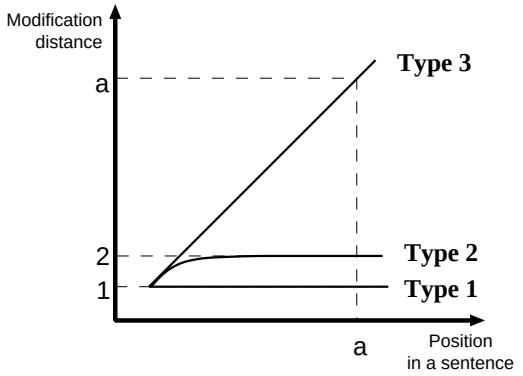

Figure 3: Tendency of the relationship between MD and position in a sentence

These are prescribed as SD types 1, 2, and 3, respectively. This shows that *bunsetsus* can be classified according to the following three types by the modification tendency.

1. The *bunsetsu* modifying the next *bunsetsu*
2. The *bunsetsu* modifying the next but one
3. The *bunsetsu* modifying the last *bunsetsu*

This result does not contradict with the statistical results of previous research[6]. Accordingly, all BT were classified into these three types. We consider that

these three types are the "estimated separation degree" (ESD), and we thus created the database shown in Table 1. ESD is thus decided by BT. We assume that this is the basic SD that Japanese native speakers estimate while reading.

Table 1: Table of Estimated Separation Degree

Bunsetsu type (48 types)	**ESD**
Particle "*no*"	1
Particle "*keredo*"	3
Particle "*shika*"	2
.....	...
Inflection "*Rentai-form*"	1
.....	...
POS "*Noun*"	2
.....	...

2.3 Correction Rules

ESD is defined as the most accurate SD. Inasmuch, however, as there are infinite expressions in a sentence, not all ESD are always correct. We assume that Japanese native speakers finally decide the *bunsetsu* SD after correcting its ESD according to the relationships of adjacent *bunsetsus*. Because human's short term memory may not be so large, we assume that people use the combined information of adjacent *bunsetsus*. To investigate whether these corrections have regularity we compared the ESD with the SD observed in natural speech. From this comparison we constructed correction rules to cancel the frequently occurring mismatches. For example, if it occurs frequently that SD becomes one when BT "ga" precedes a declinable word, we use this as a correction rule. The advantage of referring to natural speech is that we are able to construct rules that cannot be estimated by linguistic analysis. Specifically, we often observe a pitch pattern that seems to consider two *bunsetsu* as one. This case cannot be explained by linguistic theory. To resolve this point, our correction rules provide "SD type 0", which means the strongest relationship between *bunsetsus*. Accordingly, although ESD have three types –1 to 3– correction rules give four SD types of 0 to 4. We thus analyzed restricted-speech subjects as natural speech, which is done by speaking without emotional effect or prosodic intention. The reason we collected these subjects is that native speakers tend to read out a sentence in this manner if they read it for the first time. Restricted speech has relatively monotonous, but sufficiently natural, prosodic features. By our referring to restricted speech, it became clear that the F0 pattern could be modeled very simply[8]. We think this prosodic pattern shows the most fundamental feature. Under strict checking, restricted speech was recorded by a professional female narrator, and its SDs were determined. As a result of our reference to natural speech, we constructed rules that make ESD correct with high precision.

3. The SD Decision-Making Process

With these rules we constructed an SD generator. With this it is possible to construct a linguistic process for a Japanese text-to-speech synthesis system. An example of this process is shown in Figure 4.

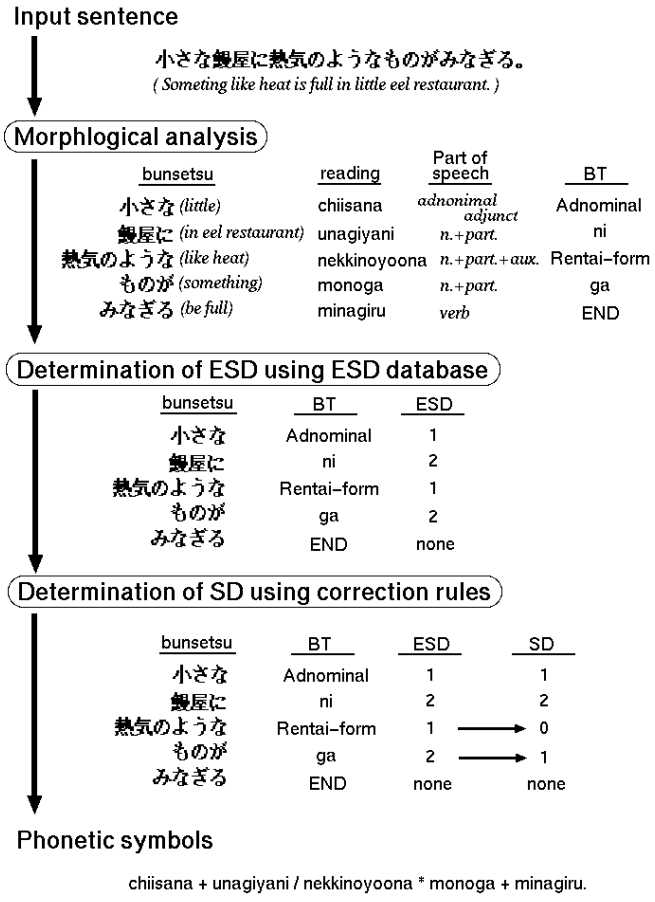

Figure 4: Example of the SD decision. Four SD types of 0,1,2 and 3 correspond to phonetic symbols of "*", "+", "/" and "|", respectively.

First, an input Japanese sentence is broken down into morphemes, parts of speech, reading, accentual type, etc. by morphological analysis. BT is confirmed simultaneously. Next, ESD is given by the ESD database. ESD is not given to the last *bunsetsu* of the sentence.

Correction rules correct the ESD by relationships with adjacent *bunsetsus*. With the shown example, correction rules correct the ESD of two *bunsetsus*, which are "*nekkinoyoona*"(like heat) and "*monoga*"(something). This is because the following correction rules are applied.

- "*nekkinoyoona*" **ESD:1 → SD:0**
  Correction Rule: *When a BT "Rentai-form" without a punctuation mark precedes a bunsetsu whose content word is a noun, and the number of Japanese moras is fewer than three, its ESD type becomes 0.*
- "*monoga*" **ESD:2 → SD:1**
  Correction Rule: *When BT "ga" precedes a declinable word, its ESD type becomes 1.*

Finally, after deciding the SD of all *bunsetsus*, this information is converted into phonetic symbols.

4. Evaluation

To evaluate this system of estimating SD, a subjective-opinion test in five grades was conducted among three native speakers. Subjects read 300 printed sentences with the SD sign, and they evaluated the SD accuracy in each sentence. (The average number of *bunsetsus* in a sentence is 7.0.) 300 sentences were prepared from diverse fields such as an editorial, a novel, and an essay. Sentences were evaluated by using the following five grades[7].

- "Natural" (score 5)
- "Somewhat unnatural but tolerable" (score 4)
- "Somewhat unnatural" (score 3)
- "Unnatural" (score 2)
- "Serious error" (score 1)

As a result of evaluation, the average score was 3.7. A breakdown of the scores is shown in Figure 5.

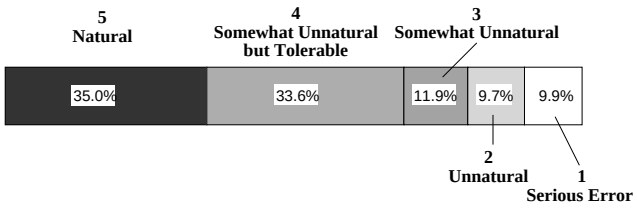

Figure 5: Subjective-opinion test result

It was found that the ratio of "Natural" and "Somewhat unnatural but tolerable" was greater than 2/3. The proportion of "Serious errors" was less than 10%, so that we thus obtained a satisfactory result.

5. Concluding Remarks

A method of estimating SD has been proposed for Japanese text-to-speech synthesis. Using this method, SD is estimated in the following manner.

1. BT is decided by the last *bunsetsu* morpheme.
2. ESD as an initial degree is decided by using the ESD database.
3. Correction rules correct ESD into SD.

Our method is far simpler than the one that strictly analyzes sentences. An absolute evaluation test of five grades was performed upon 300 sentences with prosodic symbols provided by our method. As a result the ratio of "Natural" to "Somewhat unnatural but tolerable" exceeded 2/3. The proportion of "Serious errors" was less than 10%, thus giving us satisfactory results. We concluded that the proposed method gives us the prosodic symbol without using complicated linguistic analysis. Inasmuch as we refer to SD-derived F0 patterns from natural speech, our proposed SD would be useful for generating the F0 pattern of synthesized speech. In the same way our method could be applied to devise other prosodic features. We intend to evaluate the synthesized speech created using our method.

References

[1] K.Hakoda and H.Sato, "Prosodic Rules in Connected Speech Synthesis", *Trans. of the Institute of Electronics and Communication Engineers of Japan, vol.63-D, no.9*, 1980 (*in Japanese*).

[2] J.Komatsu, T.Sakayori, S.Sasae and H.Kitagawa, "A Method of Deciding Connection Degree between Bunsetsu's for Text-to-Speech System", *Proc. of the Fall Meeting of the Acoustical Society of Japan, pp.165–166*, 1988 *(in Japanese)*.

[3] K.Suzuki and R.Teranishi, "Parsing algorithm for text-to-speech in Japanese", *Journal of the Acoustical Society of Japan, vol.44, no.5*, 1988 *(in Japanese)*.

[4] H.Kawai, K.Hirose and H.Fujisaki, "Rules for generating prosodic features for text-to-speech synthesis of Japanese", *Journal of the Acoustical Society of Japan, vol.50, no.6*, 1994 *(in Japanese)*.

[5] K.Iwata, Y.Mitome and T.Watanabe, "Pause rule for Japanese text-to-speech conversion using pause insertion probability", *Proc. ICSLP-90, pp.837–840*, 1990.

[6] H.Maruyama and S.Ogino, "A Statistical Property Of Japanese Phrase-to-Phrase Modifications", *Journal of the Mathematical Linguistic Society of Japan, vol.18, no.7*, 1992.

[7] S.Itahashi, "Guideline for text-to-speech synthesizer evaluation", *Journal of the Acoustical Society of Japan, vol.52, no.2*, 1996 *(in Japanese)*.

[8] T.Hamagami, K.Magata and M.Komura, "A Study On Japanese Prosodic Pattern And Its Modeling In Restricted Speech", *Proc. ICSLP-96*, 1996.