A left-to-right processing model of pausing in Japanese based on limited syntactic information*

Hajime Tsukada

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
tsukada@itl.atr.co.jp

ABSTRACT

This paper proposes a new model for determining where to pause in Japanese. Our model, which simulates the left-to-right processing of human speech production, incorporates the two main factors of pausing: phrase length and limited syntactic information. By taking this approach, our model can explain several types of pausing phenomena in Japanese, which cannot be explained by the naive pausing model. To prove the validity of our model, we implemented it in a text-to-speech conversion system. Through a listening test, we compared the proposed model with a naive pausing model, mainly based on phrase length. Results show that the proposed model performs well.

1. Introduction

Pausing is one of the most important factors for enhancing the naturalness of synthesized speech. Therefore, it is desirable to implement a high performance pause model in a text-to-speech conversion system. In previous studies [5, 10, 12], it was stated that pausing is strongly affected not only by phrase length but also by syntactic structure. However, in determining where to pause, it seems that humans decide this according to local information rather than global information.

For example, one can pause fairly easily, when reading aloud text seen for the first time from left to right. In this case, pausing might be used after a phrase that is independent from the next following phrase, or after a word of a lexically independent-nature such as a conjunction. We call this the promotion of pausing. In contrast, pausing might be suppressed after a phrase that modifies another phrase immediately following it. Our experience suggests that local syntactic information strongly controls pausing. Hakoda [6] experimentally pointed out that the syntactic relationship between neighboring phrases is a main factor of the syntax for pausing. Moreover, Miyazaki [11] developed a pause model based on syntactic dependency between neighboring phrases.

However, many of the previous pause models implemented in TTS conversion systems determine pause positions syntactically with little consideration given to phrase length. As a result, we often find Japanese sentences with inappropriately determined pause positions.

Moreover, pause positions vary because these positions are not strongly restricted. However, previous models have explained the appropriateness of pause positions not relatively but absolutely. Therefore, these models at times could not find an appropriate position for the next pause after a pause had been inserted into the sentence. This performance is unsuitable for a Japanese text-to-speech conversion system because Japanese punctuation marks appear at various positions in sentences where pauses tend to be inserted.

To overcome these problems, we propose a left-to-right processing model of pausing based on limited syntactic information. In this model, a pause is decided according to the length from the previously paused position, syntactic factors and the length toward the predicted next pause position. First, we explain problems of the naive model. Second, we propose our left-to-right processing model and describe factors which strongly control pausing in Japanese. Finally, an implementation of our model is explained, and the results of evaluations are shown. Results show that our proposed model performs well.

2. Problems of previous approach

2.1. Japanese syntax and prosodic phrases

Japanese minor-phrases, called *bunsetsu*, are a fundamental unit of syntax; they are single/compound content-words or postpositional phrases followed by functional words. [1] In the Japanese language, grammatical relations between phrases such as *cases* are roughly encoded in the preceding minor-phrase, by functional words or a conjugation.

In Japanese, a pitch pattern is modeled so that an accent component of an accent phrase is superposed on a phrase component of an intonational phrase. Intonational phrases are composed of several accent phrases, and most intonational phrases are bounded by a pause. In Japanese, accent phrases correspond rather closely to the minor-phrases. On the other hand, because of the structural nature in English, the previous studies [1, 2, 3] concerning English prosody had difficulty in corresponding a syntactic structure with a prosodic one. As for Japanese, however, Abnye's *chunks* [1], Bachenko's *phonological phrases* [2], and *ϕ-phrases* in Gee's work [3] are given

*This study was carried out in NTT Human Interface Laboratories as a part of a project to develop a Japanese text-to-speech conversion software system[8].

[1] In Figures 1, 2, and 3 syntactic categories are expressed functionally. Therefore, proper syntactic categories such as PP do not appear.

as the minor-phrases.

Japanese syntax is simply expressed by a binary tree, as shown in Figures 1, 2, and 3. In a sentence, a subject, object, and complement precede a verb, and the order of these elements can vary. Therefore, the order of Japanese words, roughly speaking, is inverse to that of English except for the subject-verb order.

In natural language processing in Japanese, partial syntactic structures, called dependency structures, are widely used. In this structure, only the dependency between a head minor-phrase and its sibling minor-phrase is considered. Figures 1, 2, and 3 also illustrate dependency structures under the sentences. Abney [1] defined chunks and dependencies between chunks in English. However, this kind of representation had been very commonly applied in Japanese.

2.2. Problems of naive pause model

In the naive pause model, a pause tends to be located at the right branching-point of the tree structure. In the dependency structure, this point is expressed at the boundary between adjoining minor-phrases that have no dependency. By using the syntactic dependency between minor-phrases, the syntactic factor in pausing is easy to understand.

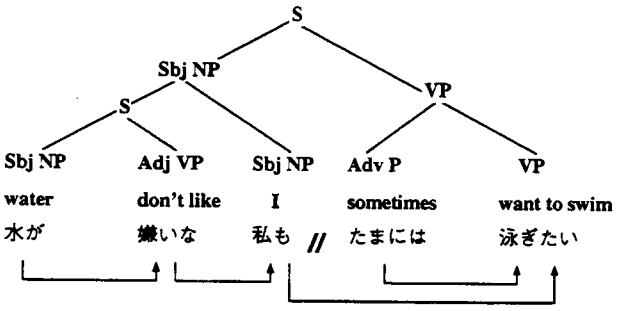

Figure 1: Example of correct pausing

However, some types of sentences cannot be handled by the naive model, as shown in Figures 2 and 3. The most relevant point for pausing in Figure 2 is point (1), but the naive model would insert a pause at point (2). In Figure 3, the most relevant position to pause in the naive model is point (4) or (1), which are between minor-phrases that have no dependency. However, the most relevant position is actually point (3) or (2).

These findings, pointed out in Figures 2 and 3, indicate that the syntactic factor does not always take precedence over the phrase length factor. The problem in Figures 2 and 3 often occurs in sentences where a minor-phrase continuously depends on the following minor-phrase, or when a short and phonologically weak minor-phrase appears as the major component of the sentence. Words such as "ことが", "ある", "ない" are examples of short and phonologically weak words. These words also have a syntactically dependent nature.

Although the previous studies [5, 10, 11] on application to Japanese text-to-speech conversion system, also incorporated the phrase length into syntactic factors, the most relevant point for pausing was decided with little consideration of phrase length factors. Consequently, there were cases of pausing before short phrases that were phonologically weak. This is quite unnatural.

Another point is that a pause position is decided not absolutely but relatively. (3) is one of the most relevant pause positions in Figure 3. However, if the reader pauses at (2), (4) would be the more relevant choice of the next pause rather than (3). Therefore, pauses should be decided according to the previous position of pausing.

Selkirk [12] modeled the appropriateness of pause position by the number of beats inserted by the syntactic conditions. The performance structures [1, 3, 4], which are proposed for combining psychology and linguistics accounts, hierarchically expresses the separation degree between sentence components with binary trees. These two types of representations are useful to express the degree of pausing. However, the phenomenon that explains how pause positions are relatively determined is difficult to understand directly by these models. Also, many of the studies [5, 6, 10, 11] concerning Japanese pausing give little consideration to the fact that pause positions are relatively determined.

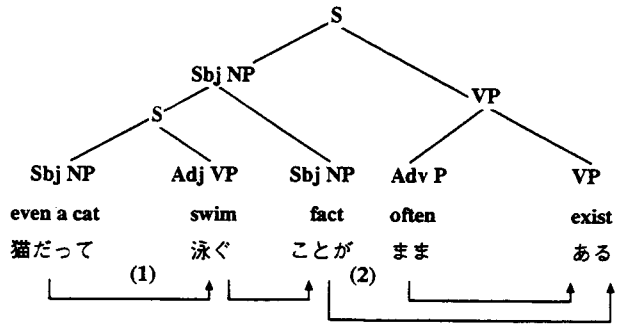

Figure 2: Simple example of incorrect pausing

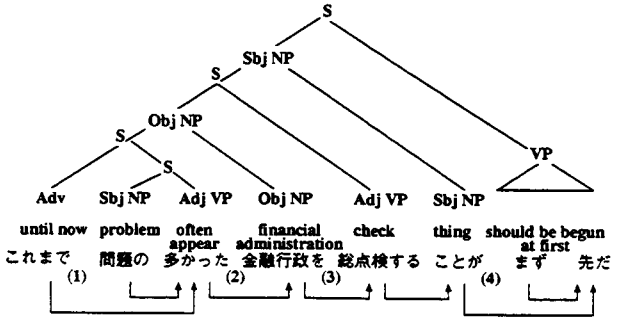

Figure 3: Complex example of incorrect pausing

3. Left-to-right processing model

To solve the problems described in the previous section, we propose a left-to-right processing model, in which pause positions are determined according to previous pause positions with consideration to the syntactic factors.

Another point is that the proposed model takes account of the scale

factor of prosodic structures. A larger prosodic structure often unites with a following small structure that is dependent. This sometimes conflicts with proper syntactic structures, as shown in the previous section. This involves the pause bounded phrase level and the accent phrase level. As for the former, the length toward the predicted next pause position is considered (see Section 3.1 (1)). As for the latter, the length of the following minor-phrase is considered (see *a structural factor* in Section 3.2).

We summarize our concept of the syntactic factors in our model here. One is that the strength of connection between minor-phrases is as follows: clause boundaries < boundaries of the sentence major categories such as subject, object, etc. < boundaries between minor-phrases that are the elements of a sentence's major categories. This idea reflects the Selkirk account [12]. Another point is that the minor-phrases that have a prominent nature tend to become independent of others.

With this model, the problems described in the previous sections are properly explained.

3.1. Default action to pause

Moving from left to right for the minor-phrase boundaries, pauses are inserted as follows:

1. If the length of the current phrase being processed exceeds a fixed value l, the phrase is paused on condition that the length of the following anticipated phrase, which might not be paused, is longer than a fixed value $\hat{l}$. The constants l and $\hat{l}$ balance the phrase length.
2. If a position being processed encounters some syntactic condition, 1 is broken. As a result, pausing is either suppressed or promoted as described in Sections 3.2 and 3.3, respectively.

In a case like that shown in Figure 3, a minor-phrase continuously depends on the following one. As a result, there is no chance for pausing. When the current phrase length being processed plus $\hat{l}$ exceeds the maximum length, a pause is inserted at the position where the connection is relatively weak under the condition constrained by l and $\hat{l}$.

Independently of the above phrase length constraint, punctuation marks are cues for pausing. In the Japanese language, punctuation marks appear at various positions in sentences. In many cases, a pause is inserted at the punctuated position. Therefore, our model inserts pauses based on punctuation marks.

3.2. Suppression factors

A syntactic factor
After an adjective or adverbial minor-phrase composes a major component of a sentence together with the following phrase, pausing is suppressed. Major components of a sentence are the subject, object, complement and verb phrase.

A structural factor
If a minor-phrase depends on another very short one that follows, the following phrase belongs to the previous one. As a result, the two minor-phrases combine into one accent phrase, although an accent phrase usually corresponds to a single minor-phrase. When this structural factor combines with the above syntactic factor, suppression is increased.

3.3. Promotion factors

The following syntactic and structural factors promote pauses.

1. After a conjunction or an interjection.
2. After a topicalized minor-phrase or prepositive adverbial minor-phrase.
3. After a minor-phrase that is at the end of a sentence.
4. Before a large structure such as an embedded clause or a coordinate structure.

4. Evaluation

To prove the validity of our pause model, we implemented it in a Japanese text-to-speech conversion system [8] and compared the naturalness of pause insertion with that of the previous system [7].

4.1. Implementation

To implement our model in a TTS conversion system, we simplified it and added some practical devices.

- We ignore degree of pausing. A pause is always inserted at the promoted position without regard to other factors such as speech rate. Also, pauses are never inserted at the suppressed position on condition that the length of the phrase being processed is shorter than the maximum length.
- The length of the following anticipated phrase in section 3.1 was approximately calculated by the string length toward punctuation marks while considering the types of characters. l and $\hat{l}$ were experimentally set to 10 and 9 morae, respectively, under about a 140 msec/mora speech rate. To be exact, these parameters should depend on the speech rate.
- In the case where minor-phrases continuously depend on the phrase immediately following them, the currently processed phrase was simply paused in terms of the maximum length without considering the following syntactic structures. This simplification harmed the accuracy of our model. However, even a human would have had difficulty inserting pauses appropriately into the sentences evaluated, while reading from left to right on the first pass.
- The dependency between adjoining minor-phrases was heuristically determined by (1) the types of functional words and conjugations in a preceding minor-phrase and (2) the type of content words in the following minor-phrase.
- Finding the large structure, described in section 3.3 (4) of the promotion factors, was not carried out. This also harmed the accuracy of our model. When a full syntactic analyzer is implemented, this factor will greatly improve the accuracy of pausing.

The previous system mainly determined pause positions by using only phrase length and punctuation marks. Fewer syntactic factors were considered than with our model even given the above limitations.

4.2. Results

We composed two sets of a text corpus made by transcribing radio news. Each text set roughly corresponded to 20 minutes of news, and the sum of both texts was 12,009 Japanese characters; this roughly corresponds to a 4,500 English word text. Our new and previous TTS conversion systems processed these texts, and generated intermediate data and synthesized speech. In the intermediate data, reading, accent phrase boundaries, the accent position in an accent phrase, and the pause position were encoded.

Two testees listened to the synthesized speech for each text set and checked for errors in each set of the intermediate data. From these data, we calculated the correct rate of pausing. After counting the accent phrases that had correct boundaries from both text sets, we selected the accent phrase boundaries from among those that were correctly tagged with a pause. The rate of correctly tagged boundaries could be calculated by dividing the latter selection by the former. In the Japanese language, the position of a minor-phrase boundary (i.e., an accent phrase boundary) is not trivial because word boundaries are not explicitly inserted, unlike English. Therefore, to concentrate on achieving validity with the pause model, those accent phrases incorrectly bounded have to be excluded.

Figure 4 shows the error rate of pausing. The error rate of our proposed model is 4.5%, and that of the naive phrase length model is 10.3%. Our proposed model decreases errors about twice as affectively as the naive phrase length model.

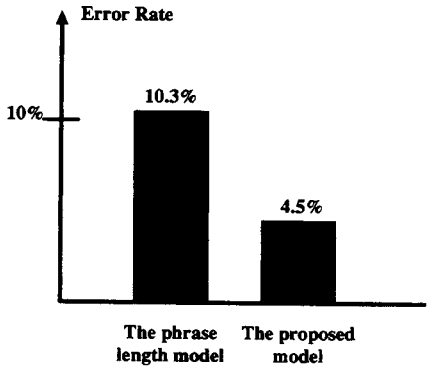

Figure 4: Error rate of pausing

5. Conclusion

We proposed a new model of pausing in Japanese, which incorporates the two major factors of phrase length and limited syntax. Our model considers the scale factors of two types of prosodic structures. One is that a short sequence of accent phrases tends to be involved with the preceding pause bounded phrase. Another is that a short minor-phrase tends to be involved with the preceding minor-phrase that depends on it, and these minor-phrases combine into one accent phrase. These involvements depend on local syntactic dependency, and are independent of the global structure. Therefore, the prosodic structures sometimes conflict with the proper syntactic structures. As a result, we could improve the naive model's ability to pause.

To prove the validity of our model, we compared the new model with the old model, mainly based on phrase length, and implemented it in the previous TTS conversion system. Results indicated that our model performs well.

Our model is still weak in explaining the degree of pausing. For this purpose, we should extend our model to a quantitative model. The stochastic methods used in [13, 9] and [6] that incorporate many factors of pausing would make good guideposts.

6. Acknowledgments

We are grateful to Miyuki Ishii of NTT-IT Co., Ltd. who evaluated our new TTS conversion system, including the proposed pause model. Also, we thank Hisako Asano of NTT Information and Communication Systems Labs. for providing a part of the results of our system evaluation. Finally, we are grateful to Masanobu Abe and Shin'ya Nakajima of NTT Human Interface Labs. for their comments on the earlier versions of this paper.

7. REFERENCES

1. Steven Abney, "Prosodic Structure, Performance Structure and Phrase Structure," Proceedings of Speech and Natural Language Workshop, 1992
2. J. Bachenko and E. Fitzpatrick, "A Computational Grammar of Discourse-Neutral Prosodic Phrasing in English," Computational Linguistics, Vol.16, No.3, 1990
3. James Paul Gee and François Grosjean, "Performance Structures: A Psycholinguistic and Linguistic Appraisal," Cognitive Psychology, Vol.15, 1983
4. François Grosjean, Lysiane Grosjean, and Harlan Lane, "The Patterns of Silence: Performance Structures in Sentence Production," Cognitive Psychology, Vol.11, 1979
5. Kazuo Hakoda and Hirokazu Sato, "Prosodic Rules in Connected Speech Synthesis (in Japanese)," Transactions of the Institute of Electronics and Communication Engineers of Japan, Vol.J63-D, No.9, 1980
6. Kazuo Hakoda, Shin'ya Nakajima, and Tomohisa Hirokawa, "A Rule for Deciding Tone Control Symbols in Text-to-Speech Conversion (in Japanese)," Technical Report of the Institute of Electronics and Communication Engineers of Japan, SP89-5, 1989
7. Kazuo Hakoda, Shin'ya Nakajima, Tomohisa Hirokawa, and Hideyuki Mizuno, "A New Japanese Text-To-Speech Synthesizer Based on COC Synthesis Method", Proceeding of International Conference on Spoken Language Processing, 1990
8. Kazuo Hakoda, Tomohisa Hirokawa, Hajime Tsukada, Yuki Yoshida, and Hideyuki Mizuno, "Japanese Text-To-Speech Software Based on Waveform Concatenation Method," American Voice Input/Output Society, 1995
9. Julia Hirschberg, "Pitch Accent in Context: Predicting Intonational Prominence from Text," Artificial Intelligence, Vol.63, pp.305–340, 1993
10. Hisashi Kawai, Keikichi Hirose, and Hiroya Fujisaki, "Rules for Generating Prosodic Features for Text-To-Speech Synthesis of Japanese (in Japanese)," Transactions of the Committee on Speech Research, The Acoustical Society of Japan, Vol.50, No.6, 1994
11. Masahiro Miyazaki, "日本文音声出力のための言語処理に関する研究 (in Japanese)," A doctoral dissertation, Tokyo Institute of Technology, 1986
12. Elisabeth O. Selkirk, "Phonology and Syntax: The Relation between Sound and Structure," The MIT Press, 1986
13. Michelle Q. Wang and Julia Hirschberg, "Automatic Classification of Intonational Phrase Boundaries," Computer Speech and Language, Vol.6, pp.175–196, 1992