ON THE RELATIONSHIP OF SPEECH RATES WITH PROSODIC UNITS IN DIALOGUE SPEECH

Keikichi Hirose and Hiromichi Kawanami
{hirose, kawanami}@gavo.t.u-tokyo.ac.jp

Dept. of Information and Communication Engineering, Univ. of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan

ABSTRACT

For the purpose of constructing prosodic rules for dialogue speech synthesis, a comparative study was conducted on speech rates between dialogue speech and read speech. Based on the generation modeling of F0 contour, we can define 4 prosodic units, such as prosodic sentence, prosodic phrase and so on. Speech rate was analyzed with respect to these units. In a prosodic sentence, dialogue speech starts with a speech rate similar to that of read speech. The speech rate then gradually increases and, after passing through the middle of the unit, decreases towards the end. Similar tendencies were also observed in lower level units, but the degree of speech rate change in a unit was smaller. Based on the above results, a preliminary rules for speech rate control were developed for dialogue speech synthesis. A hearing test showed that the developed rules could make the synthetic speech sound more dialogue-like.

1. INTRODUCTION

In most current spoken dialogue systems, speech output is generated by using text-to-speech (TTS) conversion devices or software packages commercially available. Although speech quality from these devices was recently improved very much, these are still several problems to be overcome. One of such problems is that these devices are designed to synthesize the read speech and, therefore, intonation and rhythm of synthesized speech is often too monotonous as outputs from spoken dialogue systems. From this respect, we have been conducting a comparative study on the prosody of dialogue-style and that of reading-style, to construct prosodic rules for dialogue speech synthesis. In the previous reports, we showed how fundamental frequency (F0) contours and speech rates of dialogue speech differ from those of read speech [1, 2, 3], and constructed prosodic rules through the multiple regression analysis on how various factors affecting amplitudes of phrase and accent components of F0 contours [4]. Although the rules included those on speech rate control, they were preliminary rules for a sentence without considering its syntactic structure.

In the present paper, analysis was conducted on the relationship between speech rate and syntactic structure. Although such relationship can be directly investigated by statistical methods (such as the multiple regression analysis), a large amount of speech material is required to obtain reliable results, because speech rate is subject to change also by other various factors. In view of the fact that the prosodic rules for reading style speech synthesis is available, investigations were conducted on relative mora duration of dialogue speech to read speech instead of its actual value. By doing so, basic factors related to both the dialogue and read speech will be suppressed, making the dialogue speech features clearer.

Since, as one of spoken language features, speech rate is related to the manner of human speech production, it will not show a tight relationship with syntactic structure of written language. Therefore, instead of finding relationship with the syntactic structure directly, in this paper, investigation will be conducted on the relationship with the prosodic structure, which can be clearly defined by the super-positional modeling of F0 contours [5].

2. PROSODIC STRUCTURE

The super-positional model represents a logarithmic F0 contour as the sum of phrase components and accent components, which correspond to a gradual decay from the top to the end of a phrase and a local hump realizing an accent type, respectively. Based on this model, the following 4 levels of prosodic units can be defined [6]:

1. Prosodic sentence corresponding to a phrase component, which occurs after termination of preceding phrase components. This unit is sandwiched by two long pauses (respiratory pauses) and is the highest level among the four.
2. Prosodic clause corresponding a phrase component which occurs without termination of preceding phrase components. This portion is preceded by a short pause.
3. Prosodic phrase defined similarly to the prosodic clause, but not preceded by an apparent pause.
4. Prosodic word corresponding to an accent component. This is the minimal prosodic unit.

When a unit of lower level is also a unit of higher level, it was assumed as belonging to higher level only for the

Table 1: Part of texts used for the analysis.

A: Zaooni puraNde ikitaiNdesuga.
(I'd like to go to Zao by "plan.")
B: Hai, puraNwa okimaridesuka.
(Yes, have you chosen your plan?)
A: Iie, madadesuga, nihakumikkano puraNni shitaiNdesu.
(No, not yet. But I'd like a 2 nights - 3 days plan.)
B: Kootsuuo fukunda puraNto fukumanai puraNga arimasuga.
(There are plans including and not including transportation.)
A: Fukunda hooo kaNgaeteimasu.
(I think about those including it.)
B: Zaoomadewa shiNkaNseNto basuno dochirade ikaremasuka.
(Which one will you take, Shinkansen or bus, to go to Zao?)
A: NedaNwa donokurai chigaimasuka?
(How much does the price differ?)
B: ShiNkaNseNno puraNwa ichimaNnisaNzeNeNtakakunarimasu.
(Plans by Shinkansen are 12,000 to 13,000 yen more expensive.)

current analysis. (Except for prosodic words; When a prosodic word is also a higher level unit, it is counted as belonging to both units.) Surely, there still exist larger prosodic units, such as the prosodic paragraph, they are not took into account in this paper.

3. SPEECH MATERIAL

Simulated dialogues were produced by pairs of Japanese speakers, by referring to written texts on model of dialogue between a client and an agent about ski resort accommodation and transition facilities. The same speakers also uttered in a normal reading style the individual sentences of the same texts. Each sentence was uttered in a randomized order to suppress the discourse factors. The utterances of 6 actors and 4 actresses of the Tokyo dialect were recorded, and all the recordings were submitted to Japanese listeners. Then, utterances of speaker SH, whose simulated dialogues were judged as most natural, were mainly chosen for the analysis. Recordings for two texts, consisting of 54 and 70 alternating utterances, were analyzed. These utterances are partly shown in Table 1.

In order to find out the prosodic structure for each utterance, its F0 contour was first analyzed by the analysis-by-synthesis based on the super-positional model [5]. Then pairs of dialogue and reading utterances, which have the same prosodic structure, were selected for the speech rate analysis. As the result, numbers of prosodic units included in the selected utterances became those listed in Table 2.

4. METHOD OF ANALYSIS

Speech rate analysis was conducted by measuring durations of consisting morae, after segmenting each utterance into phonemes by the forced alignment scheme using triphone HMM's (HTK Ver. 2.1) [7]. All phoneme boundaries were manually checked and corrected if necessary. As for syllables with long vowels, they were assumed to be two morae with equal durations.

The mora duration is known to be affected by various linguistic factors, viz., the consisting phonemes of the mora, its neighboring phonemes, the length of word to which

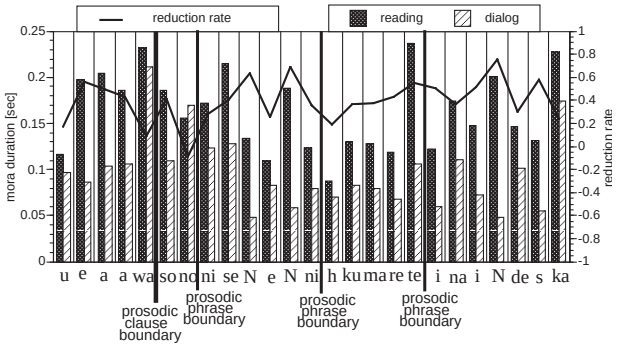

Figure 1: Reduction rates for prosodic sentence "ueaawa sono niseNeNni fukumarete inaiNdesuka? (Is the ware price included in the two thousand yen?)."

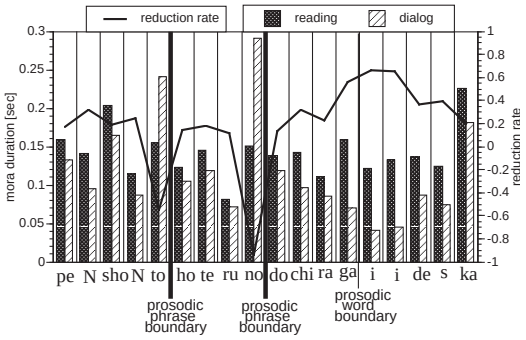

Figure 2: Reduction rates for prosodic sentence "penshoNto hoteruno dochiraga iidesuka? (Which do you prefer, a pension or a hotel?)."

it belongs, the word location in a sentence, and so on. Therefore, instead of considering absolute duration, the following reduction rate RD was defined and calculated for each measured duration [2]:

$$RD = (dur_r - dur_d)/dur_r, \quad (1)$$

where dur_r and dur_d are the duration in read style samples and dialogue style samples, respectively. When dialogue speech is uttered faster than read speech, RD takes a positive value.

5. RESULTS OF ANALYSIS

5.1. General Features

Generally, dialogue speech has higher speech rate as compared to its read speech counterpart. In a prosodic sentence, it was already revealed that the reduction rate is close to zero (speech rate of dialogue speech being close to that of read speech) at the beginning, then gradually increasing and again decreasing toward the end. In a prosodic clause or phrase, similar tendencies was observed as shown in Figure 1. There also are the cases where the unit-final mora has an extremely large duration in dialogue speech. This phrase final lengthening is clearly

Table 2: Number of samples for each prosodic unit. The numbers are shown separately for each number of morae.

	Total Number	Number of morae																
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18 or more
prosodic sentence	63	13	2	1	5	2	6	4	6	3	4	2	2	0	4	1	1	7
prosodic clause	7	2	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1
prosodic phrase	38	2	3	4	5	3	6	3	5	2	3	1	1	0	0	0	0	0
prosodic word	106	17	12	16	19	13	11	6	8	0	3	1	0	0	0	0	0	0

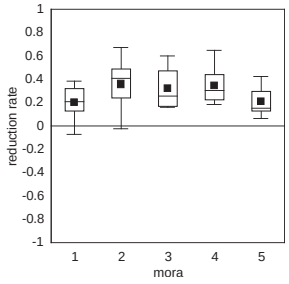

Figure 3: Average reduction rate for 5-mora prosodic sentences.

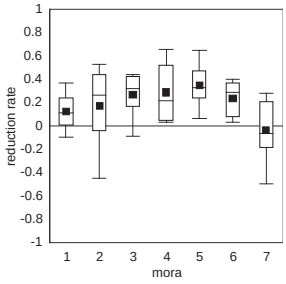

Figure 4: Average reduction rate for 7-mora prosodic sentences.

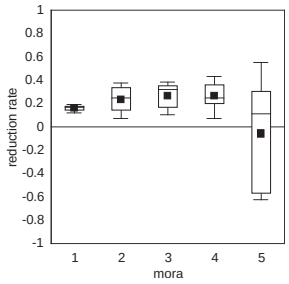

Figure 5: Average reduction rate for 5-mora prosodic phrases.

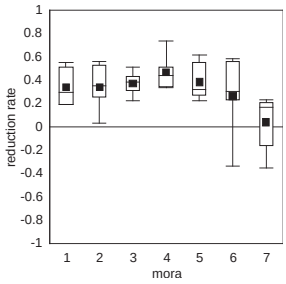

Figure 6: Average reduction rate for 7-mora prosodic phrases.

shown in Figure 2. These features on speech rate change in a prosodic unit is more or less occurring also in read speech, but the results indicate they are exaggerated in dialogue speech. As for the speech rate change in a prosodic word, no clear feature was observed, because it is obscured by differences in consisting phonemes.

Another typical feature of speech rate in dialogue speech is the shortening of mora duration in long vowels. This shortening is also observable in read speech, but its degree is larger for dialogue speech. One example is shown in Figure 2 (/ii/ part).

5.2. Detailed Features

In order to reveal the speech rate features in a prosodic units more clearly, reduction rates were averaged over several samples. Figures 3 and 4 show the averages and their standard deviations for 5-mora and 7-mora prosodic sentences respectively. Figures 5 and 6 show these for prosodic phrases. Reduction rate averages indicate similar tendency for all the figures. As for the prosodic words, reduction rates were averaged separately for their mora

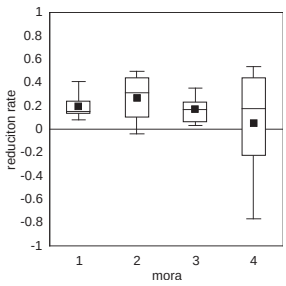

Figure 7: Average reduction rate for 4-mora prosodic words with type 1 accent.

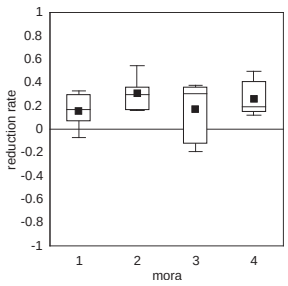

Figure 8: Average reduction rate for 4-mora prosodic words with type 0 accent.

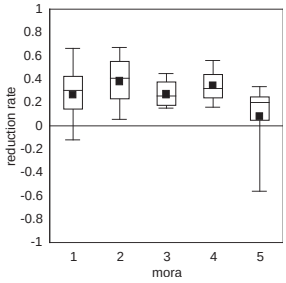

Figure 9: Average reduction rate for 5-mora prosodic words with type 1 accent.

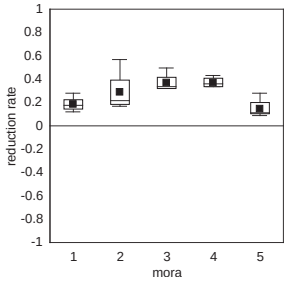

Figure 10: Average reduction rate for 5-mora prosodic words with type 0 accent.

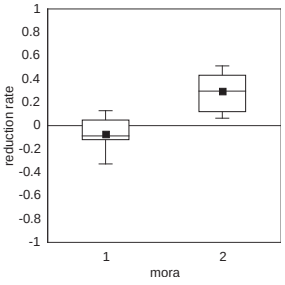

Figure 11: Average reduction rate for 2-mora prosodic word "hai (yes)."

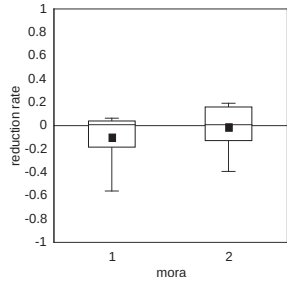

Figure 12: Average reduction rate for 2-mora prosodic word "dewa (well)."

numbers and accent types. Figures 7 and 8, for 4-mora prosodic words, and Figures 9 and 10, for 5-mora prosodic words, indicate reduction rate change in the unit similar to that observed in higher level units regardless of the accent types.

When a prosodic word consists of two morae, such as "hai (yes)" and "dewa (well)," reduction rates take rather different features form the above, as shown in Figs. 11 and 12. The results indicate that specific speech rate control is required for each short prosodic word.

6. SPEECH SYNTHESIS

Based on the results, dialogue speech synthesis was conducted in the following way by modifying prosodic rules for the formerly developed TTS system [8] into those for dialogue speech:

1. Generate prosodic symbols from linguistic information and other information of the sentence to be synthesized. Here, other information includes novelty/importance of information each word conveying, position of each word in a phrase, and so on. Prosodic symbols represent amplitudes of phrase and accent commands of the superpositional model for F0 contours. The generation rules were already explained elsewhere [4].
2. Assign appropriate duration to each mora depending on the original prosodic rules of the TTS system. Then, depending on the prosodic structures represented by the prosodic symbols, each mora duration is modified. Modification is done for prosodic sentences, prosodic phrase and prosodic words as indicated in the experimental results of the preceding section.
3. Modify mora duration further for long vowels. Also elongate the duration of phrase final mora if necessary.

Result of a hearing test indicated that the synthesized speech sounded much more dialogue-like when compared to the speech synthesized only by the above procedure 1. Although, currently, the same speech rate control is applied to prosodic units of the same level, effect of the number of consisting morae or the accent types should be taken into account. Further investigation is necessary with the increased data.

7. CONCLUSION

A comparative study on speech rates between dialogue speech and read speech was conducted. By viewing the analysis results with respect to the prosodic structures defined by F0 contours, a simple rule of speech rate change of dialogue speech in a prosodic unit was revealed. It was shown that dialogue-like speech can be synthesized by controlling each mora duration according to the results. Further studies are planned to construct prosodic rules for the dialogue speech synthesis.

We would like to acknowledge Prof. Kazuya Takeda for providing us with Japanese tri-phone HMM's.

8. REFERENCES

1. Hirose, K., Sakata, M., Osame, M., and Fujisaki, H., "Analysis and synthesis of fundamental frequency contours for spoken dialogue in Japanese," *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis, New Paltz*, pp.167-170 (1994-9).
2. Sakata, M. and Hirose, K., "Analysis and Synthesis of prosodic features in spoken dialogue of Japanese," *Proc. EURO-SPEECH95, Madrid, Vol.2*, pp.1007-1010 (1995-9).
3. Hirose, K. and Sakata, M., "Comparson of prosodic features in dialogue speech and read speech of Japanese," *Trans. IEICE, Information and Systems Society, Vol.J79-D-II, No.12*, pp.2154-2162 (1996-12). (in Japanese)
4. Hirose, K., Sakata, M. and Kawanami, H., "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features," *Proc. ICSLP Spoken Language Processing, Philadelphia, Vol.1*, pp.378-381 (1996-10).
5. Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of Acoust. Soc. Jpn., Vol.5, No.4*, pp.233-242 (1984-10).
6. Fujisaki, H., Hirose, K. and Takahashi, N., "Manifestation of linguistic information in the voice fundamental frequency contours of spoken Japanese," *IEICE Trans. Fundamentals of ECCS, Vol.E76-A, No.11*, pp.1919-1926 (1993-11).
7. Takeda, K. et. al., "Common platform of Japanese large vocabulary continuous speech recognition research: constriction of acoustic model," *IPSJ SIG Notes, 97-SLP-18-3* (1997).
8. Hirose, K. and Fujisaki, H., "A system for the synthesis of high-quality speech from texts on general weather conditions," *IEICE Trans. Fundamentals of ECCS, Vol.E76-A, No.11*, pp.1971-1980 (1993-11).