

SYNTHESIZING DIALOGUE SPEECH OF JAPANESE BASED ON THE QUANTITATIVE ANALYSIS OF PROSODIC FEATURES

Keikichi Hirose, Mayumi Sakata and Hiromichi Kawanami*
hirose@gavo.t.u-tokyo.ac.jp kawanami@gavo.t.u-tokyo.ac.jp

Department of Information and Communication Engineering, Faculty of Engineering
University of Tokyo, Bunkyo-ku, Tokyo, 113 Japan *Currently with Valeo Japan

ABSTRACT

Through the analyses of fundamental frequency contours and speech rates of dialogue speech and also of read speech, prosodic rules were derived for the synthesis of spoken dialogue. As for the fundamental frequency contours, they were first decomposed into phrase and accent components based on the superpositional model, and then their command magnitudes/amplitudes were analyzed by the method of multiple regression analysis. As for the speech rate, the reduction rate of mora duration from reading-style to dialogue-style was calculated. After normalizing the sentence length, the mean reduction rate was calculated as an average over utterances without complicated syntactic structure.

Results of the above analyses were incorporated in the prosodic rules for dialog speech synthesis. Using a formerly developed formant speech synthesizer, synthesis was conducted using both the former rules of read speech and the newly developed rules. A hearing test showed that the new rules can produce better prosody as dialogue speech.

1. INTRODUCTION

A number of text-to-speech conversion systems (TTS systems) have already been commercialized. They can generate rather good quality of speech, and, therefore, it seems that we can utilize them as speech output units in a spoken dialogue system. However, this idea includes two major problems from the viewpoint of synthesizing prosodic features. One is that prosodic rules of these TTS systems are rather simple and cannot utilize higher-order linguistic information, such as syntactic and discourse structures. Different from the case of TTS systems, in spoken dialogue systems, this kind of information is available during the process of output sentence generation. The other is that the prosodic rules were developed for read speech, not for dialogue speech. Synthetic speech from TTS systems is often too monotonous as dialogue speech. The first problem can partly be solved by analyzing read speech. Syntactic structure may affect prosodic features similarly for read speech and dialogue speech. While developing a text-to-speech conversion system, we have already constructed prosodic rules

taking higher-order linguistic factors into consideration [1]. The rules are based on a superpositional model for fundamental frequency contours (henceforth, F_0 contours) and can produce synthetic speech of high quality [2]. The second problem, on the other hand, can only be solved by analyzing dialogue speech. Although several works have recently presented the analyses of prosodic features of dialogue speech in Japanese [3, 4], the results were based on a rather rude model of F_0 contours and were lacking in the relationship with higher-order linguistic factors.

Consequently, we have been investigating prosodic features of dialogue speech, and in the previous reports, have shown preliminary results on how the F_0 contours and speech rates of dialogue speech differ from those of read speech [5, 6]. Especially, the role of a word in a dialogue was taken into account for the analysis of F_0 contours. Although the results gave us a quantitative view on the relationship between prosodic features and linguistic factors, they were not enough for the construction of rules to control the prosodic features in synthesizing dialogue speech.

In the present paper, the influence of various items on the F_0 contour was first evaluated quantitatively by the multiple regression analysis. Then, each segmental duration of speech samples was measured to show how the speech rate of dialogue speech changes in a sentence as compared to read speech. Finally, prosodic rules were derived from the results for synthesizing dialogue speech to be used as an output of spoken dialogue systems.

2. 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 (roles A and B, from now on) about ski resort accommodation and transition facilities. For a text, each pair produced at least two dialogues by changing the roles. 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 the speakers whose simulated dialogues were judged as natural dialogues were chosen for the analysis. Although several texts were used for the simulated dialogues, the results shown in this paper are based on two texts, respectively consisting of 14 and 54 alternating utterances. Utterances of speaker SH for the latter text were utilized for the analysis of F_0 contours, while those of speakers TI and YY for the former text were utilized for the speech rate analysis.

3. METHOD OF ANALYSIS

The feature parameters of an F_0 contour were extracted by the method of Analysis-by-Synthesis using the superpositional model [2]. This model is based on the assumption that, in the logarithmic scale of frequency, an F_0 contour can be decomposed into phrase and accent components, each being represented by the response of a second-order critically-damped linear system to the corresponding command. An impulse-like command is assumed for the phrase component, while a step-wise command is assumed for the accent component. The magnitude (A_p) and the timing (T_0) of each phrase command, as well as the amplitude (A_a), the onset time (T_1) and the end (T_2) of each accent command are major parameters characterizing an F_0 contour.

As for the speech rates, analysis was conducted by manually segmenting the speech waveform into morae. When the segmentation is difficult, such as the cases of elongated vowels, each mora was assumed to have the same duration. 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 it belongs, the word location in a sentence, syntactic structure and so on. Therefore, instead of considering absolute values of duration, the following reduction rate RD was defined and calculated for each measured duration.

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

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.

4. ANALYSIS OF F_0 CONTOURS

4.1. Accent Components

Dialogue speech has a general tendency in their prosodic features of having larger mean and deviation in F_0 as compared to read speech. The larger deviation was found to be ascribable mainly to the larger accent components [5]. In the previous report [6], increase in the accent command for a word of dialogue speech from its counterpart of read speech was shown to depend largely on the role of the word in discourse. The increase was also shown to depend on other

Table 1: Results of multiple regression analysis for the accent command amplitudes. r_{yi} indicates the partial correlation coefficient for item i . (Speaker SH)

	Reading					
	r_{yi}	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}
I1	0.290	0.032	-0.031	-0.001	-0.040	
I2	0.156	0.006	-0.046			
I3	0.118	-0.007	0.022			
I4	0.353	0.001	-0.081	0.027	0.080	0.051
A_{mean}	0.381					
	Dialogue					
	r_{yi}	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}
I1	0.325	0.040	-0.050	0.061	-0.107	
I2	0.152	0.011	-0.061			
I3	0.172	0.022	-0.040			
I4	0.320	0.000	-0.142	0.066	0.076	0.069
A_{mean}	0.521					

factors, such as the accent type and the position in the sentence. Taking these results into consideration, the following four items were selected for the multiple regression analysis of the accent command amplitude.

- I1 : novelty and importance of information conveyed, with four categories: (c1) new and important, (c2) new but not important, (3) repeated but still important, and (4) repeated and not important. "Important" means that the information is necessary to understand the content of the sentence and to make a reply, while "new" means that the information is not mentioned within 4 alternating utterances before.
- I2 : type of accent, with two categories: (c1) accent types with rapid F_0 downfall (D-type accents), and (c2) a flat type with no apparent downfall (F-type accent).
- I3 : position in the phrase component, with two categories: (c1) first accent component on the phrase component it is based on, and (c2) other than the first accent component.
- I4 : part of speech, with 5 categories: (c1) noun, (c2) verb, (c3) adjective or adverb, (c4) demonstrative or interrogative, and (c5) conjunction.

Within the above framework, the amplitude A_{ai} of the i th accent command is given by

$$A_{a_i} = A_{a_{mean}} + \sum_{j=1}^4 \sum_{k=1}^{c_j} a_{jk} \delta_i(j, k), \quad (2)$$

where $A_{a_{mean}}$ denotes the mean of the command amplitude through the analyzed samples, and coefficient a_{jk} indicates the influence of category k of item j on the amplitude. Table 1 summarizes these values obtained for read speech and dialogue speech by speaker SH. It also shows the partial correlation coefficient for each item. The results indicate the larger influence of I1 on accent command amplitudes for dialogue speech. It should be noted that, in the case of dialogue speech, the "importance" may increase the accent command amplitudes even for repeated words. This is not the case for

Table 2: Results of multiple regression analysis for the magnitudes of phrase commands at sentence initial. (Speaker SH)

	FIRST PHRASE COMMANDS (Role B)					
	Reading			Dialogue		
	r_{yi}	a_{i1}	a_{i2}	r_{yi}	a_{i1}	a_{i2}
I1	0.123	0.023	-0.009	0.198	0.034	-0.016
I2	0.016	0.002	-0.002	0.201	0.021	-0.030
I3	0.222	0.086	-0.008	0.235	0.078	-0.010
I4	0.229	-0.082	0.010	0.407	-0.137	0.022
I5	0.037	0.002	-0.009	0.216	0.027	-0.027
I6	0.174	-0.063	0.007	0.031	-0.011	0.001
I7	0.178	-0.082	0.005	0.291	0.096	-0.018
mean A_p		0.382			0.402	

Table 3: Results of multiple regression analysis for the magnitudes of phrase commands other than those at sentence initial. (Speaker SH)

	OTHER PHRASE COMMANDS (Role B)					
	Reading			Dialogue		
	r_{yi}	a_{i1}	a_{i2}	r_{yi}	a_{i1}	a_{i2}
I1'	0.157	0.027	-0.012	0.081	-0.018	0.006
I2'	0.209	0.011	-0.047	0.302	0.024	-0.095
I3'	0.019	-0.008	0.001	0.172	0.060	-0.007
I4'	0.367	0.141	-0.011	0.178	0.071	-0.008
I5'	0.174	-0.017	0.021	0.336	0.072	-0.027
I6'	0.157	-0.007	0.036	0.098	0.008	-0.016
I7'	0.067	-0.028	0.002	0.087	0.049	-0.003
I8'	0.018	0.006	-0.001	0.032	-0.012	0.001
mean A_p		0.237			0.277	

read speech. Although the results are totally consistent with the former results, rather low multiple correlation coefficients (0.45 for read speech, and 0.49 for dialogue speech) indicate the necessity of additional items such as the relationship with adjacent words. As for I2 and I3, further analysis should be necessary. The prosodic rules for TTS system are rather different depending on the accent type and the position in the phrase component, indicating that the samples should be divided into groups depending on I2 and I3 rather than being analyzed as a whole.

4.2. Phrase Components

As for the phrase command magnitude, the analysis was conducted separately depending on the role of the speaker (role A or role B) and on the position in a sentence (sentence initial or not). The following 7 items were selected for the phrase command at sentence initial. Each item was divided into two, viz., (c1) yes and (c2) no, categories.

I1 : belonging to an utterance which opens an FRD (fundamental routine of dialogue). FRD is a basic unit of question-and-answer dialogue defined as a pair of utterances, one for question or request and the other for response [7]. Eventually, utterances of the former type are categorized in c1, and those of the latter in c2.

I2 : containing word(s) carrying important information.

I3 : changing topics.

I4 : being a phrase which supports a subordinate conjunction.

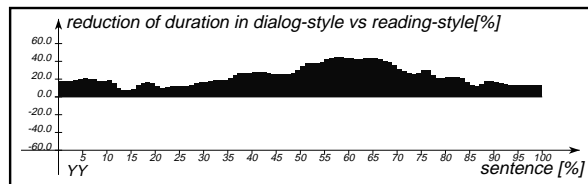


Figure 1: Mean reduction rate as a functions of position in a sentence. (Speaker YY)

I5 : 7 or less morae being included in the portion delimited by the phrase command and the following one.

I6 : being a phrase which starts with "yes" or "no."

I7 : being a phrase which ends with the sentence-final interrogative particle "ka."

While, for the phrase command not locating at sentence initial, items 1, 2, 3, 5, and 7 were also selected, which will be denoted by I1', I2', I3', I5' and I8' for this case, respectively. In addition, the following 3 items were added for the analysis.

I4' : following a phrase which supports a subordinate conjunction.

I6' : 7 or less morae being included in the portion delimited by the phrase command and the preceding one.

I7' : being preceded by a phrase which starts with "yes" or "no."

Although the results were obtained separately for the two roles, and they showed some differences among them, only those for role B are shown here. This is because the results for the agent are more important for the current purpose of constructing prosodic rules for speech synthesis in spoken dialogue systems. Tables 2 and 3 show the results of read and dialogue speech for the two cases of sentence-initial and not, respectively. As for the case of sentence-initial, larger command magnitudes were observed when opening an FRD, and, contrarily, smaller ones when starting with a subordinate conjunction. The magnitude of the following phrase command decreases/increases as a compensation. Increase in the command magnitude was also observed when the phrase contained word(s) with important information. Roughly speaking, the items showed a larger effect on the command magnitude for dialogue speech than for read speech.

5. ANALYSIS OF SPEECH RATES

In most of the analyzed sentences, the speech rate of dialogue speech was found to be close to that of read speech at the beginning of a sentence, while it starts to be faster at around one third, reaches the peak at around two thirds, and then decreases [6]. This feature can be confirmed by

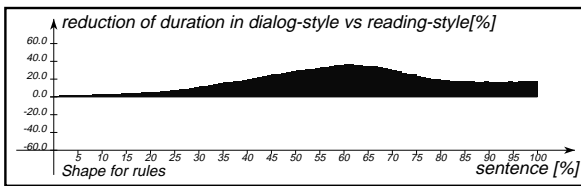


Figure 2: Shape of reduction rate in a sentence applied for the synthesis.

the mean reduction rate of Fig.1 calculated as an average over relatively short sentences (consisting of 10 to 25 morae) without complex structure in syntax, after a normalization of sentence length. Various factors, such as word length and syntactic structure, will modify the fundamental tendency.

6. SPEECH SYNTHESIS

Based on the results, prosodic rules were derived for the control of prosodic features in synthesizing dialogue speech, by modifying those previously developed for read speech of the TTS system [1]. As for the F_0 contours, we defined magnitude/amplitude of each command using the results shown in Tables 1,2 and 3. All of the four items were taken into account for accent commands, while, for phrase commands, only the items with relatively large partial correlation coefficients were taken into account. Concretely, items other than I6 were taken into account for the sentence-initial commands, and I2' and I5' for other commands. Prosodic symbols for read speech synthesis were modified so that they can represent, for each item, to which category word/phrase in question belongs. Consequently, phrase symbols with 6 digits and 2 digits were assigned for phrase commands at sentence-initial and others. Accent symbols were represented by 3 digits preceded by letter "D" or "F," which indicate the accent type being D-type or F-type, respectively. As an example, the following phonological symbols (prosodic symbols and syllabic symbols) are generated from the sentence "kootsuuo fukuNda puraNto fukumanai puraNga arimasuga?" (We have two plans, one including transportation and the other not. Which do you like?)

```
P112212 ko F111 o tsu u o A0
P12 hu D113 ku A0 n da D421 pu A0 ra n to
S3 P12 hu D113 ku ma A0 na i
D421 pu A0 ra n nga a ri ma #su nga P0 S1
```

Concerning the speech rate formerly made uniform, it was modified according to the results shown in Fig.1. Concretely, the results were first smoothed until the shape in Fig.2 was obtained. Then, this shape was used as the reduction rate to be applied on each syllable duration generated by the former rules.

A hearing test was performed to show that the newly developed prosodic rules can generate synthetic speech sounding better as dialogue speech. Using a portion of the text with 54 utterances, two versions of dialogue were arranged. In both versions, natural voice of speaker SH was used for role A, while, for role B, two versions of synthetic speech were used, one synthesized with the original prosodic rules for read speech, and the other with the newly developed rules for dialogue speech. The test was conducted for 5 native speakers of Japanese, and all of them preferred the version using the newly developed rules.

7. CONCLUSION

Quantitative analyses were conducted for the F_0 contours and speech rates of conversational Japanese. From the results, a set of prosodic rules were constructed for the synthesis of dialogue speech. Although dialogue speech synthesized based on the rules showed rather high quality, further investigations are necessary for the refinement of the rules. We are planning to increase the number of speech samples for the detailed analysis of command magnitudes/amplitudes. As for the control of speech rate, higher-linguistic factors, such as the syntactic structure, will be taken into account. Study is also planned on how to extract the role of words automatically.

8. REFERENCES

1. 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).
2. 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).
3. Kaiki, N., and Sagisaka, Y., "Prosodic characteristics of Japanese conversational speech," *IEICE Trans.Fundamentals of ECCS, Vol.E76-A, No.11*, pp.1927-1933 (1993-11).
4. Sakuta, M., Yamashita, Y., and Mizoguchi, R., "To generate various expressions of the surface sentence and the prosody in a dialogue," *IEICE Technical Report, SP95-84*, pp.21-28 (1995-12).
5. 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).
6. 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).
7. Hirose, K. and Asano, K., "Generation of speech reply in the speech response system," *Proc. EURO-SPEECH93, Berlin, Vol.2*, pp.1319-1322 (1993-9).