NEW PROSODIC CONTROL RULES FOR EXPRESSIVE SYNTHETIC SPEECH

Osamu MIZUNO, Shin'ya NAKAJIMA
NTT Human Interface Laboratories
1-1 Hikari-no-Oka Yokosuka-shi Kanagawa 239 Japan

ABSTRACT

This paper proposes new prosodic feature control rules for constructing semantic prosody control. Research was conducted into mental state tendencies using tests that examined the perceptions of the subject's sensibility to the control of synthetic speech prosody. The results showed the relationships between prosodic control rules and non-verbal expressions. Duration control reflects information processing state in spoken dialogues. Sentence final pitch contour control reflects the reliability of the information. Pitch contour dynamic range control indicates the speaker's excitement. The pitch contour control from start to peak pitch contour indicates the speaker's requirement for attention. Furthermore, for the Multi- layered Speech/Sound Synthesis Control Language(MSCL) we construct prosodic feature control commands using prosodic control rules and semantic control commands using the relationships. MSCL realizes expressive synthetic speech.

1 INTRODUCTION

We have proposed the Multi-layered Speech/Sound Synthesis Control Language(MSCL)[1]. MSCL is a synthetic speech control language in which users can describe prosodic feature control representations. MSCL has three layers as shown in Figure 1. The semantic level layer(S-layer), interpretation level layer(I-layer) and parameter level layer(P-layer). The multi-level description system enables the user to achieve simple semantic prosodic feature control with S-layer commands and to directly control prosodic features through the I-layer commands.

Semantic control commands on the S-layer are non-verbal expressions such as speaker's emotion, mental state, and situation within the dialogue. Prosodic components determined by analyzing emotional speech have been proposed[2]. This paper investigates the relationship between prosodic control and non- verbal expressions for semantic control. We propose eight prosodic control rules. These rules can be easily applied to words, phrases, and sentences. Using the prosodic control rules, we conducted two kinds of tests. At first, an association test examined the common definition of non-verbal expressions as determined from synthetic speech samples. A listening test examined the common definition and effect of semantic expressions. Through these 2 tests, we extracted the relationships between prosodic control rules and non-verbal expressions. Duration control reflects information processing state. The sentence final pitch contour control indicates the reliability of the information. The pitch contour dynamic range control reflects the speaker's excitement. The pitch contour control from start to peak pitch contour reflects speaker's intention. Furthermore, we constructed eight prosodic feature control commands on the I-layer using the eight prosodic control rules, as well as eight semantic control commands on the S-layer using the discovered relationships. We confirmed that expressive synthetic speech can be created using these MSCL commands.

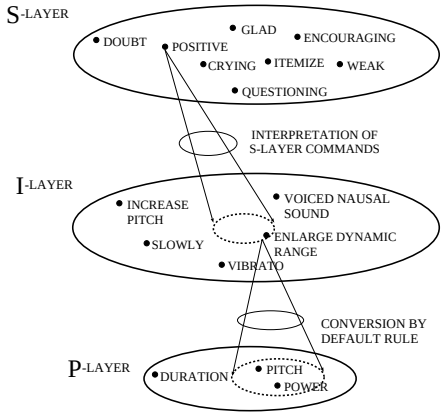

Figure 1: pattern description

There is a close relationship between a speaker's emotion and the prosodic features of his speech[3]. We investigated the relationships between the non-verbal expressions such as speaker's emotion, mental state, and situation in spoken dialogues and common prosodic features. At first, we conducted a experiment using the associa-

tion method. Five adults subject listened to synthetic speech samples and indicated what they assumed the speaker's mental state and situation were to be. Next, we inspected the results of the first experiment in a listening test. The experiment was that a subject preferred speech samples using the items based on the extracted tendencies. From the result of the experiment, we confirmed the extracted tendencies and investigated their effective usage.

2 NON-VERBAL EXPRESSION AND PROSODY

There is a close relationship between a speaker's emotion and the prosodic features of his speech[3]. We investigated the relationships between the non-verbal expressions such as speaker's emotion, mental state, and situation in spoken dialogues and common prosodic features. At first, we conducted a experiment using the association method. Five adults subject listened to synthetic speech samples and indicated what they assumed the speaker's mental state and situation were to be. Next, we inspected the results of the first experiment in a listening test. The experiment was that a subject preferred speech samples using the items based on the extracted tendencies. From the result of the experiment, we confirmed the extracted tendencies and investigated their effective usage.

2.1 EXPERIMENT 1

We created eight simple prosodic control rules for the first test. They consisted of two rules for duration control and six rules for pitch pattern control. Rule 1 shorts while Rule 2 lengthens all phoneme s to an equal degree. The pitch pattern rules are shown in Figure 2. The pitch contour is divided into three parts: section T1 extends from the beginning of the prosodic pattern of a word utterance (the beginning of the vowel of the first syllable) to the peak of the pitch contour. Section T2 runs from the peak to the beginning of the final vowel, and Section T3 covers the final vowel. The solid line indicates the original pitch contour.

Rule 3: Section T3 is given a monotonously rising pattern.

Rule 4: Section T3 is given a monotonously declining pattern.

Rule 5: The dynamic range of the pitch contour is narrowed.

Rule 6: The dynamic range of the pitch contour is enlarged.

Rule 7: Section T1 is depressed.

Rule 8: Section T1 is raised up.

These control rules are not a drastic change in terms of phrase component and accent types of the speech samples generated by rule-based speech synthesizers. They

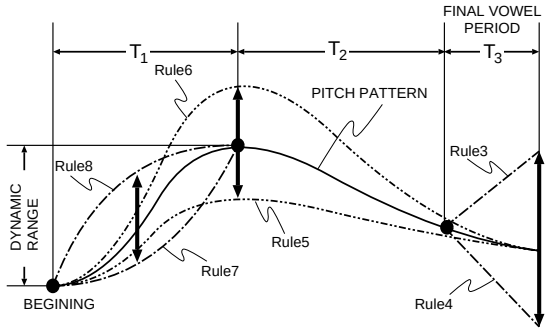

Figure 2: Pitch contour control methods

Table 1: Number of NV expressions

Subject	*Number*
subject A	117
subject B	144
subject C	152
subject D	132
subject E	196
Total	741

can be applied to words, phrases, and sentences. The speech samples were three Japanese words: "hontou"(which means "really"), "daijyoubu"(which means "all right") and "wakaranai"(which means "no understanding"). Each rule was applied by its self to each of the words so 24 speech samples were generated.

Each subject listened to an original speech sample (one of the three words in standard male voice) and then a rule-modified sample. The subject then tried to describe the speaker's emotion, mental state, and situation as understood from the modified sample. There was no restriction on what the subject could write or how much conjecture was made.

2.2 RESULTS

Table 1 shows the number of non-verbal expressions (NV expressions) discovered in the replies collected from the subjects.

Because the subjects were allowed to give free-form responses, the replies contained a wide variety of expressions, most of which were not suitable for analysis. Accordingly, we collated the replies of the subjects to discern the common relationships between 16 common non-verbal expressions and the 8 rules. For each subject, we assessed each rule (8) and non-verbal expression (16) combination as either "agree" or "disagree"; that is, the subject agreed or disagreed with the relationship. The former was given a score of 1 while the latter was scored 0. Table 2 shows the scores recorded for the 16 relationships. The maximum score for a relationship is 15. We considered the result from the spoken dialog point of view.

1. The NV expression"Speaking slowly/clearly" indicates conveying information to the listener slowly. The NV

Table 2: Dominant answers

Rule	Non-verbal expression	Score
Rule 1	a. Speaking slowly/clearly	14 (93)
	b. Thinking	10 (66)
Rule 2	c. Speaking fast	10 (66)
	d. Hurried/Urgent	7 (46)
Rule 3	e. Asking	14 (93)
	f. Worried	6 (40)
Rule 4	g.Understands / Agrees	11 (73)
	h. Convinced	6 (40)
Rule 5	i. Disappointed	12 (80)
	j. Negative	11 (73)
Rule 6	k. Positive	12 (80)
	l. Excited	10 (66)
Rule 7	m. Distrustful	9 (60)
	n. Cautious	9 (60)
Rule 8	o. Relaxed	9 (60)
	p. Irreverent	6 (40)

expression "Thinking" indicates making judgments or carefully considering the information. The common relationship of these two answers is that the speaker deals with information slowly and carefully. In other words, the relationship is viewed as deliberate information processing.

2. The NV expression "Speaking fast" indicates conveying information rapidly. The NV expression "Hurried/Urgent" indicates being quick in giving or replying to information in spoken dialogues, The common relationship of these two answers is that the speaker deals with information quickly. This relationship is viewed as fast information processing.
3. The NV expression "Asking" indicates requesting information that the speaker does not know or has doubts about. The NV expression "Worried" indicates the existence of disturbing information in the speaker's mind. The common relationship of these two NV expressions is that the speaker is cautious about the information. In other words, the relationship is viewed as indicating unreliable information.
4. The NV expressions "Understands" and "Agrees" indicate that the speaker accepts information as reliable.
5. The NV expressions and "Disappointed" indicates unhappiness with the information. "Negative" indicates that the speaker has no interest in the information or activity. The common relationship is viewed as indicating a lack of excitement about the information.
6. The NV expression "Positive" indicates a hope that the activity will occur or that the information is correct. "Excited" indicates having strong positive emotion. The common relationship is viewed as indicating excitement about the information.
7. The NV expression "Distrustful" indicates a lack of trust in the dialogue partner. The NV expression"Cautious" indicates caution. The common relationship of these two expressions is that the speaker pays attention to someone or some information in the dialogue.
8. The NV expression "Relaxed" indicates a lack of interest. The NV expression "Irreverent" indicates a feeling of carelessness to someone. The common relationship of these two answers is that the speaker does not pay careful attention to someone or some information in the dialogue.

2.3 EXPERIMENT 2

Using the above results, we conducted a listening test. The test assessed the relationships in more detail. Two words, "moshimoshi"(which means "hello") and "dousuru"(which means "how about?") were added to the original 3 words. All words were modified by the eight prosodic control rules individually to create 40 modified samples. 15 new subjects took this test. For each original sample pair, the subject was told to assign one of three assessment levels (1:agree, 2:no decision, 3:disagree) to each of the 16 NV expressions listed in Table 2.

As an example, Figure 3 shows the mean values for the word "hontou". The results basically support the results shown in Table 2. Note that each rule seems to create samples that match one of the eight pairs of NV expressions (the solid bars in Figure 3).

Thus, the relationship between the eight prosodic control rules and the eight pairs of NV expressions appear to be valid.

Figure 4 shows the result of principal component analysis if the results of the preference test. "R1" stands for "Rule 1". The distance between points indicates the effectiveness of prosodic control, in terms of semantics.

Figure 4 (b) shows that Rules 2, 3, and 6 are most effective in realizing the common semantics of "hontou", which is to ask for confirmation with some excitement or speed. Figure 4(d) shows that the rukles are widely spread for "moshimoshi" . This is reasonable because this word is used in many different ways in Japanese.

3 MSCL COMMAND CONVERSION

The prosodic feature control commands may be described at the I- layer level. It is also possible to define them using the S-layer prosodic feature control commands of MSCL. Table 3 shows examples of five S-layer commands prepared based on the experimental results and their corresponding I-layer commands.

The word in the braces `{}` is the object of the command. `[Length]` designates the duration control command. and its numerical value indicates the duration scaling factor. `[/V]` and `[/^]` designate the downward and upward controls of the pitch pattern from the start to the peak. `[~/]` and `[~/]` designate the rise and the fall of the word final pitch pattern.

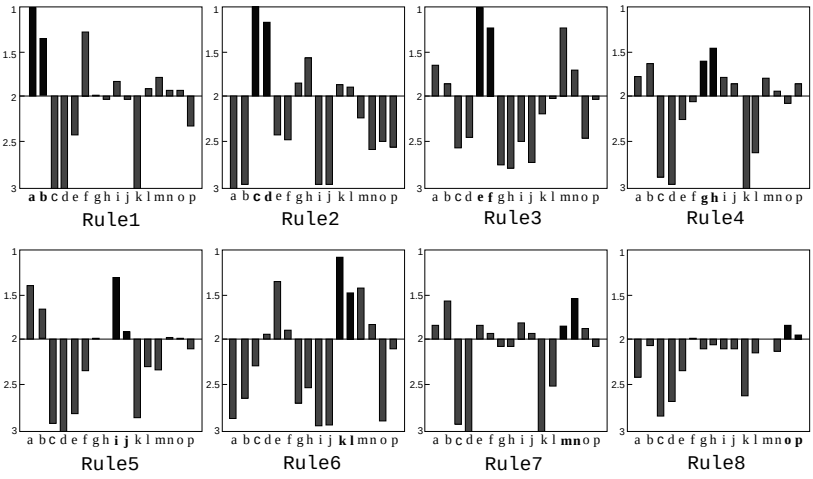

Figure 3: Mean value of listing test:"hontou"

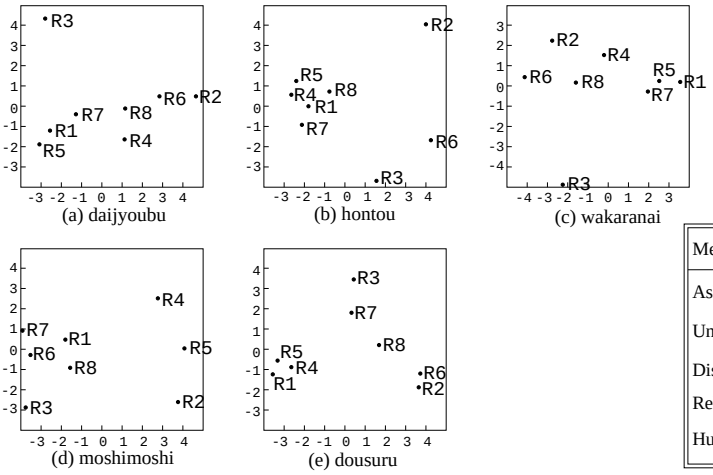

Figure 4: Result of principal component analysis

Meaning	S-layer Expression	I-layer Expression
Asking	`@Asking{honto}`	`[~/]{honto}`
Understand	`@Understand{honto}`	`[~\]{honto}`
Distrustful	`@Distrustful{honto}`	`[/v]{honto}`
Relaxing	`@Relaxing{honto}`	`[/^]{honto}`
Hurried	`@Hurried{honto}`	`[Length](0.5){honto}`

Table 3: S-layer expression to I-layer expression conversion

4 CONCLUSION

This paper proposed prosodic feature control rules and their impact in terms of the mental states they express. The rules were set as MSCL commands to realize expressive synthetic speech. Using a MSCL system, a home page browser or a dialog system can generate expressive and exciting speech.

References

[1] O.Mizuno,S.Nakajima,"A New Synthetic Speech/Sound Control Language," proceedings of ICSLP98(1998)

[2] Y.Kitahara, Y.Tohkura, "Prosodic Control to Express Emotions for Man-Machine Speech Interaction,"IEICE TRANS, Vol.E75-A, pp.155-163(1992)

[3] I.Murray, J.Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech,"Speech Communication 16, pp.369-390(1995)