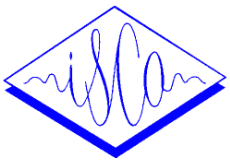

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26–29, 1998

Synthetic Speech/Sound Control Language: MSCL

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

ABSTRACT

The Multi-layered Speech/Sound Synthesis Control Language (MSCL) proposed herein facilitates the synthesizing of several speech modes such as nuance, mental state and emotion, and allows speech to be synchronized to other media easily. MSCL is a multi-layered linguistic system and encompasses three layers: and semantic level layer (The S-layer), interpretation level layer (The I-layer), and parameter level layer (The P-layer). This multi-level description system is convenient for both laymen and professional users. MSCL also encompasses many effective prosodic feature control functions such as a time-varying pattern description function, absolute and relative control forms, and SDS(Speaker Dependent Scale). MSCL enables more natural and expressive synthetic speech than conventional TTS systems. Furthermore, research was conducted into mental state tendencies using a test 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. These relationships are of use for constructing semantic prosody control.

This paper describes these functions and the effective prosodic feature controls possible with MSCL.

1 INTRODUCTION

Recent synthetic speech advances have made synthetic speech clearer and reduced mis-reading. A large number of practical text-to-speech(TTS) system have been released. However, conventional TTS systems cannot pass non-verbal expression. A spoken dialogue communicates not only verbal information but also non-verbal expression such as nuance, mental state, and emotion. These are important in passing information effectively. Conventional TTS systems offer monotonous and thus unattractive voices. If an E-mail reading system[1] use the TTS system, the monotonous voice may give listeners the wrong nuance. For multimedia contents production, synthetic speech is an important medium and significant editing flexibility is expected. For the purpose of generating expressive and flexible synthetic speech, we propose MSCL(Multi-layered Speech/Sound Synthesis Control Language). MSCL is a synthetic speech control language that has three description layers.

The first layer is the semantic layer(The S-layer). This layer is composed of a general prosodic feature control command set. This command set includes various modes of speech communication, for instance, a voice tuning command based on mental state, a voice tuning command based on speech acts, and a voice tuning command based on environment itemization. These commands can tune the synthesized voice to match the current environment. Semantic layer commands are given prosodic interpretations and broken-down into interpretation layer commands. Therefore, the semantic layer command can be viewed as a macro command.

The next layer is the interpretation layer(The I-layer). This layer has direct prosodic feature control command sets. The command sets include speech power, fundamental frequency (pitch), and duration control commands, in addition to time-varying pattern controls with detailed descriptions, and feature contour interpolation method definitions.

The last layer is the parameter layer(The P-layer). This layer includes phoneme associated parameters: pitch, power and duration.

There is an advantage to this multi-level description system. First is its support of laymen and professionals. While laymen can use simple semantic command sets such as: `@Doubt{...}`, `@Positive{...}`, and so on, professional users may use interpretation layer commands to directly control prosodic parameters, for instance, phonological analysis and synchronizing speech to other media.

MSCL also provides various kinds of control methods for specifying prosodic features. The P-layer provides time-varying pattern description commands which enable us to create dynamic contours as is possible in a GUI based system. Absolute and relative control forms leads to a reduction in the overhead of description.

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[5]. This paper investigates the relationship between prosodic control and non-verbal expressions for semantic control. We propose eight prosodic control rules. Using the prosodic control rules, we conducted an association test. The test examined the common definition of non-verbal expressions as deter-

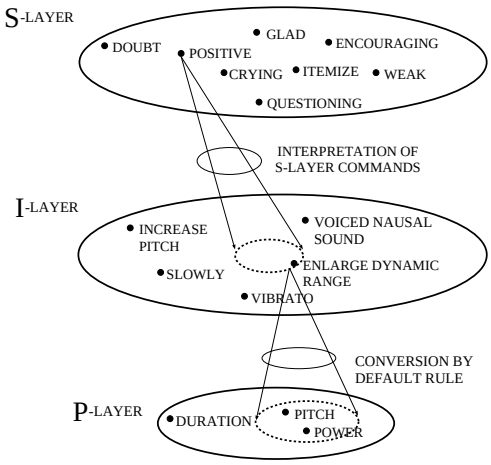

Figure 1: pattern description

mined from synthetic speech samples. Through the test, 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 attention. 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.

This paper outlines how MSCL converts the monotonous speech of TTS systems to more attractive and expressive speech, and how MSCL can express non-verbal expression through the easy control of prosodic features.

2 OVERVIEW OF MSCL

Figure 1 shows the multi-layered organization of MSCL. The highest layer is the Semantic level layer(The S-layer). The S-layer is composed of semantic prosodic feature control commands; words or phrases that directly represent non-verbal expression. The next layer is the Interpretation level layer(The I-layer). The I-layer holds prosodic feature control commands for interpreting each prosodic feature control command from The S-layer and for defining direct control of prosodic parameters of synthetic speech. The bottom layer is the Parameter level layer(The P-layer). The I-layer commands are finally converted into The I-layer command sequences by referring to a set of default rules. Default rules are prototypical values. A The P-layer description includes phoneme sequence and prosodic parameter values such as pitch frequency, power and duration of each phoneme.

This multi-layered control language provides several advantages as follows. If you desire spontaneous speech that contains non-verbal expression such as such as mental state, attitude, and understanding, you may need to control its

Command	Effect
`@Positive{`*S*`}`	Set the prosody of "positive" speech
`@Weak{`*S*`}`	Set the prosody of "weak" speech
`@Glad{`*S*`}`	Set the prosody of "glad" speech
`@Cry{`*S*`}`	Set the prosody of "crying" speech
`@Itemize{`*S*`}`	Set the prosody according to "itemized" environment
`@Doubt{`*S*`}`	Set the prosody of "doubtful" speech

"S" assert strings for speech synthesizer

Table 1: S-layer commands

prosody. If you want simple control, you can use The S-layer commands which offer semantic control. If you have expert knowledge of phonology and need to control the speech in detail, you can use The I-layer commands. MSCL supports all users; from the novice to the expert. The three layers are described in detail below.

2.1 Semantic Layer

The S-layer realizes prosody control semantically. The S-layer is composed of commands that concretely represent the non-verbal expression desired, such as the mental state, mood, intention, character–for instance, "Positive","Weak", "Glad", "Cry", "Itemize" and "Doubt" (see Figure1). Each word is followed by the mark "`@`", which indicates the prosodic feature control command of the S-layer to designate prosody control of the character string in the braces `{}` following the command. For example, the command of "Positive" enlarges the dynamic ranges of the pitch and power while the command for "Crying" shakes or sways the pitch pattern. The command "Itemize" designates the tone of reading-out items concerned and does not raise the sentence-final pitch pattern even in the case of a questioning utterance. The command "Weak" narrows the dynamic ranges of the pitch and power and shortens the duration. The command "Doubt" raises the word-final pitch and lowers the pitch average. These example commands were realized for the editing of Japanese speech. As described above, the commands of the The S-layer are used to execute one or more prosodic feature control commands of the I-layer in a predetermined pattern. An The S-layer command can be defined by the user. Thus The S-layer commands are also viewed as a sort of macro command. The S-layer permits semantic control descriptions, such as mental states and sentence structures, without requiring a knowledge of phonology. It is also possible to establish a correspondence between the commands of the S-layer and HTML, SAPI and other commands[2][3] [4]. Table 1 shows examples of usage of the prosodic feature control of the S-layer.

2.2 Interpretation Layer

The I-layer realizes detailed prosody control. This layer is

Command	Parameter	Effects
[Length](6mora){S}	Duration	Set the duration of S to 6 mora length
[Amplitude](2){S}	Power	Set the amplitude of S to doubled
[Pitch](120Hz){S}	Pitch	Set the pitch of S to 120Hz
[/-|\] {S_1|S_2}	Time-Varying pattern (raise, flatten, anchor, lower)	Set the prosodic feature of S_1 raised and flattened, set the prosodic feature of S_2 lowered
[F0d](2.0){S}	Pitch range	Set the pitch range of S to doubled

"S", "S_1", "S_2" assert strings for speech synthesizer

Table 2: I-layer commands

composed of prosodic feature control commands. These commands set not only the physical quantities of prosodic features but also time-varying pattern, accent type, and phrase component. By the use of these commands, it is possible to implement such commands as "slowly","high pitch","wide dynamic range", "vibrato","voiced nasal sound" as indicated in the I-layer command group in Figure 1. If a user specifies an The I-layer command without an argument, the command is mapped to the prosodic parameters of the P-layer using default control rules.

The I-layer commands encompass a set of symbols for specifying control of one or more prosodic parameters as control objects in the P-layer. These symbols can also be used to specify the time-varying pattern of each prosody element and a method for interpolating it. Every command of the S-layer is converted into a set of The I-layer commands. Table 2 shows examples of The I-layer commands.

Each The I-layer command is enclosed by marks '[' and ']'. The character string that designates prosody control in the braces {} follows the command. Strings for designating the I-layer commands used here will be described later on; S, S1, and S2 in the braces {} represent a character or character string of a text that is the control object to be synthesized. A short example of The I-layer MSCL text is;

```
Will you do [F0d](2.0){me} a [~/]{favor}
```

The command [F0d] doubles the dynamic range of the pitch designated by the argument (2.0). The object of control of this command is {me}. The next command [~/] raises the pitch pattern of the last vowel, in the phrase "favor."

Figure 2 shows examples of specifying the time-varying pitch pattern using The I-layer commands. The I-layer commands encompass the three symbols of '/','-', '\', and these show a rise, flattening, and declination in prosodic pattern, respectively. The upper figure shows pitch pattern modification of the Japanese word "anata" (which means "you"). The word "pitch" in the description under the figure means the declaration of pitch parameter modification. Followed the declaration, the time-varying pattern control form is specified.

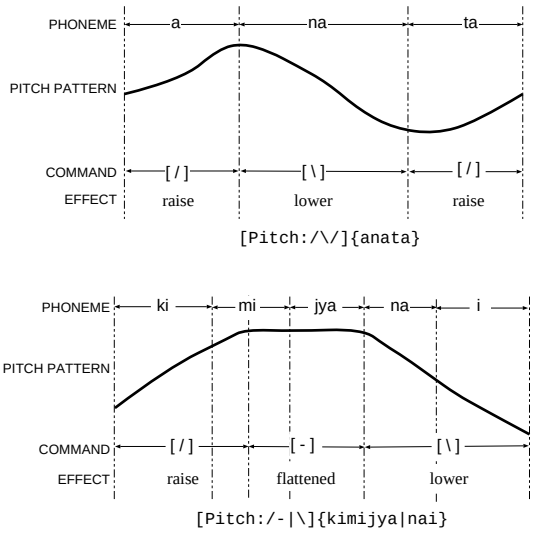

Figure 2: Example of time-varying pitch pattern control using I-layer commands

Describing the time-varying pitch pattern I more detail, the anchor symbol '|' declines partly and/or widely specification scope between a command and TTS strings. The lower figure shows pitch pattern modification of the Japanese sentence "kimijyanai" (which means "It is not your fault."). The Japanese sentence is divided into two parts between "jya" and "na" using the anchor. Pitch contour on the first part, "kimijya", is defined by the raise command '/' and flatten command '-'. Pitch contour on the second part, "nai", lowers the command '\'. Solid line in the lower figure indicates the pitch pattern generated by these commands.

MSCL includes an absolute control form and a relative form for prosodic feature control. Basically, MSCL is a synthesis-by-rule speech synthesizer and it's prosodic specification is based on the relative control form from the prosodies generated by the speech synthesizer. With the use of the relative control form, the entire synthetic speech need not be corrected and only at target places – this greatly reduces the work involved in speech message synthesis. The absolute control form makes an absolute correction to the feature. MSCL provides scaling method, Speaker Dependent Scale(SDS). 'single syllable duration', for instance, is an SDS description, and specifies the average duration of single syllable for the given speaker environment.

MSCL includes an absolute control form and a relative control form for prosodic feature control. Basically, MSCL is a synthesis-by-rule speech synthesizer and it's prosodic specification is based on the relative control form from the prosodies generated by the speech synthesizer. With the use of the relative control form, the entire synthetic speech need not be corrected and only at target places – this greatly reduces the work involved in speech message synthesis. The absolute control form makes an absolute correction to the feature.

MSCL provides scaling method, Speaker Dependent Scale(SDS). 'single syllable duration', for instance, is an SDS description, and specifies the average duration of single syllable for the given speaker environment.

2.3 Parameter Layer

The P-layer is composed of prosodic parameters that are selected and controlled by the prosodic feature control commands of the I-layer described next. These prosodic parameters are used in speech synthesis processing, i.e. controlling the pitch, power, duration and phoneme information of each phoneme. The prosodic parameters are the ultimate objects of prosody control by MSCL, and these parameters are used to synthesize speech. The prosodic parameters of the P-layer are the basic parameters of speech and have a common property that permits the synthetic speech editing technique of MSCL to be applied to various other speech synthesis or speech coding systems that employ similar prosodic parameters.

3 MSCL SCRIPT

Shown below is an MSCL form of the Japanese text "Watashi no namae wa Nakajima desu. Yoroshiku onegai shimasu" (meaning "My name is Nakajima. How do you do."):

```
[Interpolation=Linear]
[Length](8500ms)
{
   [>](150Hz, 80Hz)
   {
        [-|\|-|/](20Hz)
        {
             watashi|no|namae|wa
        }
   }
   [#](1mora)
   [/](15Hz)
   {
        [Length](2mora){Na}kajima
   }
   [\](30Hz){desu.}
   @Pray{yoroshiku onegai shimasu.}
}
```

In the above, `[Interpolation=Linear]` indicates the assignment of interpolation method. `[Length]` indicates the duration and specifies the time of utterance of the phrase. These two commands are given the same scope through the use of the corresponding braces `{}`. `[>]` represents a phrase component of the pitch and indicates that the fundamental frequency of utterance of the character string in the brace `{}` is varied from 150Hz to 80Hz. `[-|\|-|/]` shows local change of the pitch. /, -, and \ indicate that the temporal variation of the fundamental frequency is raised, flattened and lowered, respectively. Using these commands, it is possible to control the temporal-variation of the parameters. The command `{Watashi no Namae wa}`(meaning "My name"), is inserted or nested into the prosodic feature control command `[>](150Hz, 80Hz)` to change the fundamental frequency from 150Hz to 80Hz; the prosodic feature control command `[-|\|-|/]` changes the pitch locally. `[#]` indicates the insertion of a silent period in the synthetic speech. The silent period in this case is 1 mora, where "mora" is the average length of one syllable. `@Pray` is an The S-layer command of speech as in "praying".

4 PROSODY AND NON-VERBAL EXPRESSION

There is a close relationship between a speaker's emotion and the prosodic features of his speech[6]. 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. 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. From the result of the experiment, we confirmed the extracted tendencies and investigated their effective usage.

4.1 Experiment

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 lengthens while Rule 2 shorts all phonemes 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 can be applied to words, phrases, and sentences. The speech samples were three Japanese words: "hontou"(which

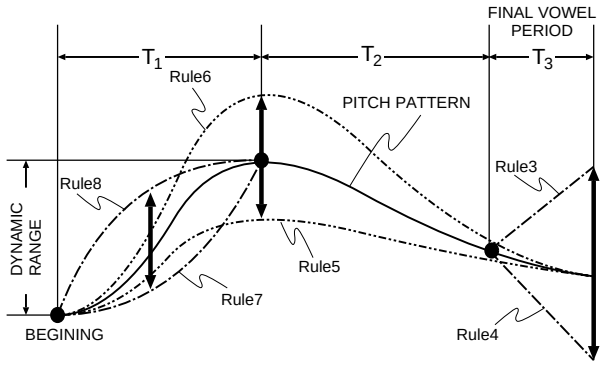

Figure 3: Pitch contour control methods

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

Table 3: Number of NV expressions

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.

4.2 Results

Table 3 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 4 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.

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)

Table 4: Dominant answers

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

tion. 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.

According to the results, the relationships between prosodic control rules and non-verbal expressions are concluded as follows: Duration control indicates deliberate or fast information processing state. The sentence final pitch contour control indicates reliable or unreliable information. The pitch contour dynamic range control indicates the speaker's excitement or lack of the speaker's excitement. The pitch contour control from start to peak pitch contour indicates paying attention or paying no attention.

4.3 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 5 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. `[~/]` and `[~\]` designate the rise and the fall of the word final pitch pattern. `[/V]` and `[/^]` designate the downward and upward controls of the pitch pattern from the start to the peak. `[Length]` designates the duration control command, and its numerical value indicates the duration scaling factor.

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 5: S-layer expression to I-layer expression conversion

5 CONCLUSION

We proposed a new synthetic speech control method MSCL. MSCL offers many advantages in creating expressive synthetic speech systems. One of the biggest advantages of MSCL is its multi-level description system which suits everybody, from laymen to professionals. The S-layer offers semantical level prosodic control commands for easy specification, and easy conversion from HTML text, Latex formated text, and SAPI formed tags. The I-layer encompass prosodic parameter commands and hence flexible time-varying pattern editing commands. MSCL offers both absolute and relative control forms to reduce the effort of description. Furthermore, we 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.

MSCL-enhanced TTS systems will make dialog systems and speech browsing systems much more expressive and friendly.

References

[1] M.Abe, K.Hakoda, H.Tsukada, "An Information Retrieval System from Text Database using Text-to-Speech," Proceedings of AVIOS'96, pp.189-196

[2] K.E.A.Silverman, J.F.Beckman, "TOBI: A standard for Labeling English Prosody," ICSLP, Vol2, pp.867-870(1992)

[3] T.V.Raman, "Audio Formatting - Making Text and Math Comprehensible," International Journal of Speech Technology,1,pp.21-31(1995)

[4] R.Sproat, P.Taylor, M.Tanenblatt, A.Isard, "A Mark-Up Language for Text-To-Speech Synthesis," Eurospeech97,pp.1747-1750(1997)

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

[6] 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)