

A NEW SYNTHETIC SPEECH/SOUND CONTROL LANGUAGE

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). The S-layer is the description level of semantics such as emotional and emphasized speech. The I-layer is the description level of prosodic feature controls and interprets The S-layer scripts to for control on I-layer level. The P-layer represents prosodic parameters for speech synthesis. 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 emotional and expressive synthetic speech than conventional TTS systems. 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 information. A spoken dialogue communicates not only verbal information but also non-verbal information 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 are two advantages to this multi-level description system. First is its support of laymen and professionals. While laymen can use simple semantic command sets such as: @Agree{...}, @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.

The second advantage is the easy conversion of many markup language scripts into MSCL scripts. The S-layer includes semantic control commands for prosodic features. For instance, HTML(Hyper Text Markup Language) texts, which are now widely used for making home-pages, can be translated to MSCL texts. The

HTML command `` signifies character enhancement. If the HTML text includes this command, the system can simply replace it with the MSCL command : `@Emphasis`. Thus, a MSCL-based TTS system can read HTML texts.

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.

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 information 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 information. 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 information such as such as mental state, attitude, and understanding, you may need to control its 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 information desired, such as the mental state, mood, intention, character-for instance, “Positive”, “Weak”, “Glad”, “Cry”, “Itemize” and “Doubt” (see

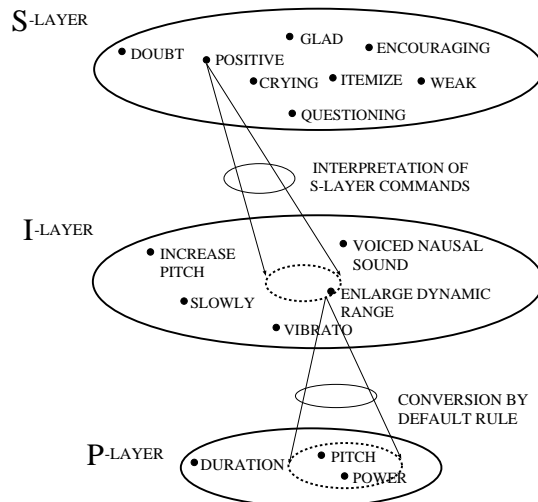


Figure 1: Multi-layered language structure overview

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 composed of prosodic feature control commands. These commands set not only the physical quantities of prosodic features but also time-varying pattern, accent type, phrase component and so on. By the use of these commands, it is possible to implement such commands as “slowly”, “high pitch”, “wide dynamic range”, “vi-

Command	Effect
@Positive{ <i>S</i> }	Set the prosody of "angry" speech
@Weak{ <i>S</i> }	Set the prosody of "weak" speech
@Glad{ <i>S</i> }	Set the prosody of "gald" speech
@Cry{ <i>S</i> }	Set the prosody of "crying" speech
@Itemize{ <i>S</i> }	Set the prosody according to "itemized" environment
@Doubt{ <i>S</i> }	Set the prosody of "doubtful" speech

"S" assert strings for speech synthesizer

Table 1: S-layer commands

Command	Parameter	Effects
[Lenght](6mora){ <i>S</i> }	Duration	Set the duration of <i>S</i> to 6 mora length
[Amplitude](2){ <i>S</i> }	Power	Set the amplitude of <i>S</i> to doubled
[Pitch](120Hz){ <i>S</i> }	Pitch	Set the pitch of <i>S</i> to 120Hz
[/- \]{ <i>S</i> ₁ <i>S</i> ₂ }	Time-Varying pattern (raise, flatten, anchor, lower)	Set the prosodic feature of <i>S</i> ₁ raised and flattened, set the prosodic feature of <i>S</i> ₂ lowered
[F0d](2.0){ <i>S</i> }	Pitch range	Set the pitch range of <i>S</i> to doubled

"S", "S₁", "S₂" assert strings for speech synthesizer

Table 2: I-layer commands

brato", "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

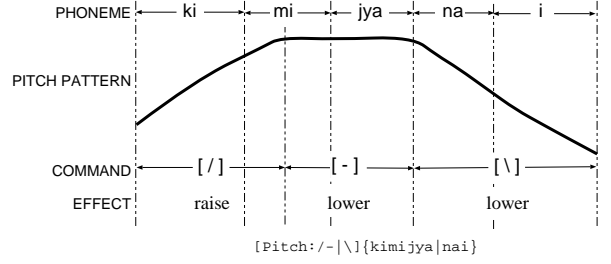
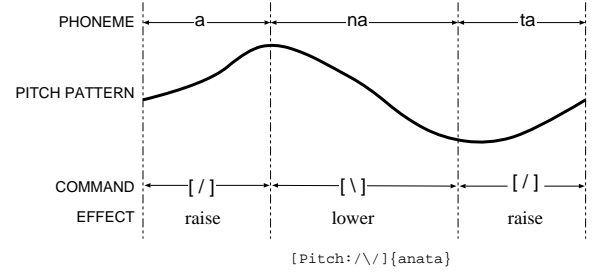


Figure 2: Example of time-varying pitch pattern control using I-layer commands

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.

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", raises 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.

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)
  {
    [pitch=SDS:-|\\|-|/](20Hz)
    {
      watashi|no|namae|wa
    }
  }
  [#](1mora)
  [/](15Hz)
  {
    [L](2mora){Na}kajima
  }
  [\\](30Hz){desu.}
  @Asking{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 150 Hz to 80Hz; the prosodic feature control command [/\\](20) 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. @Asking is an The S-layer command of speech as in "praying".

4 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 formatted 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 have proposed simple and effective prosody control rules [5].

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] O.Mizuno,S.Nakajima, "New Prosodic Control Rules for Expressive Synthetic Speech," proceedings of ICSLP98(1998)