

# A bilingual speech design tool: *Sesign*<sup>2001</sup>

Masanobu Abe, Osamu Mizuno, Tsubasa Shinozaki, Hideyuki Mizuno and Shin'ya Nakajima

NTT Cyber Space Laboratories

1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239 Japan

ave@nttspch.hil.ntt.co.jp

## Abstract

We have been developing a series of *Sesign* (speech design tools), TTS systems with the special function of manipulating prosodic parameters via a GUI (Graphical User Interface). All are intended to help the user create speech messages in a trial-and-error manner. This paper reports the following three advances in *Sesign*. (1) To extend the scope of *Sesign*, we added an American English TTS system. (2) A markup language approach called MSCL (Multi-layered Speech Control Language) is used together with the GUI-based approach. (3) We performed field trials using speech messages created by *Sesign*. One of the most successful examples is the *MyPartner* service, which informs the user of up-to-date information; the sentences generated by *Sesign* are used in combination with TTS output.

## 1. Introduction

In recent years, text-to-speech (TTS) has been used in e-mail reading systems, information retrieval systems, telephone directory systems and so on[1]. Beyond these applications, we are exploring new applications for synthetic speech. We believe that one of the most promising areas is speech message generation, including cartoon lines, messages for human-machine interfaces and so on. To achieve this goal, existing TTS systems are not powerful enough to produce conversational speech and emotional expressions. Our response has been to develop the *Sesign* series: (*Speed97*[2], *Sesign98*[3] and *Sesign99*[4]). *Sesign* (speech design tool) is a TTS system with a special function that enables prosodic parameters to be manipulated via a GUI (Graphical User Interface). It helps the user to create speech messages in the same way as word processors, drawing software, DTM(Desk Top Music) software, and CG(Computer Graphics) software support the creative activities of human beings. We have confirmed that *Sesign* is very powerful in generating dialect speech, emotive speech, conversational speech and so on. Moreover, *Sesign* has the following advantages in terms of manipulation and managing speech messages.

### (1) Low bit rate

*Sesign* can provide low bit rate speech coding; i.e., only phonetic symbols and prosodic parameters need be transmitted; approximately 800 bit/sec or less. The bit rate is much lower than that of conventional speech coders; between 2 kbit/sec and 16 kbit/sec[5].

### (2) High accessibility

*Sesign* assigns multi-layered tags such as orthographic transcriptions and phonetic transcriptions to speech. This makes it possible to easily access significant parts of a speech. For example, if the tags are orthographic transcriptions of key words, users can directly locate speech segments via the tags.

Moreover, if tags are assigned to the speech track of a video movie, users also can use the tags to locate particular parts of the movie.

### (3) Time synchronization

Another advantage of the tags is in synchronizing moving picture and speech messages. This is an important advantage when creating multi-media contents, because the user can easily identify time location by phonetic transcriptions without listening to the speech. Moreover, phonetic transcriptions make it possible to precisely synchronize the speech signal to animated lip movements.

### (4) Easy editing

*Sesign* also enables users to recycle speech fragments. To create new speech messages, it might be possible to edit existing speech messages just like text manipulation in a word processor. This kind of manipulation is useful for creating messages that contain few changes such as weather forecast messages, and traffic information messages.

This paper reports recent advances in *Sesign*. Section 2 introduces two new features; i.e., an American English TTS and MSCL. After explaining the main functions of the current *Sesign* in section 3, section 4 explains an application that uses speech messages generated by *Sesign*.

## 2. New features

### 2.1. An American English TTS

Experiments conducted over several years have found that, in Japanese, the *Sesign* approach is a powerful way to introduce TTS or synthetic speech into service systems. A motivation for integrating an American English TTS was to be able to confirm the applicability of the *Sesign* approach to other languages. The integration was easy to realize because, in terms of system components, there are few differences between a Japanese TTS and an American English TTS. Both systems use the same functions (explained in section 3) with the exception of editing phonetic transcriptions and accent types. The reasons are as follows;

(1) Japanese has phonetic symbols called Kana character. Users can easily specify Japanese pronunciation by using Kana characters; a few additional symbols are needed for devocalization and nasalization. On the other hand, English has complex mapping rules between text and phonetic symbols such as elision, reduction, contraction, linking, deletion, assimilation and so on. To specify these phenomena is hard for users with little knowledge about phonetics. Therefore, we decided not to implement a function for editing English phonetic symbols. One future task is to resolve this problem.

(2) Japanese is a tone language, and pitch accent is assigned to each phrase. This makes it easy for users to modify the speech

in a trial-and-error manner. On the other hand, English has stress accent and intonation expressed by pitch contours. We note that the TOBI system[6], which describes pitch contours, does not assign symbols phrase-by-phrase. It seems that English needs at least two layers for specifying prosodic parameters. This makes specification too complicated for users and it is better to directly manipulate prosodic parameters using the lower-level-functions of *Sesign*. Therefore, we decided not to implement a function for editing English prosody symbols, either.

The current system simply displays phonetic transcriptions for English on a screen. Syntax and meanings are shown in Table 1.

### 2.2. MSCL (Multi-layered Speech Control Language)

*Sesign*'s GUI-based approach is very powerful but has some disadvantages. To modify prosodic parameters using the GUI, users should know the effects of changing those parameters. Repeatedly modifying each phrase of speech in virtually the same way is time consuming. Considering these disadvantages, we developed an alternative way to synthesize expressive speech based on a markup language approach: the Multi-layered Speech Control Language (MSCL)[7]. Because the two approaches compensate each other, *MSCL* was integrated into *Sesign* to create a single tool that offers both advantages.

Figure 1 shows the multi-layered structure of MSCL. The first layer is the semantic layer (The S-layer). Users can write the 'semantics' or 'intention' of the message. For instance, if you want to emphasize some phrases, you simply write @Emph{...}. The S-layer command set includes various modes of speech communication such as a voice tuning command based on mental state, speech acts and environment itemization. Examples are glad, doubt, anger, sad, encouraging, negative attitude and so on. These commands can tune the synthesized voice to match the specified modes. Semantic layer commands are given prosodic interpretations and broken-down into a lower layer called the interpretation layer (The I-layer). The I-layer offers sets of direct prosodic feature control commands. The command sets include speech power, fundamental frequency (pitch), and duration control in addition to time-varying pattern control descriptions, and feature contour interpolation definitions. Table 2 shows examples of I-layer commands. The last layer is the Parameter level layer (The P-layer). The I-layer commands are finally converted into the P-layer command sequences that includes phoneme sequences associated with their prosodic parameter values such as pitch frequency, power and duration. In the P-layer level, GUI-based approach and MSCL use the identical data structures and user can use the both approaches back and forth.

### 2.3. A system block diagram

Figure 2 shows a block diagram of *Sesign*<sup>2001</sup>. The GUI modules including text, phonetic and prosodic transcription editors, are designed so as to be completely independent of the TTS engine used and permit easy extension to multi-lingual systems. *Sesign*<sup>2001</sup> has two TTS engines (English and Japanese), mainly because the two languages are quite different. Although a text file can contain both languages, no sentence can be a combination of both languages. Language identification is simply performed using the character code type, two-byte-code and single-byte-code are used for Japanese and English, respectively. The identified language is displayed in the text editor window and users can change language identity manually.

Table 1 Phonetic transcriptions for English in *Sesign*

<p><b>Syntax</b> For each syllable, the syntax is the following:</p> <p>PH PH .... PH [PITCH_ACCENT BND_TONE "WORD"]</p> <p>where PH = phoneme symbol and all of the following are optional fields: PITCH_ACCENT = { H*   L*   L+H* } BND_TONE = { L-L%   H-L%   L-H%   H-H% } "WORD" = word (in quotes) - marks end of word</p> <p>The numbers on phoneme symbols indicate stress: 1=stress 2=secondary stress 0=unstressed</p> <p>Each syllable must be terminated with []. If the syllable has none of {PITCH_ACCENT, BND_TONE, "WORD"}, then the square brackets are left empty.</p> <p><b>Example</b> Input texts; <i>Hello world.</i> Phonetic transcriptions displayed in <i>Sesign</i> window; <i>h &amp;0 l [] oU1 [H* "Hello"] w 3r1 l d [H* L-L% "world"]</i> Meanings of the phonetic transcriptions;</p> <p><i>h</i> - phoneme symbol <i>&amp;0</i> - phoneme symbol, unstressed vowel <i>l</i> - phoneme symbol <i>[]</i> - end of syllable <i>oU1</i> - phoneme symbol, stressed vowel <i>[H* "Hello"]</i> - end of word "hello", end of syllable, H* accent on syllable <i>w</i> - phoneme symbol <i>3r1</i> - phoneme symbol, stressed vowel <i>l</i> - phoneme symbol <i>d</i> - phoneme symbol <i>[H* L-L% "world"]</i> - end of word, end of syllable, H* accent on syllable, and end of phrase with L-L% boundary tone</p>
---

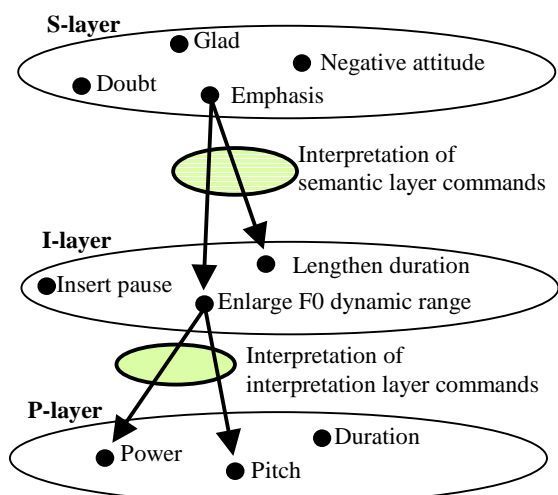


Fig. 1 MSCL

In terms of parametric information such as phoneme symbol, phoneme duration,  $F_0$  and power, the same format and data structure are commonly used for both languages. MSCL is currently available only for Japanese.

### 3. *Sesign*<sup>2001</sup>

#### 3.1. Basic operations

*Sesign*<sup>2001</sup> enables users to visually observe parameters and modify them using the GUI (Graphical User Interface), and to synthesize speech in a trial-and-error manner. Basic operations are as follows.

- Step1 : Input texts of Kana, Kanji (Chinese character) and alphanumeric characters using a text editor, or access texts created in advance.
- Step2 : Analyze the texts to obtain phonetic transcriptions, accent types, and syntax information.
- Step3 : Edit the phonetic transcriptions and accent types if needed. This function is available only for Japanese. Because readings of Kanji and accent type are usually context dependent and are difficult to estimate, this function is necessary for Japanese. To check accent types, a user can synthesize the speech.
- Step4 : Modify prosodic parameters;  $F_0$ , duration, and power. The prosodic parameters are visually displayed, and a user can modify them by mouse actions. A user can create speech in a trial-and-error manner; i.e., change prosodic parameters, then immediately synthesize and listen to the speech.
- Step5 : Store speech messages and/or their parameters.

#### 3.2. Functions for efficient production

The following are, according to our experience in developing the *Sesign* versions, the most important functions for efficient speech message production.

- (1) Mimic speech generation: While a user can create speech messages with the desired speech style using the prosody modification interface, it is sometimes difficult for beginners to create natural-sounding speech. *Sesign*<sup>2001</sup> utilizes the prosodic patterns extracted from natural speech to assist the beginner in prosodic modification. We implemented a function that automatically extracts prosodic parameters from human speech resulting in the generation of speech that mimics recorded human speech. After aligning phonemes by the use of HMM models[8], phoneme duration is determined by referring to the labels and  $F_0$  is extracted from the speech signal using the AMDF algorithm[9]. The GUI interface allows the correction of errors that occur during automatic parameter extraction.
- (2) Prosodic pattern library: Speech messages contain a number of common phrases or words. This is especially true in fixed

Table 2 I-Layer commands

Command	Effects
[Length](6mora){S}	Set the duration of S to 6 mora length
[Amplitud](2){S}	Set the amplitude of S to double
[F0d](2.0){S}	Set the pitch range of S to double
[-/\~]{S1 S2}	Set the prosodic feature of S1 raised and flattened, set the prosodic feature of S2 lowered

S, S1, S2 assert character string for speech synthesizer

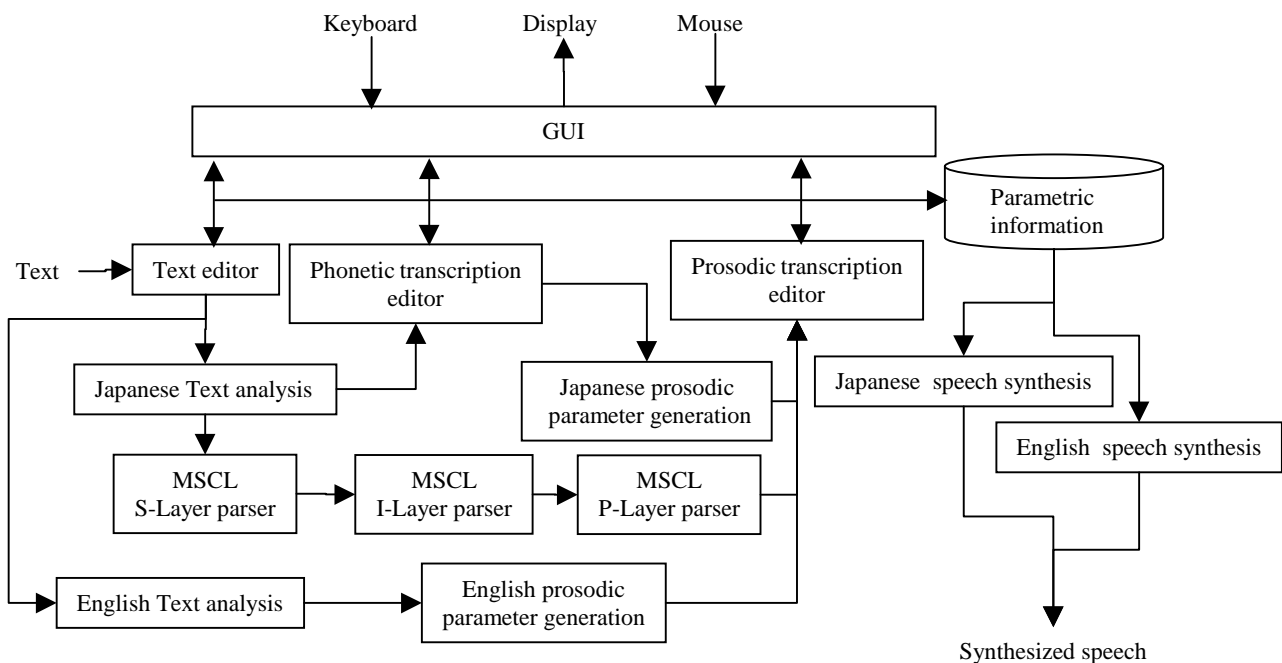


Fig. 2 A block diagram of *Sesign*<sup>2001</sup>

tasks. Given this knowledge, a user can efficiently generate speech messages by registering common prosodic patterns in the prosodic pattern library. In the registration process, associated information such as phonemes, accent types of preceding, current and following phrases, position in a sentence, a local sentence structure and so on are added to the prosodic pattern itself. When reusing a prosodic pattern, the associated information is used to locate the best prosodic pattern.

(3) Macro function: To easily modify the prosodic parameters, a set of modification procedures that are frequently used to express conversational or emotive speech are provided as macros. Examples are duration lengthening and  $F_0$  raising of the final syllable. Moreover, the history of the user's application of macros is saved for each phrase, and a new macro can be defined using this history.

(4) Voice quality and speaking style setting: To create speech messages, it is important to select the most appropriate speech quality for a message. *Sesign*<sup>2001</sup> allows users to flexibly change voice quality and speaking style. Voice quality control is performed by sampling-rate conversion and speaking style is determined by speaking rate,  $F_0$  range,  $F_0$  dynamics, and power. Users can easily define a set of parameters as a new speaker characteristic and can refer to the set by *speaker* name.

(5) Speech insertion: When generating speech it is not necessary or desirable for the message to consist of only synthetic speech. This function enables human speech and synthetic speech to be combined phrase by phrase.

(6) Plug-In interface: Experts will demand more sophisticated speech effects such as echo, delay and equalization. The external interface permits the use of external plug-ins. The specifications of the interface are open, and anyone can create and use the plug-ins.

#### 4. An example of *Sesign* application

We found that the *Sesign* approach is very effective in introducing TTS or synthetic speech into service systems. For those systems, it is not necessary for all speech messages to be synthesized from texts. For instance, some words such as time, addresses, prices, names should be changed in each transaction, while guidance messages are usually fixed. The fixed sentences are generated by *Sesign* in advance and are combined with TTS outputs. As one of the most successful examples, in this section, we describe the *MyPartner* service, which verbally passes up-to-date information to the user.

##### 4.1. Baseline framework: *WebMessenger*

*MyPartner* is constructed on the general framework called *WebMessenger*[10], which was developed to facilitate multimedia content production for Internet distribution. Key points are (1) to produce speech messages by a TTS or *Sesign*, (2) to produce moving pictures by concatenating templates from a set of moving pictures. *WebMessenger* includes TTS and the set of moving picture templates, and is installed in the user's machine as Plug-In software for an Internet browser. It recreates movies upon receiving texts or *Sesign* parameters and the IDs of moving picture templates. Therefore, the amount of data transmitted is quite small compared to sending the movie itself.

Because, in *MyPartner*, a speech agent appears in a display window and passes messages to the user, the moving picture

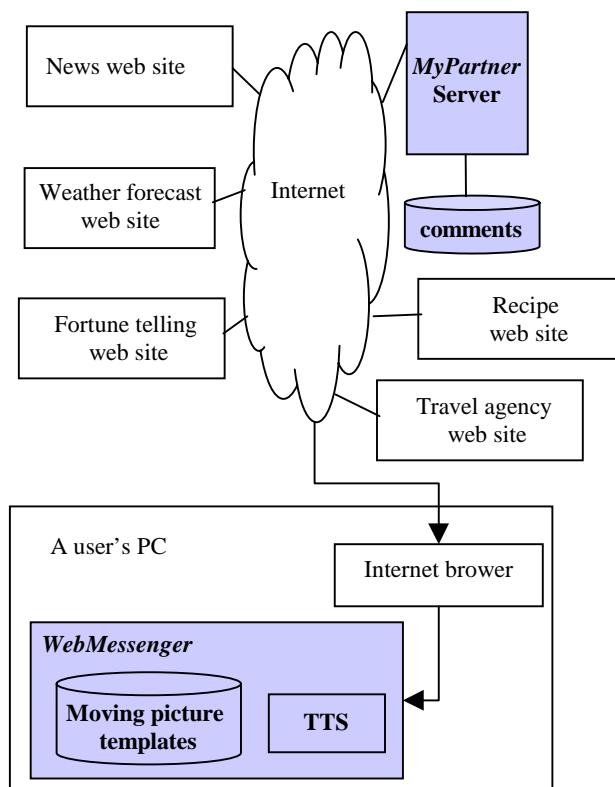


Fig. 3 A *MyPartner* service

Table 3 Announcement examples of *MyPartner* system

*Hello, Mike!*  
I got news for you from CNN sports.  
Tiger Woods won the U.S. OPEN with a new record.  
*It's amazing!*  
*Did you already know this?*

*Today's weather in Portland.*  
It will rain this evening.  
*Do not forget your umbrella, Mike!*  
*Have a nice day.*

Note: *Italics* - Sentences generated by *Sesign*.  
Underline - Sentences generated by TTS.

templates are agent gestures such as speaking, waving a hand, taking a bow, pointing, looking surprised and so on. The total number of templates is 60. After generating fixed messages using *Sesign*, the relationships between texts and moving picture templates are set. Phonetic symbols, prosodic parameters and IDs of moving picture templates are stored in a file with "pac" format and pac files are embedded into the source HTML document.

##### 4.2. *MyPartner* service

According to user preferences, *MyPartner* announces up-to-date information obtained from the Internet. Figure 3 shows a

system configuration of the *MyPartner* service. The *MyPartner* server automatically collects the latest information from the Internet at certain intervals. For example, once every hour from a news site, once every 6-hours from a weather site, and once every day from a fortune telling site, a travel agency site and a recipe site. The *MyPartner* server has information extraction programs for each site, which results in partial reading of the source HTML document. News headlines are extracted from the news sites, for example. The extracted texts are verbalized by the TTS. Moreover, for each information site, the *MyPartner* server contains comments created in advance by *Sesign*. Table 3 shows examples of *MyPartner* announcements. As shown here, sentences generated by TTS and *Sesign* are appropriately combined. These announcements are performed at intervals specified by the user, 10 minutes for instance. The usage of synthetic speech is summarized as follows.

(1)TTS is just used as an ear catcher. If the user is interested in the contents, he/she reads the whole text using an Internet browser.

(2)Sentences generated by *Sesign* are used to add friendliness to *MyPartner*. Because more than 20 sentences/expressions are prepared for each meaning and are randomly used, users develop a better feeling than they would if the same sentence were to be repeated.

## 5. Conclusions

This paper reports recent advances in *Sesign*. To integrate an American English TTS, the *Sesign* structure was redesigned to support multiple languages; its performance was confirmed. Because, in terms of phonetic and prosodic editing, we found big differences between Japanese and English, the next step is to develop a better solution for editing English speech. To create speech messages, *Sesign*<sup>2001</sup> provides both a GUI-based approach and a markup language approach. Because the two approaches provide different advantages, users can create speech messages easier than is possible with the previous *Sesign*. As the next step, we have a plan to examine cases where one approach is more effective than the other. Finally, we described one of the most successful applications of *Sesign*<sup>2001</sup>. We intend to develop bilingual systems based on the *Sesign* approach.

## 6. Acknowledgments

We are grateful to the members of the Speech Processing Department for their helpful discussions. We also thank Mr. Yamamori, the department head, for his continuous support of this work.

## 7. References

- [1] M. Abe, K. Hakoda, H. Tsukada, "An information retrieval system from text database using text-to-speech," Proc. Avios96, pp.189-196, 1996.
- [2] M. Abe, H. Mizuno, S. Nakajima, "*Speed97*: A speech creation tool for synthesizing various kinds of speech," IPSJ technical report, SLP 17-12, pp67-72, (in Japanese), 1997.
- [3] H. Mizuno, M. Abe, S. Nakajima, "*Sesign98*: Speech design tool to provide high-controllability and various expression speech signal," ASJ fall meeting, 2-P-12, pp.309-310, (in Japanese), 1998.
- [4] H. Mizuno, M. Abe, S. Nakajima, "Development of speech design tool 'SESIGN99' to enhance synthesized speech," Proc. of Eurospeech99, pp. 2083-2086, 1999.
- [5] M. Abe, H. Mizuno, S. Takahashi, S. Nakajima, "A prototype hybrid scalable text-to-speech system," Proc. Workshop on SNHC and 3D Image, pp.8-11, 1997.
- [6] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, "TOBI: A standard for labeling English prosody," Proc. ICSLP92, pp. 867-870, 1992.
- [7] O. Mizuno, S. Nakajima, "Synthetic speech/sound control language: MSCL," Proc. of the 3rd International Workshop on Speech Synthesis, pp. 21-26, 1998.
- [8] Y. Noda, Y. Yamaguchi, T. Yamada, A. Imamura, S. Takahashi, T. Matsui, K. Aikawa, "The development of speech recognition engine REX," IEICE meeting, pp. 220 (in Japanese), 1998.
- [9] M. Ross, H. Schafer, A. Cohen, R. Freuberg, H. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. ASSP, ASSP-22, pp.353-362.
- [10] M. Abe, H. Mizuno, T. Shinozaki, "*WebMessenger*: A new framework to produce multimedia content by combining synthesized speech and moving pictures in the WWW environment," Proc. of MMSp, pp. 611-616, 1999.