ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DEVELOPMENT OF SPEECH DESIGN TOOL "SESIGN99" TO ENHANCE SYNTHESIZED SPEECH

Hideyuki Mizuno, Masanobu ABE, Shin'ya Nakajima
NTT Human Interface Laboratories
1-1 Hikari-no-oka,Yokosuka-Shi, Kanagawa 239, Japan
mizuno@nttspch.hil.ntt.co.jp

10.21437/Eurospeech.1999-462

ABSTRACT

This paper introduces a new speech design tool (Sesign99) that can convert monotonous synthesized speech into a variety of speech styles. As multimedia services such as games, interactive movies and WWW home pages become more popular, more attention is being focused on the creation, management, and transmission of speech messages. Although speech synthesis-by-rule has improved with recent advances in TTS, the monotonous features of speech produced by synthesis-by-rule hamper the introduction of TTS to the application areas listed above. There is demand for software that allows even the novice to produce engaging and natural-sounding speech messages.

Keywords: TTS, multimedia contents, speech design

1. INTRODUCTION

The realization of labor saving devices and man-machine interfaces are key goals in studies of text-to-speech (TTS) systems. In fact, TTS has already been added to media conversion systems and speech interfaces for computer operation because speech can be automatically output by a TTS system and the user can change the contents of the message. However, recent reports of TTS show that while the quality of synthesized speech permits understanding, it fails to match the quality of human speech[1][2]. In particular, the prosodic characteristic of synthesized speech is monotonous and mechanical. The main problem is caused by the automatic speech production process. Recent reports have addressed prosodic modeling and speech synthesis of emotional speech[3][4]. The problem is that the range of styles created was restricted. Even if many complex rules or very large size of database are used, it is difficult to represent all speech styles. Conventional TTS systems can not offer the speech quality demanded for multimedia services such as games, interactive movies and WWW home pages. With the aim of advancing the frontier of speech applications, we propose Sesign99, a sophisticated TTS system. The advantages of speech messages produced by Sesign99 are as follows;

(1) Effective reuse: Once a speech message is created, messages with similar style can be created automatically by referring to the first message. This is useful for creating speech messages that contain few changes such as weather forecasts and traffic alerts.

(2) High accessibility: multi-layered tags are given to speech segments. This makes it possible to easily access different parts of a speech. For example, if orthographic transcriptions of key words are used as tags, users can directly locate speech segments. Moreover, if content 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) Low bit rate coding: the framework can utilize the prosodic parameters extracted from natural speech if available, that is, it can realize a low bit rate speech encoder[5]. Its low bit rate, approximately 1 kbit/sec or less, is ideal for transmitting speech messages across the Internet.

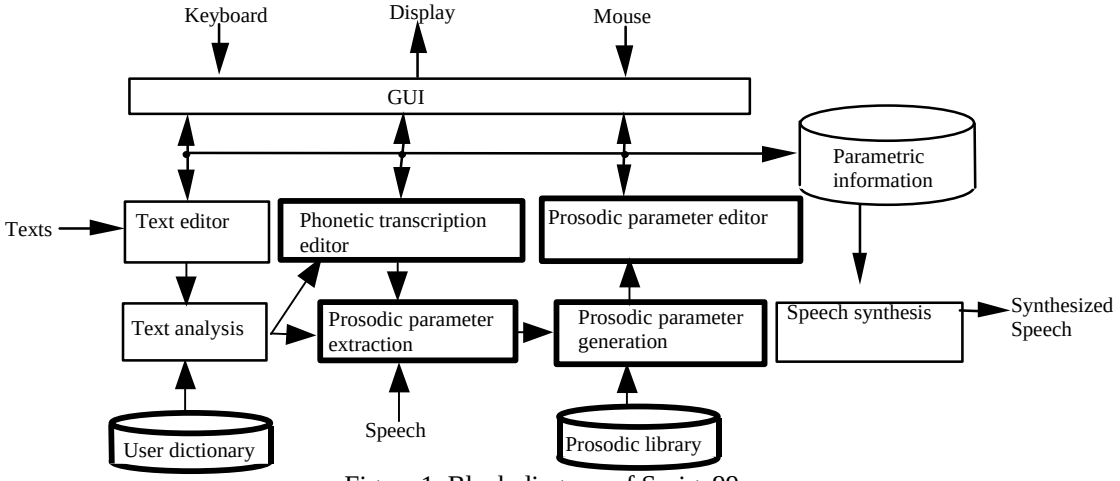


Figure 1. Block diagram of Sesign99

Section 2, briefly outlines Sesign99, while section 3 explains the four features of Sesige99. In section 4, many of the functions of Sesign99 are detailed.

2. OUTLINE OF SESIGN99

2.1 System Configuration

The system configuration is shown in Figure.1. The basic components of Sesign99, text-analysis, prosodic parameter generation and speech synthesis follow the prototype HSTTS system we introduced at IWSNHC3DI'97[5]. The newly developed or improved components are shown by the bold squares. Sesign99 currently runs on Windows95/NT.

2.2 Operation

Step1: Edit texts of Kana and Kanji (Chinese characters) using a text editor and also, if needed, phonetic transcriptions, accent types, and syntax information, all of which can be automatically obtained by analyzing the text. A screen-shot of Sesign99 main interface is shown in Figure 2.

Step2: Specify a speech style to each sentence using phonetic transcription selection speech styles. The speech style can be easily defined and named by the user. Figure 3 shows a screen shot of defining the speech style in the "dialogue display mode". The values of voice quality modification rate, power, intonation, pitch can be freely specified. All of the specified styles are displayed in the right area of main interface. A speech style can be easily changed by just clicking 'style name' and selecting another style name.

Step 3: Perform simple prosodic modification. For example, the duration of synthesized speech from each sentence can be defined, created prosodic parameters can be stored in the prosodic pattern library, and specific prosodic parameters can be applied to speech using Cut & Paste operations. Details of the above functions are described in Section 4.

Step 4: Modify prosodic parameters of synthesized speech from each sentence if necessary. The prosodic parameters are visually displayed as shown in Figure 4. The user can manipulates the prosodic parameters, power contour, F_0 contour, phoneme duration by mouse action. As shown in Figure 4, power contour, F_0 contour, speech waveform, phoneme boundary lines and phoneme symbols are displayed. The user can manipulate the power contour in the top area, the phoneme boundary in the second waveform area and F_0 contour in the third area.

Through these four manipulation processes, the user can produce the desired speech message by mouse actions in a trial-and-error manner; i.e., change prosodic parameters, then immediately synthesize and listen to the speech.

3. FEATURES

Sesign99 was designed with the aim of producing speech messages with various speech styles. The features of Sesign99 are outlined below.

3.1 Mimic Speech

Using the prosodic characteristics extracted from natural speech, stored speech style can be imitated. For example, mimicking the speech style of a famous actor or animation character is a very interesting approach. While the user can create speech messages with desired speech style using the prosody modification interface, it is not easy even for experts to create natural-sounding speech. Utilizing the prosodic characteristic of natural speech provides clues to prosodic modification and good knowledge of natural prosody to the user.

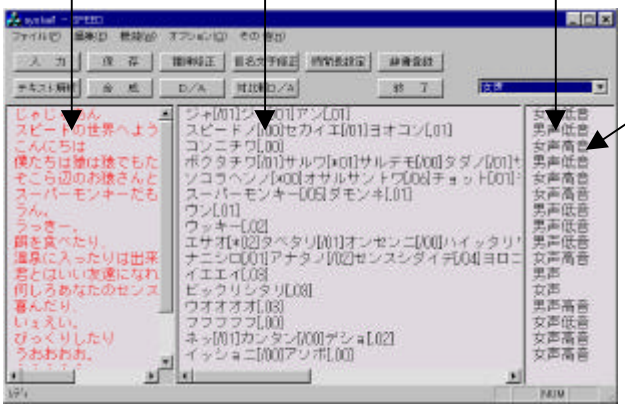


Figure 2. A screen-shot of Sesign99 main interface

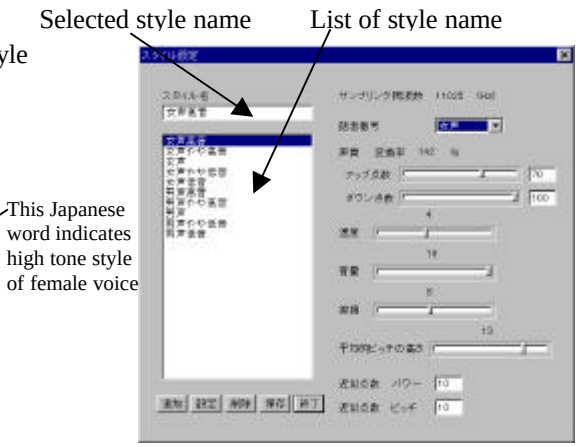


Figure 3. A screen-shot of the speech style definition dialogue

3.2 Efficient Production

The user can register created prosodic patterns in the prosodic pattern library. When reusing a speech style, speech with the desired prosody characteristics can be automatically produced by referring to a stored prosodic pattern; there is no need to re-create the prosodic characteristics. Moreover, functions that simplify the operation of prosodic modification are available. For example, basic prosodic modifying operation, i.e. extending phoneme duration or raising F_0 in the end phoneme of a sentence and etc., is provided as a macro function. Moreover, the history of user operation can be registered as a macro.

3.3 Various Speech Styles

Two basic voices, male voice and female voice, are available and voice style can be changed by using the VarioVoice algorithm[6], which is a voice conversion algorithm based on re-sampling. Because the user can set voice quality as well as prosody characteristics, power, intonation, pitch and speech speed, speech messages with desired speech styles can be produced.

3.4 Easy-to-use

Functions that allow the novice to easily control the speech parameters are introduced. For example, through the prosody modification interface, the user can modify F_0 and power contours simply by drawing the desired contour using a mouse. The following functions allow any user to produce the speech desired.

4. FUNCTIONS

The main functions of Sesign99 are described in detail below.

(1) Prosodic parameter extraction: we have already reported a prototype system that uses speech recognition and speech synthesis techniques[6]. Inputs are the speech signal and its phonetic transcription. Phoneme labeling is performed using HMM, and phoneme duration is determined by referring to the labels[7]. F_0 is then extracted from the speech signal using the AMDF algorithm[8]. Of course, these automatically extracted parameters have a lot of mistakes which degrade the quality of synthesized speech[9]. For manually correcting the errors, a GUI interface is implemented. A screen-shot of display of the error-correction interface is shown in Figure 5. By selecting and replaying parts of the speech, the phoneme boundary can be moved right/left and the F_0 contour can be freely drawn by mouse action. The appearance of extracted F_0 contour can be changed by manipulating the threshold value of the extraction method parameters.

(2) Prosodic pattern library: The desired prosody can be automatically given to the synthesized speech by referring the stored prosodic characteristics in advance.

(3) macro-function: Macro-functions for prosodic modification are provided to simplify prosodic modification. By simply selecting the macro-function, for example, we can trigger basic prosodic modification operations, i.e. extending phoneme duration or raising F_0 of the end phoneme of a sentence. Moreover, the history of user operation can be registered as a new macro-function and can be applied in the same way as system-defined macros.

(4) Speech insertion: User can combine natural speech with synthesized speech. This is efficient for tasks such as traffic alerts and weather forecast.

(5) Voice quality transformation: The VarioVoice algorithm allows the user to change voice quality. By transforming two standard voices, male and female, the speech with desired voice quality can be produced.

(6) User dictionary: for correctly performing text analysis, a user dictionary is provided in which the user can register readings and accents.

(7) Changeable approximation line: The F_0 contour and the power contour at each phoneme is represented by line approximation. The number of lines used in ap-

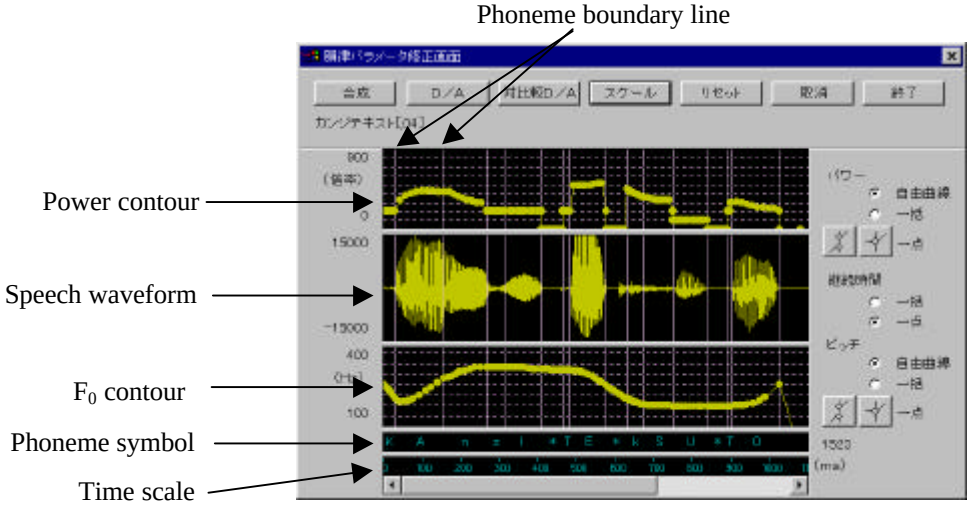


Figure 4. A screen shot of the prosodic modification interface

proximating the pattern of each phoneme can be freely changed.
(8) Cut & paste function: Prosodic parameters are handled in phonetic transcription editor. If the prosody was corrected in one sentence it is possible to reproduce the same prosody by a simple copy function.
(9) Plug-In interface: There are demands from experts for more sophisticated speech effects such as echo, delay and equalization. The external interface of Sesign99 permits the use of external plug-ins. The specifications of the interface are open and anyone can create and use the plug-ins.

3. CONCLUSION

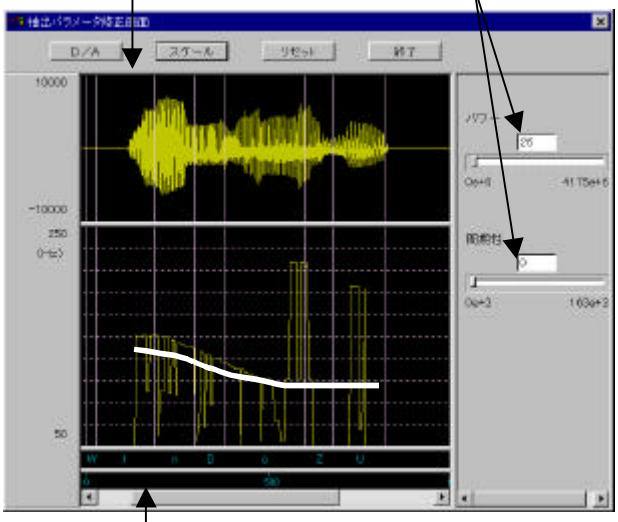
We proposed a new speech design tool that can produce engaging and natural-sounding speech messages. Sesign99 has enough power to create very natural speech messages easily. As future work, we will add new functions to make speech production even easier and create more interesting and emotional speech. Moreover, we will develop new application software to utilize the speech samples produced by Sesign99.

ACKNOWLEDGMENT

We are grateful to the members of the Media Processing Project for their helpful discussions. We also thank Mr.Yamamori, executive manager, for his continuous support of this work.

4. REFERENCES

[1] K. Hakoda, T. Hirokawa, H. Tsukada, Y. Yoshida, and H. Mizuno, “Japanese Text-to-Speech Software based on Waveform Concatenation Method,” Proc. AVIOS’95, pp. 45-54.
[2] Y.Sagisaka, ”Speech synthesis by rule using an optimal selection of non-uniform synthesis units,” Proc. ICASPP’88, pp.181-184
[3] E.Rank, H.Pirker ”Generating Emotional Speech With a Concatinative Synthesizer,” Emotion Speech, Proc. ICSLP’98, pp 671-674
[4] JM.Montenero, J.Guiterres-Arriola, Palazuelo V.Zue, ”Emotion Speech Synthesis: from speech database to TTS,” Proc. ICSLP’98, pp 923-926
[5] M.Abe, H.Mizuno, S.Takahashi, S.Nakajima,” A prototype Hybrid Scalable Text-To-Speech System,” Proc. IWSNHC3DI’97, pp.8-11
[6] M.Abe, ”A real time speech quality modification apparatus,” Proc. ASJ Spring Meeting, 3-7-6, pp.269-270(in Japanese)
[5] S. Takahashi, S. Sagayama, “Four-level Tied Structure for Efficient Representation of Acoustic Modeling,” Proc. ICASSP95, pp. 520-523.
[8] M.Ross, H.Schafer, A.Cohen, R.Freuberg, H.Manley, ”Average magnitude difference function pitch extractor,” IEEE Trans. ASSP, ASSp-22, pp.353-362
[9]M.Abe,H.Mizuno, S.Takahashi, S.Nakajima,” A new framework to provide high-controllability speech signal and the development of a workbench for it,” Proc.Eurospeech’97, pp.541-544



F_0 contour (Light line: automatic extracted F_0)
(Bold line: manual correction of F_0)

Figure 5. A screen shot of the error-correction interface