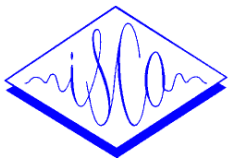

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SPEECH EDITOR BASED ON ENHANCED USER-SYSTEM INTERACTION FOR HIGH QUALITY TEXT-TO-SPEECH SYNTHESIS

Kazuo Hakoda, Tomohisa Hirokawa, and Kenzo Itoh

NTT Human Interface Laboratories
Yokosuka-shi, Kanagawa, 238-03 Japan

ABSTRACT

This paper describes a new speech editor based on enhanced user-system interaction that produces high quality synthesized speech by using an advanced text-to-speech synthesis method. A prototype system is constructed on a work station with the Open Window system. Features of the prototype are that the operator can correct the faults of the text-to-speech synthesis method and produce high quality synthesized speech from input Japanese text. System operation has been optimized by adopting a real-time synthesizer and a GUI design based on mouse operations. A system evaluation confirms that character level correction is very effective for improving synthesized speech quality. The proposed system can be used to provide voice messages for a conventional digital audio response unit at low cost.

1. INTRODUCTION

In recent years, text-to-speech synthesizers are being put to practical use in various application fields, such as telephone information services, tools for the handicapped, and voice output devices for machines[1][2]. Unfortunately, the quality of output synthesized speech remains insufficient to permit its wide use in the services needing high quality voice messages such as telephone and radio broadcasting services.

One solution is to synthesize speech using digitized voice compilation methods and this technique is used in various voice information services. However, many voice services frequently change the output messages because of changes of service content. This forces reconstruction of the digitized voice messages and makes maintenance much more expensive.

In order to solve these problems, we propose a speech editor that allows an operator to easily correct the output of a text-to-speech system, and so produce high quality synthesized speech from input Kanji characters[3]. The outstanding point of this editor design is that an operator can easily manipulate the system parameters, and reduce the operating time taken to correct and modify synthesized speech.

We construct a prototype speech editor on a work staion. System operation is optimized by adopting a real time synthesizer and a GUI design based on mouse operation. The speech parameters are graphically illustrated on a display, and can be easily modified by mouse operations. In this paper, we describe the configuration and operation of this editor. Furthermore, the results of system evaluation tests are described.

2. SYSTEM OVERVIEW

2.1 System Configuration

The block diagram of the speech editor is shown in Fig.1. Input Japanese text consists of Kanji and Kana characters. Speech control symbols such as speech speed, volume can be contained into input text.

The speech editor has two types of editing modes. One is the character correction mode, in which various character errors such as reading errors for Kanji and accentuation errors resulting from the text analysis can be corrected by editing functions. The other is the parameter modification mode. If the quality of synthesized speech is insufficient, this modification is carried out. The speech parameters such as fundamental frequency contours (pitch contours) are displayed in a window, and can be easily modified by pointing and moving the modified point by mouse.

The speech editor employs two synthesizers. One is the PC board- based real-time synthesizer developed by NTT in 1990[4]. This synthesizer can convert Japanese text into connected speech in real time and send the results of text analysis to the host machine. This real-time synthesizer is used in the text analysis of input Japanese text and voice confirmation in the character modification stage. The output strings from text analysis consist of Kana characters, accentuation symbols and phrase boundary symbols that represent the degree of connection between phrases. The synthesizer is connected to a work station via an RS-232C standard interface.

The other synthesizer yields high quality speech and is based on the waveform compilation method[5]. It allows an

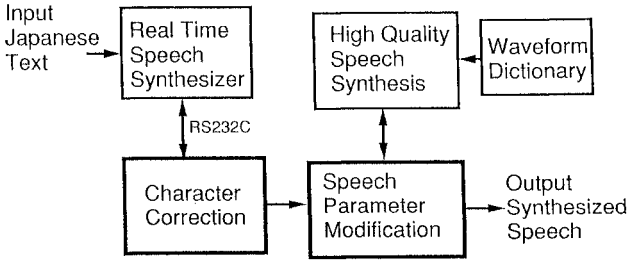

Fig.1 Block diagram for the Speech Editor

operator to modify prosody control parameters. The input format equals the output format of the text analysis stage. Synthesis is carried out on a work station. This synthesizer stores more than forty thousand segments in the waveform dictionary. The segment that most closely match the prosody and phoneme context is selected, and concatenated to produce connected speech. At this time, the five best-matched candidates for each segment are stored in memory, and used in parameter modification. Pitch parameters are set from Kana readings, accentuation and phrase boundary symbols by pitch generation rules[6]. Power and phoneme duration are set according to the phoneme environment[7]. Pitch control for waveform segments is based on the pitch synchronous overlap- add technique[8].

The quality of synthesized voice is more natural than that produced by the real-time synthesizer, but producing the speech takes five times longer.

2.2 Modification Procedure

The flow for producing voice messages is shown in Fig.2. Input Japanese text consisting of Kanji and Kana characters is created by a text editing tool of the work station. Input text is sent to the real-time synthesizer and converted into Kana characters and prosodic symbols by the text analysis function. The converted characters are then displayed in a window for character editing. The operator checks displayed characters visually, and corrects conversion errors. The operator confirms the results of error correction by listening to the synthesized speech.

After correcting the character level errors, the high quality speech synthesis is carried out. If the quality of synthesized speech is insufficient, speech parameters are modified. The speech parameters such as pitch, phoneme segment power, and phoneme duration set by rules are graphically displayed. These parameters are modified easily by pointing and moving a mouse. Changes are confirmed by listening the synthesized speech created by high quality speech synthesis. In this stage, the speech parameters are repeatedly modified until the synthesized voice has the desired quality.

3. SYSTEM OPERATION

This chapter describes the functions, utilities and tools of this speech editor in detail.

3.1 Speech Editor

The speech editor has two main windows: a character correction window and a speech parameter modification window. When the Speech Editor button is clicked, the speech editor window is opened as shown in Fig.3. Text files are illustrated as small graphical icons by clicking the READ button in the File pull down menu. When the file is selected, input text is converted into Kana characters and

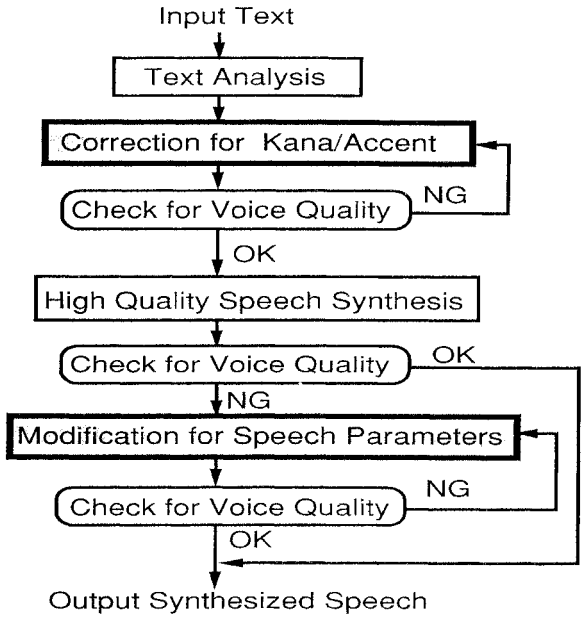

Fig.2 Flow from producing voice messages

prosodic symbols by the real-time synthesizer.

3.2 Character Correction

The Kana characters and prosody symbols of the selected text file are displayed on a character correction window as shown in Fig.3. The number of under bars represents phrase boundary connection strength[6]. One corresponds to strong connection and two bars corresponds to weak connection. More than three bars means a pause and the number of bars corresponds to pause length. The character string $S5$V10 indicates that the speaking speech rate is five(fast) and volume level is ten(loud). The over bars express the rough digitized pitch pattern. A falling position indicates the accented syllable of the phrase.

The first step of character correction is revising the Kana readings for Kanji characters, and specified Kana particles. The phrase boundary including pause is easily revised in the same way as Kana readings correction.

Next step is the correction of accentuation positioning, which is changed as follows.

The operator moves the cursor to the correct accent position, and clicks the right mouse button to display a sub-menu panel. The accentuation position is corrected by selecting the ACCENT[] item in the sub-menu. The rough pitch line for the corrected accent is then redrawn.

These correction results can be quickly and easily

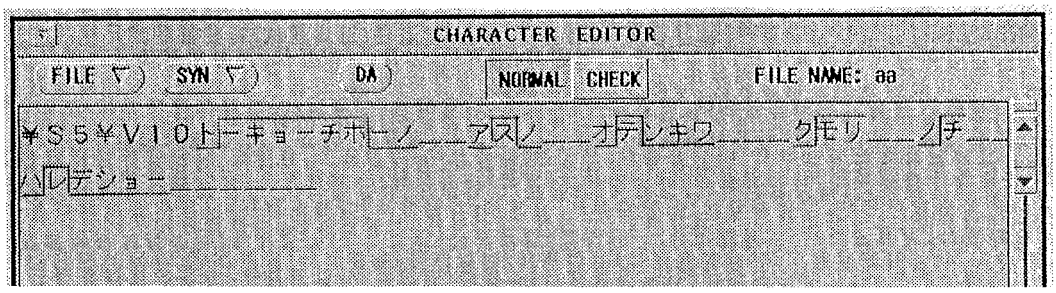

Fig.3 Character correction window

confirmed by clicking the SYN button. Synthesized speech is generated by the real time synthesizer as mentioned before.

3.3 Speech Parameter Modification

3.3.1 Speech Parameter Modification Window

After finishing character correction, the operator changes the synthesis mode from real time synthesis into high quality synthesis by clicking the NORMAL button. Clicking the SYN button now produces high quality speech synthesis. If the quality of the synthesized speech is judged insufficient, the operator can modify the speech parameters and exchange waveform segments. Speech parameter modification is carried out as follows.

At first, the operator designates the area to be modified in the character modification window by mouse operation, and clicks the right button of mouse to display the sub-menu panel. After selecting the MODIFY item, the parameter modification window is opened shown as Fig.4. The window is divided into four regions of pitch frequency, waveform, segmental power and phoneme boundary with label. This window provides several buttons, FUNCTION, SYN, DA, and DRAW PITCH. The SYN button produces high quality synthesis using the modified parameters. The DA button is used to play the high quality synthesized voice. The pull down menu of FUNCTION has several functions, SAVE&QUIT, SAVE, INIT, and QUIT. The INIT button recalls the set of last-saved parameters.

3.3.2 Prosody Parameter Modification

The pitch frequency contour is formed as a series of straight lines that connect the start and end pitch value of each phoneme. A pitch line can be changed by selecting and dragging these points vertically by the mouse. Also, pitch lines can be drawn in freehand by clicking DRAW PITCH. Original and modified pitch lines are drawn in different colors. This is very helpful in visually confirming the alterations. This type of interface is adopted for modifying power and phoneme duration. The operator can change the phoneme segment power by moving the top line of power segment vertically.

The duration of each segment can be changed by moving the segment boundary bar horizontally.

Grouping functions for pitch, power and segment duration modifications are provided. After changing the mode from the NORMAL mode to the GROUP mode, the operator designates the grouping area by mouse selection, and can change the parameters included in this area at the same time. This function is useful for wholesale changes to words or phrases.

3.3.3 phoneme segment changing

If the synthesized speech is judged to be noisy or lacking in clearness, the selected waveform segments may be inappropriate for the synthesis conditions. The playback function(DA button) can limit the playback area to waveform display window. This function is very useful for finding the location of the inappropriate segments.

Waveform segment candidates can be displayed in a sub-panel by clicking the phoneme segment area. The top five candidates with uttered phoneme context and average pitch frequency are arranged as shown in Fig.5. The operator can easily exchange segment candidates by selecting candidates from the sub-panel. These parameter modifications and segment exchanges are repeated until the desired quality is obtained.

4. SYSTEM EVALUATION

4.1 System evaluation viewpoint

In order to find the effectiveness of this editor, it is necessary to determine in what way modifications are effective for improving synthesized speech, and how much the quality of synthesized speech can be improved. We examine the relationship between the editing mode, the operating time, and the quality of synthesized speech.

4.2 The measurement of operating time

The operating time is the time taken for character correction and speech parameter modification. Four sets of text consisting of about 130 Kanji and Kana characters were used for this test.

Three subjects were selected as follows.

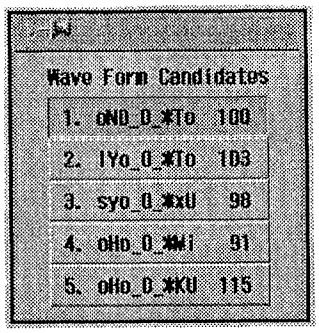

Fig.5 Waveform segment selection panel

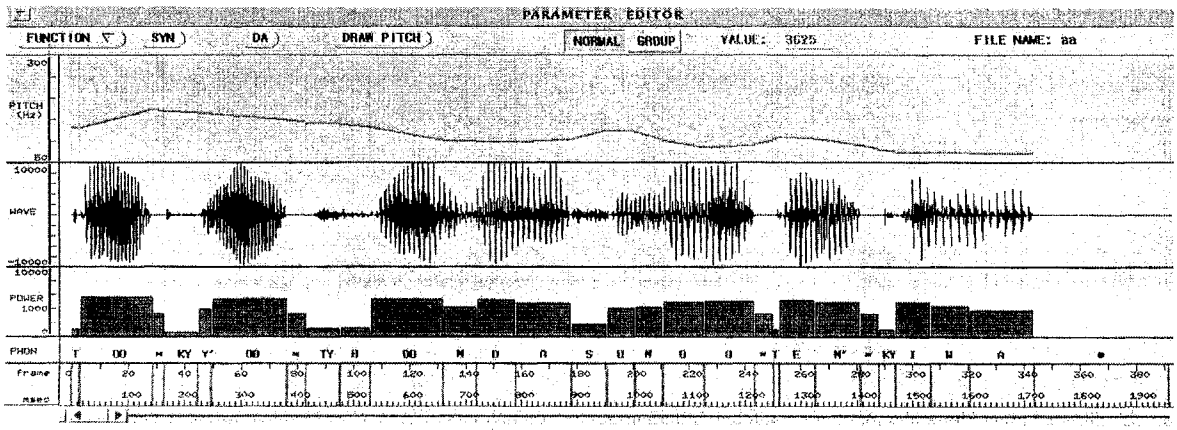

Fig.4 Speech parameter modification window

Subject A: a speech researcher, one of the developers of this editor
Subject B: an expert in this editor trained for more than one month
Subject C: a novice

The operating time taken in each editing mode was measured for each test sample. The maximum time was limited to 90 minutes for each modification mode. The measured results are shown in Fig.6. The left vertical axis means the operating time(minutes) taken to process one 1 second of synthesized speech. Each line designates the average operating time for each subject.

The time taken for parameter modification mode is no more than three to five times longer than that taken for character modification. The difference between the expert and the beginner is small for character correction and large for parameter modification. Beginner should be trained for parameter modification or receive appropriate assistance.

4.3 Hearing test for synthesized speech quality

In order to clarify the relationship between the editing modes and the quality of synthesized speech, a hearing test based on the method of successive categories was carried out. The subjects were required to score several synthesized speech samples by focusing on their satisfaction with the voice quality for telephone and radio broadcasting services. Three types of subjective voices were prepared for each text: the original synthesized speech, the re-synthesized speech after character level correction, and that after speech parameter modification.
The evaluation categories were set as follows.

5: Very satisfactory
4: Satisfactory
3: Satisfactory or unsatisfactory
2: Unsatisfactory
1: Very unsatisfactory

In this evaluation test, sixteen members of our laboratory were employed. The test results are shown in Fig.6. Each line designates the quality of the corrected and modified synthesized speech produced by the three operators.The quality of synthesized speech without the usage of this editor can not keep the minimum voice quality (point 3) which may be required in many voice services. Character correction raises the voice quality level over the minimum voice quality level.
Parameter modification can increase the voice quality further, but the degree of improvement is small compared with that of character correction. The effect of operator skill on voice quality is relative small. It is confirmed that character correction is an effective method to improve the quality of synthesized speech because it needs no special operator skills and can raise the speech quality up to the minimum level needed for many voice services within a short time.

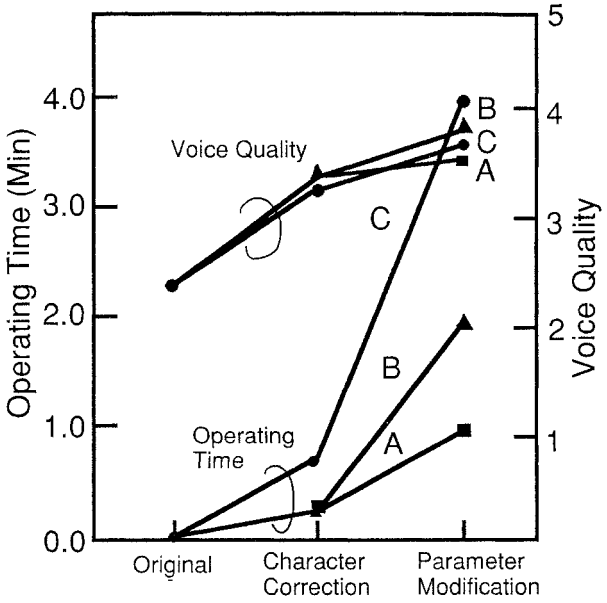

Fig.6 Results of the hearing test and measurement for the operating time

5. CONCLUSION

A new speech editor based on enhanced user-system interaction that produces high quality synthesized speech by using text-to-speech synthesis is described. This editor can produce voice messages for the conventional audio response units which need low voice message construction costs. Also, this editor can be used as a speech enhancement tools in order to provide the synthesized speech that expresses the feeling intended. The editor may help speech researchers in improving the rules applied to text-to-speech synthesis.

In the near future, the computing time taken by high quality synthesis will be reduced drastically by optimizing the waveform dictionary and the waveform synthesis process, and used to replace the conventional real-time synthesizer in the text confirmation stage.

ACKNOWLEDGEMENT

The authors wish to thank Dr. Nobuhiko Kitawaki, director of the Speech and Acoustics Laboratory, Dr. Noboru Sugamura, group leader, for their encouragement during this work. We also thank members of our group and NTT Intelligent Technology Corporation, who helped in the system evaluation tests.

REFERENCES

[1] Y.Mitomi,"Applications and a Future Prospect of Speech Synthesis",JASJ,Vol.49,No12,pp.875-880(1993) (in Japanese)
[2] T.Hirokawa,"Applications of Japanese Text-to-Speech Synthesizer",Speech Tech'89,pp.30-32(1989)
[3] T.Hirokawa, K.Itoh, and K.Hakoda,"Speech Editor based on Enhanced User-System Interaction",Proc.AVIOS'93,pp.39-45(1993)
[4] K.Hakoda, S.Nakajima, T.Hirokawa and H.Mizuno,"A New Japanese Text-to-Speech Synthesizer based on COC Synthesized Method", Proc. ICSLP90,pp.809-812(1990)
[5] T.Hirokawa, K.Itoh and H.Sato,"High Quality Speech Synthesis System based on Waveform Concatenation of Phoneme Segment", IEICE Trans, Vol.E76-A,No11,pp.1964-1970(1993)
[6] K.Hakoda and S.Sato,"Prosodic Rules in Connected Speech Synthesis", Systems, Computers,Controls,Scripta Electronica Japonica 3, Vol.11,pp.28-37(1980)
[7] K.Itoh, T.Hirokawa, and H.Sato,"Phoneme Power Control for Speech Synthesis", IEICE Trans, Vol.E76-A, No11,pp.1911-1917(1993)
[8] Charpentier,F. and Moulines,E.,"Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones",Proc.Eurospeech'89,1989