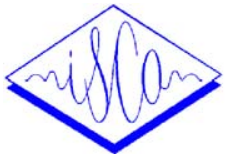

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

A NEW FRAMEWORK TO PROVIDE HIGH-CONTROLLABILITY SPEECH SIGNAL AND THE DEVELOPMENT OF A WORKBENCH FOR IT

Masanobu ABE, Hideyuki MIZUNO, Satoshi TAKAHASHI and Shin'ya NAKAJIMA
NTT Human Interface Labs.
1-1 Hikarinooka Yokosuka-Shi Kanagawa 239 Japan
Tel: +81 468 59 2547, Fax: +81 468 55 1054, E-mail: ave@nttspch.hil.ntt.co.jp

ABSTRACT

This paper proposes a new framework to enhance the access to and control of speech signals. To enhance accessibility, the proposed framework assigns multi-layered tags such as orthographic transcriptions, and phonetic transcriptions. The tags also make it possible to precisely synchronize a speech signal with animation. In terms of control, the proposed framework provides hybrid speech; combining both human speech and speech synthesis-by-rule. Its quality ranges from simple TTS (the worst case) to encoded natural speech (the best case) depending on the resources available: texts, fundamental frequency(F_0) contour, power contour, phoneme duration, and so on. To create speech messages based on the proposed framework, we developed a workbench employing speech synthesis and recognition techniques. Important features of the workbench are a powerful GUI(Graphical User Interface) with which to manipulate prosodic information and a function to synthesize speech in trial-and-error manner. An evaluation by creating speech messages shows the good performance of the workbench.

1. INTRODUCTION

Recently, multimedia contents are getting popular in our daily life, examples include electric encyclopedia, games, interactive movies and WWW home pages containing audio, video, and animation. Needless to say, speech messages are an essential part of multimedia content, and it is important that speech messages be created, used, managed, compressed, stored and transmitted in adequate ways. From this point of view, we propose a new framework to generate speech messages. There are two aspects; i.e., access from the user's side and control from the producer's side. In terms of access, digitized speech signals are now handled as stream data and are sequentially heard from begin to end. This is just like treating text strings as image data. Because text strings are stored as ASCII code, we can find a particular article by locating key words. From an analogy with text strings, the proposed framework makes it possible to handle speech messages in the same way as "ASCII code" for text strings. In terms of control, once speech messages are recorded, there is no way for producers to control the speech messages. All that they can do is to rerecord them. Moreover, it is difficult to synchronize moving pictures and speech messages. Accurate synchronization is essential in producing multimedia contents, so another aim of the proposed framework is to provide a way of flexibly controlling speech messages.

In section 2, the proposed framework is explained, in section 3, a workbench system for the proposed framework is proposed, and in section 4 the workbench system is evaluated.

2. A NEW FRAMEWORK

Once a large amount of information is available, it is important for users to get what they are interested in as quickly as possible. From this point of view, speech signals are not suitable; i.e., users have no way to directly find that part of the speech signal that tells what they want to hear. The proposed framework solves the problem by giving tags to each speech signal. For example, if 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 video movie, users also can use the tags to locate particular parts of the movie. Another advantage is that the tags make it possible for producers to synchronize moving picture and speech messages. For example, if a phonetic transcription is used for tagging and alignment between speech signal and the phonetic transcription is performed, the tags make it possible to precisely synchronize the speech signal and animated lip movements. As shown here, tags are active on several layers.

Speech uttered by human beings is perfect in terms of speech message quality, but has less controllability. On the other hand, speech synthesis-by-rule does not always have sufficient quality, but can be controlled with high flexibility. The proposed framework enables to combine both human speech and speech synthesis-by-rule and results in hybrid speech, in other words, it is an extended Text-to-Speech(TTS) that can utilize the parameters extracted from human speech if available. Its quality ranges from simple TTS (the worst case) to encoded natural speech (the best case) depending on the resources available: texts, fundamental frequency(F_0) contour, power contour, phoneme duration, and so on. Hybrid speech can be viewed as a low bit rate speech coding; i.e., only phonetic symbols and prosodic parameters should be transmitted. Another advantage is to recycle hybrid speech usage. To create new speech messages, it might be possible to edit existing hybrid 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.

The most immediate, and important features of the proposed framework is tags for speech messages that are active on several layers, and hybrid speech coding. Table 1 shows examples of speech message configurations in the proposed framework.

2. A WORKBENCH FOR THE PROPOSED FRAMEWORK

To create speech messages based on the proposed framework, we developed a workbench. The workbench has two aspects. First, the workbench is a tool to assign tags to natural speech. In this case, natural speech is stored together with tags. Second, the workbench is a tool to synthesize hybrid speech using prosodic parameters extracted from natural speech. In this case, natural speech is used only as references for prosodic

parameter generation. Another important feature of the workbench is its GUI(Graphical User Interface) to manipulate parameters; i.e., errors can be corrected in speech analysis, and prosodic parameters changed in synthesizing speech in a trial-and-error manner.

2.1 Outline of the Workbench

Figure 1 shows a block diagram of the workbench. Speech messages are created as follows.

Step1 : Input texts of Kana and Kanji (Chinese character) 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 phonetic transcriptions and accent types if needed. This function is useful for Japanese, because readings of Kanji and accent type are usually context dependent and are difficult to estimate. To check accent types, a user can synthesize the speech. A user is also allowed to start from this step; skipping Steps 1 and 2.

Step4 : Generate prosodic parameters. This step has processes. Details are explained in 2.2.

Step5 : Modify prosodic parameters; F_0, duration, and power. The prosodic parameters are visually displayed as shown in Fig. 2, and a user can modify them by mouse actions. A user can create speech in a trial-and-error manner; i.e., changing prosodic parameters, then immediately synthesizing and listening the speech.

Step6 : Store speech messages and/or their parameters.

Table 1 Example speech message configurations in the proposed framework

speech message	tag information
natural speech	orthographic transcription
hybrid speech	phonetic transcription
synthetic speech	prosodic information

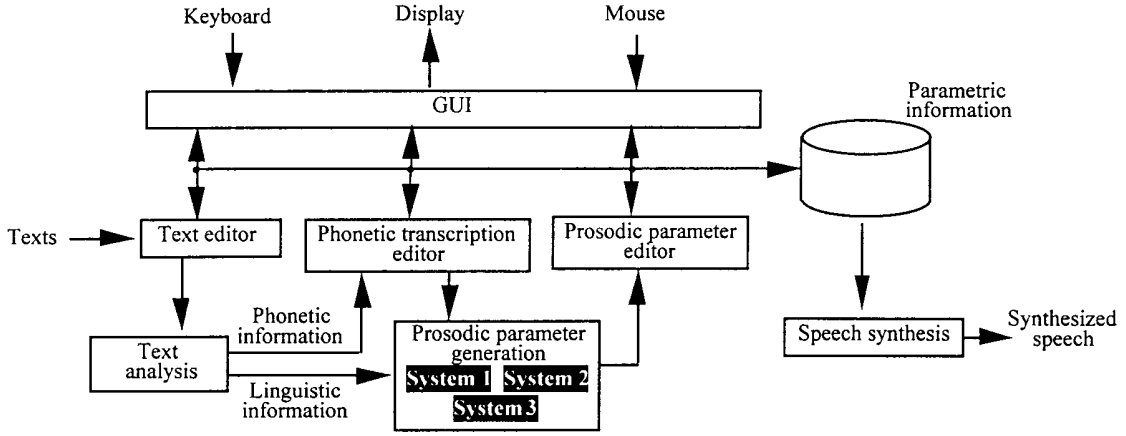

Fig. 1 Block diagram of the workbench

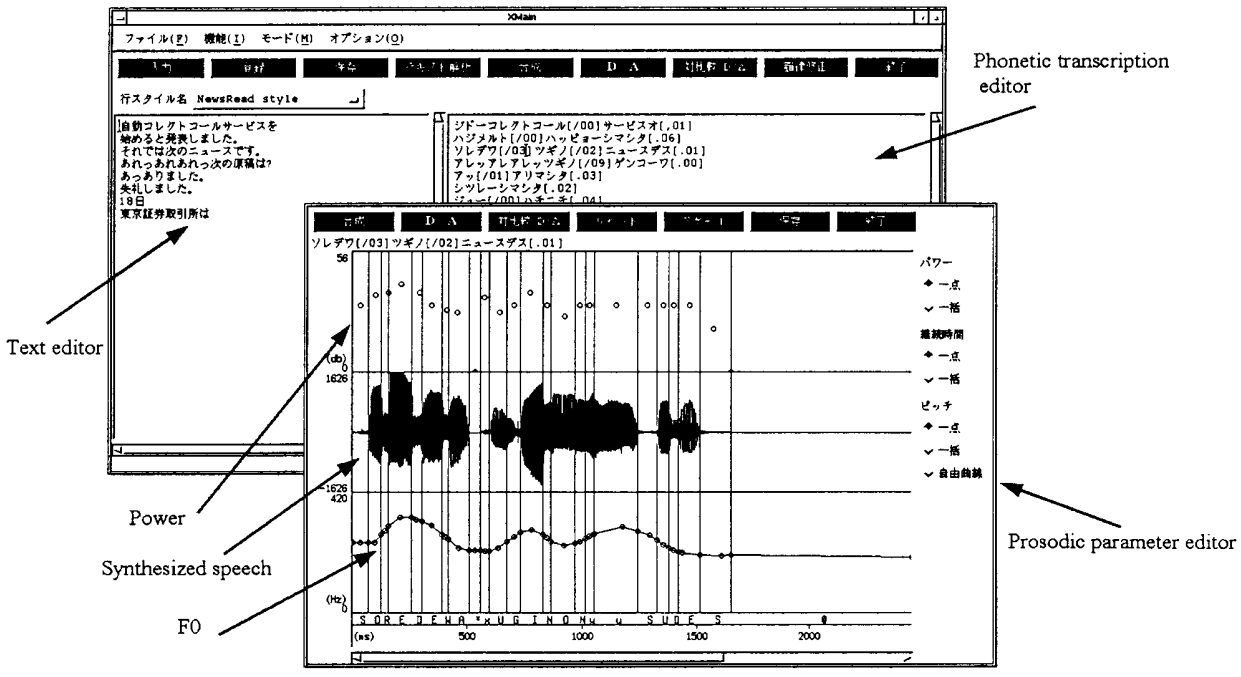

Fig. 2 Workbench display

2.2 Prosodic Parameter Generation Systems

As shown in Fig. 3, the workbench has three systems for prosodic parameter generation. System 1 is the same as the conventional TTS[1]; inputs are phonetic and linguistic information and prosodic parameters are generated by rules. In this case, prosodic parameters must be greatly changed at Step 5 to synthesize natural sounding speech. Table 2 shows the specifications of the TTS. In System 2, prosodic parameters are automatically extracted from natural speech. First, using the phonetic transcriptions generated at Step 3, natural speech is assigned phoneme labels and phoneme duration are determined by referring to the labels. Fundamental frequency(F_0) is then extracted from the natural speech; it is sampled at 3 points in every phoneme. Phoneme labeling is performed by the Hidden Markov Model (HMM)[2]. Table 3 shows the specifications of the HMM. As shown in Table 3, averaged labeling error is about 15 msec. We think this error is small enough for rough tags such as key words. However, to create speech messages, the error must be corrected at Step 5. In addition to System 2,

Table 2 TTS specifications

Speech synthesis	Waveform concatenation method
Synthesis units	Context dependent phonemes (tri-phone)
Number of synthesis units	6000
Word dictionary	100,000

Table 3 HMM specifications

Number of models	1764
Number of states	687
Number of distributions	3684
Acoustic parameters	LPC Cepstrum: 16 LPC Delta Cepstrum: 16 Delta power
Training data	45,376 words (2646 words for 16 speakers)
Performance	Errors against to hand labels: 15.3 msec /phoneme

System 3 has a GUI to correct phoneme labeling and extracted F_0. Therefore, System 3 can provide error-free prosodic parameters. The outputs are useful for synchronization between speech messages and lip movements in phoneme level. However, Step 5 is still needed for creating speech messages, because phenomena in natural speech are not always the same in the speech segments forming speech synthesis units, and prosodic parameters in natural speech are not always best for synthesizing natural sounding speech.

3. EVALUATION OF THE WORKBENCH

The workbench was evaluated in terms of speech message creation. The main purpose of the experiment was to confirm the advantages and disadvantages of the three systems for prosodic parameter generation. Two users were trained over three days to create speech messages using the workbench. Because they had three years of experience in assigning phoneme labels from spectrograms, we did not have to teach them the characteristics of speech signals. After the training, they were asked to create a set of 9 speech messages using the three systems. On the average, one message consists of 5.1 phrases, or of 52.4 phonemes. To avoid the effects caused by sentence and/or system order, the 9 messages were divided into 3 sets, and each set of messages was created using the systems in different order.

3.1 Experiment Results

Figures 4 and 5 show average time and standard deviation to create a speech messages by each prosodic parameter generation system and each sentence, respectively. Judging from the figures, while the results heavily depend on the user, System 3 required the longest time. Figure 6 shows details of the average processing time of user 1 and 2 in accessing System 3. Judging from Fig. 6 and Fig. 4, the long processing time of System 3 is caused by the correction of label and F_0 errors; about 5 minutes in total. That is, the time spent in trial-and-error synthesis was the almost the same in all three systems.

3.2 Evaluation of the Created Speech Messages

A listening test was carried out to evaluate the created speech messages. Including a version of speech synthesized by a TTS system, the total number of stimuli was 63 (9 x 3 x 2 + 9) and

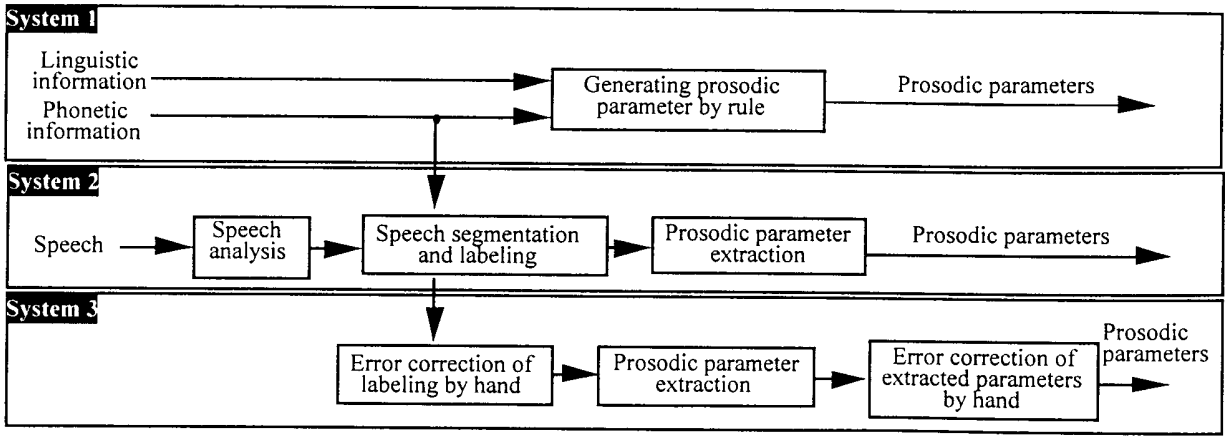

Fig. 3 Three sytems for prosodic parameter generation

they were presented to eight listeners at random. Listeners were asked to rate the quality in 5 categories; i.e., from excellent to poor. Figure 7 shows the experiment results. System 1 and System 3 can provide the highest quality. However, interestingly, the prosodic parameters of System 1 and System 3 are sometimes quite different. System 2 offers relatively poor performance. The result indicates that the initial prosodic parameters strongly influence the quality of final speech messages. In System 2, errors in automatic labeling and F_0 extraction might lead users in the wrong way.

3.3 Discussion

The sentences used in the experiments were extracted from news papers or magazines, and were independently uttered; speaker could not express context effects and emotional effects in their utterances. Therefore, there is little difference between System 1 and System 3. In an informal listening test, speech created by System 3 had closer prosodic parameters to human speech than System 1, so we think System 3 has an advantage in creating emotional speech messages.

Although System 3 can yield error-free prosodic parameters, as shown by Fig. 6, System 3 still requires trial-and-error synthesis. This is mainly because the phenomena in natural speech are not always the same as those in the speech segments used for speech synthesis units, and prosodic parameters in natural speech are not always best for synthesizing natural sounding speech. Therefore, we think the function of trial-and-error synthesis is important in the workbench.

4. CONCLUSION

We proposed a framework to enhance the access to and control of speech signals and developed a workbench based on it. In terms of creating speech messages, we think the workbench is powerful enough. As a future work, we will add new functions to achieve the final goal of the proposed framework.

ACKNOWLEDGMENT

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

REFERENCES

[1] K. Hakoda, T. Hirokawa, H. Tsukada, Y. Yoshida, and H. Mizuno, "Japanese Text-to-Speech Software based on Waveform Concatenation Method," Proceedings of AVIOS'95, pp. 45-54.

[2] S. Takahashi, S. Sagayama, "Four-level Tied Structure for Efficient Representation of Acoustic Modeling," Proc. ICASSP95, pp. 520-523.

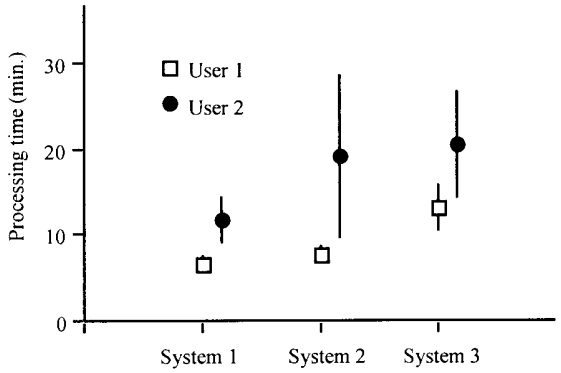

Fig. 4 Processing time of each prosodic parameter generation system

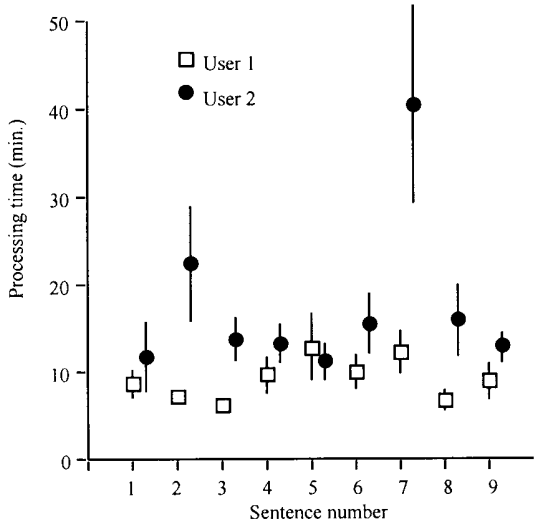

Fig. 5 Processing time for each sentence

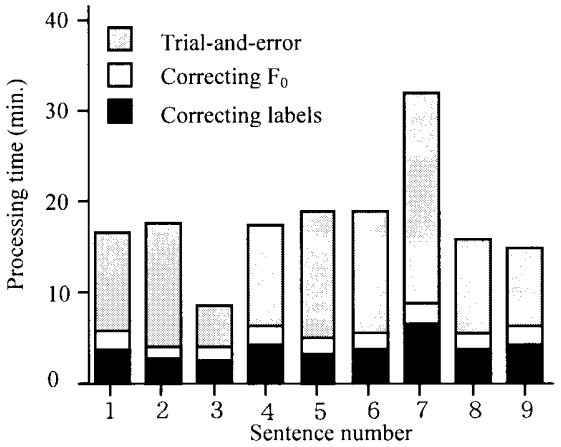

Fig. 6 Details of processing time in System 3

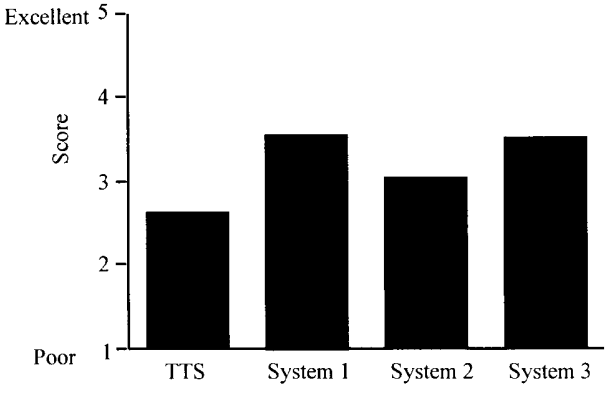

Fig. 7 Results of a listening test