DEVELOPMENT OF CAI SYSTEM EMPLOYING SYNTHESIZED SPEECH RESPONSES

Tsubasa SHINOZAKI and Masanobu ABE

NTT Human Interface Labs.
1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
E-mail: tsubasa@nttspch.hil.ntt.co.jp, ave@nttspch.hil.ntt.co.jp

ABSTRACT

This paper proposes a Computer Assisted Instruction (CAI) system that teaches students how to write Japanese characters. The most important feature of the system is the usage of synthesized speech to interact with users. The CAI system has a video display tablet interface. A user traces a pattern of a character using the tablet pen, and simultaneously his tracing is shown on the display. When the trace line is outside the pattern, the system simultaneously outputs synthesized speech to correct the errors. To design strategies for generating instructions, behavior and instruction messages of a human teacher were recorded and analyzed. One of the most interesting challenges of the system is a function that changes the "personality" of the teacher, such as a strict teacher, a friendly teacher, and a short-tempered teacher. According to the experimental results, it was confirmed that the proposed system makes it possible to convey a particular impression using synthesized speech.

1. INTRODUCTION

In recent years, synthesized speech has been used in e-mail reading systems, information retrieval systems, telephone directory systems and so on. In these systems, synthesized speech has enabled the realization of useful services with low cost. These applications are a promising area for synthesized speech. The theme discussed in this paper, however, is somewhat different, i.e., synthesized speech is used as an interface between systems and human beings [1], [2]. Examples of this kind of interface can be found in popular films such as HAL9000 in 2001: A Space Odyssey or C3PO in Star Wars. They can communicate with human beings using speech. In the films, the computers are depicted as sentient beings and could voice their emotions. Human beings could understand their feelings through their verbal expressions. As in these examples, the speech interface made it possible to communicate with the computers in a comfortable way. The speech interface in the films is fictional, but we believe this type of interaction using speech enables easy system access, and a user-friendly environment [3]. During the interactions, unlimited kinds of responses or expressions will be necessary because fixed system responses can be easily detected by human beings, and once the user realizes this, the user will lose all interest. Synthesized speech has the capability to deal with this because it can generate unlimited responses. Thus far, speech interaction shows promise as another area for synthesized speech.

This paper tries to develop a system that employs synthesized-speech interaction. In Section 2, a task is introduced to reduce the difficulty of the problem. The task is one of Computer Assisted Instruction (CAI), where a user and a system have the same goal in mind and where a user interacts with the system using scribed inputs and synthesized speech outputs. In Section 3, using the task, real speech data are collected and analyzed, and design strategies for synthesized-speech interaction are formulated. In Section 4, the strategies are implemented and experiments are performed to ensure that the system leaves a particular impression on the users using synthesized speech.

2. CAI SYSTEM "CALTURE"

In Japan as in many other countries, it is customary to exchange hand-written letters. To write well-proportioned characters is considered to be a special ability of a person. Moreover, in Japanese calligraphy, there is a certain aesthetic quality to the writing of the Japanese language. To attain proficiency in Japanese calligraphy, the first step for students, whether they are school children or foreigners who want to learn Japanese, is to learn to write well-proportioned characters through practice. Therefore, we decided to develop a system that instructs these students in handwriting by employing synthesized-speech interaction. Figure 1 shows the CAI system "**CALTURE**" (CALligraphy Training system Using voice REsponse). One feature of **CALTURE** is a video display tablet interface. A Japanese character is displayed on the video display tablet, a user traces the pattern using the pen, and simultaneously the user drawing is shown on the computer display. After several repetitions, the pattern of the character disappears and the user continues to write. Here, the synthesized-speech interaction comes into play. As the user writes the character, if it does not follow the pattern correctly, **CALTURE** outputs synthesized speech to correct the errors. Speech examples are "Right, right!", "A little bit to the left", and "No, start lower". Conversely, **CALTURE** praises the user when the character is written correctly such as outputting "Good" or "That's right! Keep it up!".

The advantages of using synthesized-speech interaction in the task are shown below.

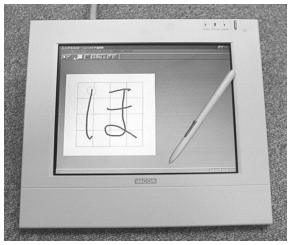

Figure 1: CAI system '**CALTURE**'.

- Practicing handwriting is a monotonous task. The synthesized-speech interaction could make practice more interesting, similar to if the user is playing a TV game.
- Audio instruction is a suitable means to quickly notify users of errors because the user's hands and eyes are occupied.

The task makes implementing synthesized-speech interaction easier for the reasons below.

- The system can track the user's writing to generate instructions. (Ex. Error repetition)
- The system can know the user's actual actions thorough pen inputs.

Finally, the essence of the task is cooperation between the system and human beings to achieve a goal. This kind of task is popular in the real world, so the result obtained from CALTURE might be applicable to other tasks.

3. ANALYSIS OF HUMAN TEACHER BEHAVIOR

3.1. Experimental procedures

To design strategies for generating instructions and to collect examples of instruction messages, an experiment was performed. As shown in Fig. 2, an operator and a subject work in separate rooms. Here, the subject is a teacher that can write well-proportioned characters. The operator traces a pattern of a character displayed on the video display tablet, and his drawing is simultaneously displayed on a CRT on the subject side. Watching the CRT, the subject instructs the operator as he writes out the pattern of the character. The experiment was setup such that the subject believes that the operator cannot see the pattern of the character, and the operator intentionally makes mistakes to collect various instruction messages. The experiment was recorded on video and audiotapes.

3.2. Experimental results

Based on the analysis of the human teacher's behavior, strategies for generating instructions are summarized as follows.

- When an error occurs at the beginning of a stroke, error correction messages are given as quickly as possible.
- Except for the above case, almost all instructions are given at the end of the stroke.
- At the beginning of the spoken instructions, interjections and demonstrative pronouns are often given to attract user attention.
- Almost all instructions using demonstrative pronoun are given with pointing gestures.

3.3. Algorithm for performing real-time indication in CALTURE

An algorithm was designed based on the results in Section 3.2. Figure 3 is a block diagram of the algorithm. The algorithm has two paths to generate messages. The first path (**VI** in Fig. 3) generates messages that attract user attention and has little correlation to the contents of the instruction itself, but must be output as quickly as possible. Examples are "Hey", "Oh!", and "No!". The second path (**II**, **III**, **IV,** and **V** in Fig. 3) generates instruction messages and incorporates the following steps. The Roman numerals in the following explanations correspond to the blocks in Fig. 3.

(1) In advance, check points of each character are prepared as scripts in **VII**. The scripts are constructed from abstract instructions for primitive strokes.

(2) In block **II**, based on the degree of error (the difference between the user's pen input and the character pattern), an appropriate abstract instruction is selected from the scripts.

(3) In block **III**, the abstract instruction is converted into actual messages. For example, when the user's drawing is shifted to the left, possible messages are "right", "right, right", "opposite direction" and so on.

These steps generate the instruction messages. The following blocks make it possible to enhance interaction between **CALTURE** and human beings.

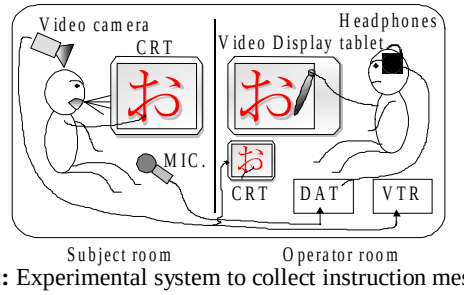

Figure 2: Experimental system to collect instruction messages.

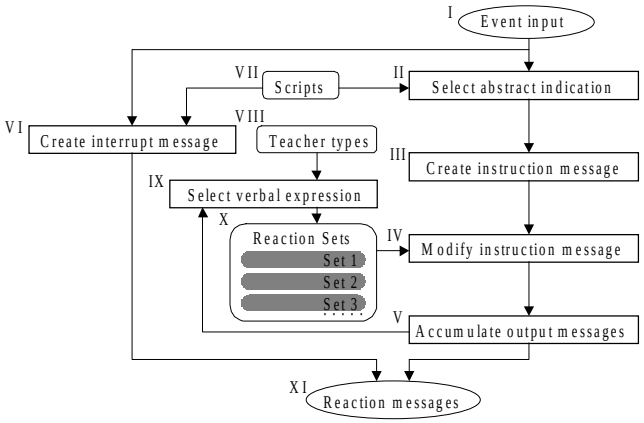

Figure 3: Block diagram of algorithm to generate instructions.

(4) Using block **VIII**, users can select the "personality" of the teacher such as strict, friendly, or short-tempered.

(5) Block **IX** generates verbal expressions that modify the instruction messages generated in step (3) to appropriately represent the "personality" of the teacher and to generate various expressions used in tracking the user's writing such as error repetitions and degree of error. These variations are performed by selecting appropriate massages or by changing prosodic parameters of the synthesized speech.

4. EXPERIMENTS TO LEAVE VARIOUS IMPRESSIONS

One of the most interesting features of **CALTURE** is the ability to change the "personality" of the teacher, or to generate various types of instruction messages. The function is important to sustain user interest. However, there is little evidence that supports the idea that synthesized speech can express these variations. This section describes experiments that confirm that users receive a particular impression from **CALTURE**'s outputs. Moreover, this section describes how different impressions are expressed using **CALTURE**.

4.1. Experimental procedures

The algorithm explained in Section 3.3 is implemented into the system. In block **X** of Fig. 3, three sets of instruction messages are prepared. The instruction messages (fourteen in total) of each set are the synthesized speech whose prosodic parameters are manually determined to express good mood, bad mood, and neutral mood, respectively (good mood [SOUND 0409_1.WAV], bad mood [SOUND 0409_2.WAV], neutral mood [SOUND 0409_3.WAV]). These three sets are tied to three internal states of **CALTURE**. By changing the internal states, **CALTURE** expresses various types of impressions. In the experiments, we controlled the parameters below.

(1) The number of states. Two vs. Three. In the case of two states, the neutral mood is omitted.

(2) The frequency of changing the states. None vs. Once every mistake vs. Once every three mistakes.

(3) The state change for mistakes. Normal vs. Opposite. Here, "Normal" and "Opposite" mean that after a user makes a mistake, a state changes from a worse or better mood state, respectively.

As shown in Table I, eleven systems in total are constructed. Nine subjects who are not familiar with synthesized speech used the systems for five minutes in their own way to experience various instruction messages of **CALTURE**. In the experiments, the areas that are accepted as correct tracing are clearly shown to the subject, so they can intentionally write an incorrect stroke.

4.2. Sensing of internal state number

For the 11 systems, the subjects are asked how many moods they can recognize from the instruction messages. Figure 4 shows the experimental results. First, it is clearly shown that the subjects discern different moods from the synthesized speech. Moreover, for the two internal states, subjects recognized the same number of moods as the number of internal states. On the other hand, subjects confused the three internal states with two states at about 50% of the time. According to the subject reports, a good mood is clearly distinguished from the other two states, but a neutral mood can easily be confused with a bad mood. The results suggest that it is difficult to express delicate mood differences using only prosodic parameters of synthesized speech. Adding particular expressions for each mood may solve this problem.

4.3. Preference for internal state changes

The subjects were asked their preference of the 11 systems in terms of like/dislike and interesting/boring. Figure 5 shows the experimental results for six systems out of the 11. Systems A and E that incorporate "opposite response" or "bad mood" are basically disliked by the subjects. Systems B and C are preferable, but are boring because they have only one state. Systems F and J, however, are not only the most preferable, but also are evaluated as interesting. The reason may be that they have several states. Judging from the results, changing system responses according to user actions is an important strategy in developing an interesting system. It is also shown that synthesized speech makes users recognize the state changes.

4.4. Impression of the instructions

To evaluate the general impression of the 11 systems, the

A	Number of states is 1	Always 'bad' mood	
B		Always 'neutral' mood	
C		Always 'good' mood	
D	Number of states is 2 'bad' mood 'good' mood	1 time	Normal
E			Opposite
F		3 times	Normal
G			Opposite
H	Number of states is 3 'bad' mood 'neutral' mood 'good' mood	1 time	Normal
I			Opposite
J		3 times	Normal
K			Opposite

Table I: Attributes of the 11 systems used in the experiment. The left column (A to K) shows identification letter of the system in this paper.

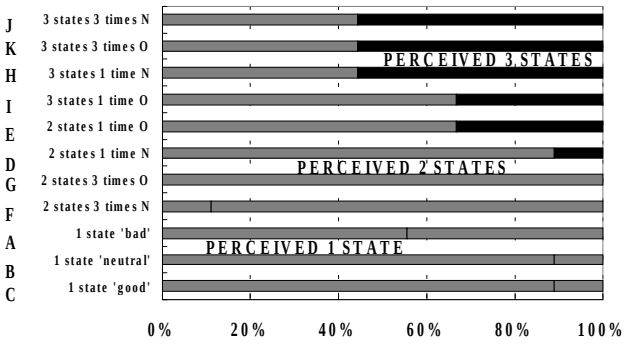

Figure 4: The number of states that subjects recognized.

subjects are asked to choose any terms from the following 14 words: patient, impatient, polite, crude, kind, awful, warmhearted, coldhearted, cheerful, gloomy, obedient, ill-natured, comical, and cool. To summarize the results, in Fig. 6, the terms are summarized into four categories: patient (patient, polite), impatient (impatient, crude), positive (kind, warmhearted, cheerful, obedient, comical) and negative (awful, coldhearted, gloomy, ill-natured, cool). Some of the results are shown in Fig. 6. The important results are summarized as follows.

1. Although Systems A, B, and C have only one state, Systems A and B can give strong impatient and patient impressions, respectively. The results indicate that the good and bad moods expressed by synthesized speech can convey a particular impression, even though the state does not change.
2. Among the systems that have multiple states, only System J is judged to be patient. The difference between Systems J and F is that System F does not have a neutral state. This implies that a neutral state is important to convey a patient impression.
3. Generally speaking, as shown in G and K, systems that output "opposite responses" leave vague impressions.

Through this, we confirmed that changing states imparts several types of impressions by synthesized speech. In all, System J is the best system, i.e., it leaves patient and positive impressions as shown in Fig. 6, and is judged to be preferable and interesting as shown in Fig. 5.

5. CONCLUSIONS

In this paper, we developed a CAI (Computer Assisted Instruction) system that teaches students how to write Japanese characters. The most important feature of the systems is the usage of synthesized speech to interact with users. Based on the analyzed results of behavior and instruction messages of a human teacher, an algorithm was proposed and implemented. One of the most interesting challenges of the system is a function to change the "personality" type of the teachers, such as a strict teacher, a friendly teacher, and a short-tempered teacher. To elucidate the validity of the function, experiments were carried out. Through the experiments, we confirmed that (1) subjects can discern different moods from the synthesized speech; (2) it is difficult to express delicate mood differences using only prosodic parameters of synthesized speech; (3) changing system responses according to user actions is an important strategy in developing an interesting system; and (4) changing states conveys several impressions by synthesized speech. Based on the above results, a CAI system can have several teacher modes. In the future, we will apply the strategies to other tasks and confirm performance.

6. REFERENCES

1. M. Abe, "Analysis of prosodic characteristics in speech advisories and their application to speech output", EUROSPEECH'95, pp. 2031-2034, 1995.
2. N. Hataoka, H. Kikuchi, "Topics on Multimodal Interfaces Which Use Speech Technologies.", Journal of IEICE Vol. 80 No. 10, pp. 1031-1035 (in Japanese), 1997.
3. T. Shinozaki and M. Abe, "Users impression for system reactions presented by synthesized speech – relationship between impression and prosodic parameters – ", Proc. ASJ Fall Meeting, pp. 229-230 (in Japanese), 1997.

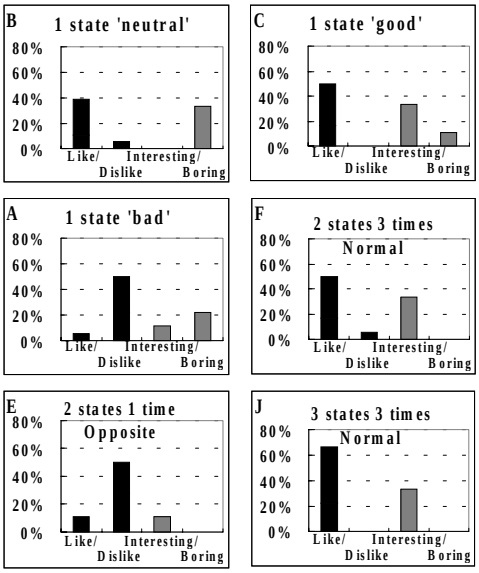

Figure 5: Preference scores of 6 systems.

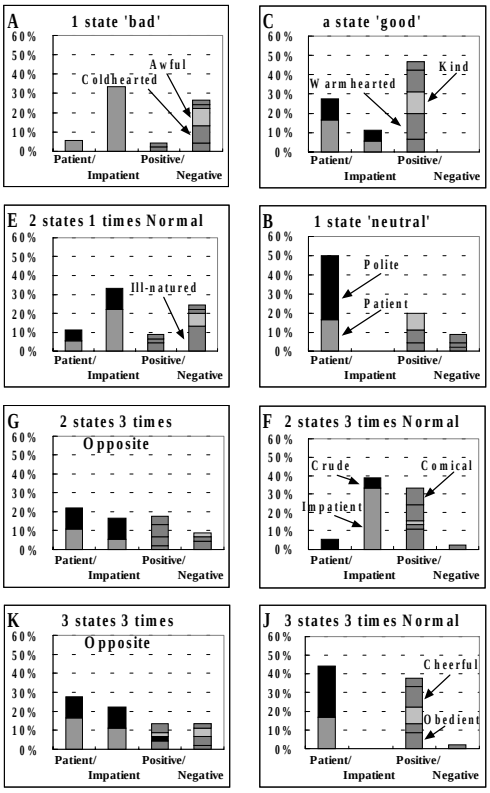

Figure 6: Impressions selected from 14 words.