VOICE-ACTIVATED HOME BANKING SYSTEM AND ITS FIELD TRIAL

Toshihiro Isobe, Masatoshi Morishima, Fuminori Yoshitani,
Nobuo Koizumi, Ken'ya Murakami

Laboratory for Information Technology
NTT DATA COMMUNICATIONS SYSTEMS CORPORATION

ABSTRACT

Speech recognition techniques are most useful when used over the phone. A telephone speech recognizer was developed and many field trials were carried out[1][2][3]. We have developed telephone speech recognition hardware for a voice-activated home banking system based on a client-server network configuration[4]. The speech recognition unit is a workstation with six boards for dealing with simultaneous multi-channel processing. The speech recognition algorithm implemented in the boards, each of which has three DSPs and an MPU, handles various tasks, such as recognizing connected digits, bank name, branch name, money amount, and confirmation for completing the service dialogs. Experimental field trials on 90 subjects showed that with proper instructions and guidance, the service task was successfully achieved in 85% of trials. We sent out a questionnaire, and one third of the subjects replied that speech recognition was useful.

1. HOME BANKING SYSTEM

The configuration of the client-server banking system is illustrated in Figure 1. A registered user need only call and talk to the system by telephone to transfer money to another bank account or to get balance information. The speech recognition unit, which can accept six calls at a time, is a workstation with six speech recognition boards. The workstation controlling speech dialogs, shown in Figure 2, is connected to the bank network by LAN and is operated by bank systems.

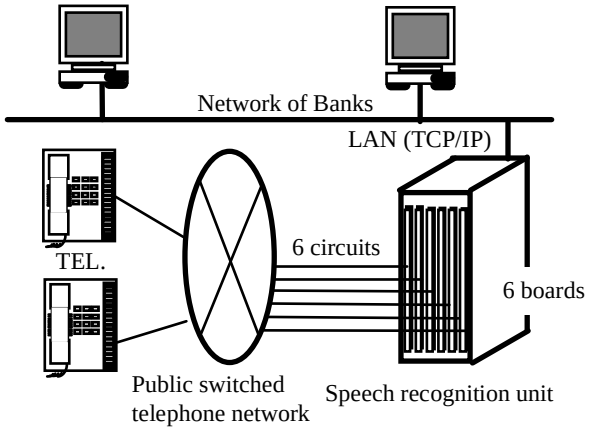

Figure 1: Outline of voce-activated banking system

"Hello! This is the telephone service center.
What kind of service do you want ?"<-"Money transfer."
"Did you say money transfer?" <-"Yes."
"Please say your account number." <-"1234567"
"Did you say 1234567 ?" <-"Yes."
*"Please say your code number." <-"****"*
"You are accepted."
"To which bank do you want to transfer ?" <-"Fuji bank"
"Did you say Fuji bank ?" <-"Yes."
"To which branch do you want to transfer ?"<-"Kyoto"
"Did you say Kyoto ?" <-"Yes."
"Please say the account number to which
you want to transfer." <-"9876543"
"Did you say 9876545 ?" <-"No."
"Please say the account number to which
you want to transfer." <-"9876543"
"Did you say 9876543 ?" <-"Yes."
"Please say the amount of money
you want to transfer." <-" 13, 000 yen."
"Did you say 13,000 yen ?" <-"Yes."
"I will send the information on your transfer
to your FAX. Please check it. Thank you."

Figure 2: Speech dialog

2. SPEECH RECOGNITION BOARD

The hardware configuration of the board is shown in Figure 3. Each board is connected to one telephone circuit and has three DSPs (TMS320C31) and one MPU (MC68040). The first DSP (DSP1) is used for telephone network control, touch tone detection, A/D conversion, feature extraction, and for calculating the likelihood of basic Gaussian distributions for the code books of tied-mixture HMMs. The second DSP calculates the output probabilities of HMM states. The last DSP computes Viterbi scores. The MPU extracts recognized words by tracking back Viterbi scores and controls all of the DSPs.

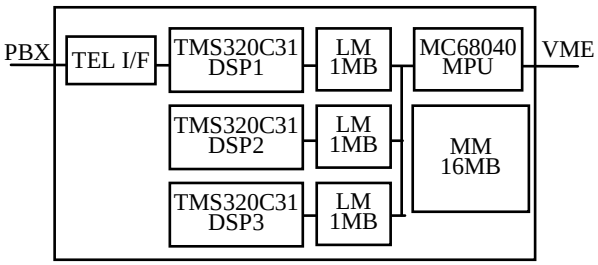

Figure 3: Speech recognition board

3. RECOGNITION ALGORITHM

We use context-dependent tied-mixture phone HMMs that provide three phone states and 501 states for all phonemes. A phone HMM consists of one state for a left context-dependent HMM, another state for a context-independent HMM, and a third state for a right context-dependent HMM (see Figure 4).

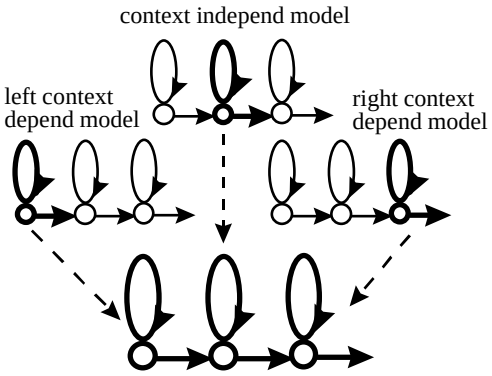

Figure 4: Phone model

In DSP1, in order to reduce the amount of calculation for code book probabilities, we have two layers of code books (Figure 5). One is a normal code book (layer 2) having 1,800 Gaussian distributions that are basic distributions of tied-mixture HMMs; the other is a small code book (layer 1) that has 64 Gaussian distributions estimated from the distributions in layer 2 by the k-mean VQ method in the Baum-Weltch algorithm. When the system receives speech, DSP1 first computes the probabilities of the 64 distributions in layer 1 and selects the one with the highest score. It then calculates the probabilities of the 500 distributions in layer 2 that are the nearest to the one selected in layer 1. The scores of the distributions not calculated in layer 2 are set to the closest distributions from layer 1.

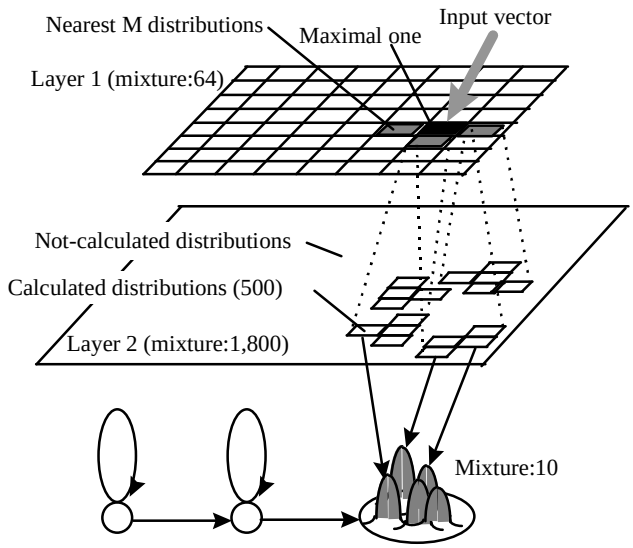

Figure 5: Tree structure of mixture

DSP2 computes the output probabilities of the HMM states by multiplying the likelihood scores from the DSP1 code book by mixture weight. DSP3, which has a finite state automaton network, computes the Viterbi score, using the beam-search algorithm.

4. SPEECH DATABASE FOR TRAINING SPEECH MODELS

To make the speech models, we collected telephone speech data from 400 males and 400 females living in seven major cities in Japan, taking into account the balance age group and region dialect[5]. The database consists of 8,000 names of banks and credit associations, 8,000 names of branches, 4,800 phrases of four connected numbers, 4,000 phrases of seven connected numbers, 4,800 phrases that represent amounts of money, and 800 sets of six words needed in banking services. We trained the speech models with half of the database, using maximum likelihood estimation, and tested the system using the other half.

5. EXPERIMENT

5.1. Benchmark test

Using the telephone speech database we measured the performance of the speech recognition board. Recognition tasks included seven connected digits, bank name, and amount of money. The results are shown in Table 1.

Task	V. size	perplexity	Rec. rate [%]
7 connected digits	12	12.0	92.4
Bank name	600	4.94	90.8
Amount of money	46	24.3	77.4

Table 1: Result of bench mark test (V. size: Vocabulary size)

5.2. System field trial

We tested the voice-activated home banking system using the public switched telephone network, with human-machine dialogs collected from 90 subjects. We asked the subjects to play-act making a bank transfer by telephone. Trials were done twice for each test. Between the first and second attempts, we presented to each subject an explanation of how to speak to the system and a sample dialog from a skilled user (Figure 6).

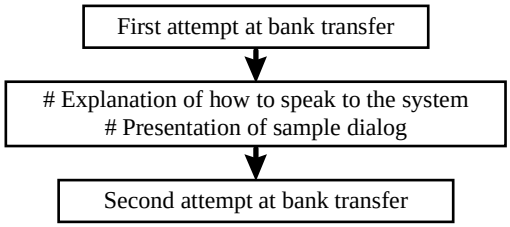

Figure 6: Field trial sequence for each subject

The results of recognition tasks for the dialogs are shown in Figures 7 through 9. In these figures, the x-axis represents the number of repeat utterances, and the y-axis represents the cumulative accuracy rate.

People get used to the system during repeat attempts, which leads to an increase in accuracy as users repeat and to convergence when the number of repetitions is equal to three, in all tasks. The explanation and the skilled user's dialogue are efficient for adapting subjects to the system, so accuracy rates for second attempts are higher than those for first attempts in all figures.

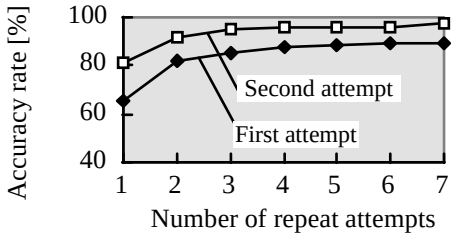

Figure 7: Accuracy rate vs number of repeat attempts in recognizing seven connected digits

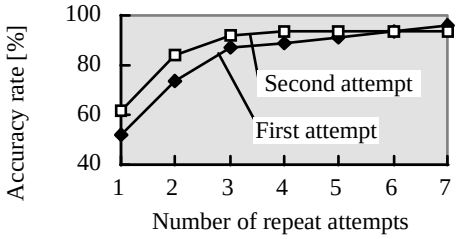

Figure 8: Accuracy rate vs number of repeat attempts in recognizing bank name

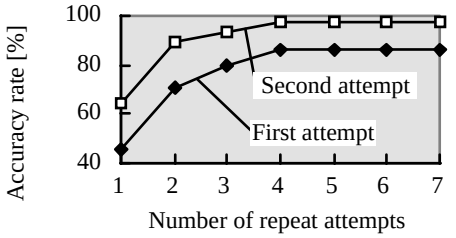

Figure 9: Accuracy rate vs number of repeat attempts in recognizing money amount

Figure 10 presents the success rate of the service. In the figure, the x-axis represents the maximum number of repetitions allowed by the system.

When users had three chances to repeat, about 85% of them completed the bank transfer service successfully. But three repeats is the maximum that users can endure.

The accuracy rate for "yes"/"no" recognition, which is the most important in controlling speech dialogs, was 89.3% in the first trial and 92.3% in the second.

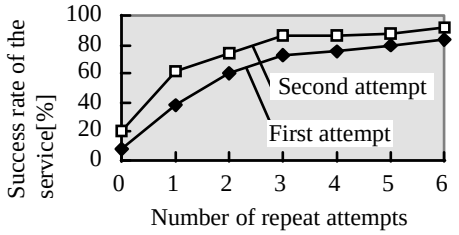

Figure 10: Success rate of money transfer service vs number of repeat attempts in each recognition task

6. QUESTIONNAIRE

We investigated how users felt about the SR (Speech Recognition) system by putting questions to the subjects after the field trial. The questionnaire was multiple choice.

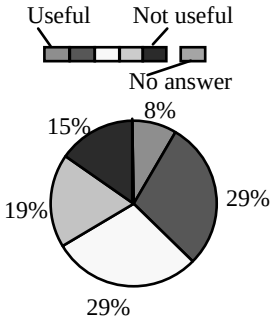

Figure 11: Is SR useful ?

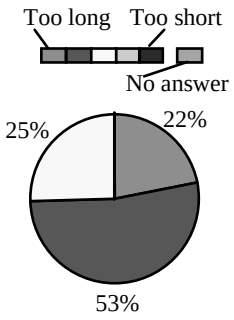

Figure 12: How about the duration of SR service ?

Figure 11 shows the response of users as to the usefulness of the SR system. About one third of them thought it was useful, and the same number of people thought it was not useful. Apart from the recognition performance, the reason was the long service duration that results from having "yes"/"no" confirmation after every item, such as bank name, account number and so on, for security in dealing with money (Figure 12).

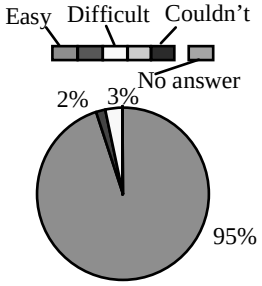

Figure 13: Is it easy to think of what to say when you are asked your account number ?

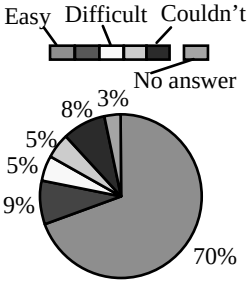

Figure 14: Is it easy to think of what to say when you are asked amount of money ?

Figures13 and 14 show that users were more at a loss for what to

say when giving responses which can be said in various ways, such as the amount of money, than they were for those responses with fewer possible variations, such as the account number. This is the reason for the low recognition rate for the money amount at less than three attempts (see Figure 9).

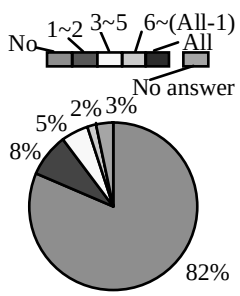

Figure 15: How many times did you say words not in the vocabulary ?

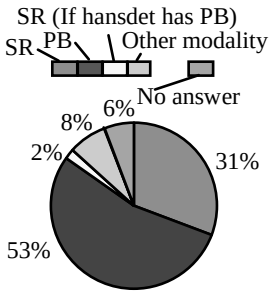

Figure 16: Do you prefer SR or PB for inputting account number and money amount ?

Figure 15 shows the percentages of people who uttered words not in the specified vocabulary.

Figure 16 shows which method users prefer, SR or PB (touch tone detection), when inputting connected digits and money amount. As a result of the popularization of PB and the secrecy of money transfer services, 51% of people chose PB.

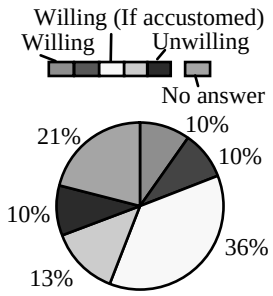

Figure 17: Are you willing to use SR as an actual system ?

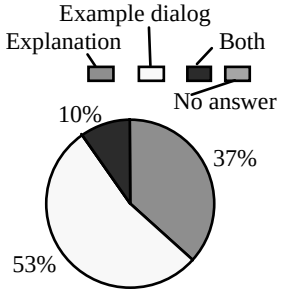

Figure 18: Which is better for support explanation or example dialog ?

About 60% of subjects answered that they would be willing to use SR for actual services on the condition that they were accustomed to the system (Figure 17). Sample dialogues made by trained speakers were preferred to explanations of how to speak to the system (Figure 18). These two figures show that getting users used to the system is an important factor for putting SR to practical use.

7. CONCLUSIONS

We have reported on our voice-activated home banking system, speech recognition board, algorithm, field trial, and how users feel about the system. In the last few years, SR techniques have reached the level of practical usefulness, but not of spontaneous speech recognition. So it is important to give users knowledge of, and help them get accustomed to, the system. Giving examples of skilled users' dialogues is highly effective for helping newly-registered customers. Both in the experiment using a telephone speech database and in the field trial, the accuracy of recognizing the money amount was less than that for other tasks. To maintain accuracy in an actual banking service requiring a high level of security, recognition of a spoken money amount can be replaced by another convenient method such as detection of a tone. SR is available for recognizing the bank name or branch name when users don't remember the code number easily.

The popularization of personal computers has enabled customers to use various modes of service like home banking. We are going to continue system trials and establish the most effective way to use speech recognition techniques with a human-machine interface.

8. REFERENCES

1. G. Ortel, "Observed long-term changes in customer calling in a telephone application using automatic speech recognition", Proc. EUROSPEECH *95, pp273-276, Sep. 1995.
2. M. Lennig and G. Bielby, "Directory assistance automation in Bell Canada: Trial results," Proc. Workshop on IVTTA, pp9-13, Sep. 1994.
3. S. Yamamoto, K. Takeda, N. Inoue, S. Kuroiwa and M. Naitoh, "A voice-activated telephone exchange system and its field trial," Proc. Workshop on IVTTA, pp21-26, Sep. 1994.
4. T. Isobe, M. Morishima, F. Yoshitani, N. Koizumi and K. Murakami, "Voice-activated home banking system," Proc. Workshop on Automatic speech recognition, pp163-164, Dec. 1995.
5. T. Isobe and K. Murakami, "Telephone speech data corpus and performances of speaker independent recognition system using the corpus," Proc. Workshop on IVTTA, pp101-104, Sep. 1994.

電話音声認識装置

Sound File References:

[a265s1.wav]