AN EXPERIMENTAL BIDIRECTIONAL JAPANESE/ENGLISH INTERPRETING VIDEO PHONE SYSTEM USING INTERNET

Shoji Hiraoka[1], Masakatsu Hoshimi[1], Kenji Matsui[2], and Jean-Claude Junqua[3]

[1]Matsushita Research Institute Tokyo, Inc., 3-10-1 Higashimita Tama-ku, Kawasaki, 214 Japan.

[2]Central Research Laboratories, Matsushita Electric Industrial Co. Ltd., Kyoto, Japan.

[3]Speech Technology Laboratory, Panasonic Technologies Inc., Santa Barbara, California, U.S.A.

ABSTRACT

In this paper we report on an experimental bidirectional Japanese/English interpreting video phone system using Internet. We particularly emphasize the motivation for this work, the task, and the experiments conducted. Using in house technology developed both in Japan and in the United States, we demonstrated an Internet home shopping application where an American shop assistant and a Japanese customer engaged in task-directed dialogues, using their native languages. The experiments showed that when users are familiar with the application language, a natural interaction can be obtained.

1. INTRODUCTION

These days, the way we do business is evolving very fast. There is continuous pressure to increase productivity as well as quality and customer service. Growing expectations for reponsiveness and personalized services is changing the culture and operation of many industries. There are also new business models which emerge from the way in which organizations and people work together and more generally communicate with each other using new technologies. These include in particular, teleworking, collaborative product development, and virtual meetings. Recently, the use of Internet contributed to the expansion of communication and enabled the development of new services. Internet is becoming progressively one of the main ways to deliver information on demand.

As the use of Internet is growing, the number of people who communicate and speak different languages is increasing rapidly. Recently, new services such as hotel and airline reservations, travel information, etc. are becoming available all over the world through the Internet. Already these services are commonly accessed from home. When these services will become widely available, telephone will constitute an important way of accessing the information provided.

New Internet services are often attractive. However, it can be cumbersome for a common (non-expert) person to use these services because of the language barrier or other difficulties related to user interface complexity.

Given this vision of the future, Matsushita decided to develop an experimental system which aimed at translating one language into another in a specific domain utilizing natural conversations. The following sections of this paper briefly mention some related work, present an overview of the selected task and report on the experiments done. At the end of the paper, we discuss what we learned from these experiments and provide some of our own perspectives.

2. BACKGROUND

Related published work in speech translation and multilingual communication includes systems and research prototypes developed with close co-operation among ATR [1], AT&T [2], Carnegie Melon University and the University of Karlsruhe [3] and other systems at NEC [4], and Siemens AG.

Recently, in the context of the Vermobil project (e.g. [5]) a speech to speech translation system for two persons who want to find a date for a business meeting has been developed. A special characteristic

of this work is that it assumes that both participants have a least a passive knowledge of English which is used as intermediate language.

In 1989, Matsushita developed a speech translation system which was aimed at non-expert users in the domain of travel information. However, this system was a one-way translation system (Japanese to English). In contrast, the work described in this paper is an example of a bidirectional Interpreting Video Phone (IVP) system which has been designed to allow a user to shop from home via Internet where the stores can be located anywhere in the world. The idea is that a customer using his own language, a relatively free interaction, the Internet, and a machine translation system, can have a natural interaction with a shop assistant who might be speaking a different language and be located half way across the world.

3. THE TASK

A Japanese speaking customer browsing through a mock electronic catalog initiated and engaged in a real-time dialogue with an English speaking sales person at the U.S. site. Both the customer and the salesperson could see and hear each other on their PC screens. Their dialogues included continuously spoken sentences from a finite list of sentences that comprised a shopping task. A synopsis of the task is presented in Figure 1.

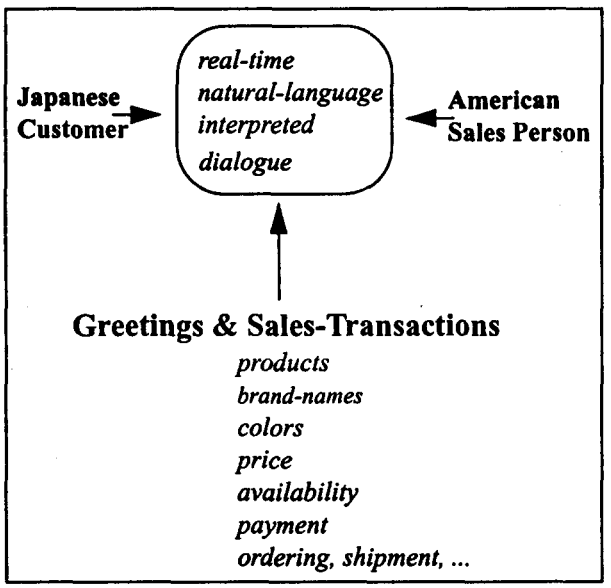

FIGURE 1. The Task.

Figure 2 contains a short sample skit. Lines labeled by [E] indicate sentences spoken in English by the salesperson. Lines labeled by [J] are the customer's sentences uttered in Japanese and then translated to English.

[E] Hello. Welcome to Panasonic luggage shop.

[E] How may I help you?

[J] *May I see a sports bag?*

[E] What size are you looking for?

[J] *I would like to see the largest one that you have.*

[... the salesman brings a sports bag and shows it through the camera ...]

[E] This is popular these days.

[E] How do you like it?

[J] *Could you please turn the bag around?*

[... the salesman turns the bag around in front of the camera ...]

[J] *Thank you. What should I do to purchase it?*

[E] We'll e-mail the order form in Japanese right away.

[J] *How long will it take for the bag to get here?*

[E] We will ship it by air-mail.

[E] You should have it by the beginning of next week.

[J] *Thank you very much.*

[E] Thank you for choosing Panasonic.

FIGURE 2. A Sample Dialogue Skit.

The salesperson first hears a Japanese sentence [not shown in Figure 2] through the video phone, and then hears the synthesized translation. The salesperson responds in English, sees the recognition result on his/her PC screen and then confirms by hitting a key. The transmitted sentence is then heard in the Japanese synthesizer's voice in Japan. The salesperson also hears the Japanese synthesizer voice through the video phone and therefore knows the customer has heard the response. Table 1 summarizes the different steps involved during one dialogue turn.

The dialogues are constrained by a fixed list of natural conversational sentences which cover most

aspects of a sales-transaction, covering questions and answers about products, brand names, colors, availability, price, payment, ordering, shipments, greetings, good-byes, etc., We were able to design natural sentences covering this limited domain using fewer than 500 words for both Japanese and English.

1	*Voice of the customer (received also at the shop assistant site)*
2	*Confirmation of the recognition results*
3	*Translated synthesis voice received at the shop assistant site*
4	*Voice of the shop assistant*
5	*Confirmation of the recognition results*
6	*Translated synthesis voice received at customer site*

TABLE 1. Steps involved in a dialogue turn.

4. OVERVIEW OF THE TECHNOLOGY

The key software components of the IVP system are: English and Japanese continuous speech recognizers, English and Japanese text-to-speech synthesizers, English-to Japanese and Japanese-to-English language translators, and network communication software. Two separate communication media are used: ISDN phone lines for the audio/visual link; Internet for socket based communication between local recognition and remote translation/synthesis processes. The dialogue scenarios used during the demonstrations contained up to 1162 English and 3000 Japanese sentences which covered a customer/salesperson interaction task for an electronic on-line catalog shop for luggage and tote bags.

Table 2 and Table 3 summarize the characteristics of the speech recognizers and synthesizers used in the experiments. These software components were developed in-house at the authors' laboratories and are proprietary technologies.

To give an idea of the performance of the recognition technology, the English recognizer obtained a 97% word accuracy on a laboratory recognition experiment using 200 English sentences similar to that used in the experiments.

More details about the technology used in the experiments can be found in [7].

The translation of text from Japanese to English, and from English to Japanese was carried out at the Japanese site using an example based translation system [8]. With this method, an input sentence is translated by pattern matching similar examples that are recorded in the translation knowledge database. The translation knowledge database consists of three main components: (1) word dictionary, (2) sentence pattern dictionary, (3) actual translation examples.

English ASR	Japanese ASR
Speaker-independent	*Speaker-independent*
Continuous speech	*Continuous speech*
Dictionary: 438 words	*Dictionary: 500 words*
Automatic speech detection	*Push-to-talk*
Phoneme-based continuous density HMM	*Model Speech Recognition Method (see [6])*
BNF-style grammar	*word-pair like grammar*

TABLE 2. Main characteristics of the English and Japanese recognizers.

Hybrid system (combination of formant synthesis and wave concatenation)
Male/Female switch
Customized lexicon
Software only version

TABLE 3. Common characteristics of the text-to-speech synthesizers.

With our example-based machine translation system, dialogue sentences that are not easily expressible by grammar rules can simply be registered as actual examples. The ability to easily incorporate changes to the task-dialogue in this way is very valuable for experimental tasks such as ours.

5. EXPERIMENTS AND DISCUSSION

The IVP electronic shopping prototype has been shown to live audiences in Japan and in the U.S. at three exhibits lasting a total of nine days. A total of 10

male and female speakers participated to the experiments. The demonstrations have been very valuable in demonstrating the feasibility of an interpreting video phone system using currently available recognition, synthesis, translation, and communication technologies, and in promoting interest in using speech technologies. The experimental nature of the demonstrations contributed to its success.

The real challenge of building the IVP system has been integrating all the independent technologies for recognition, synthesis, translation, and communication to facilitate a real-time and natural interpreted dialogue while using reasonable and standard hardware resources. To a great extent our goals have been met. Table 4 summarizes the average subjective delay 1) observed during the experiments and 2) measured on the different components used. This table shows that the average delay between the time a sentence is spoken and heard at the other end is between 2 and 3 seconds. Even with this delay, our experience shows that the communication is still acceptable. The customer and the shop assistant could carry a natural conversation. This is partially due to the fact that the two people communicating can see each other while they are talking and can hear the other language through the video phone. Furthermore, the fact that the users know that a machine takes care of the translation task also makes this delay acceptable. Sometime we experienced 1 to 10 minute delays due to Internet which required us to pause the on-going skits until transmission has cleared. Otherwise, the reliability of the Internet was high, and frequency of restarts due to network or software crashes was low.

Internet	*Typical=0.8-1.2 secs.* *Occasional 1-10 mins.*
ASR	*Typical=0.3-0.8secs.*
TTS, Translation	*Typical=0.1-0.3 secs.*
Confirmation step	*Typical 0.3-1.0 secs.*

TABLE 4. Average delay for the different components.

In our experiments, we also found that it is important to hear the other party's voice as well as the transmitted text string to be synthesized. Even if there is a delay between them, the user needs to know what is happening on the other side.

We should keep in mind that in these experiments the speakers were familiar with the language used in the dialogues. Future enhancements of the system should include the use of spontaneous speech with a larger vocabulary size (in the order of 1000 words) and more flexible language modelling and translation.

6. CONCLUSIONS AND PERSPECTIVES

An experimental bidirectional Japanese/English Interpreting Video Phone system was developed and demonstrated to live audiences in an Internet home shopping simulation task. These experiments showed that using current state-of-the-art technology, it is possible to have a natural dialogue between experienced users.

For developing a system of this kind, it is necessary to involve different laboratories in different countries. In the near future we want to extend these collaborations and develop a practical system which can incorporate other languages in addition to Japanese and English.

ACKNOWLEDGEMENTS

We would like to acknowledge the continuous support received from Brian Hanson, Katsuyuki Niyada and the team members in the different laboratories whose efforts made of this project a reality.

REFERENCES

[1] Morimoto, T., et al., "*ATR's Speech Translation System: ASURA*", Proc. EUROSPEECH, pp.1291-1294, 1993

[2] Roe, D.B., et al., "*Efficient Grammar Processing for a Spoken Language Translation System*", Proc. ICASSP, Vol.1, pp.213, 1992.

[3] Suhm, B., et al., "*JANUS: Towards Multi-Lingual Spoken Language Translation*", Proc. ARPA Workshop on Spoken Language Technology, Austin, TX, V.1,pp.221-226, 1995.

[4] Hatazaki, K. et al., "*INTERTALKER: An Experimental Automatic Interpretation System Using Conceptual Representation.*" Proc. ICSLP, pp.393-396, 1992.

[5] Wahlster, W. "*Vermobil - Translation of Face-to-Face Dialogues*", Proc. Fourth Machine Translation Summit, Kobe, Japan, 1993.

[6] Miyata, M. et al., "*Speaker-Independent Speech Recognition Using Sub-Word Units of Model Speech Uttered by a Small Number of Speakers*", IEICE Technical Report SP91-83, 1991.

[7] Karaorman, M. et al., "*An Experimental Japanese/ English Interpreting Video Phone System*", Proc. ICSLP, pp.1676-1679, 1996.

[8] Sato, S. "*Example-Based Translation*", Journal of Information Processing Society of Japan, Vol.33, No.6, pp. 673-681, 1992.