AN EXPERIMENTAL JAPANESE / ENGLISH INTERPRETING VIDEO PHONE SYSTEM

Murat Karaorman, Ted H. Applebaum, Tatsuro Itoh[†] , Mitsuru Endo[†], Yoshio Ohno[†], Masakatsu Hoshimi[†], Takahiro Kamai[‡], Kenji Matsui[‡], Kazue Hata, Steve Pearson, Jean-Claude Junqua

Speech Technology Laboratory, Panasonic Technologies, Inc.,
3888 State Street, Santa Barbara, California 93105 USA

[†] Matsushita Research Institute Tokyo, Inc.
[‡] Central Research Laboratories, Matsushita Electric Industrial Co. Ltd.

ABSTRACT

In this paper we report on the architectural design issues and experiences gained while building and demonstrating an experimental interpreting video phone (IVP) system. The IVP system has been demonstrated in an internet home shopping simulation simultaneously before live audiences in Japan and the U.S. An American shop assistant and a Japanese customer engaged in task-directed dialogues, using their native languages. In addition to their direct audio/visual contact by ISDN video phone, each participant heard a translation of the remote speaker's utterances in a synthetic voice in real-time.

Each site used a medium-size vocabulary, a continuous speech recognition system and a text-to-speech synthesis (TTS) system for the local language. Recognition results were transmitted over the internet to the remote site, where the corresponding translated sentence was spoken by TTS in the listener's native language. All of the speech and language processing software components of the system were independently developed proprietary technologies of the authors' laboratories which were integrated using commercially available hardware and communication media. Difficulties encountered in developing the system, the accommodations which were made, and other experiences gained through the process are reported in this paper.

1. INTRODUCTION

An interpreting video phone system (IVP) has been built by integrating the speech and language processing technologies developed at the authors' research laboratories. The final system has recently been shown at three different exhibitions to large audiences for a combined duration of 72 hours (nine eight-hour days.) The demonstrations involved a live communication link between a show site in Japan and the U.S. site at Speech Technology Laboratory at Santa Barbara, California. During these demonstrations, a Japanese speaking customer, at the site in Japan, engaged in a real-time dialogue with an English speaking sales person in the U.S. The dialogues used during the demonstrations include continuously spoken sentences from a finite list of sentences that comprises a shopping task. Each site used a medium-size vocabulary, speaker-independent continuous speech recognition system, a high quality text-to-speech synthesis system, and an example-based language translation system. Two independent communication media connected the two sites: an ISDN phone line for the audio/visual link; Internet for exchange of recognition text and task initiation protocol (shown in **Figure 1**.)

Related published work in speech translation and multilingual communication includes systems and research prototypes developed with close co-operation among ATR [9], AT&T [10], Carnegie Melon University and the University of Karlsruhe [11] and other systems at NEC [12], and Siemens AG. While it is difficult to make a direct comparison between IVP and these systems, we share similar views on the difficulty of recognition and translation involving natural spoken language, and the need for robust, intelligent translation systems.

We first give a detailed description of the system's components, and then relate how various difficulties were resolved in the design and integration of the IVP system, and in the design of the task and dialogues. We also relate experiences gained while using the system during the live demonstrations.

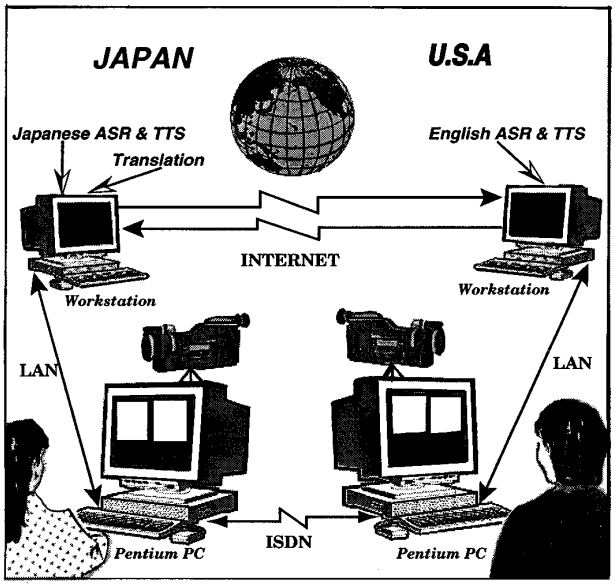

Figure 1 IVP System Architecture.

2. SYSTEM OVERVIEW

The key *software* components of the IVP interpreting video phone 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. The key *hardware* components are: 100 Mhz Pentium PCs equipped with a commercial video-conferencing board (VIVO) and accompanying software; Sun SS-20 Unix workstations; microphones; speakers; video-camera. 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. The English task-dialogue used a dictionary of 438 words while the Japanese task-dialogue used a 500 word dictionary.

All of the software components have been developed in-house at the authors' laboratories and are proprietary technologies. All of the hardware and communication media components are standard and commercially available systems.

3. AUTOMATIC SPEECH RECOGNITION

The English and Japanese automatic speech recognizers (ASR) have been developed independently from the IVP task and use entirely different technologies. Certain adaptations made to the ASRs for task integration are explained in Section 8.

3.1. English ASR

The English ASR system is a speaker-independent, medium vocabulary, continuous speech recognition system with perceptually real-time operation [1]. It is implemented as two concurrent processes on Unix workstations: a time-synchronous front-end acoustic analysis process and a recognition search process. The front-end process detects the beginning and end of the speech utterance, performs spectral analysis, and does acoustic normalization to reduce the effect of irrelevant sources of variation (such as changes of background noise, microphone or individual voice characteristics) [3]. The second process, which is also time-synchronous, searches forward through the acoustic features received from the front-end, guided by a language model which consists of a pronunciation dictionary, word duration model and context-free task-grammars. Speech detection runs continuously, so that it is not necessary to "push-to-talk".

The acoustic model is based on Hidden Markov Models (HMM) of phone units using continuous Gaussian mixture densities. The phone units are as defined in the TIMIT database. Each phone is modeled by a three state hidden Markov model; each state is modeled by a Gaussian mixture density, requiring a total of 1071 component densities.

The language model consists of a dictionary, word duration data and task-grammars. The dictionary used in the demonstration contained 438 words and short phrases, with a total of 582 transcriptions (pronunciations). Task grammars may be entered in Backus-Naur form and may be switched at run-time. Three task grammars were prepared for the demonstration, generating languages of 78, 239 and 1162 sentences respectively. The recognition result is typically obtained within 0.3 seconds after the actual end of utterance for the mentioned task-grammars.

3.2. Japanese ASR

The Japanese ASR system is also a speaker-independent, medium vocabulary, continuous speech recognition system with perceptually real-time operation. It is organized as a two-pass system. The first pass builds a lattice for every word at every frame. Upon end of utterance the second pass searches backward for the best path through the lattice, via the A* algorithm [8]. The solution is obtained immediately after end-point detection, as the second pass uses the matching paths calculated by the first pass. The search grammar is based on word-pair co-occurrence frequencies at all distances.

The acoustic model is based on Model Speech Recognition Method using the similarity vectors as feature parameters [6]. In this method, sub-word templates trained with a small number of speakers yield high recognition rates in speaker independent recognition.

The language model is statistical, based on the word co-occurrence model [7]. The word co-occurrence model learns the task-language by counting all word-pairs, including those that are far apart within a sentence, and storing the distance between the words in each word-pair. Changes to the task-language can easily be incorporated into the language model by changing the sentences data and re-training.

4. TRANSLATION

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 [4]. 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.

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. SYNTHESIS

Two different and independently developed text-to-speech (TTS) synthesizers are used, one for Japanese [5] and another for English [2]. Both TTS systems are highly intelligible, hybrid systems combining formant-synthesis method and wave-concatanation methods. The TTS synthesizers can be dynamically switched between a male and female voice.

Both TTS synthesizers run in real-time on PCs without additional hardware support and provide a user-lexicon capability.

6. COMMUNICATION

Two separate communication channels are used: ISDN phone lines for the audio/visual link; Internet for socket communication between the recognizer and the remote translator/synthesizer process.

PCs equipped with a commercial video-conferencing system, VIVO, are connected via ISDN (INS Net-64) phone link to provide audio/visual contact. On both sites' PC screens two windows display both the local and the remote camera's view. The picture and audio quality of the system is satisfactory but not very high due to the limited bandwidth offered by the ISDN phone link.

Textual translation data is transmitted between the two sites using Internet and IP-stream sockets. Two independent one-way stream connections are used. The English ASR system transmits English ASCII-text to the English-to-Japanese machine translator in Japan using one of the stream connections. The Japanese ASR system directly invokes the Japanese-to-English translator at the Japanese site and uses the other connection to transmit the translated English ASCII-text to the TTS synthesizer at the U.S. site. A daemon-based protocol is developed for system start-up and possible restarts in order to recover from a crash. In this arrangement each site has an independent outbound IP-stream to send data, and therefore both of the recognizers can operate completely asynchronously. This design also increases robustness by allowing each site to be able to restart only their side's ASR in case of a communication or software failure.

7. DIALOGUE TASK

We designed an electronic-catalogue shopping task to demonstrate the IVP system. During these demonstrations, at a show site in Japan, 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. **Table 1** 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. The salesperson first hears a Japanese sentence [not shown in the table] 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.

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.

[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 plese 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 you 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.

Table 1 A Sample Dialogue Skit

8. EXPERIENCES

In this section we report about our experiences gained from: (1) designing and building the interpreting video phone system, (2) conducting the live demonstrations.

8.1. Design of IVP

Naturalness of the human-to-human dialogue and translation system during the live interaction between the Japanese and English speakers has been a key design objective. Naturalness as an objective has several specification implications. All software components need to be real-time and be able to operate in a medium noise environment. We found it unnatural and error-prone to always use a switch in order to talk to the microphone. Therefore, a robust and adaptive voice-activated speech detection mechanism is developed as part of the front-end of the English ASR.

To meet tight real-time constraints has been a key design objective to ensure smooth and natural dialogue. Based on our initial experiments we have determined 2–3 seconds as an acceptable delay on the time taken from the instance a listener hears the end of the speaker's actual utterance through the video phone until the listener begins hearing the translated sentence locally from the TTS synthesizer. This time constraint includes the transmission time over Internet and other protocol delays. Our timing and reliability experiments with network transmissions resulted in our choosing the more reliable, connection-based stream sockets which gave us typical packet delivery times in the 0.8 to 1.2 seconds range (excluding occasional long delays.) Since we did not have immediate control over the Internet and network delays, we focused on optimizing and streamlining the ASR, translation and TTS components.

To achieve real-time response, the English ASR system was re-

designed to work frame synchronously with its front-end. Recognition process starts as soon as the first speech frame is detected and runs concurrently with the front-end. In most cases recognition result is immediately available as soon as the last frame designating end-of-speech gets delivered by the front-end process.

In order to keep task recognition accuracy high, while using a voice activated front-end, we added a confirmation step which involves hitting a key in order to transmit the recognized sentence. In case of misrecognition the speaker repeats the same sentence until the sentence is successfully recognized. While this step adds a non-trivial time delay, since the speaker needs to see and verify the recognition result by hitting a key, it has helped the speaker by increasing confidence and eliminating the need to correct recognition errors due to incomplete or out-of-grammar sentences, long pauses, voice-activation errors, etc., We were able to still meet the 2–3 second real-time constraint by improving the combined ASR, translation and TTS processing time to under 1 second.

In order to keep English ASR recognition accuracy high we used two strategies: (1) add additional alternative pronunciations for critical words in the dictionary, (2) combine short words into subphrases and introduce the subphrases as new words to the dictionary. While both of these approaches add additional complexity to the task-grammar and add to recognition time, for our application's task size the speed loss was unnoticeable.

8.2. Experiments and Demonstrations

The IVP electronic shopping demonstrations have been shown to live audiences in Japan and U.S. at three exhibits lasting a total of nine days. We also had many days of experiments at various stages of the IVP system's development for testing and fine-tuning. A total of ten male and female speakers participated in the experiments and final demonstrations. At the experimentation stage we continually enhanced the dialogues and fine tuned system settings. The confirmation-step was added after realizing that it helped for a smoother dialogue by eliminating the need to send "correction" messages in case of misrecognitions. The likelihood of misrecognition is higher in a dialogue setting than in an isolated sentence recognition task because the speaker may get distracted and say sentences that are outside of the task-grammar. Incomplete sentences, long pauses between words, repetitions, false starts, restarts, or omissions are typical in human dialogues and still easily understandable by humans, however, extremely difficult to model using machines. Additionally, the voice detection is harder and recognition accuracy is negatively affected in the presence of showroom noise. Adding the confirmation step gave the speaker greater control and confidence over the system's accuracy and reduced stress.

During the shows we had 8 hours of continuous demonstrations each day, and experienced recognition-to-synthesis times of 2–3 seconds quite consistently, however, about twice a day we would experience 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, at less than once per day.

9. CONCLUSION

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.

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. We were able to use a closed task dictionary and grammar, and use high-end PC and workstation equipment. The speakers were familiar with language used in the dialogues. Future enhancements for the IVP should include futher research on more flexible language modelling for recognition and translation of spontaneous speech with open vocabulary and grammar rules and more advanced speaker and microphone adaptation techniques.

We would like to acknowledge the continuous support received from Brian Hanson, Shoji Hiraoka, Antoine Lefloch, Philippe Morin, and Michael Galler throughout this project.

REFERENCES

1. Zhao, Y. "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units," *IEEE Trans.SAP,* pp. 345-361, 1993.
2. Moran, H., K. Hata, and S. Pearson, "The Use of Sampled Consonants for Improved Intelligibility in Formant Synthesizers," *JASA,* Pt. 2, 2816, 1994.
3. Zhao, Y., "An Acoustic-Phonetic Based Speaker Adaptation Technique for Improving Speaker-Independent Continuous Speech Recognition." *IEEE Trans. SAP* 2:380–394, 1994.
4. Sato, S. "Example-Based Translation", *Journal of Information Processing Society of Japan*, Vol.33, No.6, pp. 673–681,1992.
5. Kamai, T., Matsui, K. "Hybrid Synthesis Method using Prewindowed Waveform Segments", *Proc. Spring Meet. Acoust. Soc. Jpn., 3-4-7, pp287-288, 1995.*
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.*
7. Endo, M., et al., "A Study on Sentence Recognition Technique Using Linguistic Constraints Between Separate Word Pairs and Between Consecutive Word Pair", *Proc. Spring Meet. Acoust. Soc. Jpn., 3-P-8, pp177-178, 1995.*
8. Endo, M., et al., "A Study on Fast Algorithm for A* Search with Cooccurrent-Word Model", *Proc. Autumn Meet. Acoust. Soc. Jpn., 2-2-11, pp59-60, 1995.*
9. Morimoto, T., et al.,"ATR's Speech Translation System: ASURA", *EUROSPEECH, pp.1291–1294, 1993*
10. Roe, D.B., et al., "Efficient Grammar Processing for a Spoken Language Translation System", *ICASSP 1992, Vol.1, pp.213.*
11. Suhm, B., et al.,"JANUS: Towards Multi-Lingual Spoken Language Translation", *ARPA Workshop on Spoken Language Technology, Austin, TX, 1995, V.1,pp.221–226*
12. Hatazaki, K. et al.,"INTERTALKER: An Experimental Automatic Interpretation System Using Conceptual Representation." *ICSLP 1992.*