CORBA-BASED SPEECH-TO-SPEECH TRANSLATION SYSTEM

Rainer Gruhn, Koji Takashima, Atsushi Nishino, Satoshi Nakamura

ATR Spoken Language Translation Res. Labs.
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
E-mail: {rgruhn,taka,anishino,nakamura}@slt.atr.co.jp

ABSTRACT

We describe the new implementation of a speech-to-speech translation system at ATR Spoken Language Translation Research Labs (SLT). We use the architecture standard CORBA (Common Object Request Broker Architecture) to interface between a speech recognizer, translation system and TTS engine. Various input types are supported, including close-talking microphone and telephony hardware.

1. INTRODUCTION

We have developed a new speech-to-speech translation system platform using CORBA architecture. The interface was named "**s**peech and **m**ultimodal **i**nterface for multi**l**ingual **e**xchange (SMILE)". Previously, ATR had implemented the speech-to-speech translation system ATR-MATRIX [1, 2], as a research prototype. It consists of the speech recognition subsystem ATRSPREC, the language translation subsystem TDMT, the speech synthesis subsystem CHATR and a main controller. As from early 2000 the new laboratory of ATR SLT was started with new goals, not only the subsystems, but also the software interfacing them needed to meet higher demands, such as higher flexibility and modularity. Instead of simply modifying the old software, we decided to implement a new structure following an established international interface standard. After examining different approaches, CORBA was chosen.

This paper first discusses previous work in the field of speech recognition based service systems. The basic philosophy and properties of CORBA are explained. The structure of the new system, application examples and future steps are given.

2. EXISTING ARCHITECTURAL APPROACHES

2.1. ATR-MATRIX

The previously developed ATR-MATRIX has a modular though fixed structure. The modules (ATRSPREC, TDMT, CHATR) are started and scheduled by a main controller, each module is managed by a local controller. Communication between the modules and the main controller follows the self-defined SipPacket protocol. Figure 1 shows the structure of the previous system. We decided to discard ATR-MATRIX and implement new interface software because of the following problems:

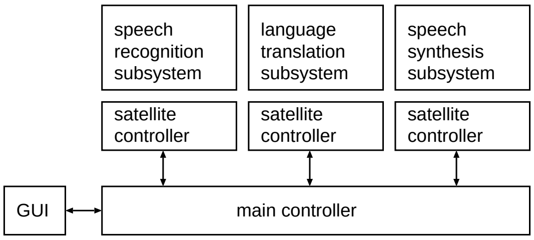

Fig. 1. Previous system architecture

- insufficient modularity
- complicated controller software, making ATR-MATRIX hard to modify
- the system is not easy to install or handle for inexperienced users
- the modules cannot be distributed over the network
- one-user one-directional system, i.e. although ATR-MATRIX is a system meant to translate dialogs, only one conversation side can be translated per computer.
- non-standard interface protocol

Recent papers about architectures focus on man-machine dialog systems rather than on speech-to-speech translation. But the basic problems and architectural approaches are similar. Widely used system architectures are in particular the DARPA Communicator and CORBA.

2.2. DARPA Communicator

The "DARPA Communicator", also known by the name of its reference implementation "Galaxy", is a modular architecture developed at MIT [3]. It was designed for speech recognition based information retrieval systems. It consists of a central hub which interfaces between servers like audio server, dialog manager, database etc. The hub behaviour is defined through a set of rules implemented in a special scripting language. The US Department of Defense Advanced Research Projects Agency (DARPA) encouraged American research facilities and companies to adapt their systems to the Communicator architecture with funding regulations, thereby making it a de facto standard in the USA.

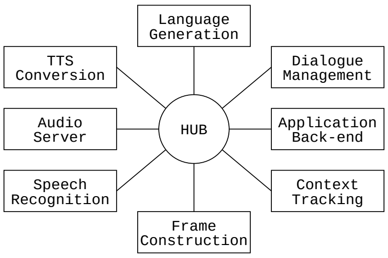

Fig. 2. The DARPA Communicator

The decision not to follow the Communicator standards at ATR was made for several reasons.

- neither specifications nor the reference implementation were open to the public at the time SLTs future development plans were discussed (they have been published in the meantime, though).
- a self-made implementation is familiar and therefore easier to modify according to changing demands. There is no need for external support.
- the Communicator is a standard which is exclusively used for dialog systems, there are no other applications.
- Communicator and CORBA follow similar basic concepts, such as a client-server achitecture, therefore could be made compatible with special interfaces.

2.3. CORBA

CORBA [4] is an object oriented architecture introduced by the Object Management Group (OMG). Using the standard protocol IIOP (Internet Inter-ORB Protocol), modules can interoperate even if they are

- distributed over a network,
- in different programming languages,
- on different platforms.

CORBA encapsulates servers as classes, clients access them using the interface definition language IDL. Servers and clients communicate using IIOP protocol network connections. There is no single central hub; each client and server contains CORBA stubs and skeletons, i.e. interfaces for parameter transmission.

Examples for speech recognition based systems using CORBA include ISIS [5], a multilingual spoken dialog system for financial transactions, and the speech recognition based medical recording system presented in [6]. CORBA is a well established industry standard. CORBA interfaces are available for many programming languages, including C++, Java, Perl and Python.

3. TARGETS

While aiming for a multiple use system, we selected a telephone-based translation system as first target application. Two users in a face-to-face situation access the system using cellular or standard telephones. SMILE provides bidirectional translation.

Besides following a standard architecture, there are other requirements for an ideal speech translation system.

- ability to distribute system: ATRs first speech-to-speech translation system Asura could only run distributed over several computers, ATR-MATRIX's components had to be on one machine. A CORBA based system allows both.
- one system (on one computer) should allow bi- or multidirectional translation.
- multimodal input data: not only speech from microphone, but also text, speech files and video input should be supported.
- a WWW or other network-based interface.
- support multiple programming languages.
- dynamic system control.
- dynamic configuration change.
- logging of all data must be easy.
- the (graphical) user interface must be easy to understand.

4. SYSTEM ARCHITECTURE

Figure 3 gives an overview of the new system structure. The servers include a speech recognition server, text-to-text translation, speech synthesis and additionally CORBA services. CORBA services are administration servers that are

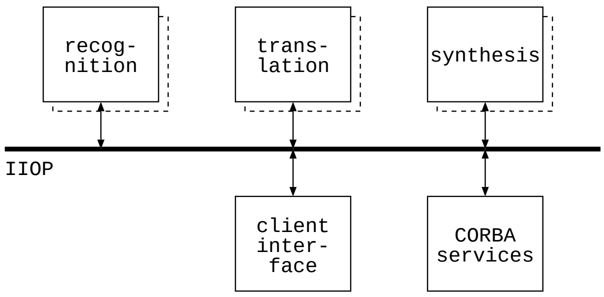

Fig. 3. Overview of the new SMILE system prototype

part of the CORBA distribution, such as a naming server to bind a name to an object reference, a security server for authorization and encryption etc. A client module interfaces to GUI and input data. The client interface requests services from the servers, controls the data flow and returns the results to the user.

The splitup into servers allows both the systems application as speech-to-speech translation system as well as an easy change to a human-machine dialog system (if a dialog manager is given). New servers can be quickly included into the system as a standard IDL based interface library is provided to easily wrap any new program.

The system flow and server selection is completely driven by a client program. SMILE can be used by more than one client at the same time. Therefore, unlike a system with a central hub and central flow control, the SMILE servers can be used for completely different tasks at the same time. For example, one client driving the servers as translation machine and a second client using the servers for a dialog system can run simultaniously.

5. APPLICATIONS

SMILE is designed to be easily configurable for many different applications. The speech-to-speech translation task are hotel reservation and general travel related conversations, including shopping, sightseeing,
transportation and simple medical expressions. Conversation languages are English and Japanese, Mandarin Chinese will be added soon. SMILE is designed to be multimodal and support various types of user interfaces. For example, to support different kinds of speech input, we insert a recognition server for clean speech and an additional recognition server for telephone speech.

5.1. Telephone-based Translation Service

A very simple telephone interface has been developed earlier [7] using only minimal hardware. To be able to implement additional features and to enhance the user interface, we now use a CTI board to access the telephony hardware. For a translation system, dialog control is not possible or necessary. Only in an initial phase the user is prompted to tell the system the dialog partner's telephone number which is then called by the system. After that, the system will wait for speech input, recognize, translate and then play the result to the cellular phone of the dialog partner. As a face-to-face situation is assumed, only limited feedback is needed, such as the information that the server is currently working on the recognition. This can be provided e.g. with a signal tone. Additional feedback can be provided by displaying the recognition and translation results on a PDA, as previously shown in [8]. Most modules in this systems run on the Linux operating system, others on Windows NT or CE. CORBAs interoperability supports such distribution.

5.2. C-Star 3

The international C-Star 3 speech-to-speech translation research project [9] also focuses on recognition of telephone speech. Each site develops their own human language translation (HLT) systems, ATR uses the new CORBA-based SMILE. The interface between those HLT servers is provided with the communication switch (CS), a socket-based broadcasting system similar to internet chat. It transmits recognition and translation results as well as control messages to all connected sites. The common interface to the computer telephony hardware has been called Mediator. It initially prompts a user to input the dialog partners language and telephone number, then sends speech data to the recognizer and plays speech generated by the text-to-speech system. Figure 4 shows the architecture of the C-Star 3 system architecture.

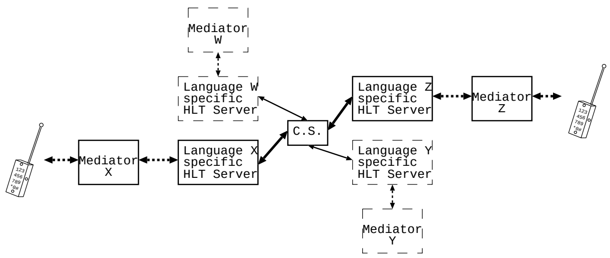

Fig. 4. The C-Star 3 architecture with multiple telephone interfaces (Mediator) and human language technology (HLT) servers

6. EVALUATION

The previous speech-to-speech translation system ATR-MATRIX has been evaluated for naive users [10] and showed a score of about 500 in the TOEIC, a standard test of English as a foreign language. Such an end-to-end evaluation has not been undertaken for the SMILE-based system yet. The current application for SMILE is a phrasebook-type translation service task which is currently developed at ATR [11]. The vocabulary size is 21329 unique words for Japanese and 19173 for English. Current recognition rates for this task are shown in Table 1. The English acoustic model was trained on the WSJ database, the Japanese on the ATR phoneme balanced sentences database. Training and evaluation are based on clean read speech recorded with a close-talking microphone at 16 kHz sampling frequency. The test set for English consists of 2288 sentences uttered by 44 speakers, the Japanese test set contains 510 utterances uttered by 20 speakers. Work on telephone speech acoustic models is still in progress, with acoustic models being generated from training data filtered by a codec or by piping, i.e. playing to a speaker and then recording with a cellular phone.

Table 1. Acoustic model evaluation

Language	English	Japanese
WER	17.4	14.7

7. FUTURE WORK AND CONCLUSION

As the server interface software has just been finished, many future steps are possible. One development effort will focus on the user interface part, we are currently working on a telephone-based speech-to-speech translation service system. To support a high number of languages, ATR cooperates with research institutes in seven countries in the C-Star 3 project. The second new user interface type is a PDA in a client-server scenario.

A different target is to include a dialog management module and start developing a human-machine communication system such as travel information retrieval or ticket and hotel reservation.

This high variation of possible future applications shows the high flexibility of the CORBA approach. SMILE can also be used for a dialog system, or for a translation system with dialog control.

8. ACKNOWLEDGMENTS

The authors would like to thank Dr. Seiichi Yamamoto, President, ATR SLT, for giving us the opportunity to do research at ATR, and also all members of the technical support group (TSG) for support and valuable discussions.

9. REFERENCES

[1] B. Reaves, A. Nishino, and T. Takezawa, "ATR-MATRIX: Implementation of a speech translation system," in *Proc. Acoust. Soc. Jap.*, Spring 1998, pp. 53–54.

[2] T. Takezawa, T.Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX," in *Proc. ICSLP*, 1998, pp. 957–960.

[3] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," in *Proc. ICSLP*, Sydney, 1998.

[4] OMG, "Common Object Request Broker Architecture (CORBA)," http://www.corba.org/, 1990.

[5] H. Meng, S.F. Chan, Y.F. Wong, T.Y. Fung, W.C. Tsui, T.H. Lo, C.C. Chan, K. Chen, L. Wang, T.Y. Wu, X. Li, T. Lee, W.N. Choi, Y.W. Wong, P.C. Ching, and H. Chi, "ISIS: A multilingual spoken dialog system developed with CORBA and KQML agents," in *Proc. ICSLP*, Beijing, 2000.

[6] D.F. Rosenthal and J. Mayo, "Using corba and speech recognition for structured medical reporting," in *Proc. AMIA 1997 Symposium*, Nashville, 1997.

[7] R. Gruhn, H. Singer, H. Tsukada, M. Naito, A. Nishino, A. Nakamura, Y. Sagisaka, and S. Nakamura, "Cellular-phone based speech translation system ATR-MATRIX," in *Proc. ICSLP*, 2000, pp. IV–448–451, 03068.

[8] H. Singer, R. Gruhn, M. Naito, H. Tsukada, A. Nishino, A. Nakamura, and Y. Sagisaka, "Speech translation anywhere: Client-server based ATR-MATRIX," Tech. Rep. SP99-121, IEICE, 1999.

[9] C-Star 3, "Consortium for Speech Translation Advanced Research (C-Star): C-Star Experiment," http://www.c-star.org/, July 2001.

[10] F. Sugaya, T. Takezawa, A. Yokoo, Y. Sagisaka, and S. Yamamoto, "Evaluation of the ATR-MATRIX speech translation system with a pair comparison methodbetween the system and humans," in *Proc. ICSLP*, 2000.

[11] E. Sumita, "Example-based machine translation using dp-matching between work sequences," in *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, 2001, pp. 1–8.