

ATR-MATRIX: IMPLEMENTATION OF A SPEECH TRANSLATION SYSTEM

Ben Reaves, ©Atsushi Nishino, Toshiyuki Takezawa
(ATR Interpreting Telecommunications Research Laboratories, Kyoto 619-0288)*

1. INTRODUCTION

ATR-Matrix¹ is a fully automatic speech-to-speech language translation system developed at ITL. Unlike its predecessor ASURA², ATR-Matrix is designed for spontaneous speech input; and it's much faster. This paper describes its implementation for 2 operators, applied to the Hotel Reservation Task, from Japanese to English.

2. DESIGN

The basic design consists of three subsystems and a Satellite Controller for each, and a Main Controller as shown in Figure 1.

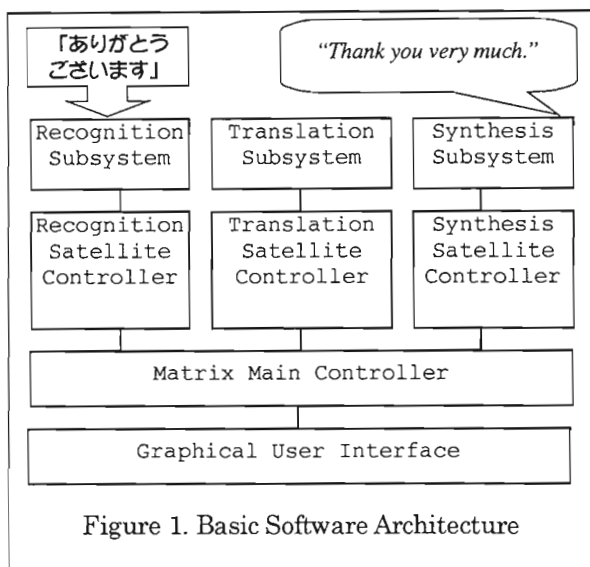


Figure 1. Basic Software Architecture

The recognition subsystem consists of SPREC^{3,4}, the translation subsystem consists of TDMT⁵, and the synthesis subsystem consists of CHATR⁶. Each Satellite Controller encapsulates the knowledge for its subsystem, so that the Main Controller can interact with them in a uniform way, using a standard packet message format.

SPREC is a speaker independent recognizer designed for spontaneous speech input. For ATR-Matrix, we added a module to SPREC that converts its output from a Lattice to N-Best format and (1) marks the end of each sentence within an utterance (Sentence Splitting⁷), and then (2) at the end of each sentence, if the pitch is rising, tags it as a question to be translated appropriately (Question Detection).

The Translation Satellite Controller sends each of the utterances included in the N-Best list from SPREC to TDMT. The first utterance that TDMT can translate is captured and sent back to the Main Controller as the translation result.

Before the Synthesis Satellite Controller sends the

translation result to CHATR, it selects the speaker, male or female, depending on the sex of the operator; this information comes from the acoustic model set (one for male and one for female) that gave the best match from SPREC.

The Main Controller has access to all major data paths, and sends relevant information to the Graphical User Interface (GUI). Based on user input from the GUI, it controls the subsystems via the appropriate Satellite Controllers. It also logs activity on all major data paths, for later analysis of system behavior.

Each Satellite Controller handles details so that the Main Controller can efficiently perform the switching and GUI interaction. This separation of Main Controller and Satellite Controllers facilitates complete substitution of a major subsystem. Separation into individual Unix processes allows us to test each subsystem and its Satellite Controller individually. The basic configuration of ATR-Matrix can be changed by rewriting only the Main Controller.

We developed a 2-user version of ATR-Matrix by modifying the Main Controller to accept commands for turn-taking from the GUI, described below, and to accept a 2nd recognizer identical in all respects to the first but running on a separate host workstation with separate audio input.

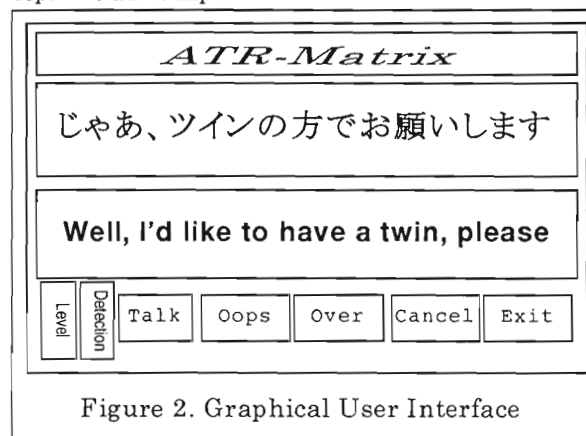


Figure 2. Graphical User Interface

The large letters of the upper windows (logo, recognition result, translation result) are for easy viewing by the audience. The smaller lower windows and buttons are for the operator: "Talk" to begin talking, "Over" to pass the turn to the other party. Real-time display of input level and speech detection state is essential for operator feedback. The "Oops" and "Cancel" buttons are reserved for use with dialogue control.

The timing diagram in Figure 3 is a rough measurement of packet arrival times for a typical example: 「宿泊の予約をお願いします」 translated to "I'd like to make a reservation for a room, please."

*音声翻訳システム「ATR-MATRIX」の実装方法。リープス ベン, 西野敦士, 竹澤寿幸 (ATR 音声翻訳通信研究所)
ben@itl.atr.co.jp anishino@itl.atr.co.jp takezawa@itl.atr.co.jp

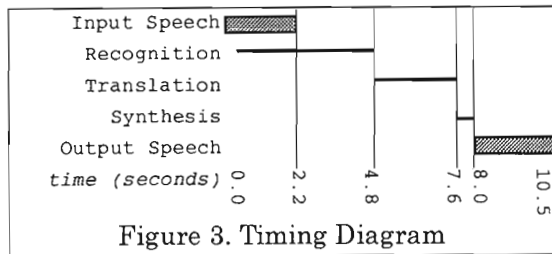


Figure 3. Timing Diagram

Recognition appears to have the longest processing time, but we believe that most of it is due to delays in the data acquisition hardware and driver. This figure shows the real delay that the user sees.

SPREC calculates the feature vectors and performs the forward search while the operator is speaking. Display on the GUI comes a little later but is not shown here because it does not delay the main data path. Passing data through the Satellite Controller and Main Controller is too fast to be shown in the figure (this is a benefit of writing the controllers in C++ rather than a Rapid Prototyping language as we did the GUI). Translation requires 2.8 seconds in this case. After CHATR spends a short time concatenating the output utterance and sending it to the output device, the sound arrives from the loudspeaker. The total response time of ATR-Matrix is about 5.8 seconds for this example. During this time, written information and partial translations are displayed, one by one, so it does not appear to be slow.

3. IMPLEMENTATION ISSUES

Allowing some direct connections between subsystems would decrease the traffic concentration through the Main Controller, but it would also create synchronization problems. We opted for the safety of central control over the speed of direct connections.

Running this as an online demonstration revealed many issues that are not obvious when each subsystem is demonstrated in isolation.

First is the importance of Streaming Speech Detection. SPREC's EPD module is a streaming speech detector, able to detect the start of speech within about 50ms but the detection of the end is much longer: almost 1 second. If SPREC's forward search detects a long match with a pause model, then it may detect the end of speech before EPD does, thus greatly reducing the response time. If EPD or the search decides that what was detected was not speech, nothing is output, and SPREC continues waiting for the operator to speak.

To avoid SPREC triggering unexpectedly, we have configured SPREC to disable its audio input when an utterance has been recognized. The operator must enable it with the Talk button in the GUI. This gives the operator more control over the system, but it does require use of the hands.

A second issue is error handling. If TDMT can not translate any of the output from SPREC, then the Synthesis Satellite Controller commands CHATR to choose a Japanese female speaker and say the Japanese equivalent of, "Please repeat." We choose Japanese because this should be fed back to the operator (Japanese), not to the audience (English).

A third issue is feedback to the operator. The current audio input level and state of the speech detection are indispensable to the operator. These are all on the GUI screen, located near the operator's face.

Software and hardware resource usage:

| | |
|--------------------|---|
| Workstation | DEC Alphastation 500/500 |
| Runtime Memory | 210 MB real + 203MB virt. |
| Disk space | 1.6 Gigabytes |
| Operating System | Digital Unix v3.2G |
| Microphone | Sennheiser 410 |
| Audio Input | DAT-Link+, 12KHz, 16 bit |
| Audio Output | internal, 16KHz, 16bit |
| Installed Software | gcc, python, Tcl/Tk, Allegro Common Lisp v4.3, Rogue Wave C++ Class Libraries |

4. WHAT'S NEXT?

Work is in progress on an English-to-Japanese version of ATR-Matrix. More work is also planned on Dialogue Control.

We are also working on improvements to the Question Detection. It is currently implemented not as a subsystem but as a SPREC Module. Work is also in progress on a faster and more stable pitch-tracking algorithm for use in Question Detection. We are also working on algorithms for detecting and passing more complex prosody-based information for more natural and intelligible output speech.

ATR-Matrix is a starting point for new areas of research and development. ATR-Matrix is a real system in which we can implement the result of ongoing research at ITL. A comprehensive evaluation of the whole ATR-Matrix system, from speech input to speech output, will shed light on those areas needing work.

5. ACKNOWLEDGEMENTS

The development of ATR-Matrix has been fun. All research departments cooperated fully with each other, with ITL management, and with the Technical Support Group Development Team, which designed and implemented ATR-Matrix, and wrote all of the controllers. We would especially like to thank this team's members: Toshio Ban, Hiroaki Tagawa, Kouji Takashima, and Takeshi Matsuda. We also thank ITL President Seiichi Yamamoto for the support to bring it all together.

¹ T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, "A Japanese-to-English Speech Translation System: ATR MATRIX", *IPSI*, March 1998.

² 森元暹、田代敏久、竹澤寿幸、永田昌明、谷戸文廣、浦谷則好、鈴木雅実、菊井玄一郎 "音声翻訳システム(ASURA)のシステム構成と性能評価," *情報処理学会論文誌* Vol. 37, No. 9, pp. 1726-1735.

³ 山本博史、シンガー ハラルド、リーブス ベン、匂坂芳典 "日英音声翻訳システム「ATR MATRIX」における音声認識部分の構造と制御方法," *本音講論*.

⁴ 内藤正樹、政瀧浩和、シンガー ハラルド、塚田元、匂坂芳典 "日英音声翻訳システム ATR-MATRIX における音声認識用音響-言語モデル," *本音講論*.

⁵ O. Furuse, J. Kawai, H. Iida, S. Akamine, K. Kim, "Multi-lingual Spoken-Language Translation Utilizing Translation Examples," *3rd Natural Language Processing Pacific Rim Symposium (NLPRS-95)*, December 1995.

⁶ N. Campbell, "CHATR: A High Definition Speech Re-Sequencing System," *Proceedings of the 3rd ASA/ASJ Joint Meeting* (1996) pp. 1223-1228.

⁷ 竹澤寿幸、森元暹 "発話単位の分割または接合による言語処理単位への変換," pp. 19-24, *情報処理学会研究会資料* (1997) SLP-18-4.