# A JAPANESE-TO-ENGLISH SPEECH TRANSLATION SYSTEM: ATR-MATRIX

*Toshiyuki Takezawa, Tsuyoshi Morimoto†, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida,*
*Fumiaki Sugaya, Akio Yokoo and Seiichi Yamamoto*

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
E-mail: {takezawa, morimoto, sagisaka, nick, iida, sugaya, ayokoo, s-yama}@itl.atr.co.jp
†currently with Department of Electronics Engineering and Computer Science, Fukuoka University, Fukuoka, Japan

## ABSTRACT

We have built a new speech translation system called ATR-MATRIX (ATR's Multilingual Automatic Translation System for Information Exchange). This system can recognize natural Japanese utterances such as those used in daily life, translate them into English and output synthesized speech. This system is running on a workstation or a high-end PC and achieves nearly real-time processing. The current implementation of our system deals with a hotel room reservation task/domain. We plan to develop a bidirectional speech translation system, i.e., Japanese-to-English and English-to-Japanese. We also plan to develop multi-language output functions from ATR-MATRIX (Japanese-to-English, German and Korean) for the international joint experiment of C-STAR II (Consortium for Speech Translation Advanced Research).

## 1. INTRODUCTION

We have built a new speech translation system called ATR-MATRIX (ATR's Multilingual Automatic Translation System for Information Exchange). This system can recognize natural Japanese utterances such as those used in daily life, translate them into English and output synthesized speech. This system is running on a workstation or a high-end PC and achieves nearly real-time processing. Unlike its predecessor ASURA [1], ATR-MATRIX is designed for spontaneous speech input, and it's much faster.

Recently, there have been many projects on speech-to-speech translation [2, 3]. *Verbmobil* [2], which is one of the major research projects, in Germany, adopts a combined method of deep and shallow processing. JANUS [3] is another major research project that adopts an interlingua-based language translation method. In contrast to those works, we adopt a cooperative integrated language translation method. Moreover, ATR-MATRIX has features such as a personalized speech synthesis based on dynamic speaker selection in speech recognition.

Section 2 gives an overview of the system. Section 3 describes the key features of three major subsystems in our system: speech recognition, language translation and speech synthesis. Section 4 describes more features for dealing with spontaneous speech. Section 5 discusses implementation issues such as speech detection. Section 6 describes a preliminary system evaluation. Section 7 offers discussion and describes future works. Section 8 gives our conclusions.

## 2. SYSTEM OVERVIEW

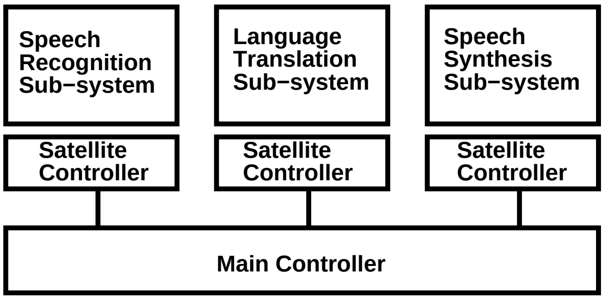
Speech Recognition Sub−system
Language Translation Sub−system
Speech Synthesis Sub−system
Satellite Controller
Satellite Controller
Satellite Controller
Main Controller

**Figure 1. System configuration**

Figure 1 shows the system configuration. This system consists of a speech recognition subsystem, a language translation subsystem, a speech synthesis subsystem, and a main controller. Each subsystem is connected to the main controller via each satellite controller. Each satellite controller encapsulates the knowledge for its subsystem so that the main controller can interact with them in a uniform way by using a standard packet message format. The current implementation of our system deals with a hotel room reservation task/domain.

## 3. KEY FEATURES

### 3.1. Real-time speech recognition using speaker-independent phoneme-context-dependent acoustic models and a language model of variable-order $N$-gram

Speech features are widely different between speakers such as males or females and phoneme-contexts. Therefore, we have proposed a statistical method (ML-SSS) [4] to make speaker-independent phoneme-context-dependent acoustic models. Using this method, we have prepared speaker-independent phone models for males and females separately.

We have also proposed a language model of variable-order $N$-gram [5], which is a compact language model to deal with various expressions in spontaneous speech. We realized real-time processing by an effective search method based on a word-graph [6].

The vocabulary size of the speech recognition subsystem is about 2,000 words, which is almost enough for one task/domain such as hotel room reservation except for the problem of proper nouns such as human names.

### 3.2. Robust language translation to deal with speech recognition results

Our language translation subsystem can deal with various expressions in spoken language because it uses not only sentence structure but also examples such as translation pairs

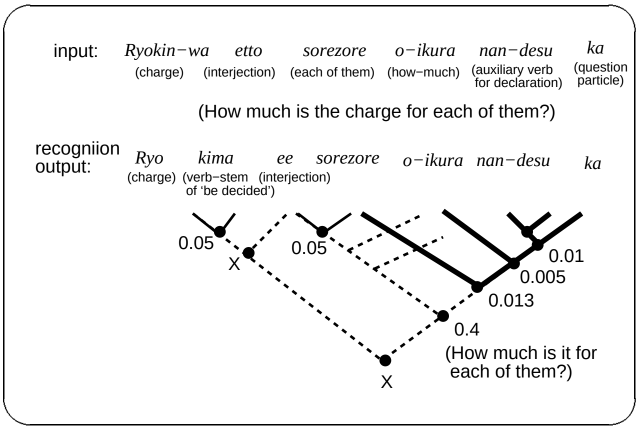

Figure 2. Example of partial translation

[7]. Furthermore, we have introduced a partial translation mechanism for accepting speech recognition results that include recognition errors [8]. We adopted two heuristics:

(1) Similar constituents to translation examples are preferred. We use semantic distance based on translation pairs represented by patterns, e.g., the upper threshold is set to 0.2.

(2) Larger constituents are preferred. We use the number of word sequences in the constituent, e.g., the lower threshold is set to 2.

Figure 2 shows an example of this partial translation method. In this example, utterance of "*Ryokin-wa*," which means "charge," is miss-recognized as "*Ryo kima*," which consists of a word that means "charge" and another word from a verb-stem of "be decided". The structure of "*Ryo kima*" would not be made much larger. This hypothetical structure would be pruned because the lower threshold of the number of word sequences in the constituent is set to 2. The semantic distance corresponding to "*ee sorezore o-ikura nan-desu ka*" would be 0.4. This hypothetical structure would be pruned because the upper threshold of semantic distance is set to 0.2. Finally, a constituent of "*sorezore o-ikura nan-desu ka*" would be selected and equivalent English, such as "How much is it for each of them?", would be generated.

The vocabulary size of the language translation subsystem from Japanese to English is about 13,000 words, which cover almost all of our bilingual travel conversation database [9, 10]. The vocabulary used for our speech recognizer is a subset of this vocabulary.

### 3.3. Personalized speech synthesis

Personalized speech synthesis is essential for a realistic speech-to-speech translation system. Since the current configuration of our system has male and female acoustic models, the CHATR [11] speech synthesis subsystem outputs male or female voices (Fig. 3). It is easy to improve our system for more speakers because unneeded models can be pruned quickly due to the efficient beam-search in the speech recognition process (Fig. 4).

## 4. MORE FEATURES FOR SPONTANEOUS SPEECH

The utterance units that serve as input to a speech translation system for handling spontaneous speech are not always sentences. However, the processing units of language

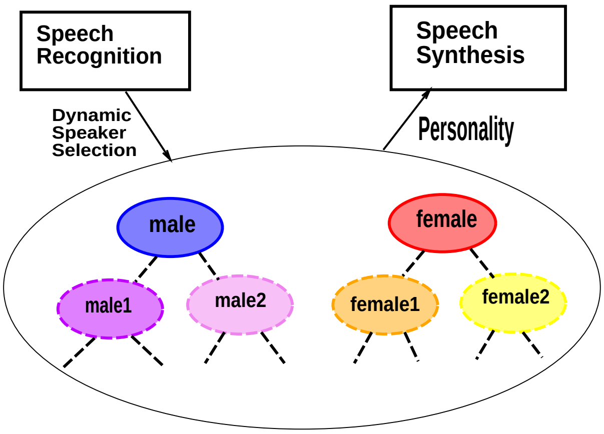

Figure 3. Personalized speech synthesis

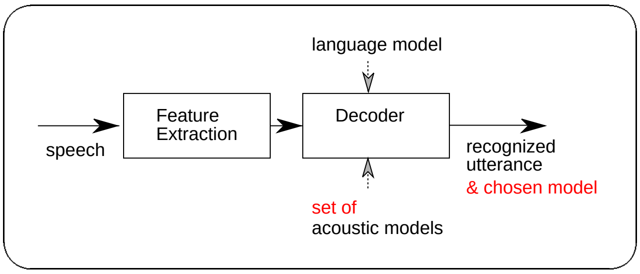

Figure 4. Dynamic speaker selection and effective search in speech recognition

**Table 1. Utterance division into meaningful chunks**

| | Recall | Precision |
|---|---|---|
| Statistical model | 88.6% | 65.7% |
| Statistical model with heuristics | 97.7% | 99.4% |

translation are sentences. Since we do not have enough knowledge about sentences in spoken languages, we use the term "meaningful chunks" instead of sentences. According to our bilingual travel conversation database [9, 10], utterance units often need to be divided into several meaningful chunks. We have proposed a method of transforming utterance units into meaningful chunks based on pause information and the $N$-gram of fine-grained part-of-speech subcategories [12]. Table 1 shows a summary of preliminary experiments. The statistical model refers to two words before the current position and one word after the current position. The heuristics that we have introduced are as follows.

- If a conjunctive postpositional particle does not follow an interjection, we set a boundary to the position.
- If a conjunctive postpositional particle follows an interjection, we do not set a boundary to the position.
- We do not set a boundary to the position between a finished form of auxiliary verb and a sentence final postpositional particle.

In spontaneous conversational speech, sentence-final prosody information sometimes conveys question information instead of the Japanese sentence-final particle "*ka*". A prosody extraction function enables us to generate the equivalent English "Are rooms available?" instead of "Rooms are available." from the Japanese utterance "*Heya wa aite-masu (↗)*". A (↗) mark indicates that sentence-final prosody is high.

## 5. IMPLEMENTATION ISSUES

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

The first is the importance of streaming speech detection. Our end-point detection (EPD) module is a streaming speech detector able to detect the start of speech within about 50 ms, but the detection of the end is much longer: almost 1 second. If the forward search in our speech recognition subsystem detects a long match with a pause model, then it may detect the end of speech before EPD does, thus greatly reducing response time. If EPD or the search decides that what was detected was not speech, nothing is output, and our speech recognition subsystem continues waiting for the operator to speak.

A second issue is error handling. If our language translation subsystem cannot translate any of the output from the speech recognition subsystem, then our main controller commands the speech synthesis subsystem to choose a Japanese female speaker and say the Japanese equivalent of "Please repeat." We chose 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 speech detection are indispensable to the operator. All of this information is placed on the graphical user interface (GUI) screen near the operator's face.

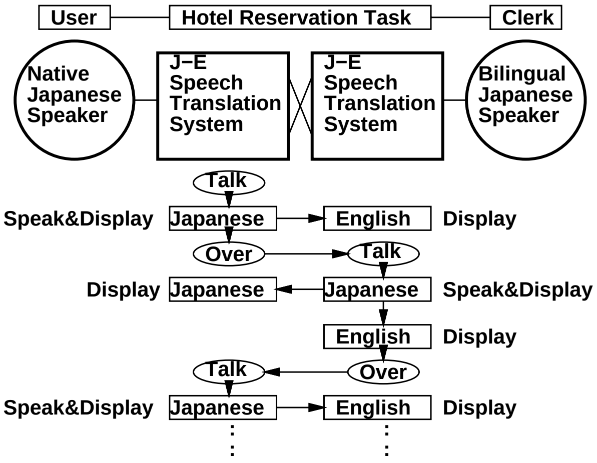

Figure 5. Dialog test process

## 6. PRELIMINARY SYSTEM EVALUATION

### 6.1. Concepts

The state-of-the-art technology of speech recognition and translation cannot avoid errors. To solve this problem, some systems, like human-machine speech dialog systems, use deep knowledge processing. However, in our tests we do not use intelligent knowledge processing to mediate conversation but instead give users many chances to check and retry. This means that the speaker can check the result of speech recognition and speak again if errors are unacceptable, and if the other party cannot understand the result of translation, he/she can ask the necessary questions. Users can make necessary interventions in the system at intermediate levels and try to improve the final result. To make this scheme run efficiently, interactions must cycle as fast as possible. We have tuned the system parameters, and for both recognition and translation we can obtain the same processing time as the utterance length. The system has some time lag due to sequential architecture, but questionnaires and interviews revealed that this time lag does not irritate speakers.

### 6.2. Dialog Test Process

Our current system supports a hotel reservation task conducted between a user and a clerk. In the test setting, we cast people who are unfamiliar with the system as users and those who are familiar with it as clerks. Thus the users do not have expertise in the system while the clerks are experts in it, being researchers, designers and programmers in the project. The users speak in Japanese, while the clerks read the translated English texts and speak in Japanese (Fig. 5). In this test, we do not test speech synthesis. We give users and clerks scenarios that explain the mission and the meaning of proper nouns.

### 6.3. Preliminary Results

When speakers start a dialog, they use complex expressions with hesitations, which makes the recognition difficult and results in many recognition and translation errors. In the test, we allow retries so the speaker may try initially similar expressions a few times. In the retries, the determined speakers tend to use other simpler expressions to achieve tasks; gradually, the speakers start to use simpler expres-

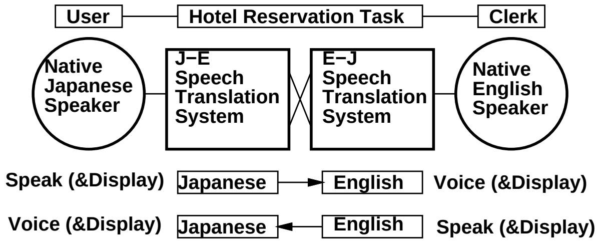

Figure 6. Future test process

sions to make ATR-MATRIX perform well. For the hotel reservation task, task achievement rate is roughly 60% or more on the users' side and 80% on the clerks' side.

We are now analyzing the dialog structures for the number of retries and utterance length within turns. The initial results show that retry is an important factor, and that through the retries the speakers can learn the system performance and control the output quality.

## 7. DISCUSSION AND FUTURE WORKS

We showed that ATR-MATRIX can achieve a score of 60% or more for multiple language real-time conversations in hotel reservation tasks when retries are allowed. We plan to also evaluate Japanese/English bi-directional ATR-MATRIX without controlling speakers' turns (Fig. 6). Finally, we would like to utilize these results to build successful systems such as the international joint experiment of C-STAR II (Consortium for Speech Translation Advanced Research).

## 8. CONCLUSIONS

This paper reports a new speech translation system called ATR-MATRIX. We are now carrying out further system evaluation. We plan to develop a bidirectional speech translation system, i.e., Japanese-to-English and English-to-Japanese. We will conduct much more research on understanding of utterance conditions such as prediction of next utterances for speech recognition and disambiguation for the generation of target languages. We also plan to develop multi-language output functions from ATR-MATRIX (Japanese-to-English, German and Korean) for the international joint experiment of C-STAR II.

## ACKNOWLEDGMENTS

The authors wish to thank all members of ATR Interpreting Telecommunications Research Laboratories for their contributions to building our system.

## REFERENCES

[1] Tsuyoshi Morimoto, Toshiyuki Takezawa, Fumihiro Yato, Shigeki Sagayama, Toshihisa Tashiro, Masaaki Nagata and Akira Kurematsu: "ATR's Speech Translation System: ASURA," *Proc. of EuroSpeech '93*, pp. 1291–1294 (1993).

[2] Thomas Bub, Wolfgang Wahlster and Alex Waibel: "Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation," *Proc. of ICASSP '97*, pp. 71–74 (1997).

[3] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld and Puming Zhan: "JANUS-III: Speech-to-Speech Translation in Multiple Language," *Proc. of ICASSP '97*, pp. 99–102 (1997).

[4] Mari Ostendorf and Harald Singer: "HMM Topology Design Using Maximum Likelihood Successive State Splitting," *Computer Speech and Language*, Vol. **11**, No. 1, pp. 17–41 (1997).

[5] Hirokazu Masataki and Yoshinori Sagisaka: "Variable-Order $N$-gram Generation by Word-Class Splitting and Consecutive Word Grouping," *Proc. of ICASSP '96*, pp. 188–191 (1996).

[6] Toru Shimizu, Hirofumi Yamamoto, Hirokazu Masataki, Shoichi Matsunaga and Yoshinori Sagisaka: "Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graph," *Proc. of ICASSP '96*, pp. 145–148 (1996).

[7] Hitoshi Iida, Eiichiro Sumita and Osamu Furuse: "Spoken-Language Translation Method Using Examples," *Proc. of COLING '96*, pp. 1074–1077 (1996).

[8] Yumi Wakita, Jun Kawai and Hitoshi Iida: "Correct Parts Extraction from Speech Recognition Results Using Semantic Distance Calculation, and Its Application to Speech Translation," *Proc. of ACL/EACL Workshop on Spoken Language Translation*, pp. 24–31 (1997).

[9] Tsuyoshi Morimoto, Noriyoshi Uratani, Toshiyuki Takezawa, Osamu Furuse, Yasuhiro Sobashima, Hitoshi Iida, Atsushi Nakamura, Yoshinori Sagisaka, Norio Higuchi and Yasuhiro Yamazaki: "A Speech and Language Database for Speech Translation Research," *Proc. of ICSLP '94*, pp. 1791–1794 (1994-09).

[10] Toshiyuki Takezawa, Tsuyoshi Morimoto and Yoshinori Sagisaka: "Speech and Language Databases for Speech Translation Research in ATR," *Proceedings of the First International Workshop on East-Asian Language Resources and Evaluation (EALREW) — Oriental COCOSDA Workshop '98 —*, Tsukuba, Japan, pp. 148–155 (May 1998).

[11] Nick Campbell: "CHATR: A High-Definition Speech Re-Sequencing System," *Proc. of ASA/ASJ Joint Meeting*, pp. 1223–1228 (1996).

[12] Toshiyuki Takezawa and Tsuyoshi Morimoto: "Transformation into Language Processing Units by Dividing or Connecting Utterance Units," *IPSJ SIG Notes*, 97-SLP-18-4, Vol. **97**, No. 101, pp. 19–24 (1997) *(in Japanese)*.