ATR-MATRIX: A Speech Translation System from Japanese to English

Akio Yokoo, Toshiyuki Takezawa, Yoshinori Sagisaka, Nick Campbell,
Hitoshi Iida and Seiichi Yamamoto
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, JAPAN
E-mail: {ayokoo, takezawa, sagisaka, nick, iida, s-yama}@itl.atr.co.jp
Telephone: +81 774 95 1355
Facsimile: +81 774 95 1308

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 can run on a workstation or a high-end PC and achieves almost real-time processing. The current implementation of our system can deal with a hotel room reservation task/domain. Our future plans are to develop a bidirectional speech translation system, that is, a system to handle Japanese-to-English and English-to-Japanese, and to make multi-language outputs from ATR-MATRIX possible (Japanese-to-English, German, and Korean) for the international joint experiments 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 can run on a workstation or a high-end PC and achieves almost real-time processing. Unlike its predecessor ASURA [1], ATR-MATRIX is designed to handle spontaneous speech inputs, and is much faster.

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

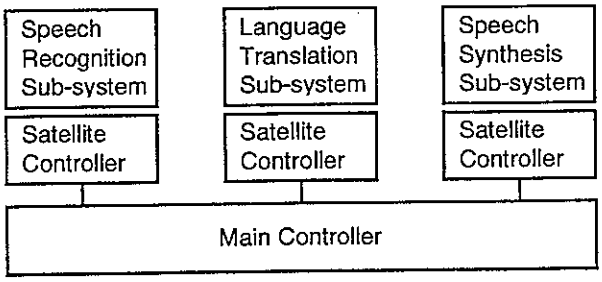

Figure 1. System construction

Section 2 describes the system overview. Section 3 describes the key features of the three major sub-systems (i.e., speech recognition, language translation, and speech synthesis) in our system. Section 4 describes additional features enabling us to deal with spontaneous speech. Section 5 describes implementation issues such as speech detection. Section 6 describes future works. Finally, section 7 is the conclusion.

2. System overview

Figure 1 shows the construction of the system. This system consists of a speech recognition sub-system, a language translation sub-system, a speech synthesis sub-system, and a main controller. Each of the sub-systems is connected to the main controller via a satellite controller. Each of the satellite controllers encapsulates the knowledge for its sub-system, so that the main controller can interact with all of them in a uniform way, using a standard packet message format. The current

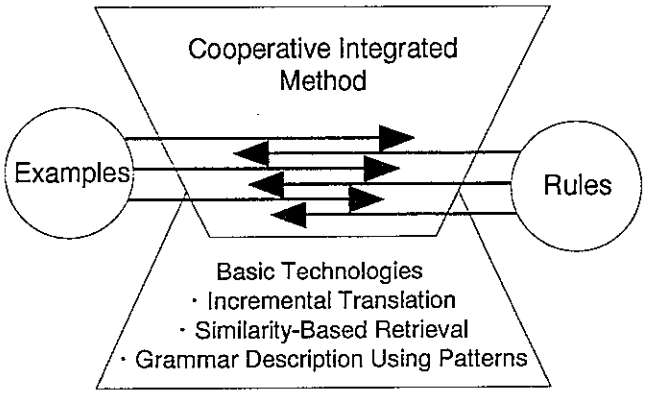

Figure 2. Cooperative Integrated Translation

implementation of our system can deal 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 variable-order *N*-gram language model

Speech features widely differ among speakers, e.g., males, females, phoneme-context, etc. Therefore, we have proposed a statistical method (ML-SSS) [4] in order to make speaker-independent phoneme-context-dependent acoustic models. Using this method, we have already prepared speaker-independent phone models for males and females separately.

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

The vocabulary size of the speech recognition sub-system is about 2,000 words, that is, almost enough for one task/domain such as a hotel room reservation task/domain (excluding the problem of proper nouns like human names).

3.2 Robust language translation to deal with speech recognition results

We have established, through comparative experiments, that example-based translation methods are the most effective in handling a wide variety of natural speech translation problems. These problems include:

- fragmental utterances;
- ungrammatical/everyday/metaphorical expressions;
- ambiguous/omitted particles in Japanese utterances;
- ambiguous dependency structures in English.

Our Example-based Machine Translation [7] integrates both examples and rules in a common framework, and is now able to easily cope with the translation problems we have endeavored to solve since the beginning of the project (Figure 2). When complex sentences are translated, the closest examples are retrieved from the database, dependencies among component words are analyzed, and at the same time, translation equivalents are assembled.

In much the same way as people learn a foreign language by the use of source-to-target expression pairs, our Cooperative Integrated Translation produces equivalent utterances by synchronization of the source and target language structures. The source language is analyzed by Constituent Boundary Parsing to match target language expressions.

Cooperative Integrated Translation has the following advantages:

(1) reliance on the existence of efficient parsing algorithms
(2) the ability to handle various linguistic phenomena with the help of translation examples selected by a simple best-first mechanism

Constituent Boundary Parsing is performed using bottom-up and left-to-right chart parsing techniques. Best-first syntactic and semantic similarity enables flexibility in this approach. For instance, a sentence like "A cheap and clean room would be good." can be translated with the help of similar examples like "A cheap and clean election campaign is good." This incremental robust parsing technique can even handle ungrammatical phenomena, such as derivation in metonymical relationships, and significantly reduces the explosion of structural ambiguities.

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

(1) Similar constituents to translation examples are preferred. We use semantic distances based on translation pairs represented by patterns, e.g., the upper bound 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 bound threshold is set to 2.

Figure 3 shows an example of this partial translation method. In this example, the utterance of *"Ryokin-wa"* (which means "charge") is mis-recognized as *"Ryo kima"* (which consists of a word that means "charge" and another word of a verb-stem that means "be decided"). The structure of *"Ryo kima"* is not made much larger. This hypothesis structure is pruned because the lower bound 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"* is 0.4. Because the upper bound threshold of the semantic distance is set to 0.2, this hypothesis structure is pruned. Finally, a constituent of *"sorezore o-ikura nan-desu ka"* is selected and the equivalent English "How much is it for each of them?" is generated.

The vocabulary size of the language translation sub-

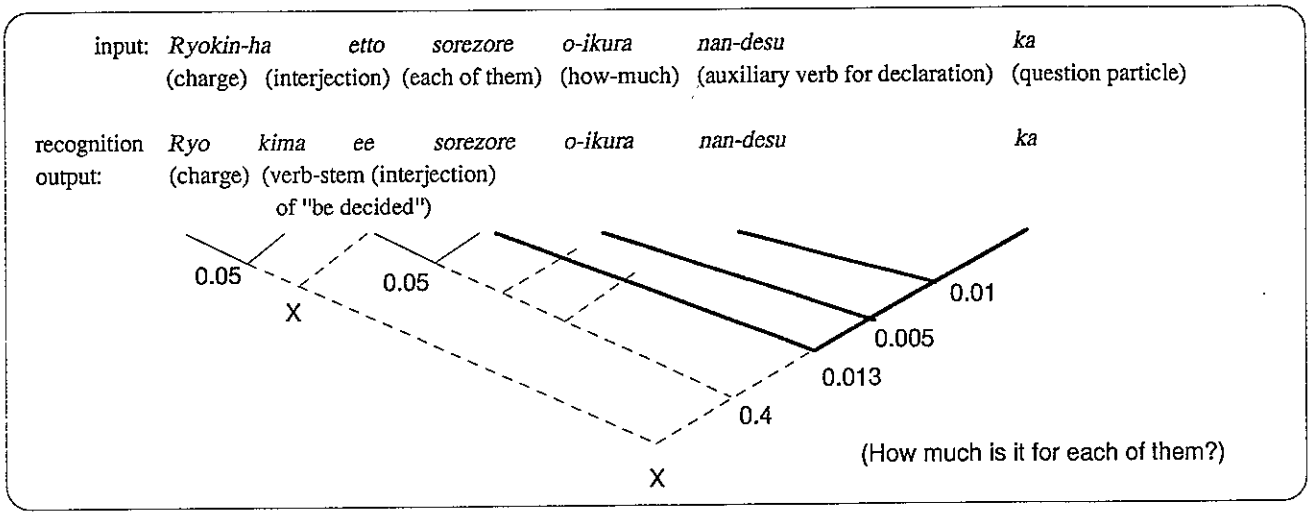

Figure 3. An example of a partial translation

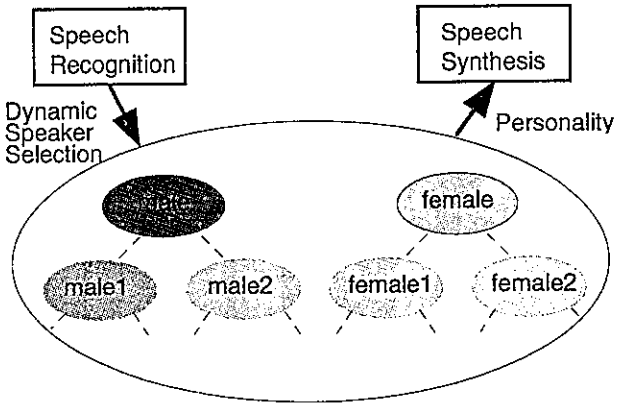

Figure 4. Personalized speech synthesis

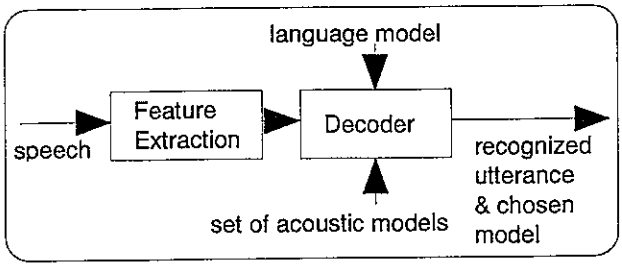

Figure 5. Dynamic speaker selection and effective search in speech recognition

system from Japanese to English is about 13,000 words, and this amount covers almost our entire bilingual travel conversation database [9]. The vocabulary for our speech recognizer is its subset.

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 [10] speech synthesis sub-system produces outputs in a male or female voice (Figure 4). It is easy for our system to be enhanced to accommodate more speakers, because unnecessary models are pruned quickly due to the efficient beam-search in the speech recognition process (Figure 5).

4. Additional features for dealing with spontaneous speech

The utterance units that serve as input to a speech translation system handling spontaneous speech are not always sentences. However, the processing units of language translation are sentences. Since we do not have enough knowledge about the sentences of spoken languages, we use the term "meaningful chunks" instead of sentences. According to our bilingual travel conversation database [9], utterance units often need to be divided into several meaningful chunks. We have proposed a method of transforming from utterance units to meaningful chunks based on pause information and the *N*-gram of fine-grained part-of-speech subcategories [11].

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

5. Implementation Issues

Running this system online for demonstration, evaluation, and assessment purposes revealed many issues that were not obvious when each sub-system was 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 a

speech segment within about 50 ms; the detection of the end, however, is much longer (almost one second). If the forward search in our speech recognition sub-system detects a long match with a pause model, then it may detect the end of the speech segment before EPD does. Should this happen, the response time would be greatly reduced. If EPD or the search decides that what was detected was not speech, nothing is output, and our speech recognition sub-system continues waiting for the operator to speak.

The second issue involves error handling. If our language translation sub-system cannot translate any of the outputs produced from the speech recognition sub-system, then our main controller commands the speech synthesis sub-system 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), and not to the audience (English).

The third issue involves feedback to the operator. The current audio input level and speech detection state are indispensable to the operator. They are available on the graphical user interface (GUI) located near the operator's face.

6. Future Works

Our plans are to develop a bidirectional speech translation system, that is, a system able to handle Japanese-to-English and English-to-Japanese, and to carry out system evaluation and assessment. Much more research is planned on understanding utterance situations, e.g., prediction of next utterances in speech recognition and disambiguation for the generation of target languages. We also plan to make multi-language outputs from ATR MATRIX possible (Japanese-to-English, German, and Korean) for the international joint experiments of C-STAR II (Consortium for Speech Translation Advanced Research).

7. Conclusion

We showed a new speech translation system called ATR-MATRIX. This system can recognize natural Japanese utterances such as those used in daily life, translate them into English, and output synthesized speech. This system can run on a workstation or a high-end PC and achieves almost real-time processing. The current implementation of our system can deal with a hotel room reservation task/domain.

Acknowledgment

The authors wish to thank all the members of ATR Interpreting Telecommunications Research Laboratories, for their contributions in building the proposed 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 Gavalda, Torsten Zeppenfeld, and Puming Zhan: "JANUS-III: Speech-to-Speech Translation in Multiple Languages," *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] Nick Campbell: "CHATR: A High-Definition Speech Re-Sequencing System," *Proc. of ASA/ASJ Joint Meeting*, pp. 1223-1228 (1996).

[11] 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)*.