A New Evaluation Method for Speech Translation Systems and a Case Study on ATR-MATRIX from Japanese to English

Toshiyuki Takezawa
ATR Interpreting Telecommunications Research Laboratories. Kyoto, Japan

Fumiaki Sugaya
ATR Interpreting Telecommunications Research Laboratories. Kyoto. Japan

Akio Yokoo
ATR Interpreting Telecommunications Research Laboratories. Kyoto. Japan

Seiichi Yamamoto
ATR Interpreting Telecommunications Research Laboratories. Kyoto. Japan

Abstract

ATR-MATRIX is a multi-lingual speech-to-speech translation system designed to facilitate communications between two parties of different languages engaged in a spontaneous conversation in a travel arrangement domain. In this paper, we propose a new evaluation method for speech translation systems. Our current focus is on measuring the robustness of a language translation sub-system, with quick calculation and low cost. Therefore, we calculate the difference between the translation output from transcription texts and the translation output from input speech by a dynamic programming method. We present the first trial experiment of this method applied to our Japanese-to-English speech translation system. We also provide related discussions on such points as error analysis and the relationship between the proposed method and translation quality evaluation manually done by humans.

1 Introduction

In the coming 21st century, the global demand for communications between speakers of different languages is expected to grow. According to recent surveys [1,2],automatic speech translation will probably come into practical use for the general public by around 2010-2020. The ideal speech translation system is expected to automatically and instantly convey the speaker's ideas in the language of the listener. Such a system is also expected to allow the speaker to speak naturally without having to give any special consideration to the translation system. However, the expectations of such an ideal speech translation system are still hampered by a mountain of technical difficulties.

NEC conducted a demonstration of a prototype speech translation system at the Telecom '83 exhibition. ATR Interpreting Telephony Research Laboratories was established in 1986 to conduct research on basic technologies for speech translation. ASURA [3] is one of the research results of ATR Interpreting Telephony Research Laboratories. This system can recognize well-formed Japanese utterances in a limited domain, translate them into both English and German, and output synthesized speech. The ASURA system was used for the International Joint Experiment on Interpreting Telephony with participants from Kyoto. Japan (ATR). Pittsburgh. USA (Carnegie Mellon University [4]), and Munich, Germany (Siemens and the University of Karlsruhe) in January 1993 [3].

Many projects on speech-to-speech translation began at that time [5, 6, 7]. SRI International and Swedish Telecom developed a prototype speech translation system that could translate queries from spoken English to spoken Swedish in the domain of air travel information systems [5]. AT&T Bell Laboratories and Telefonica Investigacion y Desarrollo developed a restricted domain spoken language translation system called VEST (Voice English/Spanish Translator) [6]. In Germany. *Verbmobil* [7] was created as a major speech-to-speech translation research project. The *Verbmobil* scenario assumes a native speaker of German and a native speaker of Japanese. Both speakers possess at least a basic knowledge of English. The *Verbmobil* system supports them by translating from their mother tongues, i.e.. Japanese and German, into English.

In 1995, KDD developed a prototype of a Japanese-to-Korean speech translation system [8] and carried out an international experiment with Korea Telecom (KT) and the Electronics and Telecommunications Research Institute (ETRI) in Korea. This KDD system can also recognize well-formed Japanese utterances in a limited domain, and translate them

into Korean.

The ATR Interpreting Telecommunications Research Laboratories was set up to meet expectations for an efficient speech translation system, that is to say. to develop technologies with which natural colloquial conversations can be easily translated. Our research activities were started in March 1993. following the lead of ATR Interpreting Telephony Research Laboratories. Recently, a Japanese-to-English speech translation system called ATR-MATRIX (ATR's Multilingual Automatic Translation System for Information Exchange) was built [9]. This system can recognize natural Japanese utterances such as those used in daily life, translate them into English, and output synthesized speech.

In this paper, we propose a new evaluation method for speech translation systems. Subjective evaluations were adopted in almost all early projects. Needless to say. subjective evaluation methods are ideal for obtaining the overall performance of speech translation prototypes. The cost of such evaluation methods, however, is very high and the turn-around time is very long. Our current focus is on measuring the robustness of a language translation sub-system, with quick calculation and low cost. Therefore, we calculate the difference between the translation output from transcription texts and the translation output from input speech by a dynamic programming method. We present the first trial experiment of this method applied to our Japanese-to-English speech translation system. We also provide related discussions on such points as error analysis and the relationship between the proposed method and translation quality evaluation manually done by humans.

In section 2. we give an overview of our Japanese-to-English speech translation system: ATR-MATRIX. In section 3. we report on the new evaluation method with a preliminary result for ATR-MATRIX and related discussions. Section 4 presents our conclusions.

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

ATR Interpreting Telecommunications Research Laboratories (ATR ITL) recently built a speech translation system called ATR-MATRIX [9]. This system can recognize natural Japanese utterances such as those used in daily life, translate them into English, and output synthesized speech. The system runs on a PC (or a workstation) and achieves nearly real-time processing. Unlike its predecessor ASURA [3], ATR-MATRIX is designed for spontaneous speech input, and it is much faster.

There have recently been many projects on speech-to-speech translation [4, 7]. *Verbmobil* [7], a major research project in Germany, adopts a combined method of deep and shallow processing. JANUS [4] is another major research project that adopts an interlingua-based language translation method. In contrast to these works, ATR-MATRIX adopts a spoken language translation method using not only sentence structures but also examples such as translation pairs [10, 11]. Moreover. ATR-MATRIX has features such as personalized speech synthesis based on dynamic speaker selection in speech recognition.

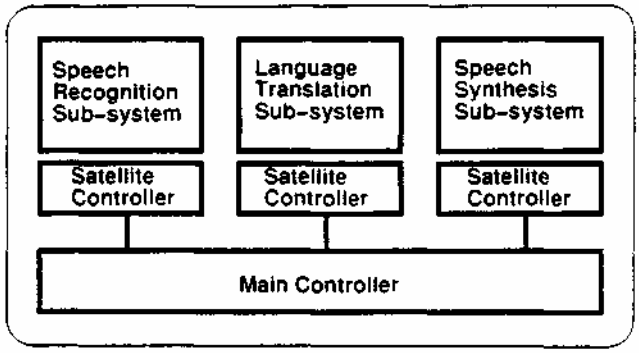

Figure 1: System configuration

2.1 System Overview

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

2.2 Spontaneous Speech Recognition

Speech features differ widely between speakers, such as between males and females, and phoneme-contexts. Therefore. ATR ITL proposed a statistical method (ML-SSS) [12] to make speaker-independent phoneme-context-dependent acoustic models. In addition, we separately prepared speaker-independent phone models for males and females using this method.

ATR ITL also proposed a language model of a variable-order N-gram [13], which is a compact language model to deal with various expressions in spontaneous speech. Real-time processing has been achieved by an effective search method based on a word-graph [14]. Dynamic speaker selection has also been achieved by using an efficient beam-search.

The speech recognition sub-system contains approximately 13,000 words, which is enough for one task/domain such as hotel room reservations (except for the problem of human names and other proper nouns).

2.3 Robust Language Translation Dealing with Speech Recognition Results

ATR ITL has established, through comparative experiments, that example-based translation methods are the most effective for handling a wide variety of natural speech translation problems. Our Example-based Machine Translation [10, 11] integrates both examples and rules in a common framework. When complex sentences are translated, the closest examples are retrieved from the database, and the dependencies among the component words are analyzed while translation equivalents are assembled.

Furthermore, a partial translation mechanism for accepting speech recognition results that include recognition errors has been introduced [15]. Two heuristics were adopted:

(1) An upper threshold is set for the semantic distance between translation pairs. Source expressions for which a matching target expression cannot be found within the threshold will not be translated.

(2) A lower threshold is set for the length of matching word sequences. If no word sequences of sufficient length can be found to match satisfactorily, then the expression will not be translated.

Figure 2 shows an example of this partial translation method, in which the first threshold is set to 0.2 and the second to 2. In the input utterance. the section *"ryokin-wa"* (which means "charge, fee") is mis-recognized as *"ryo kima",* which consists of a noun meaning "charge" and the verb stem of *"ki-maru",* meaning "be decided". Our partial translation mechanism would decline to expand the sequence *"ryo kima",* because it would be unable to find a word sequence long enough to satisfy the threshold of 2. For the remainder of the utterance in the example. the semantic distance corresponding to "*ee sorezore o-ikura nan-desu ka"* would be 0.4. This hypothetical structure would be pruned for exceeding the upper threshold for the semantic distance, and as a result, the constituent *"sorezore o-ikura nan-desu ka"* would be selected for translation. The equivalent English, "How much is it for each of them?" could then easily be generated.

The language translation sub-system from Japanese to English contains approximately 13,000 words, covering almost all of our bilingual travel conversation database [16, 17].

2.4 More Features for Dealing with 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 translations are usually sentences.

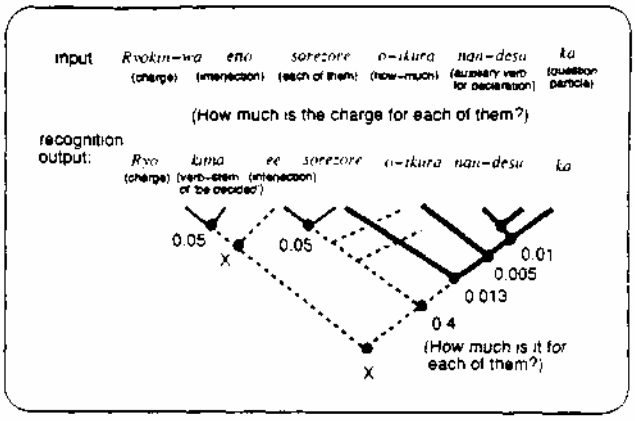

Figure 2: Example of partial translation

Since we do not have enough knowledge about sentences in spoken languages, we use the term "language processing units" instead of sentences. From a study of our bilingual travel conversation database [16, 17]. we have found that utterance units often need to be divided into several language processing units. We have proposed a method for transforming utterance units into language processing units based on pause information, the N-gram of fine-grained part-of-speech subcategories, and a few heuristics [18].

3 System Evaluation

3.1 Evaluation Method

Our current focus is on measuring the robustness of a language translation sub-system, with quick calculation and low cost. Therefore, we calculate the difference between the translation output from transcription texts and the translation output from input speech by a dynamic programming method. *Accuracy,* which is a kind of similarity measure, is calculated by the following equation.

$$Accuracy = \frac{Total - Sub - Ins - Del}{Total} \quad (1)$$

where *Total* is the total number of words in the translation outputs from the transcription texts. *Sub* is the number of substitution words comparing the translation outputs from the transcription texts and from the speech inputs. *Ins* is the number of inserted words comparing the translation outputs from the transcription texts and from the speech inputs, and *Del* is the number of deleted words comparing the translation outputs from the transcription texts and from the speech inputs.

3.2 Evaluation Experiment

First, we carried out a spontaneous speech translation experiment using our bilingual travel conversation

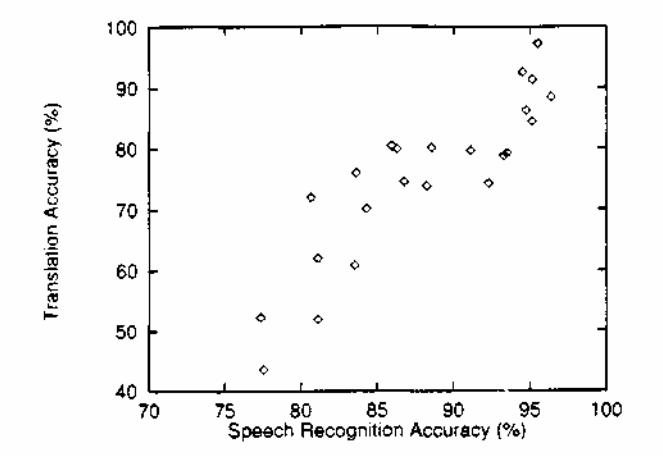

Figure 3: Relationship between spontaneous speech recognition and translation accuracy

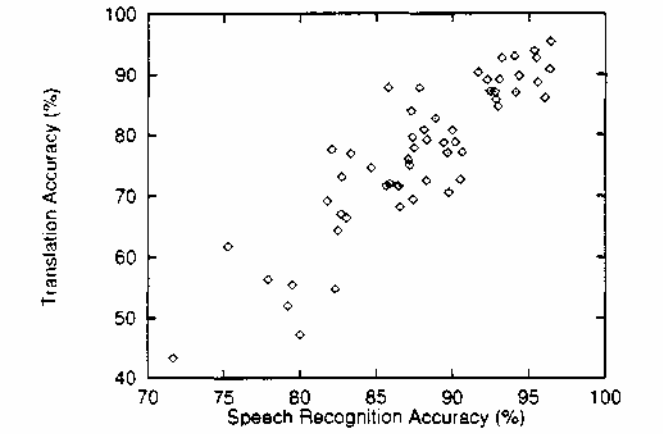

Figure 4: Relationship between read speech recognition and translation accuracy

database [16, 17]. We selected 23 conversations from the database as a test set. These 23 conversations are open for both the speech recognition sub-system and the language translation sub-system. Figure 3 shows the results. Each plot indicates the average score for one Japanese native speaker in a conversation.

Next, we carried out a read speech translation experiment of conversational texts. The transcription texts were the same as the above-mentioned 23 conversations for the test set. Fourteen female speakers and 13 male speakers read two conversational texts from the transcription files of the 23 conversations. Figure 4 shows the results. Each plot indicates the average score for one Japanese native speaker in a conversation.

Table 1 shows a summary of the first experiment of our system evaluation. This table indicates the total average of the spontaneous speech translation experiment shown in Fig. 3 and the total average of another translation experiment using the read speech of conversational texts shown in Fig. 4.

According to Fig. 3 and Fig. 4, there seems to be a correlation between the speech recognition accuracy and proposed translation accuracy.

3.3 Error Analysis

Table 2 shows an analysis of influence of one speech recognition error into a translation. The number in parentheses indicates the number of samples. There are 54 samples having one speech recognition error in one utterance in the spontaneous speech translation experiment in Fig. 3. Since Table 2 includes one-word error cases, we can find the relationship between the speech recognition error and translation output straightforwardly.

In the following, we show some typical examples, which were selected from the spontaneous speech translation experiment.

There are 22 samples (40.7%) having no error in the translation output, but having one speech recognition error in the utterance. An example is shown in the following.

The Japanese noun *"goro"* is mis-recognized as another Japanese noun *"kurai"* in example (1). In the context of this utterance, the meanings of *"goro"* and *"kurai"* are almost the same such that the output English sentences are the same.

Example (1)

Transcription:	Gogo san ji *goro* ni naru to omoi masu
Recognizer output:	Gogo san ji *kurai* ni naru to omoi masu
English output from transcription:	I think it will be around three pm
English output from recognizer output:	I think it will be around three pm
Result:	Speech recognition one error, translation no error

In example (2). the Japanese particle *"wo"* is deleted in the speech recognition output. The verb in the output English is changed from *"make"* to *"have."*

Table 1: Summary of the first experiment of our system evaluation

	Speech recognition accuracy	Translation accuracy (Proposed method)
Spontaneous speech	87.7%	76.1%
Read speech of conversational texts	91.8%	83.3%

Table 2: Analysis of influence of one speech recognition error into a translation

Spontaneous speech translation experiment		
Error word count in translation	Ratio	Breakdown
0	40.7% (22)	filled pause (9), prefix (4), sentence final expression (4), particle (2), different expressions of *kanji/kana* (1), content word (1), *de* (1)
1	20.4% (11)	human name (2), number (2), particle (2), day/day of the week (2), content word (1), filled pause (1), special case (1)
2	14.8% (8)	content word (2), money (2), potential verb (2), particle (1), filled pause (1)
3	7.4% (4)	particle (2), content word (1), adverb (1)
4	3.7% (2)	particle (1), content word (1)
5	5.6% (3)	particle (2), content word (1)
6	1.9% (1)	*de* (1)
8	1.9% (1)	content word (1)
11	1.9% (1)	content word (1)
13	1.9% (1)	*de* (1)
Total	100.0% (54)	

Example (2)

Transcription:	Anoo o heya no yoyaku *wo* o negai shi tai no desu keredomo
Recognizer output:	Anoo o heya no yoyaku o negai shi tai no desu keredomo
English output from transcription:	I'd like to *make* a reservation for a room
English output from recognizer output:	I'd like to *have* a reservation for a room
Result:	Speech recognition one error, translation one error

In example (3), the part-of-speech (POS) of the Japanese word *de* is different between the transcription and the speech recognition. The former is a case particle *de*. The latter is an auxiliary verb. Since the dependency structure is different, the surface strings also differ from each other. However, the information may be conveyed in this example.

Example (3)

Transcription:	Soreto kurejittokaado wa masutaakaado *de* 5 2 7 9 3 9 2 0 2 4 6 9 0 0 9 8 desu
Recognizer output:	Soreto kurejittokaado wa masutaakaado *de* 5 2 7 9 3 9 2 0 2 4 6 9 0 0 9 8 desu
English output from transcription:	and *it's* five two seven nine three nine two zero two four six nine zero zero nine eight *to the credit card by master card*
English output from recognizer output:	and *the credit card is master card* five two seven nine three nine two zero two four six nine zero zero nine eight
Result:	Speech recognition one error, translation 13 errors

Errors concerning particles, content words, and Japanese word *de* tend to make different surface strings in English.

3.4 Discussions

We plan to use three major evaluation methods for our speech-to-speech translation project, as shown in Fig. 5. The first one is a subjective evaluation from the transcription texts, which is done manually by humans. The purpose of this evaluation method is to obtain the translation quality of our language translation sub-system [11]. The second one is the method reported in this paper. The purpose of this method is to obtain the robustness of our language translation subsystem for accepting speech recognition outputs. The combination of these two evaluation methods indicates the overall system performance. The third one is an end-to-end evaluation using a bi-directional speech translation system between English and Japanese [19]. The purpose of this evaluation method is to measure the task achievement, such as the reservation of a hotel room, as well as to research human factors in speech translation systems.

Table 3: Results of translation quality evaluation manually by three humans

Spontaneous speech translation	
Rank	Average
(A)	40.2%
(A)+(B)	61.9%
(A)+(B)+(C)	77.0%

Table 4: Relationship between proposed translation accuracy and translation quality

Spontaneous speech translation	
Rank	Average
(A)	94.3%
(B)	83.8%
(C)	75.8%
(D)	44.9%

Table 3 shows the translation quality evaluation results for the same test set of spontaneous speech translations shown in Fig. 3. This test set is open for both the speech recognition sub-system and the language translation sub-system. The output utterances were manually graded by experienced native speakers. Each utterance was assigned one of four ranks for translation quality: (A) Perfect: no problem in both information and grammar; (B) Fair: easy-to understand with some unimportant information missing or flawed grammar; (C) Acceptable: broken but understandable with effort; (D) Nonsense: important information has been translated incorrectly. Since the "acceptability" is the sum of the (A), (B), and (C) ranks, about 77% acceptability was achieved in JE spontaneous speech translation.

Table 4 shows the average ratio of the proposed translation accuracy corresponding to each of the ranks in Table 3. Although the test set is open for both the speech recognition sub-system and the language translation sub-system, the proposed translation accuracy tends to indicate the translation quality. Therefore, the proposed translation accuracy is not only the measure for robustness of the language translation sub-system but may also become a rough score of the translation quality with quick calculation and low cost.

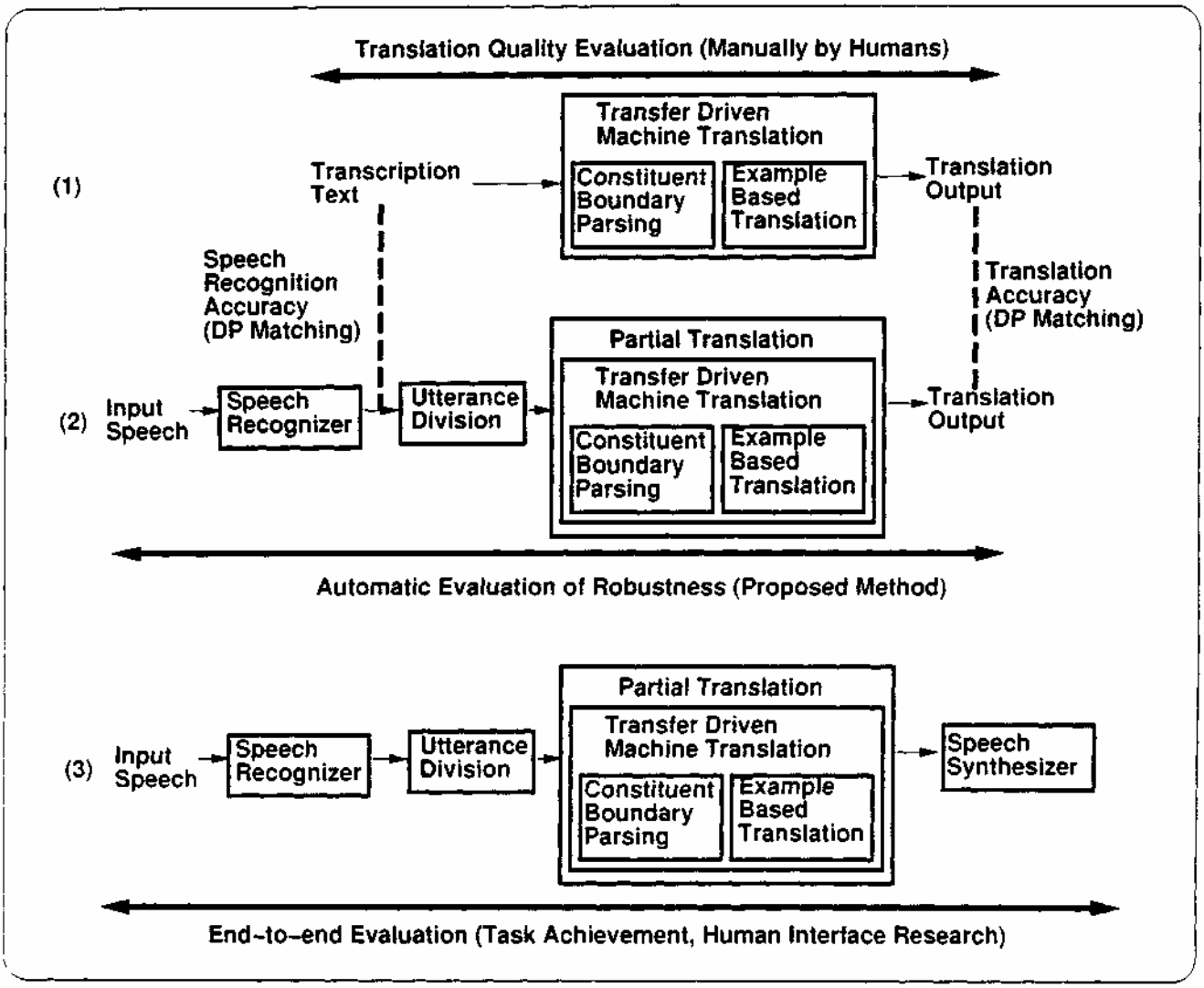

Figure 5: Three major evaluation methods for our speech-to-speech translation project

4 Conclusions

We have proposed a new evaluation method for speech translation systems. Our current focus is on measuring the robustness of a language translation subsystem, with quick calculation and low cost. Therefore, we calculated the difference between the translation output from transcription texts and the translation output from input speech by a dynamic programming method. We presented the first experiment of this method applied to our Japanese-to-English speech translation system. For further information on ATR-MATRIX. please access the following location on our Web site:

http://www.itl.atr.co.jp/matrix/

Acknowledgments

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

References

[1] Ministry of Posts and Telecommunications. http://www.mpt.go.jp/policyreports/japanese/telecouncil/tsusin/koudojouhou/FS5.html *(in Japanese).*

[2] The Institute for Future Technology, http://www.iftech.or.jp/hiroba/chronotab/chrono12.html *(in Japanese).*

[3] Morimoto, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M., and Kurematsu, A. (1993). "ATR's Speech Translation System: ASURA". In *Proceedings of EUROSPEECH '93,* pp. 1291-1294.

[4] Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavaldà, M., Zeppenfeld, T., and Zhan, P. (1997). "JANUS-III: Speech-to-Speech Translation in Multiple Language". In *Proceedings of ICASSP '97,* pp. 99-102.

[5] Rayner, M., Bretan, I., Carter, D., Collins, M., Digalakis, V., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Pulman, S., Price, P., and Samuelsson, C. (1993). "Spoken Language Translation with Mid-90's Technology: a Case Study". In *Proceedings of EUROSPEECH '93,* pp. 1299-1302.

[6] Roe, D.B., Moreno, P.J., Sproat, R.W., Pereira, F.C.N., Riley, M.D., and Macarrón, A. (1992). "A Spoken Language Translator for Restricted-domain Context-free Languages" . In *Speech Communication.* Vol. 11. pp. 311-319.

[7] Bub, T., Wahlster, W., and Waibel, A. (1997). "Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation". In *Proceedings of ICASSP '97.* pp. 71-74.

[8] Suzuki, M., Inoue, N., Yato, F., Takeda, K., and Yamamoto, S. (1995). "A Prototype of a Japanese-Korean Realtime Speech Translation System". In *Proceedings of EUROSPEECH '95,* pp.1951-1954.

[9] Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A., and Yamamoto, S. (1998). "A Japanese-to-English Speech Translation System: ATR-MATRIX". In *Proceedings of ICSLP '98* pp. 2779-2782.

[10] Mima, H., Furuse, O., Wakita, Y., and Iida, H. (1997). "Multilingual Spoken Dialogue Translation System Using Transfer Driven Machine Translation". In *Proceedings of Machine Translation Summit VI,* pp. 148-155.

[11] Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K., and Shirai, S. (1999). "Solutions to Problems Inherent in Spoken-language Translation: the Approach of ATR-MATRIX". In *Proceedings of Machine Translation Summit VII. (to appear).*

[12] Ostendorf, M., and Singer, H. (1997). "HMM Topology Design Using Maximum Likelihood Successive State Splitting". In *Computer Speech and Language.* Vol. 11. No. 1, pp. 17-41.

[13] Masataki,H., and Sagisaka, Y. (1996). "Variable-order N-gram Generation by Word-class Splitting and Consecutive Word Grouping". In *Proceedings of ICASSP '96,* pp. 188-191.

[14] Shimizu, T., Yamamoto, H., Masataki, H., Matsunaga, S., and Sagisaka, Y. (1996). "Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graph". In *Proceedings of ICASSP '96,* pp. 145-148.

[15] Wakita, Y., Kawai, J., and Iida, H. (1997). "Correct Parts Extraction from Speech Recognition Results Using Semantic Distance Calculation, and Its Application to Speech Translation". In *Proceedings of ACL/EACL Workshop on Spoken Language Translation,* pp. 24-31.

[16] Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N., and Yamazaki, Y. (1994). "A Speech and Language Database for Speech Translation Research". In *Proceedings of ICSLP '94,* pp. 1791-1794.

[17] Takezawa, T. (1999). "Building a Bilingual Travel Conversation Database for Speech Translation Research". In *Proceedings of the 2nd International Workshop on Easl-Asian Language Resources and Evaluation — Oriental COCOSDA Workshop '99 —*, pp. 17-20.

[18] Takezawa, T. (1999). "Transformation into Language Processing Units by Dividing and Connecting Utterance Units". In *Proceedings of EUROSPEECH '99. (to appear)*.

[19] Sugaya, F., Takezawa, T., Yokoo, A., and Yamamoto. S. (1999). "End-to-end Evaluation in ATR-MATRIX: Speech Translation System between English and Japanese". In *Proceedings of EUROSPEECH '99. (to appear)*.