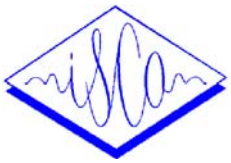

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

END-TO-END EVALUATION IN ATR-MATRIX: SPEECH TRANSLATION SYSTEM BETWEEN ENGLISH AND JAPANESE

Fumiaki Sugaya Toshiyuki Takezawa Akio Yokoo Seiichi Yamamoto
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
sugaya@itl.atr.co.jp

10.21437/Eurospeech.1999-534

ABSTRACT

ATR Interpreting Telecommunications Research Laboratories developed ATR-MATRIX speech translation system, which translates both ways between English and Japanese, enough to hold natural on-line real-time conversations. Using this system we started an end-to-end evaluation of a speech translation system through a dialog test with naive speakers who are not involved in system development and not familiar with a speech translation technology. This paper explains the speech translation design concept, evaluation system overview, evaluation procedure and some interesting results observed in the test. Finally after concluding we will mention our future plan.

1. INTRODUCTION

With great progress in speech translation technology [1,2], ATR ITL has achieved a state where a PC supports natural on-line real-time conversations between speakers of different languages. The ATR-MATRIX speech translation system [3,4] translates both ways between English and Japanese, quickly enough to hold a realistic conversation in the hotel reservation task domain. Since on-line real-time conversation through a speech translation system is a new medium for people, we need to study how people communicate using ATR-MATRIX with acoustic model, language model, and meta communication. And, to improve the performance of the system, we started end-to-end evaluation of the speech translation system. This paper explains the speech translation design concept, system overview, evaluation procedure and results. On the system overview, this paper focuses on technical features that are new in the past year, including hands-free operation, bi-directional operation.

2. SPEECH TRANSLATION DESIGN CONCEPT

Even state-of-the-art technology of speech recognition and translation cannot eliminate errors. To handle this problem, in our design we targeted making a responsive system with the processing delay as the same of speech duration, so that users can interact with the system as easily as possible and whenever he/she likes from each side. In this scheme the speaker can check the result of speech recognition or another party's responses and speak again if errors are intolerable, and if another party can't understand the result of translation, he/she can ask questions. Users can make interruptions to the system in intermediate levels and have a chance to improve the final result.

We have adjusted the system parameters, and we achieved the combined delay of recognition and translation about equal to the utterance length. Some parts of the system have time lag due to sequential architecture, but questionnaires and interviews show this time lag does not bore and frustrate speakers.

3. SYSTEM OVERVIEW AND EVALUATION PROCEDURE

3.1. System Overview

ATR-MATRIX integrates software subsystems: SPREC [5,6,7] consisting of real-time speech recognition using speaker-independent phoneme-context-dependent acoustic models, a language model of variable-order N-gram and sentence splitting [8]; TDMT [9,10] featuring robust language translation to deal with speech recognition results; and CHATR [11] for personalized speech synthesis.

Figure 1 shows a block diagram of ATR-MATRIX in the bi-directional configuration. To handle the interaction in a uniform manner between the main controller and the subsystems, we put a satellite controller where subsystems' local knowledge are so encapsulated that subsystems' modifications don't affect Maincontroller. Communication between controllers is in a packet format. The Main Controller is essentially a switch that controls the basic data flow of the system. Each subsystem' s satellite controller transforms the data from received packets to the data format and control signals necessary for each subsystem.

We use one host computer for each party: a Japanese speaker and an English speaker. Two host computers are connected through 10M LAN. The host computer for the Japanese side runs a Japanese recognizer, a Japanese-to-English translator, and a Japanese synthesizer. Table 1 shows system's specification. Figure 2a shows the Graphical User Interface (GUI) for Japanese speakers. J-tagged sentences are Japanese recognition results. An E-tagged sentence shown in italic is a translation result sent from the English host. This sentence is, however shown just for explanation. In the test the Japanese synthesizer in the Japanese host outputs the translation result, which is not shown in the GUI. The English host runs a complementary function. Sentences with the same tag like J001 in Figure 2a and Figure 2b are corresponding source and target sentences.

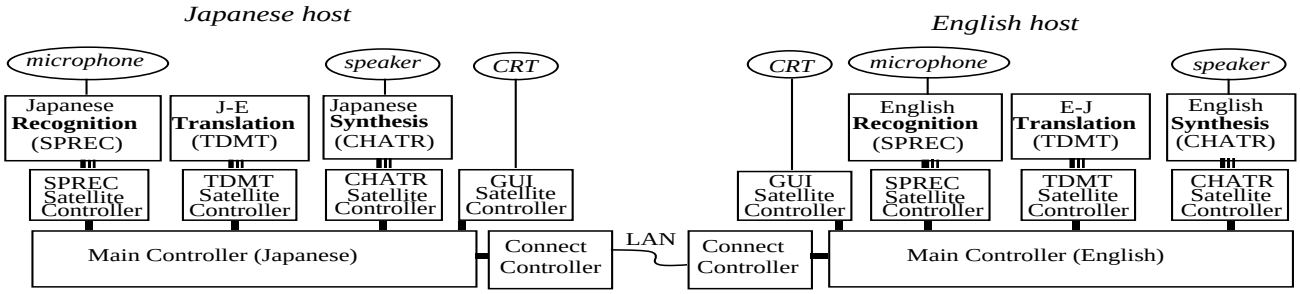

Figure 1 Basic design for bi-directinal configuration

Table 1 Specification of ATR-MATRIX

	Japanese	English
Speech recognition(SPREC)		
Lexicon	3,000 words	1,000 words
Language model	variable-order N-gram	
Translation(TDMT)		
Scheme	Transfer-driven machine translation	
Dictionary	13,000 words	8,000 words
Speech synthesis	CHATR	
Sentence splitting	Used	NA
Host computer		
CPU	Alpha (600MHz)	Dual Pentium II (450MHz)
Memory	250MB	

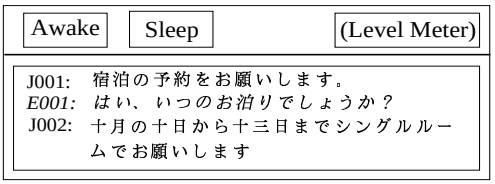

Figure 2a Japanese side

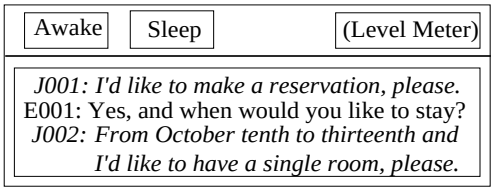

Figure 2b English side

Figure 2 The graphical user interface

Each party can see the other party with a commercial TV conference system. We disabled voice transmission function in the TV conference system, so speakers can't hear another party's direct voice. Each speaker can interrupt whenever they like to speak. It means that ATR-MATRIX does not use utterance management with prevailing "talk" and "over" buttons' operation. Instead of push-to-talk operation, streaming speech detectors are always active during system operation to trigger the speech translation processing.

3.2. Evaluation Procedure

Our test is simulating a conversation between an English speaker and a Japanese traveler who wants to reserve a hotel in English. In the test, we cast non-expert English speakers for the hotel clerk side and, for the guest side, Japanese speakers who are unfamiliar with the system. We tested 3 sessions for each person in the guest side. The number of guest speakers is 5 people. In the first session, we gave the Japanese speaker operating instructions: how to use GUI buttons. We started the second session after just a short break without guidance. In the third session, we gave guidance where we helped speakers to improve speech recognition performance, presenting how to speak using 11 sentences unrelated to the test. For English speakers, we gave essentially the same operational instructions and guidance as Japanese ones. In each session we gave guests and hotel clerks basic and simple hotel reservation task scenarios which only explain the mission, and list allowable proper nouns, reservation condition and guest ID information. After each session we collected a questionnaire. In table 2, subjective scores on speaker's satisfaction and its meaning are shown. At the end of the final session we interviewed speakers. After the test, we transcribed conversation data and analyzed from many perspectives including speech recognition and perplexity. To check the translation acceptability of the whole system, including intermediate errors, we check eleven items indispensable for the hotel reservation task and, from that, score a task achievement rate. The items consist of hotel name, customer's name, phone no., number of guests, date of stay, number of nights, room type, charge, credit card no., valid date, and check-in time.

Table 2 Question on speaker's satisfaction

5	speaker can achieve completely the task
4	speaker can achieve acceptably the task
3	speaker achieve the task without satisfaction
2	speaker can barely achieve the task
1	speaker can not achieve the task

4. THE EVALUATION RESULTS

4.1. Perplexity and Task Achievement

When speakers start a dialog, they initially use complex expressions and show hesitations that make the speech recognition difficult and result in many recognition and translation errors. In the test we allow retries, so the speaker may try initially similar expressions a few times. The determined speakers tend to use other simpler expressions to achieve the tasks, and gradually the speakers start to use simpler expressions that make ATR-MATRIX perform well. Figure 3 shows perplexity vs. session no. Perplexity decreases through the progress of sessions. Average perplexity for the first, second and third sessions are 33.03, 30.41 and 26.98. In the third session, perplexity is reduced by 18.3 % compared with the first session. This reduction is corresponding to the simpler expressions observed in conversation.

The task achievement rate was roughly 90% on the guest side. The bad scores belong to connected number speech recognition in the topic of phone no. and credit card. Since the number recognition is an indispensable topic for the hotel reservation task, we severely scored 0 point if even one digit error occurred in a sentence. The severe scoring mainly accounts to the bad score. The other reason is that perplexity of the sentences including connected digit recognition is very large as 43.80 while perplexity for other sentences exclusive of connected digit expressions is 26.03. In this test we did not use any dedicated techniques for number speech recognition. Many published connected recognition techniques [12] will, however improve the performance that will hopefully leave speakers less frustrated.

4.2. Features On Transcription

The speakers' utterances shorten to 6.1 words per sentence in the test conversation, while ATR's SLDB [13,14] averages 10.3 per sentence. We must study the reason for the shorter utterance, but it apparently improves the system performance. Figure 4 shows speech recognition performance for read text. Figure 5 shows translation performance for correct text. Using these performance the recognition rate is expected to improve by 2.2% and translation rate gains 10.5% for these shorter sentences, compared to SLDB's. The difference of utterance length also indicates the importance of collecting DB by using a real speech translation system like ATR-MATRIX.

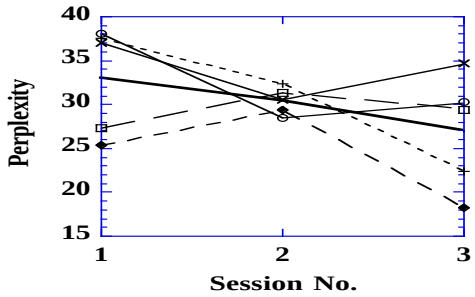

Figure 3 Perplexity

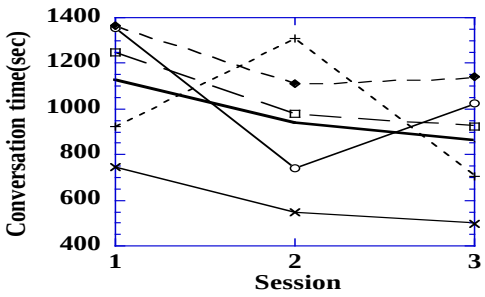

Figure 6 Conversation time

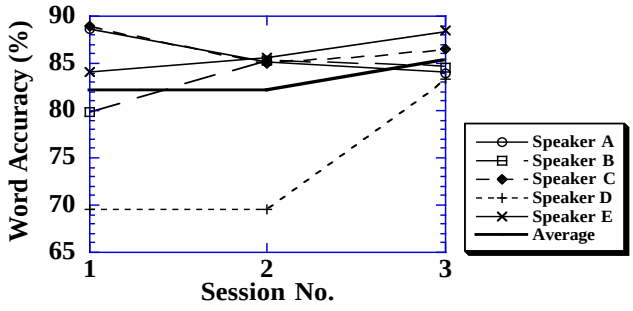

Figure 7 Speech recognition rate

4.3. Conversation Time

The conversation time shown in Figure 6 shrinks as 18 minutes 48 seconds for the first session, 15:38 for the second and 14:19 for the third. In the third session, speaker could complete efficiently the task with the rate of 76.2% than in the first session.

4.4. Speech Recognition Rate

Figure 7 shows the speech recognition rate for the conversation test. The average word accuracy for all utterances varies as 82.2, 82.1 and 85.4% for the first, second and third sessions. In Figure 7, the one speaker with the lowest recognition rate shows drastic improvement in speech recognition from one session to the next. The other speakers can be clustered into one group of higher recognition rate. If the lowest speaker's sample is removed from averaging sampling group, the average recognition rate stays almost the same at 85.4%, 85.3% and 85.9%. This result means multiple sessions and guidance is helpful for speakers with low speech recognition rate, while for speakers with high speech recognition rate do not show remarkable improvement.

In the test, speakers say similar expressions if the recognition results are not acceptable for them. Since these retries occur in the difficult expressions for the speech recognition, the total recognition rate probably decreases due to repetition of failed utterances. The average word accuracy for successful utterances, where failure utterances leading to successful utterances are deleted, is 82.9%, 82.8% and 88.4% for the first, second and third sessions. If the worst speaker is removed from the averaging sampling group, the recognition rate is 86.1%, 85.1% and 89.5%. This shows that speakers with higher recognition rate can also improve the speech recognition rate in the successful utterances if they have guidance, while speakers with low speech recognition rate can improve the speech recognition rate through the sessions and with the guidance.

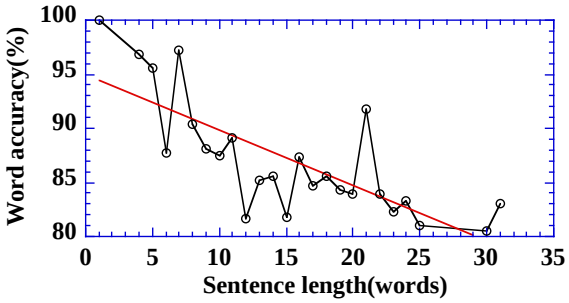

Figure 4 Speech recognition rate vs. sentence length

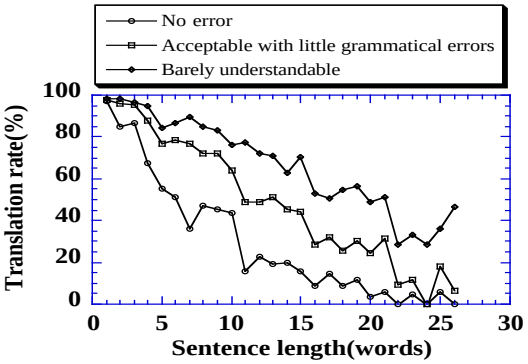

Figure 5 Translation rate vs. sentence length

4.5. Subjective Score

Table 3 shows subjective scores and other summarized results. The subjective score in the third session is 3.8 which means speakers could mostly achieve acceptably the task. In summarized data in Table 3, session by session, the subjective score, the speech recognition error rate (=100% - word accuracy), the perplexity, and the conversation time decrease dramatically. The improvement rate is not trivial. Through the test, members in ATR could give a helpful and useful guidance to speakers. It is, however so labor intensive. Probably an automatic guidance system will help speakers, especially speakers with low speech recognition rate use effectively the speech translation system.

Table 3 Subjective score and summary of results

Session	Subjective Score	Recognition Error(%)	Perplexity	Conversation Time(sec)
1	3.4	17.58	33.03	1127.8
2	3.8	17.49	30.41	937.8
3	3.8	14.02	26.98	859

4.6. TV conference System

In the preparatory test without TV conference system, we observed many collisions in which speakers interrupt each other and then stop speaking. With this evaluation test, we observed that speakers could avoid collisions with the help of TV conference system. The questionnaire also shows that the TV conference system helps to put speakers relaxed.

5. CONCLUSION AND FUTURE WORK

We showed that ATR-MATRIX achieves a score of 90 % for the basic hotel reservation task in the user side. And we also show some interesting and useful results observed in the test. (1) Multiple sessions are helpful for speakers with low speech recognition rate. Speakers showing high recognition rate in the first session don't show any improvement. Guidance is, however helpful for both types of speakers. (2) The utterance length shortens in the communication with the speech translation system that makes ATR-MATRIX tend to perform well. (3) The learning and training help speakers to use the system efficiently. (4) Introducing TV conference system indicates to help speakers to efficiently and comfortably use the system.

In this evaluation, the domain is limited to basic hotel reservation; in the next evaluation test we are planning to enlarge the domain to cover all travel conversation with the vocabulary size of 13,000 Japanese words including claim handling.

ACKNOWLEDGEMENT

We would like to thank every department which provided us software modules and all members to support our work.

REFERENCES

[1] 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).

[2] 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 Language," *Proc. of ICASSP '97*, pp. 99-102 (1997).

[3] Ben Reaves, Atsushi Nishino, Harald Singer, Ken Fujisawa, Setsuo Yamada, Fumiaki Sugaya, Toshiyuki Takezawa, Akio Yokoo, Seiichi Yamamoto, "ATR-MATRIX: A Speech Translation System Between English and Japanese," *Proc. of 58 th IPSJ Convention*, Spring 1999.

[4] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida. F. Sugaya, A. Yokoo, S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX," *Proc. ICSLP 1998*, pp. 2779-2782.

[5] 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).

[6] 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).

[7] 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).

[8] Toshiyuki Takezawa: "Transformation into Language Processing Units by Dividing and Connecting Utterance Units," *Eurospeech99* (1999-09)

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

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

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

[12] Hisashi Kawai, Norio Higuchi: "Recognition of Connected Digit Speech in Japanese Collected over the Telephone Network," *Proc. of ASA/ASJ Joint Meeting*, pp. 1223-1228 (1996).

[13] 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).

[14] 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).