Assessing Degradation of Spoken Language Translation by Measuring Speech Recognizer's Output against Non-native Speakers' Listening Capabilities

Toshiyuki TAKEZAWA, Keiji YASUDA,
Masahide MIZUSHIMA, and Genichiro KIKUI
ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
{toshiyuki.takezawa, keiji.yasuda, masahide.mizushima, genichiro.kikui}@atr.jp

Abstract

We propose a new evaluation method of spoken language translation by measuring a speech recognizer's output against non-native speakers' listening results. The proposed method consists of two kinds of measurements. One is a comparison of listening capabilities by measuring the speech recognizer's output against the listening results of non-native speakers. The other is an assessment of the degradation of machine translation (MT) by measuring an MT output from the speech recognition result against MT outputs from non-native speakers' listening results. We show that the change in speaking style degrades not only spoken language translation by machine but also the English listening capability of Japanese native speakers.

1 Introduction

A machine translation (MT) system for speech-to-speech translation must accept an automatic speech recognizer's output. However, state-of-the-art speech recognition by machine cannot avoid errors.

Basically, a human language is not the mother tongue of machines. As for humans it is easy to listen to other humans if the language spoken is their mother tongue. However, if the language is not their mother tongue, it is generally difficult to listen to others, particularly when there exist differences in speaking styles or environmental noise levels.

Some factors may cause severe degradation of automatic speech recognition by machine but no damage to a native speaker's listening capability. If such factors caused severe degradation of a non-native speaker's listening capability, it would be meaningful to compare automatic speech recognition by machine against the non-native speaker's listening capability. Moreover, in a conversational interface such as speech-to-speech translation, automatic speech recognition is usually used in combination with application systems such as MT. In such situations, we have to consider not only word accuracy but also the degradation of application systems such as MT.

In this paper, we propose a new evaluation method of spoken language translation by measuring a speech recognizer's output against non-native speakers' listening results. The proposed method consists of two kinds of measurements. One is a comparison of listening capabilities by measuring the speech recognizer's output against the listening results of a non-native speaker. The other is an assessment of the degradation of MT by measuring an MT output from the speech recognition result against MT outputs from non-native speakers' listening results.

At ATR, we have been collecting bilingual spoken dialogues for speech-to-speech translation research (Takezawa and Kikui, 2003; Takezawa and Kikui, 2004). From these accumulated data, we selected several types of test sets for evaluation. Using these test sets, we collected English transcription data by many Japanese natives who have various levels of English language skill. We assume that the English language skills of Japanese natives can be estimated by TOEIC (Test of English for International Communication) (TOEIC, 2005), which is one of the most commonly used English tests in Japan. We show that changes in speaking style degrade not only spoken language translation by machine but also the English listening capabilities of Japanese native speakers. We attempt to estimate the TOEIC score of a system by using an appropriate relationship of regression analysis.

This paper is organized as follows. Section 2 gives a comparison of listening capabilities, then Section 3 presents our assessment of the degradation of MT. Section 4 offers some discussions and mentions related works. Finally, we give our conclusions in Section 5.

Utterances including fillers	
Monolingual dialogues (SDB/TRA)	29.4%
Human-aided dialogues (SLDB)	16.3%
Machine-aided dialogues (MAD3)	13.8%
Machine-aided dialogues (MAD4)	6.2%

Table 1: Evidence of changes in speaking style

2 Comparison of Listening Capabilities

2.1 Characteristics of Test Sets

The test sets used in this paper were selected from MAD (MT-Aided bilingual Dialogues), which is data of dialogue spoken between English and Japanese speakers (Takezawa and Kikui, 2003; Takezawa and Kikui, 2004). We conducted several data collection experiments by changing various conditions. Among them, we use two kinds of test sets, called MAD3 and MAD4. Since speech-to-speech translation systems are currently under development, we employed human typists instead of speech recognition systems to collect high-quality data for research on MAD3 and MAD4. User instructions for MAD3 were different from those for MAD4 (Takezawa and Kikui, 2004). In MAD3, we employed instructions such as "one utterance must be made within ten seconds" so that the speaking style of users would be a rather relaxed one. In MAD4, we instructed users to speak briefly and concisely so that the speaking style of users would be a rather tense one. Table 1 shows the rate of utterances including fillers as one of the parameters of a change in speaking style. For reference, monolingual dialogues (SDB/TRA) (Takezawa et al., 2004) and bilingual dialogues through human interpreters (SLDB) (Takezawa et al., 2004) are also shown.

According to Table 1, utterances including fillers of human-aided dialogues are fewer than those of monolingual dialogues. The characteristics of MAD3 are similar to those of human-aided dialogues, but MAD4 has fewer utterances including fillers.

Table 2 shows an overview of the test sets. The average utterance length of MAD4 is also shorter than that of MAD3. Experiments using MAD3 and MAD4 were both conducted in the same room so that the environmental noise level and recording conditions would be the same. The tasks given to the dialogue participants of both MAD3 and MAD4 were also the same.

	Speakers	Utt.	Words	Length
MAD3	6	504	5,709	11.33
MAD4	12	502	4,694	9.35

Table 2: English test sets

	Word acc.	Perplexity	OOV rate
MAD3	77.9%	55.3	0.65%
MAD4	86.4%	39.8	0.05%

Table 3: English speech recognition results

2.2 Automatic Speech Recognition

For the English speech recognition system, we employed ATRASR, which is a speech recognition system developed by ATR (Itoh et al., 2004). For the experiment's language model, we employed a multi-class composite bigram trained by the ATR corpus for the travel domain (Kikui et al., 2003). Table 3 shows the experimental results of MAD3 and MAD4. The word accuracy of MAD3 is different from that of MAD4 because the values of test set perplexity and OOV (out-of-vocabulary) rate differ.

2.3 Experimental Results

Using these two test sets, MAD3 and MAD4, we collected English transcription data by Japanese native speakers who have various levels of English language skill. We assume that the English language skills of Japanese natives can be estimated by TOEIC (TOEIC, 2005). For each test set, we collected transcription data from 21 subjects. The range of their TOEIC scores was between the 300s and 900s; every 100-point range included three subjects. There were no duplications between subjects of MAD3 and MAD4.

The subjects were asked to listen to English speech and then type its transcription onto computers. They could listen to each utterance twice. Because they were allowed to use neither a dictionary nor a spell checker, the data contained various numerical writing styles, typos and so on. We calculated the accuracy of the data with the same tool used for calculating the speech recognition accuracy.

Figure 1 shows the results of the MAD3 English listening experiment, while Fig. 2 shows the results of the MAD4 English listening experiment. Each plotted point in these figures represents a subject. Regression lines are also shown in the figures. Table 4 provides the corre-

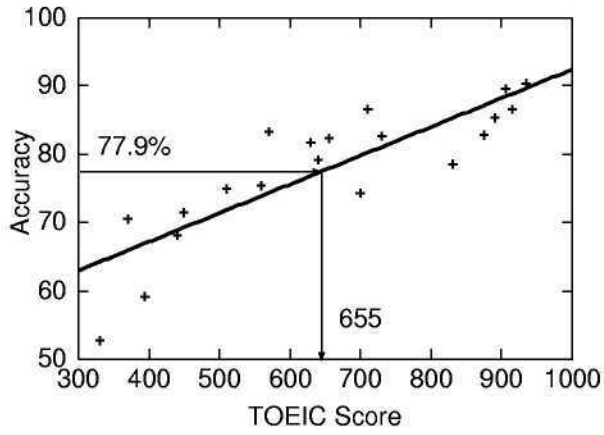

Figure 1: MAD3 English listening experiment

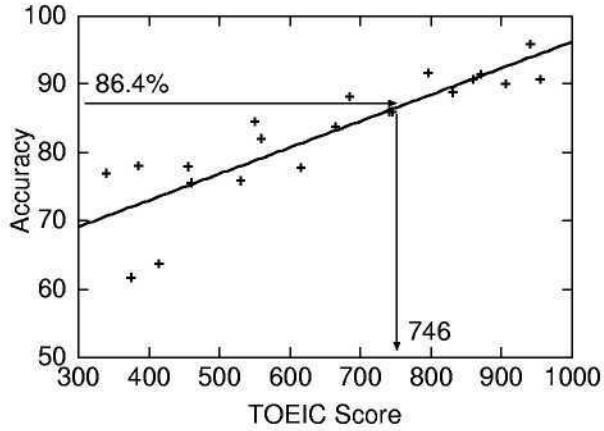

Figure 2: MAD4 English listening experiment

lation between TOEIC score and transcription accuracy.

According to Figs. 1 and 2, for humans the accuracy of MAD4 is higher than that of MAD3. According to Table 3, for machines, the accuracy of MAD4 is again higher than that of MAD3. Using this relationship, we try to estimate the TOEIC score of the speech recognition system through the regression line. Table 5 presents the estimated TOEIC score of the system.

	Correlation
MAD3	0.848
MAD4	0.868

Table 4: Correlation between TOEIC scores and accuracy

	Estimated TOEIC score
MAD3	655
MAD4	746

Table 5: Estimated TOEIC scores achieved by the system

	Confidence interval
MAD3	±81
MAD4	±82

Table 6: Confidence interval of the estimated TOEIC scores

2.4 Confidence Interval of System's TOEIC Score

We assume the following regression expression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (i = 1, 2, \cdots, n), \quad (1)$$

where X_i is the TOEIC score of each subject i, Y_i is the listening capability of the subject i, ε_i is the error term, and n is the number of subjects.

In addition, we assume that the error term (ε_i) satisfies the following conditions:

(a) $E(\varepsilon_i) = 0$,

(b) $V(\varepsilon_i^2) = \sigma^2 \qquad (i = 1, 2, \cdots, n)$,

(c) $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \qquad$ if $i \neq j$.

According to the previous study (Sugaya et al., 2001), under the above conditions the standard deviation of the system's TOEIC score (σ_{TOEIC}) is

$$\sigma_{TOEIC} = \left|\frac{\sigma}{\beta_1}\right| \sqrt{\frac{1}{n} + \frac{(C_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}} \qquad (2)$$

where C_0 is the system's TOEIC score, and $\bar{X}$ is the average of the human non-native subjects' TOEIC scores.

Using a t-distribution, the confidence interval (CI) of the system's TOEIC score with confidence coefficient $1 - \alpha$ (in this study, we employ 0.01 for α) is given by

$$CI = [C_0 - I, C_0 + I], \qquad (3)$$

$$I = \sigma_{TOEIC} \times t(\frac{\alpha}{2}; n - 2). \qquad (4)$$

Table 6 shows the confidence interval for the estimated TOEIC scores.

2.5 Consideration

If we assume that the English language skills of Japanese native speakers can be estimated by TOEIC, the correlation of TOEIC scores and the accuracy of English dictations is relatively high and good enough for experiments: A test set that makes it relatively difficult for a machine to recognize speech also makes it relatively difficult for a non-native speaker to listen to speech.

Results show that the estimated TOEIC score of the system with MAD3 is lower than that with MAD4. This may suggest that factors such as changes in speaking style more severely degrade a machine's recognition than a human non-native speaker's listening skill.

3 Assessing Degradation of MT

3.1 Preparing Experimental Data

In the previous experiment we evaluated results based on word accuracy. In conversational interfaces such as speech-to-speech translation, automatic speech recognition is usually achieved by combining the interface with an application system such as MT. In such situations, we have to consider not only word accuracy but also the degradation of application systems.

We employed MT as an application system and conducted an experiment. We used SAT, which is a statistical MT system developed by ATR (Watanabe et al., 2002). In order to conduct our MT experiment, we prepared data with sufficient quality for input. As mentioned above, the transcription data by Japanese native speakers contains various numerical writing styles, typos and so on. We fixed the writing styles of numbers and modified typos to the level of sufficient quality for input into the SAT system. After refining the data of MAD3 and MAD4, we conducted an experiment using these refined data.

3.2 Experimental Results

Using the refined data, we calculated word error rates (WER), using a calculation tool for evaluating MT, not one for speech recognition. Figure 3 shows the results for MAD3 and MAD4. Each plotted point in the figure represents a subject, and the regression lines are also present. Table 7 shows the correlations between TOEIC scores and WER of subjects, WER of the speech recognition system, the estimated TOEIC score of the system, and the confidence interval. The speech recognition accuracy in Table 3 is calculated by considering not only surface forms but also part-of-speech (POS) information, whereas the WER in Table 7 is calculated by considering only surface forms.

	MAD3	MAD4
Correlation	−0.858	−0.890
WER of the system	0.156	0.0889
Estimated TOEIC score	721	785
Confidence interval	±81	±80

Table 7: Results of English listening experiment using refined data

	MAD3	MAD4
Correlation	−0.600	−0.798
mWER of the system	0.642	0.562
Estimated TOEIC score	132	750
Confidence interval	±477	±109

Table 8: Results of MT experiment using refined data

Using the refined data, we conducted an MT experiment, preparing fifteen reference translations for each utterance and calculating the value of multi-reference word error rate (mWER). Figure 4 shows the results for MAD3 and MAD4, where each point in the figure represents a subject. The regression lines are also shown. Table 8 shows the correlations between TOEIC scores and mWER of MT outputs from subjects' transcriptions, mWER of MT output from the speech recognition results, the estimated TOEIC score of the system, and the confidence interval.

3.3 Consideration

Regarding the English listening experiments, the difference in the estimated TOEIC scores based on the refined data of MAD3 and MAD4 (Table 7) is smaller than that based on the original data of MAD3 and MAD4 (Table 5). The confidence intervals of these results are almost the same (Table 6 and Table 7). The absolute values of correlation and estimated TOEIC score in Table 7, which are based on refined data, are higher than those of the previous experiment shown in Tables 4 and 5, which are based on original data.

As for the MT experiments, the absolute value of correlation of MAD4 is relatively high, and the mWER of the MAD4 system is much

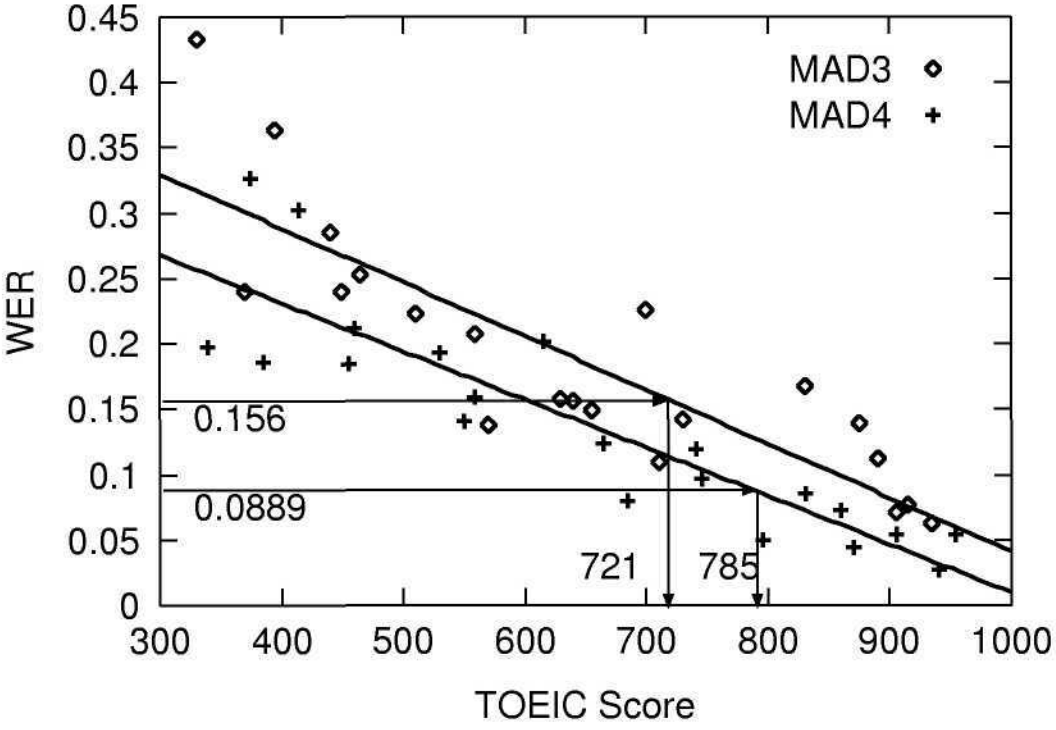

Figure 3: English listening experiment using refined data

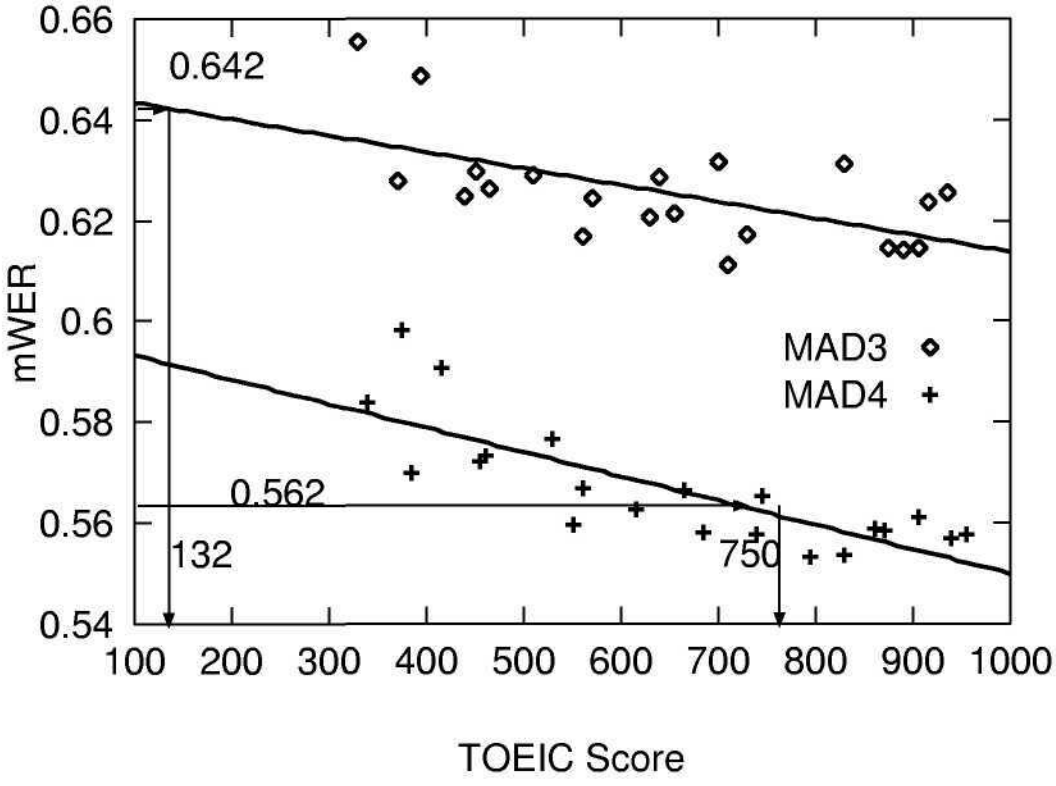

Figure 4: MT experiment using refined data

better than that of the MAD3 system (Table 8). This suggests that the MT system at ATR tends to translate expressions in MAD4 better than those in MAD3.

According to Table 7, the estimated TOEIC score obtained by the MAD4 listening experiment is 785 with a confidence interval of ±80. According to Table 8, the estimated TOEIC score by the MAD4 MT experiment is 750 with a confidence interval of ±109. Consequently, these values for MAD4 are almost the same. The speech recognition accuracy of MAD4, which is more than 85% when considering both surface forms and POS information and more than 90% when considering only surface forms, is intuitively good and may be considered sufficient for spoken language translation.

On the other hand, the estimated TOEIC score obtained by the MAD3 listening experiment is 721 with a confidence interval of ±81, while the estimated TOEIC score by the MAD3 MT experiment is 132 with a confidence interval of ±477. The degradation of the MAD3 MT experiment is thus very large. The speech recognition accuracy of MAD3, which is less than 80% when considering both surface forms and POS information and about 85% when considering only surface forms, is intuitively not so good and may be considered inadequate for spoken language translation.

According to (TOEIC, 2005), a subject whose TOEIC score is more than 730 can satisfactorily communicate in English. From the English listening experiments, we determined that the estimated TOEIC scores based on the refined data of MAD3 and MAD4 are nearly equal, or more than 730 (Table 7). If the speaking style was a rather tense one like that of MAD4, the recognition accuracy of both non-native speakers and the machine might be relatively good, and the degradation levels of MT from both non-native speakers and the machine would be comparable. If the speaking style was a rather relaxed one like that of MAD3, the recognition accuracy of both non-native speakers and the machine might be relatively poor, and the degradation level of MT from the recognizer's output would be larger than that from the non-native speakers' results.

The MT system itself might be able to deal with expressions in MAD4 better than those in MAD3 because both the correlation and mWER of the MAD4 system were better than those of MAD3 as shown in Table 8.

3.4 Summary and Implication

We obtained reasonable results for MAD4 experiments but found that the degradation of MAD3 experiments were very large. As Table 2 shows, the average utterance length of MAD4 is shorter than that of MAD3. As shown in Table 3, the values of test set perplexity and OOV rate of MAD4 are smaller than those of MAD3. As Table 1 indicates, MAD4 has fewer utterances including fillers than MAD3. These are evidence of the change in speaking style between MAD3 and MAD4. Such differences were caused by the instructions to users when collecting dialogue data. In MAD3, we employed instructions such as "one utterance must be made within ten seconds" so that the speaking style of users would be a rather relaxed one. In MAD4, we instructed users to speak briefly and concisely so that the speaking style of users would be a rather tense one.

Human non-natives tended to type the transcriptions to the extent to which they could listen and understand. However, a machine tends to output the recognition result according to the utterance length. Thus, transcriptions by human non-natives tend to become shorter than the recognition output by machine if the test set becomes more difficult to hear and understand; the MAD3 test set is considered to be more difficult than the MAD4 test set. There may also be many word deletion errors in MAD3 for human non-natives while there may be many word substitution errors in MAD3 for the machine. Such types of mis-matches may be one of the reasons for the degradation of the MAD3 test set compared to the MAD4 test set. Future work will include quantitative analysis of this.

4 Discussions and Related Works

4.1 Discussions

There have been many research and development activities devoted to spoken language processing technologies and systems. It is usually difficult, however, to compare one particular system with another. To overcome this difficulty, researchers and engineers have recently organized international workshops to study evaluation using common data, such as (IWSLT, 2004). Since the participating institutions use common data for training and testing, they can compare their evaluation results with each other at these meetings. However, it is usually very difficult for non-experts to understand the evaluation results and the information

obtained from such comparisons.

Being able to measure a spoken language processing system by comparing it with a human's language skill would be useful for non-experts to understand the current level of technology. Users of a speech-to-speech translation system could thus easily understand its performance level if we could say, for example, that the estimated TOEIC score of the English speech recognition system or the speech translation system from English to Japanese was 730. Consequently, non-experts could expect such a system to be very useful for Japanese with TOEIC scores below 730, a range that encompasses the great majority of Japanese people.

The MT experiments of MAD3 and MAD4 suggest that recognition accuracy is still an important factor in developing a system that can reduce the degradation of MT.

4.2 Related Works

Some research has been conducted to measure machine output against human results (Sugaya et al., 2001; Yasuda et al., 2005). Sugaya et al. proposed an evaluation method for a speech translation system from Japanese to English by paired comparisons of machine output against human results (Sugaya et al., 2001). Based on that system, Yasuda et al. proposed an automatic method by using an appropriate relationship of regression analysis (Yasuda et al., 2005). The method proposed in this paper may be considered its application to automatic speech recognition, in other words, a speech translation system from English to Japanese; however, the proposed method considers not only a comparison of listening capabilities but also the degradation of MT according to the recognizer output.

5 Conclusions

We proposed a new evaluation method of spoken language translation by measuring a speech recognizer's output against non-native speakers' listening results. The proposed method consisted of two kinds of measurements. One was a comparison of listening capabilities by measuring the speech recognizer's output against listening results of non-native speakers. The other was an assessment of the degradation of MT by measuring an MT output from the speech recognition result against MT outputs from non-native speakers' listening results. We demonstrated that the change in speaking style degrades not only spoken language translation by machine but also the English listening capability of Japanese native speakers. Assuming that the English language skills of Japanese native speakers could be estimated by TOEIC, we attempted to estimate the TOEIC score of a system by using an appropriate regression analysis relationship. This approach is expected to be useful for non-experts to understand the current level of spoken language translation technology.

6 Acknowledgments

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology (NiCT) entitled, "A study of speech dialogue translation technology based on a large corpus."

References

G. Itoh, Y. Ashikari, T. Jitsuhiro, and S. Nakamura. 2004. Summary and evaluation of speech recognition integrated environment ATRASR. In *Autumn Meeting of the Acoustical Society of Japan*, volume I, pages 221–222.

IWSLT. 2004. International workshop on spoken language translation — Evaluation campaign on spoken language translation —. ICSLP Satellite Workshop.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc. 8th European Conference on Speech Communication and Technology*, pages 381–384.

F. Sugaya, T. Takezawa, A. Yokoo, and S. Yamamoto. 2001. Proposal of an evaluation method for speech translation capability by comparing a speech translation system with humans and experiments using the method. *IEICE Trans. Inf. & Syst.*, J84-D-II(11):2362–2370.

T. Takezawa and G. Kikui. 2003. Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. In *Proc. 8th European Conference on Speech Communication and Technology*, pages 2757–2760.

T. Takezawa and G. Kikui. 2004. A comparative study on human communication behaviors and linguistic characteristics for speech-to-speech translation. In *Proc. 4th International Conference on Language Resources and Evaluation*, pages 1589–1592.

T. Takezawa, G. Kikui, A. Nakamura, Y. Sagisaka, and S. Yamamoto. 2004. Spoken lan-

guage corpora development at ATR. In *Proc. 18th International Congress on Acoustics*, pages 401–404.

TOEIC. 2005. Test of English for International Communication. http://www.toeic.com/.

T. Watanabe, K. Imamura, and E. Sumita. 2002. A statistical machine translation based on hierarchical phrase alignment. In *Proc. 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 188–198.

K. Yasuda, F. Sugaya, T. Takezawa, G. Kikui, S. Yamamoto, and M. Yanagida. 2005. An objective method for evaluating speech translation system: Using a second language learner's corpus. *IEICE Trans. Inf. & Syst.*, E88-D(3):569–577.