Evaluation of a Simultaneous Interpretation System and Analysis of Speech Log for User Experience Assessment

Akiko Sakamoto, Kazuhiko Abe, Kazuo Sumita and Satoshi Kamatani

Knowledge Media Laboratory,
Corporate Research & Development Center,
Toshiba Corporation
akiko7.sakamoto@toshiba.co.jp

Abstract

This paper focuses on the user experience (UX) of a simultaneous interpretation system for face-to-face conversation between two users. To assess the UX of the system, we first made a transcript of the speech of users recorded during a task-based evaluation experiment and then analyzed user speech from the viewpoint of UX.

In a task-based evaluation experiment, 44 tasks out of 45 tasks were solved. The solved task ratio was 97.8%. This indicates that the system can effectively provide interpretation to enable users to solve tasks. However, we found that users repeated speech due to errors in automatic speech recognition (ASR) or machine translation (MT). Users repeated clauses 1.8 times on average. Users seemed to repeat themselves until they received a response from their partner users.

In addition, we found that after approximately 3.6 repetitions, users would change their words to avoid errors in ASR or MT and to evoke a response from their partner users.

1. Introduction

This paper focuses on user experience (UX) of our simultaneous interpretation system ([1], Figure 1), which is a variation of a speech-to-speech translation (S2ST) system.

The goal of this paper is to assess whether users are satisfied with the whole conversation process when they use the simultaneous interpretation system and to evaluate whether the system provides interpretation of a quality sufficient for users to obtain information from speakers of other languages.

To assess the UX, we analyzed the transcription of recorded speech during a task-based evaluation experiment. The simultaneous interpretation system consists of several modules: automatic speech recognition (ASR), sentence boundary detection (SBD), machine translation (MT), and user interface (UI). However, from the viewpoint of a user, the whole system is one application. This is why we chose a task-based evaluation experiment when trying to assess UX.

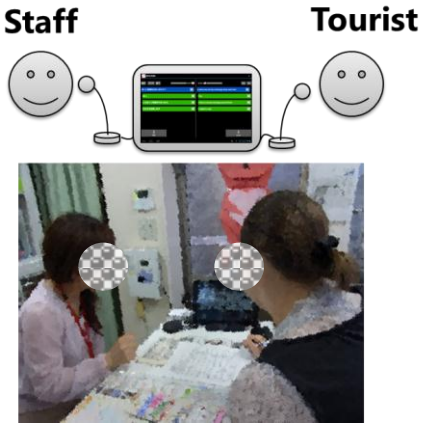

Figure 1*: Our simultaneous interpretation system and users*

Section 2 introduces related work. Section 3 introduces the system that we developed and used for the evaluation experiment. Section 4 describes the evaluation experiment. In section 5, we analyze a transcript of speech recorded during the evaluation experiment and also explore some methods to detect whether users are satisfied with the whole experience of using our system. Section 6 provides a summary of this paper.

2. Related Work

Many studies have targeted S2ST ([2], [3], and [4]). In the early stage of S2ST technology studies, systems were restricted to certain topics and speech styles. Recently, systems that can incrementally interpret utterances have been developed ([5], [6]). Some of them are commercially available [8]. Some complex applications are targeted by S2ST systems, such as lecture interpretation [9].

Most previous studies of S2ST systems have evaluated these systems in terms of recognition, translation accuracy and time efficiency. For example, one simultaneous interpretation system reportedly shortened by 20% the time needed for interpretation

without an accompanying decrease in quality [7].

When developing a simultaneous interpretation system, it is important to evaluate the precision of the interpretation and its time efficiency. In addition, it is important to consider the experience of users during actual use of the system.

Many systems implicitly expect that users will speak rather clearly and fluently. However, those users who are interested in receiving information (e.g., information about shopping), rather than in conversation with the other speaker, do not pay much attention to learning how to use the system. We observed this habit in the conversation of users during task-based evaluation.

Because simultaneous interpretation systems will soon be put to practical use, it is important to pay attention to the UX for the system. It has not been sufficiently discussed what kind of support and UX the system provides. There are few reports on the UX for simultaneous interpretation systems. Here, we focus on the number of repetitions of speech. In the experiment that we discuss in section 4, users repeated similar utterances until the ASR system recognized their speech correctly or until the other speaker responded. We also counted how many times a user would repeat something before changing the spoken words to avoid ASR or MT errors and obtain correct interpretation results and a response from the other user. This means that errors in the ASR or MT system interrupt conversation and decrease user satisfaction.

This paper discusses the UX of the simultaneous interpretation system as measured by repetition of qualitatively identical speech. This paper proposes a guiding principle for developing a practical system of simultaneous interpretation. We developed our own simultaneous interpretation system and evaluated it in terms of conversation goal achievement. We also transcribed speech recorded during the task-based experiment and analyzed how the users spoke.

3. System Architecture

We introduce our simultaneous interpretation system here to clarify the experimental conditions. The simultaneous interpretation system comprises ASR, SBD, MT, and UI components. Figure 2 illustrates the simultaneous interpretation process. The server side engines of the ASR, SBD, and MT components communicate with the UI application, which works as a client terminal through the Internet.

First, the system recognizes the user's spontaneous speech, segmented by 200 ms of pause,

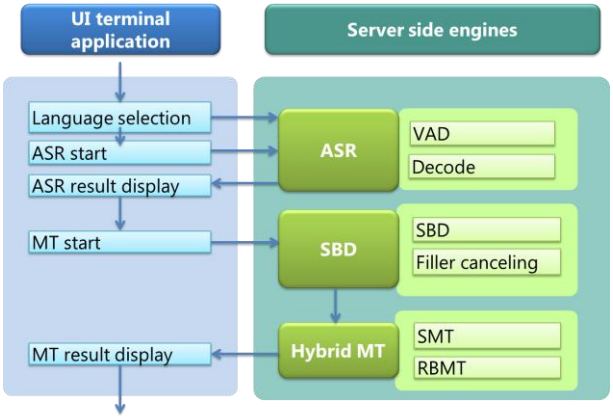

Figure 2*: Schematic diagram of speech production*

Figure 3*: Schematic diagram of speech production*

and then the system continuously outputs a transcribed text. Second, the client terminal UI application gathers several speech segments and sends them to the SBD module. Segments are gathered only when the pause between them are shorter than 500 ms. The SBD module detects a sentence boundary to split the text into segments suitable for translation. Next, the SBD module examines each segment to see whether it needs to be translated. Segments are translated in the order of their speech. This procedure enables the system to start the MT process without waiting for the end of the whole speech by a speaker and to interpret users' utterances after only a short delay for the original user's utterance. In addition, when a user presses a button for text-to-speech (TTS), the TTS engine synthesizes a voice sound for the translation result.

Figure 3 shows an example of the process. The original speech "Excuse me, I lost <pause> a bag at the train station" contains a pause longer than 200 ms between "lost" and "a." Therefore, the ASR engine regards them as separate speech segments of "excuse me i lost" and "a bag at the train station". Next, the

UI application gathers these ASR results and sends them for SBD. The SBD module examines the whole string "excuse me i lost a bag at the train station" and finds a boundary suitable for translation. In the example, SBD found a boundary between "me" and "lost." The system finally outputs the interpretation result for "excuse me" and "i lost a bag at the train station." The rest of this section briefly introduces ASR, SBD, MT and UI, in that order.

3.1. ASR

To achieve accurate speech recognition under noisy environmental conditions, we carefully select the acoustic features for voice activity detection [10] and acoustic modeling [11]. The language model is trained with a large-scale text corpus collected from the web and a bilingual corpus that we developed for the travel domain.

The ASR dictionary contains 200,000 Japanese words and 30,000 English words. These entries are selected according to frequency of appearance in the corpus. In addition, we registered words specific to Kawasaki City in Kanagawa Prefecture, Japan (e.g., names of sightseeing spots, transport facilities, etc.), where we conducted the experiment described in section 4.

We configure the ASR module to output a recognition result for every speech section separated by a 200 ms pause. Because of variety in user speech style, the speech segments processed by ASR are not always appropriate for translation. We introduce an SBD method to provide input text for MT.

3.2. SBD

Among the many works on SBD, [12] is to our knowledge the newest report on SBD for simultaneous interpretation systems. The authors there prepare parallel corpora and create a phrase table using a statistical MT (SMT) tool. They realize SBD by using the phrase table.

In contrast, our SBD is realized by a rather simple process. We first prepared monolingual corpora for Japanese and English. For Japanese, we set sentence boundaries by references to a set of manually developed rules; for English, we regarded punctuation as indicative of boundaries. Next, we used CRF++ [13], a machine-learning tool based on conditional random fields, and created a discrimination process to find sentence boundaries. Through these processes, we obtained monolingual SBD modules for three languages. For Japanese, we added a rule-based filler detector, and sentences that consist of only fillers are deleted as semantically null.

3.2.1. Detection model

Sentence boundaries are detected in two steps. In the first step, the system performs morphological analysis on the results from ASR and obtains word segmentation and also part-of-speech (POS) tags on Japanese and English. Then, fillers and other redundant parts are removed using simple pattern matching to POS.

In the second step, machine-learning-based classifiers detect sentence boundaries. Sentence boundary detection is treated as a labeling task for each word [14]. We prepare spontaneous speech corpus in which words at the beginning of a sentence have "B" labels and other words have "I" labels. We use CRF++ [13] and create a discrimination model for the labeling. For the learning features, we use the surface form of two morphemes before and after each morpheme for Japanese and English.

3.2.2. Training corpus

To create Japanese and English sentence boundary detectors, we used two different corpora: for Japanese, 140,000 sentences from "Corpus of Spoken Japanese (CSJ) [15]", and for English, 110,000 sentences from WIT3 [16] data including transcriptions of TED talks.

These corpora do not contain any tags denoting a suitable unit for translation. We regarded a punctuation mark as a boundary marker in English. For Japanese, we regarded a clause to be a suitable unit for translation [17] and prepared simple rules to find clause boundaries in the training corpus.

3.2.3. Detection performance

We evaluated precision and recall of boundary detection on test sets. The test sets had been ideally segmented into 244 Japanese sentences and 1664 English sentences. We regarded punctuation as definitive segment boundaries. Table 1 shows detection accuracy. In this table, we calculate the precision and recall values as follows:

$$\text{Precision} = \frac{\text{No. of correctly estimated sentence boundaries}}{\text{No. of estimated sentence boundaries}}$$

$$\text{Recall} = \frac{\text{No. of correctly estimated sentence boundaries}}{\text{No. of periods in original corpus}}$$

□

Table 1: *Segment detection accuracy*

	Precision	Recall	F-value
Japanese	0.739	0.672	0.705
English	0.720	0.809	0.763

3.3. MT

3.3.1. Forest-driven rule-based MT

Rule-based machine translation (RBMT) has been used in commercial systems for a long time. A well-developed RBMT engine outputs a better translation and covers a larger domain than other types of systems. However, commercial MT systems are usually designed for use on grammatically written language, and they sometimes fails to process ungrammatically spoken language.

We introduce a forest-driven parsing mechanism ([18], Figure 4) into RBMT. It parses input sentences by generalized LR parsing, which can accept ungrammatical chunks by using an original context-free grammar to capture the clause structure and deal with various ambiguities. The parser then generates possible syntax structures as a forest and transfers the best structure to the target language structure according to syntactic and semantic preferences.

3.3.2. Hybrid MT

SMT can generate natural translation results for restricted and specific domains. RBMT, however, can translate an input sentence robustly, but the result sometimes lacks fluency.

We viewed these strengths and weaknesses as complementary, and so we used SMT and RBMT engines together to form a hybrid MT engine. Specifically, when the probability of an SMT result falls below a specified threshold, the RBMT result is selected instead as the final result of the hybrid MT engine [18]. This engine selection is made for each segment produced by SBD.

We used phrase-based SMT [19]. For Japanese-English and English-Japanese SMT, we trained the engine with a travel domain corpus consisting of 220,000 sentence pairs developed by ourselves and 20,000 sentence pairs distributed by the Advanced Language Information Forum [20].

3.3.3. Translation quality

We evaluated engines both automatically and manually (Table 1). We used the IWSLT 2004 corpus [20] as a test set. For automatic evaluation, 500 sentence pairs were used; the first 100 of these

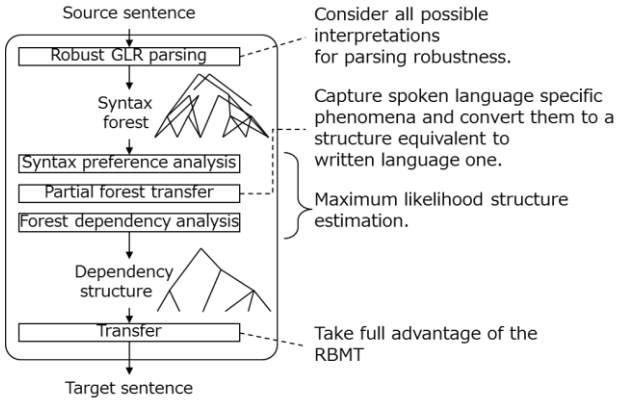

Figure 4*: Process flow of forest driven RBMT*

Table 2: *Detailed Translation Quality (data of IWSLT)*

		Adequacy	Fluency	BLEU	RIBES
E	RBMT	3.93	3.69	20.64	0.575
	SMT	3.90	4.12	33.97	0.650
	Hybrid	4.01	3.89	28.54	0.631
J	RBMT	4.15	3.94	22.21	0.755
	SMT	4.25	4.29	34.28	0.807
	Hybrid	4.30	4.25	32.27	0.790

sentence pairs were used for manual evaluation.

We used BLEU [21] and RIBES [22] for automatic evaluation. We also manually evaluated fluency and adequacy metrics [23]. Table 2 shows the evaluation results. We assumed that adequacy of manual translation reflects correctness of meaning, and we chose the hybrid engine for our simultaneous interpretation system.

3.4. UI

We developed a translation system whose user interface runs on a tablet with the Android operating system. In the task-based assessment, a "host" and a "guest" share a terminal display and communicate with each other through the system.

Figure 5 shows the user interface. A user starts speaking after pressing the "speak" button. While the user continues to speak, it is not necessary to hold the button. When the user presses the button a second time, the system processes it as an explicit signal that speech is concluded.

Until the speech recognition result is finalized, a recognition candidate is shown in gray. When the translation result is finalized, the system displays the ASR and MT text. In Figure 6, the speak button for the English speaker is placed on the right hand side, and the button for the Japanese speaker on the left.

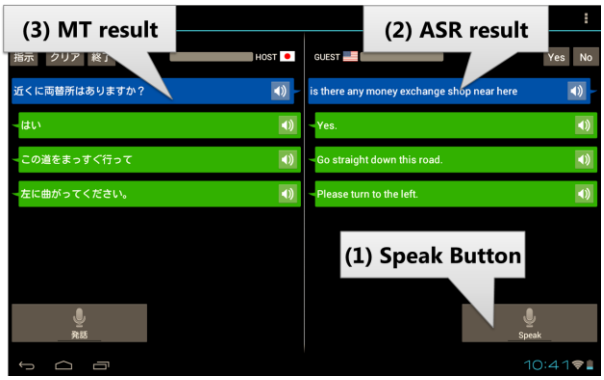

Figure 5: *User interface of Client Application*

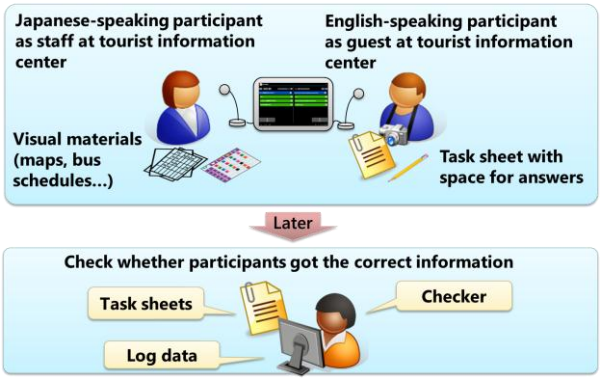

Figure 6: *Experiment situation and the evaluation process of Solved Task Ratio*

For interpretation from English to Japanese, the English speaker presses the speak button (1) and says something, such as "Is there any money exchange shop near here?" After this, the ASR result "is there any money exchange shop near here" is shown on the display (2). Then, the MT result "近くに両替所はありますか [*Chikaku ni ryougaejo wa arimasu ka*]" is shown (3). For Japanese to English, the speak button, ASR result, and MT results are on the opposite side.

4. Task-based Evaluation Experiment

We conducted a task-based evaluation experiment in the Toshiba Customer Service Evaluation Center. This experiment is in addition to a previous evaluation experiment conducted in a tourist information center in Chiba City in Chiba Prefecture, Japan [1]. In this section, we discuss the parts of this prior experiment that relate to the analysis in section 5.

4.1. Tasks

The tasks in the evaluation experiments were as follows. We prepared these tasks on the assumption that the conversation is being held in a tourist information center. The previous experiment [1] was conducted in Chiba City. This additional experiment was held in Kawasaki City in Kanagawa Prefecture. Therefore, we modified some of the tasks to make them appropriate to Kawasaki City. We added 2 tasks to the 8 tasks in [1], and now we have the following 10 travel tasks.

(1) Ask whether you can book any local tours here.
(2) Ask whether you can get to Tokyo Disneyland by train without changing trains.
(3) Ask how much the fare is from Kawasaki Station to Hamamatsucho Station by train.
(4) Ask how to get to a money exchange shop near here.
(5) Now you would like to know the bus route and its schedule in Kawasaki City. Ask how you can get this information.
(6) Ask what is the best souvenir from Japan. Ask about its features and how to get to a store where you can buy it.
(7) Ask your partner to recommend a sightseeing spot and how to get there. Decide whether you will go according to your interest.
(8) Imagine what you would like to try in Japan and ask where you can experience it around here.
(9) Ask how to get downtown from here. Assume that you will have dinner there or go shopping.
(10) You lost your bag on the train. Ask what you should do to find it.

Table 3: *English Speaking Participants*

English Speaking Participant	Sex	Years in Japan	Place of Birth
A	M	3	Los Angels
B	F	3	Hawaii
C	F	3	Arizona
D	M	3	California
E	M	3	South Carolina

Table4: *Japanese Speaking Participants*

Japanese Speaking Participant	Sex	Place of Birth
A	F	Okayama
B	F	Kanagawa
C	F	Tokyo
D	F	Kanagawa
E	M	Tokyo

4.2. Participants and collected data

The data collected for the analysis in section 5 includes conversation logs and transcriptions of five English-speaking participants (Table 3) and of five

Japanese-speaking participants (Table 4). The labels A to E were given to the five pairs of people who had conversations through the system.

4.3. Solved Task Ratio

The solved task ratio indicates the proportion of tasks achieved out of all tasks. In this paper, we focus on 45 tasks for which speech was successfully recorded. Of these, 44 tasks were solved. Therefore, we had a solved task ratio of 97.8%.

5. Analysis of UX

The solved task ratio confirms that our simultaneous interpretation system can almost always help users to obtain information from speakers of a different language. However, we would like to ascertain whether users were satisfied with the whole process of conversation through our system. In other words, we would like to find a way to assess the UX of our simultaneous interpretation system.

5.1. UX for our system

It would be ideal if users would say each thing only once and this speech would be perfectly interpreted by our system. However, since ASR, SBD, and MT do not perform perfectly, users sometimes need to repeat themselves until the partner speaker can understand the interpretation result and respond. It is clear that less frequent repetition is preferable; however, we would still like to determine how many repetitions users will tolerate before experiencing stress. In other words, we would like to know what level of performance is needed so that our system does not put stress on users.

5.2. Statistics from transcript and system log

To assess the UX of the conversation process, we transcribed the 45 conversations from the evaluation experiment and manually analyzed them.
Since spoken language includes parts smaller than clauses, we define here the relationship between "speech," "clause," and "intention of the clause." A "speech" indicates the words from a transcript of the users' voices, terminated by a pause of 200 ms. When spoken slowly, one clause will spread into several speeches, so we manually detected a clause chunk by hand from the transcription. For example, as shown in Figure 7, when a user says, "I want to go," and pauses for 200 ms before saying "on a tour," the speaker uttered two "speeches" but only one "clause." We recorded 1330 speeches during the 45 conversations and manually chunked the speech into

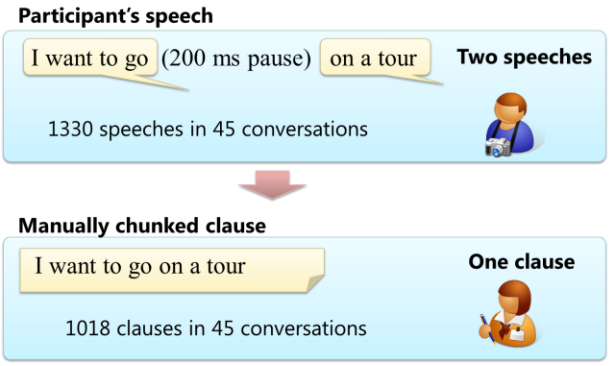

Figure 7*: unit of "a speech sound" and "an utterance"*

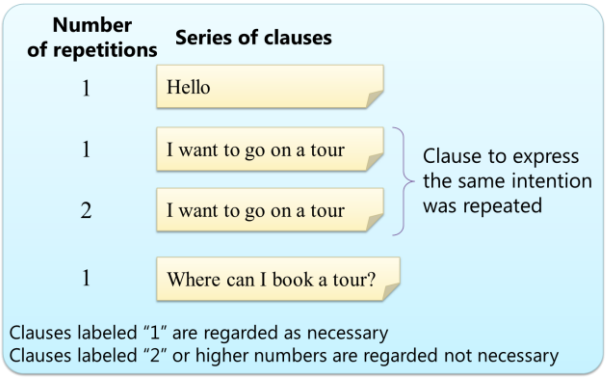

Figure 8*: an example of repeated utterances*

Table 5: *Change of intention after repeated failure of interpretation*

Number of repetition	Transcription of utterances	ASR result
1	Where can I eat Yakiniku?	where can i am eat your key to do it
2	What is a good Yakiniku restaurant?	what is a good jockey to restaurant
-	OK. Where can I get great Sushi?	ok our can i get great sushi

clauses. This gave 1018 clauses in the 45 conversations. The "intention of a clause" indicates the intended meaning of a clause.

5.3. Repeated clauses

We counted how many times clauses were repeated before being understood by the partner speaker. Figure 9 illustrates how we counted the number of repetitions for each clause. In the example, utterances of the same letter are regarded as repetition to express the original intention of the speaker. In this analysis, a question asked by the partner speaker to clarify an unclear interpretation result caused by an interpretation error is also regarded as a repeated utterance.

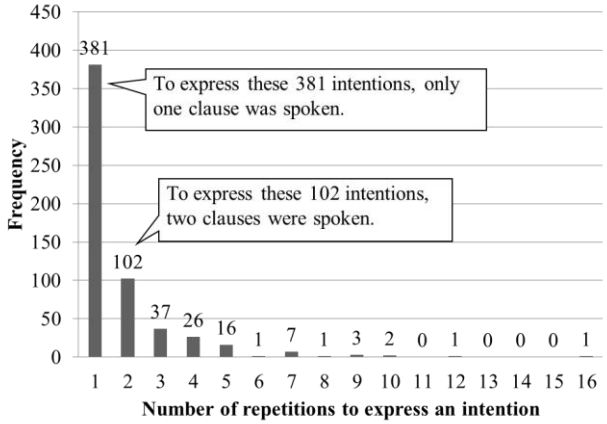

Figure 9: *Number of repeated clauses for 578 intention*

Figure 10 shows the number of intentions that were expressed through multiple, distinct clauses or through more than two repetitions. We found that 381 intentions were expressed through a clause without repetition; 102 intentions were expressed through a clause repeated once. The total number of intentions across the 45 conversations was 578.

To assess whether the number of repetitions was too large, we used another measure. As shown in Table 5, the speaker originally wished to eat "yakiniku," which is a Japanese-style grilled meat. However, the word "yakiniku" was not recognized well and so was not interpreted to get the response from the partner speaker. The speaker changed to asking about "sushi" instead; this was successfully recognized and interpreted, and the partner speaker responded. The speaker did not return to the original intention of "yakiniku" again. In this example, an ASR error caused the interpretation error, but in some other cases, the ASR succeeded and MT caused an interpretation error.

In the 45 conversations, there were 6 intentions that were changed due to repeated utterances. The speaker changed intentions after an average of 3.6 interpretation errors (as indicated by lack of response from the partner speaker).

6. Conclusions

We introduced our simultaneous interpretation system for face-to-face conversation between two people, and we also analyzed the transcription of the speech and the system log in the experiment. This new version of our system has a revised SBD module. In the new system, several speeches are first combined together and then the system finds a suitable unit for translation.

We also evaluated the system by a task-based experiment. The evaluation experiment showed a solved task ratio of 97.8% across 45 tasked-based conversations. However, we found that users repeated each utterance 1.8 times on average.

From analysis of the transcripts and the system log, we found that after approximately 3.6 interpretation errors, users would change what they said to avoid interpretation error and receive a response from the partner user. For future work, we would like to improve our system to reduce user speech repetition.

7. References

[1] A. Sakamoto et al., "Development of a Simultaneous Interpretation System for Face-to-Face Services and Its Evaluation Experiment in Real Situation," In *Proc. Machine Translation Summit XIV*, Nice, France, 2013, pp.85-92.

[2] A. Waibel et al., "JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies," In *Proc. ICASSP'91*, Toronto, 1991, pp.793-796.

[3] F. Metze et al., "The NESPOLE! speech-to-speech translation system," In *Proc. HLT 2002*, San Diego, CA, 2002.

[4] W.Wahlster, "Verbmobil: translation of face-to-face dialogs," In *Proc. 3rd European Conf. on Speech Communication and Technology*, Berlin, 1993, pp.29-38.

[5] S. Matsubara and Y. Inagaki, "Incremental Transfer in English-Japanese Machine Translation," *IEICE TRANSACTIONS on Information and Systems*, Vol.E80-D, No.11, pp.1122-1130, 1997.

[6] S. Bangalore et al., "Real-time Incremental Speech-to-Speech Translation of Dialogs," In *Proc. NAACL-HLT 2012*, Motreal, 2012, pp.437-445.

[7] H. Shimizu et al., "Constructing an Automatic Simultaneous Interpretation System using Simultaneous Interpretation Data," In *Proc. The 2013 Autumn Meeting of the Acoustic Society of Japan*, Toyohashi, 2013, pp.59-62.

[8] NTT docomo, 2012, *NTT DOCOMO to Introduce Mobile Translation of Conversations and Signage*, Available: http://www.nttdocomo.co.jp/english/info/media center/pr/2012/001611.html

[9] C. Fügen , A. Waibel, M. Kolss, "Simultaneous translation of lectures and speeches," Machine Translation, 21, pp.209-252, (2007).

[10] H. Ding et al., "Comparative evaluation of different methods for voice activity detection,"

In *Proc. Interspeech 2008*, Brisbane, 2008, pp.107-110.
[11] M. Nakamura et al., "Evaluation of Group Delay based Features in Noisy Environments," In *Proc. The 2012 Spring Meeting of the Acoustic Society of Japan*, 2012, Yokohama, pp.947-952.
[12] G. Neubig et al., "A method for deciding translation timing in speech translation considering reordering between languages," In *Proc. The 2013 Autumn Meeting of the Acoustic Society of Japan*, Toyohashi, 2013, pp.55-58.
[13] Y. Liu et al., "Using Conditional Random Fields For Sentence Boundary Detection In Speech," In *Proc. of the 43rd Annu. Meeting of ACL*, Ann Arbor, MI, pp.451-458, (2005).
[14] T. Kudo, 2005, *CRF++: Yet Another CRF toolkit*, Available: https://code.google.com/p/crfpp/
[15] K. Maekawa et al., "Spontaneous Speech Corpus of Japanese," In *Proc. LREC 2000*, Athens, 2000, pp.947-952.
[16] M. Cettolo et al., "WIT3: Web inventory of transcribed and translated talks," In *Proc. EAMT 2012*, Trento, 2012, pp.261-268.
[17] K. Takanashi et al., "Identification of "Sentence" in Spontaneous Japanese – Detection and modification of clause boundaries –," In *Proc. SSPR 2003*, Tokyo, 2003, pp.183-186.
[18] S. Kamatani et al., "Hybrid Spoken Language Translation Using Sentence Splitting Based on Syntax Structure," In *Proc. Machine Translation Summit XII*, Ottawa, 2009.
[19] H. Wang et al., "The TCH Machine Translation System for IWSLT 2008," In *Proc. IWSLT 2008*, Waikiki, HI, 2008, pp.124–131.
[20] Y. Akiba et al., "Overview of the IWSLT04 evaluation campaign," In *Proc. IWSLT 2004*, Kyoto, 2004, pp.1-12.
[21] K. Papineni et al., "BLEU: a method for automatic evaluation of machine translation," In *Proc. the 41st Annu. Meeting of ACL*, Sapporo, 2002, pp.311-318.
[22] H. Isozaki et al., "Automatic Evaluation of Translation Quality for Distant Language Pairs," In *Proc. EMNLP 2010*, Cambridge, MA, 2010, pp.944-952.
[23] P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between European languages," In *Proc. the HTL-NAACL Workshop on Statistical Machine Translation*, New York, NY, 2006, pp.102-121.