Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study

Kotaro Hara
Department of Computer Science
University of Maryland
College Park, MD 20742 USA
kotaro@cs.umd.edu

Shamsi T. Iqbal
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
shamsi@microsoft.com

ABSTRACT

Language barrier is the primary challenge for effective cross-lingual conversations. Spoken language translation (SLT) is perceived as a cost-effective alternative to less affordable human interpreters, but little research has been done on how people interact with such technology. Using a prototype translator application, we performed a formative evaluation to elicit how people interact with the technology and adapt their conversation style. We conducted two sets of studies with a total of 23 pairs (46 participants). Participants worked on storytelling tasks to simulate natural conversations with 3 different interface settings. Our findings show that collocutors naturally adapt their style of speech production and comprehension to compensate for inadequacies in SLT. We conclude the paper with the design guidelines that emerged from the analysis.

Author Keywords

Multilingual communication; Spoken language translation; Automatic speech recognition; Machine translation

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]: Computer-supported cooperative work

INTRODUCTION

Language barrier is the primary challenge for effective cross-cultural communication in education, health care, and business collaboration [6,14,15]. In the U.S., for example, roughly 40 million foreign-born individuals constitute the total population [19], and many of whom do not use English as their first language. Despite the increasing need for a solution, language remains "the biggest barrier to intercultural collaboration" [26].

Though human interpreters can bridge the communication gap between people with different languages, the service is accessible only to a privileged few because of its cost and availability. For example, international organizations like the United Nations pay a freelance interpreter over $600 per day [27]. As a result, only a handful of spoken encounters between humans are interpreted [7].

Automatic spoken language translation (SLT) has been expected to address the problem since its first appearance in the 1980s [21,22]. Decades of research has advanced the state-of-the-art, and some claim that it is usable in limited domains [7,12]. Despite the increasing interests in SLT, however, little research has paid attention to how people interact with the added complexity of speech recognition, translation, and speech synthesis. Consequently, we know little about how the system can be designed to support this complex exchange while still maintaining the flow of a conversation.

To better understand user experience of SLT systems, we conducted two sets of formative studies. We invited 8 pairs of French speakers and German speakers, and 15 pairs of English speakers and German speakers to hold cross-lingual conversations using our translator application. We collected and analyzed survey responses and interview recordings. Our main research questions were: what are the overall impressions on quality and challenges in using the translator tool? Do people adapt the way they use the system, and, if so, how? What interface components facilitate/disrupt their conversations?

Our findings suggest that, while SLT systems do not produce perfect results due to limitations in state-of-the-art speech recognition and machine translation, people naturally adapt to imperfect translation systems by changing the way they speak and comprehend speech. However, recurrent translation problems can cause frustration. Problems also arise from lack of information about when to speak and when to wait. Other issues necessitate a mix of speech and text to correctly convey an intended message, which then has a higher probability of being interpreted correctly.

This paper offers design guidelines for SLT systems with a focus on HCI. In the following sections, we first provide background on simultaneous interpretation, speech recognition, and machine translation technologies. Next, we review prior work on computer-mediated communication. Finally, we describe the user studies and present the quantitative and qualitative results.

TRANSLATION VS. INTERPRETING

The difference between "translation" and "interpreting" is blurred in computer science literature. "Translation" often refers to transfer of meaning from text to text, whereas

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3145-6/15/04...$15.00
http://dx.doi.org/10.1145/2702123.2702407

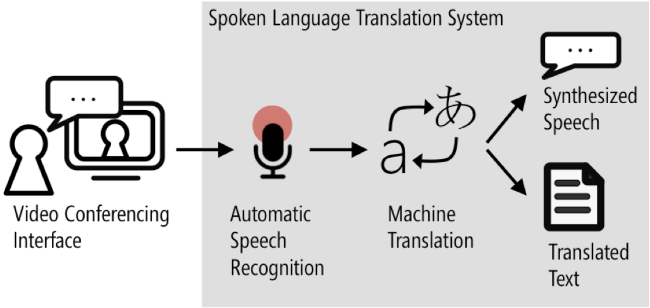

Spoken Language Translation System
Video Conferencing Interface
Automatic Speech Recognition
a
あ
Machine Translation
Synthesized Speech
Translated Text

Figure 1. We used a translator application that has an interface similar to typical video conferencing tools (*e.g.,* Microsoft Skype, Google Hangout), but with a *spoken language translation (SLT)* system on background. In general, a translation system comprises *automatic speech recognition (ASR)*, *machine translation (MT)*, and *text-to-speech (TTS)* components, so does our system.

"interpreting" often refers to conversion from speech to speech [10]. Historically, computer science scholars use "translation" indiscriminately regardless of medium [7]. Our translator application performs simultaneous interpreting (*i.e.,* interpreting in nearly real-time). Unless otherwise mentioned, we use the term "interpretation" and "translation" interchangeably to indicate the automatic simultaneous interpretation of spoken language.

RELATED WORK

Spoken Language Translation

Outcome of decades of research on SLT technology is starting to enable people to overcome language barriers. The technology generally comprises 3 components: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) (Figure 1). ASR transcribes human speech in real-time [5,12]. MT translates transcribed speech into another language [7,11,16]. Though ASR and MT are not perfect yet, they are expected to improve over time [7]. Translation can then be delivered in a form of synthesized speech and/or by text. Despite the extensive work on development and evaluation of translation quality, there is limited research on how people interact with the technology [26].

Cross-cultural Computer-mediated Communication

ASR and MT not only empower monolinguals, but they can also benefit greater audience such as second language speakers [9,17,23]. Research shows that MT enables non-native speakers to produce more ideas in brainstorming tasks with native speakers in text-based conversation [23]. Previous work reported benefits of presenting real-time transcription of English conversation between native and non-native speakers: changes in speaking behavior of native speakers (*e.g.,* better enunciation) and textual information both help non-native speakers to better comprehend conversation [9,17]. The work, however, only discusses the effect of each component in cross-cultural communication, and does not discuss the effect of SLT as a system.

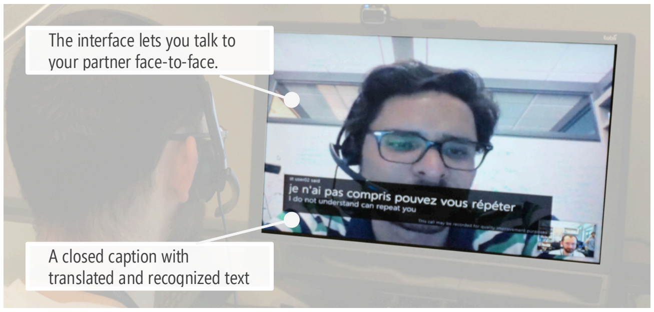

The interface lets you talk to your partner face-to-face.
je n'ai pas compris pouvez vous répéter
I do not understand can repeat you
A closed caption with translated and recognized text

Figure 2. The user interface of the translator application. Users can hear each other's voice. Users can choose to display transcribed text and translated text in a closed caption. Users can also choose to turn on TTS, so they can hear translated speech.

Effective MT-mediated communication requires solutions for hurdles posed by the technology [8,25,26]. Yamashita *et al.* found that MT-mediated chat imposes extra difficulties of building common ground between multiple parties due to mistranslation and peoples' lack of awareness thereof [25,26]. This not only hinders effective mutual understanding, but also reduces perceived quality of communication [8].

The aforementioned work, however, primarily focuses on text-based interaction, and limited HCI research exists for conversation with SLT. The most relevant to our work is the evaluation of NESPOLE! by Constantini *et al.*, in which the authors evaluated success rate of negotiation tasks using a multi-modal speech-to-speech translation system [3]. Though they uncovered effect of issues such as word sense ambiguity in interaction, they did not investigate important characteristics such as adaptation in speech production. Their push-to-talk interface also makes it hard for us to understand the difficulty of turn-taking in MT-mediated natural conversation. Thus, our goal is to reinforce the findings from the previous work and contribute to extend the body of HCI research on MT-mediated conversation.

SYSTEM DESCRIPTION

We used a video conferencing system with real-time SLT to evoke characteristics of MT-mediated conversation. The translator application allows two users to communicate face-to-face remotely like modern videoconference clients (*e.g.,* Microsoft Skype, Google Hangout). As a user speaks, the system recognizes and translates his/her utterance. Translation is then conveyed to another user via synthesized audio (*i.e.,* TTS) and/or text in a closed caption (CC), just like any other SLT systems (Figure 1).

Users can choose to turn CC or TTS on or off (in the study sessions, the experimenter controlled the setting). Note that a user's speech can be heard by their partner. Both parties could hear TTS. CC was displayed at the bottom of the interface (Figure 2). It displayed both transcribed text and the corresponding translation. CC was updated when the next utterance got translated and was ready to be displayed, similar to news captions and movie subtitles.

STUDY

Our goal was to understand how using the translator system in an interlingual conversation affects the interaction between two collocutors speaking in two different languages.

Participants

Potential users of an SLT system include not only pairs of monolingual speakers without a common language, but also pairs of second language learners who share a common second language where one side has much better understanding of one language. To study a diverse usage scenario, we conducted two studies with different groups of participants. The first group in the study consisted of French-German (Fre-Ger) pairs who spoke either French or German, but not both, so both sides had limited to no knowledge of each other's language. The second group in the study consisted of English-German (Eng-Ger) pairs who spoke either English or German as their primary language. This group represented a situation where both sides know a common language, but one side hesitates to speak in his/her second language and would prefer communicating in their native language. All participants could speak English reasonably well.

Participants in both groups were recruited via cold-calling, listserv, and word-of-mouth in rolling basis. Groups such as local English/German/French speaking meet-ups and schools were emailed and called. We also emailed employees of the authors' organization. While most pairs of participants did not know each other, four pairs in Fre-Ger pairs were colleagues or friends. Although participants were convenientce samples of local residents of any gender/age, we believe this did not negatively affect the validity of the study given its explorative nature. We considered reporting the results of two studies separately, but we decided to describe them in parallel to compare responses from the two groups.

Experimental Design

The study was a 3x3x2 (*Interface* x *Round* x *Language*) mixed design in-lab study. *Interface* condition (*i.e.,* (CC, TTS) = (on, on), (on, off), and (off, on)) and *round* (round 1 to 3) were within-subject factors; and *Language* (French and German, or English or German) was a between-subject factor. Changing between three *interface* settings—TTS only, CC only and both—allowed us to see the tradeoffs between the naturalness of the exchange (speech to speech), speed of exchange (speech to text), and reinforcement of the same information through two modalities (speech and text). *Round* was defined as a set of three task trials, and each of them included one of the three interface conditions. We considered *round* as a separate variable in our analysis as we wanted to see how user experience and adaptation change with increasing exposure to the system.

Task

Each pair of participants was asked to work on a story telling task, and each participant conversed in their own language (Video Figure). The objective was to collaboratively generate a fictional story. This task was inspired by prior work (*e.g.,* conversation in a second language with ASL [9]) and adapted from a conversation study in [13]. We did not ask participants to follow a strict turn taking protocol, but we suggested them to take turns so both parties could contribute to story formation.

Starting Sentence	
English	Olivia was practicing for the dance-off.
French	Olivia s'entraine pour le concours de danse.
German	Olivia trainierte für den Tanzwettbewerb.

Word List							
English	swim	pig	plane	dance	smile	drink	doctor
French	natation	cochon	avion	danse	sourire	boisson	médecin
German	schwimmen	Schwein	Flugzeug	Tanz	Lächeln	Getränk	Arzt

Table 1. An example of starting sentence and a word list. English sentence and key words were translated into French and German.

Our task allowed free form speech, just like a normal conversation. However, we provided a starting sentence to initiate a task and so a conversation was more controlled than a random conversation (Table 1). To make it easier for participants to keep a conversation going, we provided 7 keywords relevant to the starting sentence, which participants could optionally use in the conversation. The English starting sentences and keywords were translated into German and French by translators hired from oDesk (www.odesk.com). Each task consisted of different starting sentences and corresponding keywords.

This task was appropriate for our study as it has important properties that map to normal conversations. (i) Building coherent narrative for explaining one's thought and sharing knowledge is a common and important task in daily life. (ii) The task requires effective conversational turn taking. (iii) The task has a clear objective without being restrictive in the required answers. To create starting sentences and keywords, we gathered inspiration from a web site that lists words for family games. We used the "easy" sets for Catchphrase, Charades, and Pictionary on the web site of http://www.thegamegal.com/printables/.

Procedure

Each study session comprised (i) task introduction, (ii) a practice task in English, (iii) 9 conversation tasks, and (iv) a final story telling task in English (*i.e.,* without translation) as a baseline. On arrival to the study location, each participant was asked to fill out an informed consent form and was provided a brief overview of the study by a member of the research team. Then, they were given a practice conversation task in English in order to familiarize themselves with task itself without using the translator tool.

We then placed participants in two separate rooms to have interlingual conversation using the translator application. Each participant was asked to wear a headset. A researcher sat in one of the two rooms and initiated a video call using a computer in the room, and a participant in the other room was instructed to respond to the call via the interface. Once the call was established, the researcher asked a participant in the same room to start the conversation using a staring sentence. The researcher stopped the conversation at a natural stop point after 2-4 minutes.

We exposed participants to all the interface settings in story telling tasks (*i.e.,* CC, TTS, and CC&TTS). Our goal was to observe and explore participant variation in how they use the application and obtain their opinions on interface condition preference (*e.g.,* do they prefer to have TTS or not). One round consisted of 3 story telling tasks using different interface settings. To minimize ordering effects, we permuted the orders of exposure. We repeated the tasks for 3 rounds. After each task and round, participants were asked to fill out a survey.

After the 9 main tasks were completed, we asked participants to work on another story telling task in English to assess how easy/difficult the task was when there was no real-time translation involved. The session ended with a post study questionnaire asking about their overall impressions and preference of conditions. A brief interview was held in which participants could suggest what interface elements they thought would have helped improve the interaction.

Data and Measures

In addition to in-situ observation, we collected the following data to conduct post-hoc analysis.

Survey responses. To analyze perceived usability of the system (*e.g.*, how successful a conversation was) and participants' preference towards interface settings, we asked participants to fill out surveys after each task, each round and at the end of a session. The surveys in each trial and round had Likert scale questions. Each survey had an optional free-text response space where participants could comment on, for example, their experience, suggestions for interface improvements, and specific problems with ASR and MT.

Conversation log. The translator tool collected conversation logs with the following data: (i) transcribed text, (ii) translated text, and (iii) a record that synthesized audio were sent to clients. All events were time stamped, and computation time was recorded where appropriate. However, we lost logs for 30 out of 72 (41.7%) Fre-Ger conversations and 47 out of 135 (34.8%) Eng-Ger conversations due to technical issues. We thus limited ourselves from drawing conclusions from these data.

Video recording and transcript. We also recorded and transcribed each session with a camcorder positioned over the user's shoulder in order to perform a post-hoc video observation. This allowed us to understand users' behaviors and interaction issues through the observation.

DESCRIPTIVE STATISTICS

We invited two groups of pairs to our in-lab study: 8 Fre-Ger pairs (16 participants, 6 female, *Avg. age*=32.9, *SD*=13.2) and 15 Eng-Ger pairs (30 participants, 15 female, *Avg. age*=46.0, *SD*=13.2). Participants self-reported their language proficiency in 5 point Likert scale where 1 is "not proficient at all" and 5 is "very proficient" (Table 2). All participants except for one English speaker had experience in using a video conferencing tool (*e.g.,* Skype), and the most common use case was conversation with friends followed by business, family, school, and game.

	French-German (N=16)		English-German (N=30)	
	French	German	English	German
Average Age (yr)	31.3 (13.5)	34.5 (13.6)	49.4 (14.2)	42.7 (11.7)
English Proficiency	4.8 (0.5)	4.9 (0.4)	4.9 (0.3)	4.7 (0.5)
German Proficiency	1.8 (0.9)	4.9 (0.4)	1.3 (0.6)	5.0 (0.0)
French Proficiency	4.8 (0.5)	1.5 (0.8)	1.1 (0.4)	1.1 (0.3)
Application use cases	Friends (40), business (18), family (12), school (8), game (2)			

Table 2. Participants' language proficiency and experience in using a video conferencing tool. All participants rated proficiency for their own language more than proficient. All participants had experience in using some video conferencing tool except for one English speaker in the Eng-Ger group.

On average, storytelling tasks in Fre-Ger group and Eng-Ger group lasted 2.6 min (*SD*=0.4 min) and 2.7 min (*SD*=0.4 min) respectively.

We first quantitatively analyzed participants' survey response from each task and each round to grasp how participants perceived the usability of the system and how (if ever) they adapted to using SLT. We then moved on to results from a content analysis to further investigate how participants felt while using the translator application.

STATISTICAL ANALYSIS

We used mixed models to analyze the participants' response. We first transformed the ordinal Likert scale data with aligned rank transformation to satisfy the normality assumptions [24]. The results were then analyzed by 3x3x2 (*Interface* x *Round* x *Language*) mixed model (*i.e.,* restricted maximum likelihood model)[1]. We used mixed models instead of traditional repeated measure ANOVA to incorporate subject variability. A critical value $\alpha = 0.05$ was used to assess significance. We omit descriptions of insignificant interactions and main effects, unless there was possible trend (*i.e., p*<0.10). Contrast tests with pairwise comparisons were protected against Type I error using a Bonferroni adjustment.

Overall Interface Preference

We asked participants to rank three *Interface* conditions (*e.g.,* (CC, TTS) = (on, off)) from "1" (least preferred) to "3" (most preferred) based on which one they liked/disliked after each round. We allowed ties in the ranking (*e.g.,* people could score everything as 1).

[1] There were only 2 *Interface* levels CC and CC&TTS for one question that asked about subjective speech recognition accuracy. Thus we used 2x3x2 mixed model for analysis.

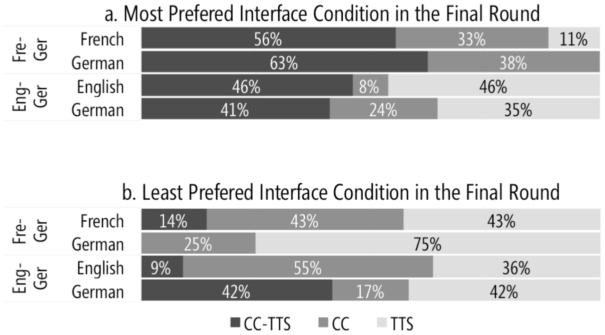


Figure 3. Participants' interface setting preference distribution after the final round of a study session. (a) About a half of participants liked CC&TTS. There were very few participants who liked TTS in the Fre-Ger group, but relatively more participants liked TTS in the Eng-Ger group. (b) Many disliked TTS only condition in the Fre-Ger group. Noticeably more German participants disliked CC&TTS condition.

There were no interaction effect or main effect of *Round* suggesting overall preference did not dramatically change over time. Given this, the preference from a single round provided a good picture of which interface condition was preferred by which language group. Therefore, we show the demography of the most preferred and least preferred interfaces for the final round in Figure 3.

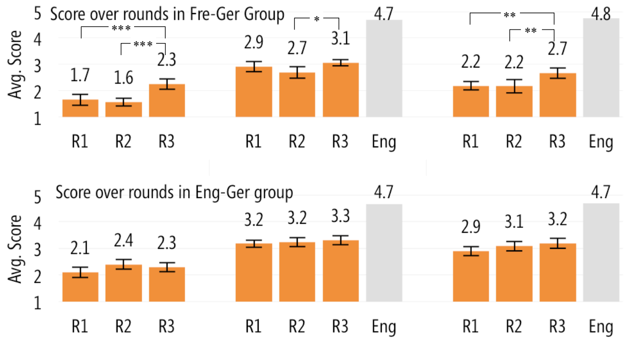
In the Fre-Ger group, there was a significant interaction effect between *Interface* and *Language* ($F_{2,112}$=6.18, p<0.01) suggesting that interface preference differs between French and German speakers. Though we cannot strongly argue with the presence of interaction, it seems both disliked TTS only condition, which suggests the presence of a closed caption is crucial for this group (Figure 3).

Significant main effects of *Language* ($F_{1,28}$= 4.38, p<0.05) and *Interface* ($F_{2,224}$=6.967, p<0.01) were observed in the Eng-Ger group. Some English speakers ranked all conditions as "1", and thus the overall ranking of English speakers was significantly lower than that of German speakers. It indicates the quality of the experience was equality unsatisfactory for all *Interface* conditions for these English speakers. CC&TTS was significantly preferred compared to CC (p<0.05) and TTS (p<0.01).

In general, CC&TTS was most preferred by participants (Figure 3.a), but some disliked it. Noticeably more German speakers disliked CC&TTS compared to other participants, most likely because their speech overlapped with synthesized speech more frequently than others as discussed in the next section.

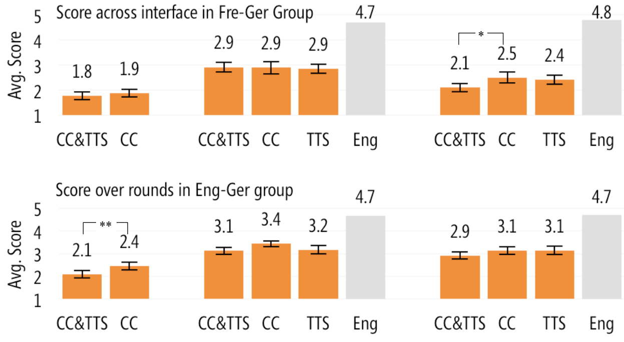
Subjective Usability Rating

After each trial, we administered a survey with questions on the usability of the system. Each question was answered in a form of 5-point Likert scale, where 1 was "strongly disagree" and 5 was "strongly agree."



(a) Speech Recogn. Quality (b) Turn Taking Difficulty (c) Overall Conversation Quailty

Figure 4. Change in average ratings for perceived usability over three rounds of a session. Likert scaling ranged from 1 to 5, where 5 is more positive. Scores for turn taking difficulty and overall conversation quality were compared to the scores from English conversation. Error bars represent standard errors. Statistical significance: *** p < 0.001; ** p < 0.01; * p < 0.05.



(a) Speech Recogn. Qual. (b) Turn Taking Difficulty (c) Overall Conversation Quailty

Figure 5. Average usability ratings for each interface condition. Likert scaling ranged from 1 to 5, where 5 is more positive. Perceived speech recognition quality was not assessed for tasks with the TTS only condition and English tasks because of the absence of a closed caption. Error bars represent standard errors.

Perceived quality of speech recognition. Automatically recognized speech was displayed in a closed caption in CC and CC&TTS conditions. Participants rated their agreement to: *"there were no noticeable errors in transcription."* Note that TTS condition was excluded from this analysis as closed caption was not displayed in this condition (*i.e.,* participants had no mean to assess the speech recognition quality). We observed a significant main effect of *Round* ($F_{2,70}$=10.774, p<0.001) to perceived quality of speech recognition in the Fre-Ger group. Contrast tests suggested that participants felt speech recognition quality in third round was significantly better than first (p<0.001) and second rounds (p<0.001), suggesting their speech style adjusted to the system. Figure 4.a shows the mean of raw scores for each round.

In the Eng-Ger group, we observed possible trend ($F_{2,140}$=2.622, p=0.076) of *Round* (Figure 4.a) as well as a significant main effect of *Interface* ($F_{1,140}$=9.197, p<0.01) on subjective speech recognition quality (Figure 5.a). Pairwise comparisons showed that participants felt the speech recognition quality of second round better than first round.

The results show that participants became used to the nuances of the system and learned *how they should speak* so the system could recognize their utterance. Note, however, the quality of the transcription in either group was far from perfect. We found that problems such as proper noun misrecognition and dialect act misclassification lowered the subjective scores; we discuss these issues further in the next section.

Participants in the Eng-Ger group felt that transcription quality in CC condition was significantly better than CC&TTS (p<0.01). A possible reason for the difference could be distraction in speech due to overlap with synthesized speech and resulting premature speech recognition.

Turn taking difficulty. We asked if participants felt turn taking was easy. The responses of participants in the Fre-Ger group showed a significant main effect of *Round* on turn taking difficulty ($F_{2,112}$=3.123, p<0.05). Pairwise comparisons showed that participants thought turn taking in third round was easier than second round (p<0.05). This suggests that participants became better at taking turns as they worked on more tasks.

We observed a significant main effect of *Interface* in the Eng-Ger group ($F_{2,224}$=3.434, p<0.05), but no difference between rounds. Participants in this group tend to prefer CC over CC&TTS (p=0.069) and TTS conditions (p=0.062) with regards to turn taking. Because TTS takes additional time for completion, participants, especially German speakers who understood their partner without translation, may have felt it was redundant and disrupted effective turn taking.

Figure 4.b shows the scores for turn taking difficulty in each round, and Figure 5.b shows the scores grouped by the interface condition. In both figures, we also show the turn taking difficulty scored in the English task that was performed at the end of the session. The results show that turn taking was harder in translation-mediated conversation compared to a normal remote conversation.

Overall quality of conversation. For each story telling task, we asked if participants agreed to a statement "*overall, we could hold a successful conversation.*" We found significant main effect of *Round* on overall success of a conversation in Fre-Ger pairs ($F_{2,112}$=6.275, p<0.01), and a possible trend in the Eng-Ger group ($F_{2,224}$=2.615, p=0.075). The responses from participants in Fre-Ger group also showed significant difference in rating between *Interface* conditions ($F_{1,112}$=3.676, p<0.05).

The contrast tests showed that third round was significantly more successful compared to first round (p<0.01) and second round (p<0.01) in the Fre-Ger group. Similarly, there was a trend that third round was more successful compared to first round (p=0.058) in the Eng-Ger group. Though some participants noted that they got used to using the application in early stage of a study session, the result suggest that it takes about two rounds (six tasks, which is approximately 16 min of conversation) to improve conversation quality.

Pairwise comparison revealed that participants in the Fre-Ger group rated perceived success higher in CC condition over CC&TTS condition (p<0.05) (Figure 5.c). Unexpectedly, we also observed a potential trend in preference of TTS condition over CC&TTS (p = 0.058). This was surprising, because the result contradicts with the overall interface preference, which was leaning towards favoring CC&TTS (Figure 3.b). This suggests that even conversation success is rated lower, participants wished to have both CC&TTS for as much information as possible to comprehend their partner's message.

Even though the quality of conversation was not as good as that of a remote conversation in English (Figure 4.c), we noted that, in the 2nd and 3rd rounds of the Eng-Ger group, the overall conversation quality was rated slightly above the neutral value of 3. However, the perceived conversation quality of the Fre-Ger group remained below 3 even after three rounds. In the next section, we discuss factors that may have affected the conversation quality.

Section Summary

The results suggest that despite limitations of the translator system, people do adapt their speaking styles as they get more familiar with these limitations. The results also suggest diverse preference in interface settings, highlighting the need of multimodal presentation (audio and text) of utterances as a fail-safe technique. Next, we contextualize the results by analyzing users' comments.

CONTENT ANALYSIS

To better understand *why* participants were satisfied or dissatisfied with the experience, we looked into their interview responses and free-text answerers in surveys. Interviews were recorded and transcribed by authors. Interview transcripts and survey responses were coded using a content analysis method; recurring themes were extracted in open coding step, then researchers developed a coding frame and combined similar themes to organize and identify findings [2]. Table 3 shows the final coding frame and number of participants who mentioned each theme. The remainder of this section will explain each theme in more detail together with exemplar quotes that represent participants' feelings.

Speech Recognition Errors

Among different types of speech recognition errors (*e.g.,* [18]), some were mentioned more frequently by participants in the current study: the system's inability to recognize *proper nouns* and *statement vs. question* classification.

	Fre-Ger	Eng-Ger	Overall
	(N=16)	(N=30)	(N=46)
Speech recognition errors			
Proper nouns	10 (62.5%)	16 (53.3%)	26 (56.5%)
Statement vs. question	2 (12.5%)	6 (20.0%)	8 (17.4%)
Translation errors			
Grammar and word order	6 (37.5%)	12 (40.0%)	18 (39.1%)
Idiomatic and colloquial expressions	6 (37.5%)	5 (16.7%)	11 (23.9%)
Comprehending imperfect messages			
Deciphering translated text	6 (37.5%)	11 (36.7%)	17 (37.0%)
Information from original speech	5 (31.3%)	10 (33.3%)	15 (32.6%)
Speech production and adaptation			
Closed caption as a feedback	10 (62.5%)	10 (33.3%)	20 (43.5%)
Adapting the way of speaking	14 (87.5%)	25 (83.3%)	39 (84.8%)
Repeat and rephrase	12 (75.0%)	17 (56.7%)	29 (63.0%)
Forgiving	6 (37.5%)	9 (30.0%)	15 (32.6%)
Difficulty in Turn Taking			
Latency	5 (31.3%)	15 (50.0%)	20 (43.5%)
Overlapping speech	4 (25.0%)	15 (50.0%)	19 (41.3%)

Table 3. Challenges identified using a translator tool coded from the user responses and number of participants who mentioned each theme.

Proper nouns. More than half of participants (56.5%) mentioned that the system did not recognize proper nouns (*e.g.,* a name of a person). For example, when a participant was asked to start a conversation with a starting sentence of *"Ava was planning a trip for the weekend,"* the name "Ava" could not be recognized.

"'Ava', there was no way it was getting 'Ava' no matter how many times I tried. So in German, when you say 'but,' it sounds very similar. So that's what it was picking up" – French-German Pair 4, German.

Unsuccessful recognition caused two problems: (i) it hindered participants from mutually agreeing on how to refer to a subject/object, and (ii) false recognition got translated into a different word(s) and changed the meaning of a speech. The issue, a case of an *out-of-vocabulary* problem [1,4], was severe primarily because of lack of effective adaptation techniques. Unlike other types of speech recognition errors (*e.g.,* incomplete utterance [18]), users could not fix the problem by repeating, rephrasing, or using simpler words.

Statement vs. question. Multiple participants (17.4%) mentioned that the speech recognition system did not correctly identify whether the utterance was a statement or a question. One participant noted:

"I found it really hard not to see or hear whether the translation is just making a statement or asking a question. And this was a total block in a conversation if you cannot make this decision"– E-G Pair 12, German.

The way one responds can change depending on the *act* of speech (*i.e.,* statement or question). For example, a response to "William likes tea" and "William likes tea?" could differ. Classifying a speech act is an open area of research [20] and there is no prescribed solution for a system to use.

Translation Errors

Translation errors affected quality of communication, and severity may have differed between interface conditions.

Grammar and word order. Some participants (39.1%) pointed out grammar and word orders were off. More interestingly, a participant noted that it affected a conversation more severely for the TTS condition.

"Because [...] German sentence structure is so different from English, [translated] sentence wouldn't make any sense when [TTS] said it, but when you see it on the screen, you can kind of reorganize the words in the way it supposed to go. And then it makes sense"– F-G Pair 3, German.

Trying to comprehend grammatically incorrect speech burdened one's working memory, but presence of a closed caption seemed to help. Some participants said "deciphering" a closed caption was their way of understanding their partner. This partially explains participants' preference towards CC and CC&TTS conditions (Figure 3).

Idiomatic and colloquial expressions. A participant's utterance did not get translated as they intended when it contained idiomatic and colloquial expressions. The challenge was mentioned by 23.9% of participants.

"Problem with me was that if you use a translator, it conveys the literal meaning but not the cultural one. Sometimes when you say some sentence in your own language, you can't translate it right way. In English, you need to find a different way to say that sentence" – F-G Pair 4, French.

People seemed to adapt and find workarounds to the problem where the system did not translate expression with non-literal meaning. For example, they rephrased and described their speech when their partner seemed confused. Though the problem may not be a showstopper, it certainly affected the quality of communication.

Comprehending Imperfect Messages

Although comprehending erroneous translation was hard, participants managed to understand their counterparts.

Deciphering translated text. Participants preferred settings where closed caption was available (Figure 3), and 37% noted that they could understand the meaning from text even there were some errors.

"[...] even if there are a lot of words that are switched in translation, it was still clear and the conversations made sense" – F-G Pair 5, French.

Because closed caption was a key for comprehending a message, its persistence on a screen was of great importance. Some participants complained about the design choice of the closed caption interface; when a person made consecutive utterances, the system removed a translated sentence and refreshed a closed caption with a new translation. Sometimes this happened too quickly that one could not finish reading the previous translation. One said with frustration:

"I disliked that the… translation sometimes, it flashes, and it's gone" – F-G Pair 8, German.

Information from original speech. 32.6% of participants mentioned that being able to hear their partner's voice was helpful. Original speech (i) maintained the feeling of talking to a human, and (ii) it allowed a participant to pick up information that was lost in translation.

"[Having partner's original voice] humanizes the relationships as well, because if you only get the robotic voice, I think it just cuts you off from the person you are engaged with. Whereas you talk, then it feels much more human relationship" – F-G Pair 1, French.

Another participant mentioned that original speech of their partner compensated for translation errors. For example, they could pick up basic words that were similar across languages. It even allowed participants to pick up proper nouns that were lost in speech recognition—a problem mentioned by more than half of participants.

"I know a couple of words 'nein' and 'ja', and a couple of other things from German. [...] I could pick up like 'Space Needle.' And there were a couple of words ... that I could pick up and help guide, like 'ok, is this what you mean?' as opposed to 'I don't know'" – E-G Pair 6, English.

Speech Production and Adaptation

To hold a successful conversation, participants adjusted the way they speak and how they interact with the system.

Closed caption as a feedback. A transcription of participant's own utterance helped them to see whether the system correctly recognized their utterance. Many of participants (43.5%) noted that it allowed them to adjust the way they spoke so that the system could recognize them well (*e.g.,* by speaking clearly).

"[Closed caption is useful] because you can see the mistakes better. Because people pronounce things differently than ... than it should be pronounced. If you see it, it's much easier to fix mistakes" – E-G Pair 4, German.

When both CC and TTS were on, recognized speech was displayed when TTS started. So when there were speech recognition errors, participants knew that translation was not going to make sense to their partner due to cascading errors. Participants identified a limitation of the design of the current system, which does not allow them to abort TTS. One participant said:

"I was seeing my own recognition in French, I started laughing because what was going to be sent was wrong. I saw very bad recognition. I knew I couldn't do anything, so I just had to wait for ... getting sidetracked by what you got or sometimes conversation totally getting fallen apart" – F-G Pair 7, French.

Adapting the way of speaking. Majority of participants (84.8%) mentioned that they adapted to the system and changed the way they spoke to accommodate the system. Main adaptation techniques included slowing down, enunciating, speaking louder, choosing simpler words, and constructing a simple sentence structure.

German: "I spoke more clearly, and a little bit slow." English: "Enunciate a little bit. Because we all have our own slung and we have a tendency to run words together, and software may not pick up" – E-G Pair 7, English & German.

However, sometimes speaking too slowly backfired. ASR uses a pause between speeches as a cue to segment an utterance; if there is too much of a pause, it misclassifies an intended pause as an end of a sentence consequently translating a premature utterance.

"I may have tried to speak slowly and clearly. Because [ASR] seemed bad. But it didn't do any good. Or it made the matter worse. [...] it didn't realize it was a pause (when he stopped between words to pronounce each word clearly)" – E-G Pair 11, English.

Repeat and rephrase. When their partners seemed lost, many participants (63.0%) found that not only repeating, but also rephrasing was a way to effectively fix the conversation.

"In normal conversation, you wouldn't just repeat it, but you would try to say it in a different way" – E-G Pair 8, German.

To request for repetition or rephrase, a listener used non-verbal cue as well as explicitly asking. One participant mentioned that she tried to convey that translation did not make sense via gesture and facial expressions.

Forgiving. Though they had to conform to the system and adjust the way they speak, about a third of participants (32.6%) said they did not feel uncomfortable because it was expected.

"My sentence became shorter, I pronounced more clearly, I was speaking ... I mean, if I were speaking to someone who's not a native German, and if I didn't have a translator, then I would also slow down, obviously. It was a similar experience to ... accommodate the machine" – G-F Pair 5, German.

This aligns with the behavior of people who hold inter-lingual conversation with a help of human simultaneous interpreter; when all parties are aware of the communication situation, people may be more cooperative and adapt their way of speech [10].

Turn Taking

System latency and overlap between participants' utterance and TTS distracted turn taking.

Latency. Many participants (43.5%) noted that latency, due to computation time of speech recognition, machine translation, and speech synthesis, disrupted the communication. Some also noted that waiting for TTS to finish speaking was cumbersome.

"The audio is slow and cumbersome, and, it kind of got in the way of the flow. I like the text better" – E-G Pair 4, English.

Some participants raised this as a reason for disliking CC&TTS condition. They knew what their partner had said from translated text, and did not want to wait for synthesized speech to finish playing.

Overlapping between speech and TTS. Nearly half of participants (41.3%) mentioned that their speech overlapped with translated TTS, which made it hard for them to speak.

> *"My partner and I would begin to speak then the translation from the previous statement would catch up, translation and speaker would then all be talking at the same time" – E-G Pair 4, English.*

Speech overlap was caused by two reasons: (i) speaker's disfluency and (ii) forgetting the presence of TTS. The system treated speaker's disfluency (*e.g.,* "um…" in a sentence) as an end of a sentence and it prematurely started translating and synthesizing speech. Sometimes participants forgot the presence of TTS; a participant started speaking when they understood what their partner said before hearing the translation. As a result, their speech collided with TTS that was played slightly later. This was most prominent in German speakers in the Eng-Ger group because they understood their partners' English without translation.

Section Summary

The analysis suggests that speech recognition errors, translation errors, and difficulties in turn taking decreased the usability of the system. To deal with the imperfect system, participants adjusted the way that they speak, and comprehended messages not only from translation but also from CC and original speech. Some adaptation techniques were only available in some interface setting (*e.g.,* deciphering CC). Some problems were more severe because of the lack of adaptation techniques demanding some fallback options to mitigate the problems.

DESIGN GUIDELINES

We discuss design guidelines that emerged as a result of the analysis. Though they are not meant to be exhaustive, we believe they cover fundamental requirements for designing typical SLT systems.

Support Users' Adaptation for Speech Production

It is unlikely that MT and ASR will perform perfectly in the near future. However, system designers can help improve user experience by designing their application to support users' adaptation for speech production. The translator application suffered from incorrectly translating sentences with complex word structures and idiomatic expressions. Incorporating feedback mechanisms to let users know when translation is unsuccessful would allow them to effectively repeat and rephrase their utterance.

Participants found closed caption useful for identifying speech recognition errors and subsequently adjusting the way that they spoke. Nevertheless, because there were no clear indications of *why* the recognition failed, participants tried various adaptation techniques. Typical speech recognition software produces a *confidence* score for transcription accuracy and the information can be presented along with other information. Presenting *why* it had low confidence to its users allows them to better adjust the way they speak. For example, a user can slow down their speech if the system shows low speech recognition confidence with fast utterance speed.

Offer Fallback Strategies with Non-verbal Input

We found that some speech recognition problems were more severe due to the lack of effective correction techniques. System designers should consider offering alternative input methods to provide fallback options for problems such as *proper noun* recognition and *statement vs. question* classification. Allowing editable text-based input on top of speech input would mitigate the problem with ASR's out-of-vocabulary problem.

Support Comprehension of Messages

Closed captions helped participants comprehend imperfect translation; however, some were frustrated with it not persisting on a screen. Providing a history of translation on a screen could help one to comprehend a long message. Interface designers should consider how to show old transcriptions/translations on limited display of real estate.

Participants mentioned that their partners' original speech could be used to extract information lost during translation. For example, it allows one to pick up a proper noun in a sentence and conveys unspoken *feeling* in it. At the same time, some people may find it redundant. As there would be diverse preference for hearing/not hearing original speech, we recommend system designers to allow users to easily adjust the volume of original speech.

Support Users' Turn Taking

Increasing awareness of the system state would benefit turn taking. People can use information about *what* the system is currently doing to determine when they can safely speak and avoid speech overlap with TTS. Showing when recognition and translation are occurring is also useful as users can see if the system is passing premature utterance to the machine translation component or not.

Preference for having or not having TTS was polarized; some preferred having it to gain additional information, and others mentioned that it disrupted turn taking. To accommodate participants with different preferences, we suggest designers to make it easier to turn TTS on and off. Other potential design directions to reduce turn taking cost would be: (i) make TTS smarter to avoid it speaking while a user is speaking; (ii) simplify isolating TTS from the original users' voice (*e.g.,* ducking the volume of original voice, 3D audio effect to simulate different audio source position); (iii) allow users to abort ASR and MT to avoid immature translation due to speech collision.

LIMITATIONS AND FUTURE WORK

Our German participants in Eng-Ger group had reasonable English proficiency. This may have impacted their experience with the tool as they were able to comprehend what their partner said without the translation. However, we feel that studying this population is important as some of our potential users could be those who understand the other language, but prefer speaking in their own language.

Though we compared subjective usability scores between conditions with translation and without translation (*i.e.,* in English), the better control setting would have been a conversation with a human interpreter. We could not set up this control setting because participants were recruited in rolling basis and, as a result, we could not hire a human interpreter in a timely manner (in fact, a lack of no flexibly available human interpreter is one motivation for developing SLT). Future work should involve a human interpreter and compare the usability of SLT with quality of experience with human-interpreter-mediated conversation.

CONCLUSION

We have investigated the effect of spoken language translation on user experience in interlingual conversation. The statistical results from our study with 8 French-German pairs and 15 English-German pairs (a total of 46 participants) illustrated evidence of participants' adaptation and suggested preference of diverse interface setting. Content analysis contextualized *how* participants adapted to produce and comprehend speech, and *why* participants preferred one interface setting to another. The design guidelines based on the findings will contribute to the better design of spoken language translation systems.

ACKNOWLEDGEMENTS

We thank the Skype Translator team at Microsoft for many discussions and the feedback on this research.

REFERENCES

1. Bazzi, I. and Glass, J.R. Modeling Out-Of-Vocabulary Words For Robust Speech Recognition. Doctoral Dissertation. 2000.

2. Berg, B.L. and Lune, H. *Qualitative research methods for the social sciences*. Pearson Boston, 2001.

3. Costantini, E., Pianesi, F., and Burger, S. The added value of multimodality in the NESPOLE! speech-to-speech translation system: an experimental study. *IEEE ICMI*, (2002), 235–240.

4. Creutz, M., Hirsimäki, T., Kurimo, M., et al. Morph-based Speech Recognition and Modeling of Out-of-vocabulary Words Across Languages. *ACM TSLP. 5*, 1 (2007), 3:1–3:29.

5. Deng, L. and Huang, X. Challenges in Adopting Speech Recognition. *Commun. ACM 47*, 1 (2004), 69–75.

6. Feely, A.J. and Harzing, A.-W. Language management in multinational companies. *IJCCM 10*, 2 (2003), 37–52.

7. Fügen, C., Waibel, A., and Kolss, M. Simultaneous Translation of Lectures and Speeches. *MT 21*, 4 (2007), 209–252.

8. Gao, G., Xu, B., Cosley, D., and Fussell, S.R. How Beliefs About the Presence of Machine Translation Impact Multilingual Collaborations. *Proc. of CSCW* (2014), 1549–1560.

9. Gao, G., Yamashita, N., Hautasaari, A.M.J., Echenique, A., and Fussell, S.R. Effects of Public vs. Private Automated Transcripts on Multiparty Communication Between Native and Non-native English Speakers. *Proc. of CHI* (2014), 843–852.

10. Gile, D. *Theoretical components in interpreter and translator training*. . John Benjamins Publishing Company, 2009.

11. Hamon, O., Fügen, C., Mostefa, D., et al. End-to-end Evaluation in Simultaneous Translation. *Proc. of ECACL* (2009), 345–353.

12. He, X. and Deng, L. Speech-Centric Information Processing: An Optimization-Oriented Approach. *Proc. of the IEEE 101*, 5 (2013), 1116–1135.

13. Janssen, C.P., Iqbal, S.T., and Ju, Y.-C. Sharing a driver's context with a caller via continuous audio cues to increase awareness about driver state. *J. of E. Psychology*, (2014).

14. Kale, E. and Syed, H.R. Language barriers and the use of interpreters in the public health services. A questionnaire-based survey. *PEC 81*, 2 (2010), 187–191.

15. Morita, N. Negotiating Participation and Identity in Second Language Academic Communities. *TESOL Quarterly 38*, 4 (2004), 573–603.

16. Nakamura, S., Markov, K., Nakaiwa, H., et al. The ATR multilingual speech-to-speech translation system. *IEEE Trans. on ASLP 14*, 2 (2006), 365–376.

17. Pan, Y., Jiang, D., Picheny, M., and Qin, Y. Effects of Real-time Transcription on Non-native Speaker's Comprehension in Computer-mediated Communications. *Proc. of CHI* (2009), 2353–2356.

18. Prasad, R., Kumar, R., Ananthakrishnan, S., et al. Active error detection and resolution for speech-to-speech translation. *IWSLT*, (2012), 150–157.

19. Singer, A. Contemporary Immigrant Gateways in Historical Perspective. *Daedalus 142*, 3 (2013), 76–91.

20. Stolcke, A., Coccaro, N., Bates, R., et al. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Comp. Ling. 26*, 3 (2000), 339–373.

21. Takezawa, T., Sugaya, F., Yokoo, A., and Yamamoto, S. A new evaluation method for speech translation systems and a case study on ATR-MATRIX from Japanese to English. *Proc. of MTS.* (1999), 299–307.

22. Waibel, A., Jain, A.N., McNair, A.E., Saito, H., Hauptmann, A.G., and Tebelskis, J. JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. *ICASSP*, (1991), 793–796 vol.2.

23. Wang, H.-C., Fussell, S., and Cosley, D. Machine Translation vs. Common Language: Effects on Idea Exchange in Cross-lingual Groups. *Proc. of CSCW* (2013), 935–944.

24. Wobbrock, J.O., Findlater, L., Gergle, D., and Higgins, J.J. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. *Proc. of CHI* (2011), 143–146.

25. Yamashita, N., Inaba, R., Kuzuoka, H., and Ishida, T. Difficulties in Establishing Common Ground in Multiparty Groups Using Machine Translation. *Proc. of CHI* (2009), 679–688.

26. Yamashita, N. and Ishida, T. Effects of Machine Translation on Collaborative Work. *Proc. of CSCW* (2006), 515–524.

27. AIIC-UN Agreement rates (viewed Sept 2014). http://aiic.net/page/275/aiic-un-agreement-rates-last-update-july-2014/lang/1.