Effects of Video and Text Support on Grounding in Multilingual Multiparty Audio Conferencing

Andy Echenique[1,2], Naomi Yamashita[2], Hideaki Kuzuoka[3] & Ari Hautasaari[2]

[1]University of California, Irvine
Department of Informatics
Irvine, California USA
echeniqa@uci.edu

[2]NTT Communication Science Labs
Kyoto, JPN
naomiy@acm.org
ari.hautasaari@lab.ntt.org

[3]University of Tsukuba
Faculty of Engineering
Information and Systems
Tsukuba, Ibaraki, JPN
kuzuoka@acm.org

ABSTRACT

With computer-mediated communication (CMC) tools allowing collaborations to span the globe, teams can include multiple collaborators located in different countries. Previous research shows how audio communication supplemented by video conferencing or text transcripts improves conversation grounding between native speakers (NS) and non-native speakers (NNS) in one-on-one multilingual collaborations. This research investigates how supplemental cues (video or real-time text transcripts) support NNSs' participation in multiparty audio conferences. We implemented a collaborative grounding task with triad groups of NS and NNS to investigate possible effects. We found that NNSs' task accuracy dropped significantly between video+audio trials. By comparison, NNSs' ability to understand common ground increased over trials in the text transcripts+audio condition. Our results demonstrate the difficulties of common ground establishment for NNS in multiparty collaborations and how the development of supporting tools for multilingual audio conferencing can aid NNSs' communication ability.

Author Keywords: Computer-Mediated Communication; Audio Conferencing; Multilingual Communication; Non-Native Speakers

ACM Classification Keywords

H.5.3 Group and Organization Interface: Computer-supported cooperative work

INTRODUCTION

Audio conferencing is among the most frequently used communication tools in global business and social interactions. It offers a convenient and cost-effective way for multiple collaborators located in different countries and time zones to communicate and contribute to decision-making processes.

Although audio conferencing tools connect distant collaborators, non-native speakers (NNS) experience difficulties when participating in multiparty audio conferences. In particular, audio conferencing tools challenge NNSs' ability to follow the conversation and reply [25]. Imperfect audio conditions (reverberations and extraneous noise) during an audio conference also limit NNSs' ability to perceive speech [14,15]. Furthermore, when NNS try to compensate for the missed information, their ability to think about current conversational content is likely to decline, resulting in an impaired ability to respond [21]. These problems become prominent in multiparty communication with mostly native speakers (NS) because the conversation can move forward rapidly while NNS are left behind [25]. It is therefore necessary for system designers to understand and lessen the burden imposed on NNS in audio conferencing.

The aim of this paper is to investigate what supplemental cues might better support NNSs' participation in multilingual, multiparty audio conferences. Our study is motivated by two sets of previous research. One is how common ground negotiation between NS and NNS improves with the addition of video feed to an audio channel [23]. The second is the use of real-time text transcripts and audio communication and its support of NNS comprehension [17]. Although a text transcript of an audio feed may be redundant for native speakers, it may help the NNS recover from missed information and cues by allowing them to view the conversation in text format. For example, NNS in East Asian countries perform better in reading tasks compared to listening tasks, as the education systems focuses heavily on reading comprehension [21].

Previous research therefore leads us to the following research question: Does adding video or real-time text transcripts to audio conferencing assist non-native speakers negotiate common ground communication with native speakers? With previous studies already identifying how adding text transcripts and video can improve communication between native and non-native English speakers when compared to audio only, we aim to compare these two supplemental communication media. We further

CABS'14, August 20–22, 2014, Kyoto, Japan.
Copyright © ACM 978-1-4503-2557-8/14/08...$15.00.
http://dx.doi.org/10.1145/2631488.2631497

investigate how these added cues affect the grounding process between native and non-native speakers in a multiparty group.

We conducted a laboratory experiment investigating two communication media: audio+video and audio+real-time text transcripts. Twelve groups of NS and NNS participants (each group consisting of two NS and one NNS) participated in a tangram-matching task designed to investigate the grounding process of common references among participants.

Building off previous literature supporting the addition of supplemental communication media to audio channels in multilingual collaboration [17, 23], our results suggest that text support (real-time transcripts) in audio conferencing helps NNS retain and repair common ground between the NS and NNS during repeated referential communication. Task results while using video support exhibited a degradation of common ground during continued collaboration. Although both video and text-transcript support have been studied in previous literature, this research will identify how these technologies mediate three person groups and grounding for NNS.

We will discuss the prior research, describe our experimental design, and results. Finally, we conclude with a discussion of our findings and draw design implications for the development of tools to support multilingual audio conferencing.

RELATED WORK

NNS Communication Difficulties during CMC

Computer-mediated communication (CMC) imposes a variety of challenges for non-native speakers (NNS) [16,19,20]. Multiple parallel processes, such as speech recognition, foreign language production, recovering from missed conversational context, and intensive thinking can overwhelm NNS [15,18]. CMC, such as audio conferencing, is susceptible to extraneous noise from multiple sources, which make hearing some utterances difficult and can further impact NNSs' to follow the conversation [14]. In addition to low quality audio signals, problems with participants rapidly advancing the conversation without checking others' understanding also hinder NNS performance [25].

Previous work has attempted to improve audio conferencing for NNS [17,23]. In multiparty settings (more than two collaborators), previous research has also tried to reduce the cognitive load placed on NNS by providing him/her with additional processing time [25] or providing supplemental cues such as text transcripts [7].

Additional Communication Media to Support NNS

Previous research demonstrates how adding additional communication media to audio improves collaboration for NNS. Veinott et al. (1999) found that NNS pairs (in this instance indicating teams of English as a non-primary language or live in and English speaking country for more than 4 years) can establish common ground more efficiently with video+audio compared to only audio [23]. This finding is further significant given that no difference was seen for NS pairs. NNS pairs also had fewer miscommunications with video+audio compared to only audio [22]. These studies support the use of video, and the non-verbal cues they provide, as assisting NNS during CMC.

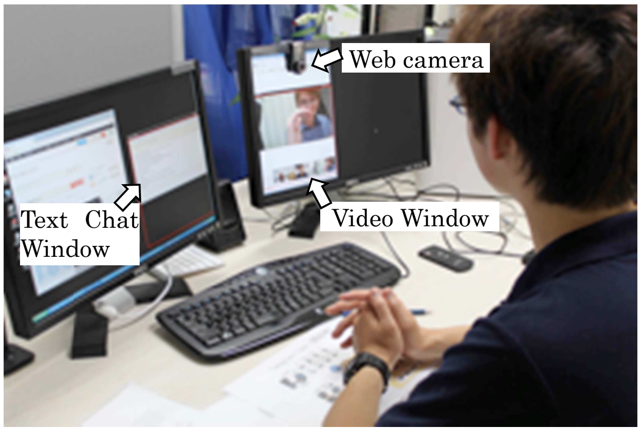

Figure 1. Dual monitor setup with video and text interface during pre-experimental introductions

Augmenting audio communication with additional text transcripts can further aid NNS during computer-mediated communication. Pan et al. (2009) studied the effects of adding text transcripts to audio and audio+video recordings to NNS in a non-interactive setting [17]. Their results show that adding transcripts improved comprehension in both conditions, although performance between audio+video+text transcripts and audio+text transcripts did not differ significantly. Extending these results into an interactive setting, Gao et. al. (2014) investigated how speech-to-text transcripts affected multiparty communication between NS and NNS [7]. Text transcripts increased the NNSs' comprehension, yet the necessity of reading lengthy transcripts with errors imposed a significant burden on NNS. Additionally, imperfect speech-to-text transcript accuracy strongly influenced the comprehension of NNSs.

In sum, prior research suggests that adding video or text transcript to multilingual audio conferencing can improve communication for NNS. Video can provide NNS with non-verbal cues that assist communication. Text transcripts can also increase comprehension, but are susceptible to poor accuracy rates and can place a cognitive load on NNS. Our research extends this previous knowledge by comparing both these supplemental media in a multiparty setting. This comparison furthers our understanding of how CMC mediates multilingual communication and improves or deteriorates NNSs' resulting collaborative capabilities.

Grounding in Computer-Mediated Communication

Grounding is the process by which conversational participants attempts to establish a common, shared understanding [9]. One way to examine peoples' grounding processes is to examine referential communication where speakers and addressees work together to establish common ground on something in the environment [5]. Once speakers and addressees agree on the perspective included in a common referent (the thing being described), this mapping between the perspective (reference) and the object (referent) indicates that participants have established common ground.

The process of grounding illustrates how the refinement of ideas and perspectives helps collaborators decide on a common language for communication. Once these references are agreed upon, the supporting concepts used to narrow down their meaning are no longer explicitly mentioned. An example is how longer, broader descriptions used to reference an object are later shortened to simple words or phrases when referring back to it. This process is known as lexical entrainment [1]. Studies on referential communication describe how conversational participants entrain towards an expression by abbreviating their referring expressions in repetitive trials. Given the critical nature of conversation grounding in collaboration, this is our core method of evaluating common ground establishment among participants during and after group interaction.

Previous research suggests that participants develop different strategies to effectively build common ground depending on the information available in the medium they are using [26]. NNS who are not fluent in their non-native language have different communication needs than NS, and their grounding process may differ between different media conditions. To enhance multilingual collaborations, it is vital to understand entrainment on a common reference within a group across different communication media.

METHODS

Overview

We investigated how the use of either video+audio (Video condition) or text transcripts+audio (Text condition) impacts conversation grounding between native English speakers (NS) and non-native English speakers (NNS) in a multiparty (triad) collaborative setting. We used a within subjects laboratory experimental design, with each group performing both Video and Text conditions

Participants

Each group consisted of three participants: two native English speakers (NS) and one non-native English speaker (NNS). The NNS in this experiment were Japanese native speakers. None of the NNS participants lived in an English speaking country for more than 2 years. We required all NNS to have a minimum TOEIC[1] English proficiency test score of 550. Overall, 12 groups participated (24 NS and 12 NNS) in total.

Task and Experiment Design

Tangram-matching tasks are frequently used to study common ground establishment in laboratory settings (e.g., [26]). In a tangram-matching task, participants are instructed to arrange an identical set of tangram figures (black polygon silhouettes) into matching orders. In our study, we assigned one participant the role of Leader and gave them a set of numbered figures in a predetermined order during each trial. The remaining participants, Followers, were given the same figures in random orders (i.e. each tangram was assigned a different number and serial order than that of the Leader) (Figure 2). We instructed the Leader to assist the Followers in matching the tangrams with the same numbers as are on his/her sheet.

Participants performed two tangram-matching trials during each condition to examine lexical entrainment of common references. Task sheets used between the first and second trial had an identical, but differently ordered, set of tangrams. A NS was always assigned as the Leader and a Follower (NS Follower). The NNS are always the second Follower (NNS Follower). Once assigned, roles stay consistent for the entirety of the experiment. We counterbalanced condition order across groups, with six groups performing the Video condition first and six groups completing the Text condition first. Participants would thus perform two tangram-matching trials in two conditions for a total of four tangram-matching trials performed over the course of the study.

In the Video condition, we provided a video feed showing each participant's face and upper torso to the other participants (Figure 1, only the right-side screen was used and left-side screen was turned off). Participants were not allowed to use the video feature show their task sheets or illustrations to their collaborators. In the Text condition, we provided a text chat window where participants can type direct transcripts and keywords during the task (Figure 1, right-side screen was turned off and only the left-side screen was used).

In the Text condition, we asked participants to type down the keywords or essential parts of their own utterances in a text window. Previous research demonstrates how imperfect text transcripts can increase NNS burden and impair comprehension and performance [7]. In order to maximize the positive effects of text support on NNS grounding and comprehension, we opted to use this participant-entered text input rather than imperfect transcripts, even though it may impose a burden on the NS participants. We believe that ASR and keyword extraction techniques may be used as a substitute for human-entered

[1] TOEIC: Test of English for International Communication (http://www.ets.org/toeic). TOEIC score of 550 is about the average level of Japanese university students (TOEIC Program 2012 Data Analysis)

transcripts in the future when they reach sufficiently high accuracy. Thus, our experimental setting mimics future technological systems where speech-to-text software and conversation semantic analysis becomes highly sophisticated.

Materials and Equipment

Tangram-matching task sheets. During each trial, participants received a paper task sheet with 10 tangrams. Each sheet included the same tangrams, but in a different serial order. Only the Leader's tangram sheet contained a number above each tangram (Figure 2).

Video and Text interface. We seated each participant in front of a dual monitor setup (Figure 1) in separate rooms. The right screen displayed the video conferencing interface and the left displayed the text interface. The experiment organizers turned on only the relevant screen for each experimental condition (Video: right screen only, Text: left screen only).

We used Google Hangouts [2] to transmit audio in both conditions and video in the Video condition. We positioned the web cameras so that only participants' upper torso and face were captured. Each participant's video feed was visible on the bottom of the screen. As they talked, each participant's video feed was displayed in the large window on Google Hangouts interface.

Google Talk was used during the Text condition. The text chat window was adjusted to same size as the video window (Figure 1).

Tangram Common Reference Survey. After performing each tangram-matching task, we handed all participants a separate sheet of paper with a picture of each tangram used during the task. We asked participants to write down the common references and/or keyword used during the task. Each participant completed the Survey individually without access to any materials including notes and chat logs.

Procedure

We divided the study into three portions: self-introductions, the first condition and the second condition. At the beginning of the experiment, we turned on both displays and participants stated their name, where they are from, and typed their name into the text chat. Following the self-introductions, the experiment organizers turned off the monitor not relevant to the first condition.

In both the Video and Text conditions, participants completed an initial training trial with five tangrams to familiarize themselves with the tangram-matching task and the communication media. After the training task, the participants completed the first task of the main tangram-matching task with 10 tangrams (Trial 1). There was no time limit set for the task completion.

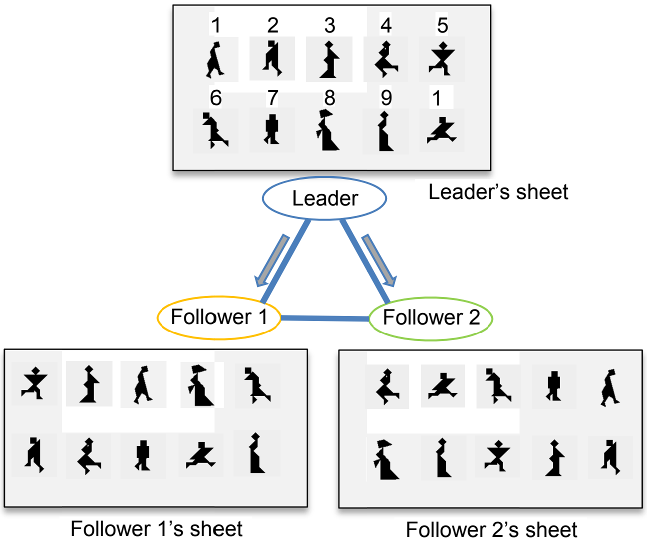

Figure 2. Diagram of experimental material setup.

After the first trial, we asked participants to complete a Tangram Common Reference Survey. After writing down the common references, the experiment organizers distributed the tangram-matching sheets for Trial 2. For trials 1 and 2, the same tangrams were used, but in a different serial position on the page and with different numbers. After completing the second trial, the participants again filled in a common reference survey.

Once two trials were finished in the first condition, the experiment organizers switched the displays and distributed the training task for the second condition. The remainder of the experiment followed the same pattern with different tangram-matching sheets (different tangrams) used for each condition. Each condition thus consists of two trials, which we label according the communication media and which trial order it is. Thus, Video is split between Video 1 and Video 2 and Text contains Text 1 and Text 2. The tangram-matching sheets assigned to each condition were counterbalanced.

Measures

The measures used will be detailed and discussed as they relate to our findings.

Task performance. We measured each Follower's task performance by how accurately they completed each tangram-matching task (i.e., the number of correctly matched tangrams according to the Leader's tangram sheet). This score acts as an indication of how successfully common ground was achieved between the Leader and Followers.

Non-verbal gestures. We videotaped the conversations with the same web camera used in the Video condition, which was placed between both screens. We analyzed the video data to understand of how participants used non-verbal gestures during the grounding process in a multilingual multiparty group. For non-verbal gestures in

[2] http://www.google.com/hangouts/

this study, we focused on hand gestures and body gestures directly relevant to the task and used to create common ground between participants.

NNSs' Missed References The Tangram Common Reference Survey responses indicated whether participants share a common reference (or expression/word) to identify each tangram during each trial. This is accomplished by comparing each participant's referring expressions of the same referent (tangram). As we are interested in NNSs' difficulties during the lexical entrainment process, we focused on instances where both NS used the same common reference yet the NNS did not. First, we excluded the tangrams for which NS Followers and Leaders did not have the same common reference. From this corpus, we calculated the percentages of the tangrams that the NNS did not report the same reference. This measure represents the percent of missed references the NNS failed to acknowledge, represented by a number between 0 and 1. 1 indicates 100% of the NNSs' Tangram Common Reference Survey references did not match those of the NS and 0 indicates that NNS missed no common references.

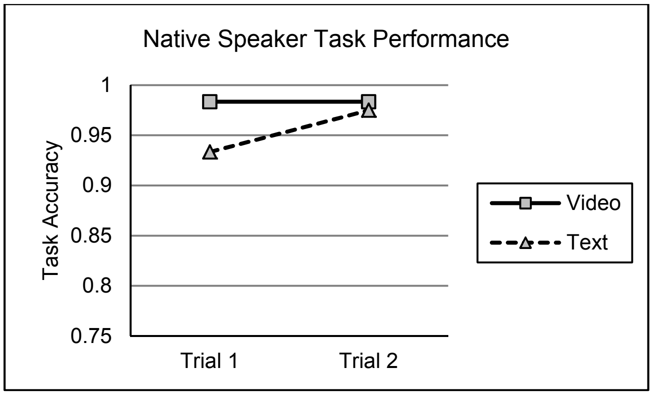

Figure 3. Native speaker task accuracy (1 indicating 100% task accuracy)

RESULTS

Our results measure how lexical entrainment and grounding between two NS (a leader and a follower) and NNS (a follower) is impacted by the two conditions: audio+video (Video) and audio+text transcripts (Text).

First, we present the task performance results, followed by an analysis of the non-verbal gestures and NNSs' Missed References, which speak to the lexical entrainment of common references. These results indicate how supplementary communication channels (Video or Text) support NNS in multiparty, multilingual communication.

Task Performance

We scored each Follower's tangram-matching sheet on how accurately participants assigned each tangram the same number as on the Leader's sheet. These scores range from 0 to 1, with 1 representing a perfect match and 0 indicating that no tangrams were assigned the same number. We used a Wilcoxon Signed Ranked test to compare our results due to the nonparametric nature and within subject design of our study.

The results for the NS follower indicate a strong ceiling effect. As Figure 3 shows, Trial 1 performance on both the Video and Text conditions were close to perfect, with an average accuracy of 0.98 on Video 1 and 0.93 on Text 1. These averages were not statistically different (p = 0.18, Z = -1.3). Trial 2 scores also indicated a strong ceiling effect, with Video 2 performance averaging at 0.98 and 0.98 for Text 2. Differences between these scores were also not statistically significant (p=0.32, Z = -1.0).

We compared Trial 1 and Trial 2 performance to investigate the possible interaction between the continued use of a communication channel and common ground establishment. Comparisons between Video 1 and Video 2 demonstrated no statistical difference (p = 1.0, Z = 0.00). A similar trend was present between Text 1 and Text 2, with no statistical difference observed (p = 0.10, Z = -1.6). These results indicate that NS had little difficulty with the task and that the supplemental communication media had little effect on their task performance.

For the NNS, communication media did have a noticeable effect on task performance (Figure 4). Although there was no difference when comparing Video and Text trials of the same order, comparison between Trials 1 and Trials 2 did indicate one of the communication media as detrimental during continued use.

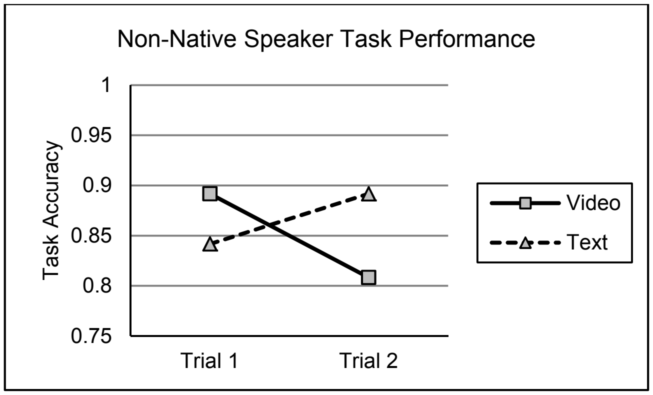

Figure 4. Non-native speaker task accuracy (1 indicating 100% task accuracy)

Comparisons between NNSs' Video and Text resulted in no statistical difference between their performances: Video 1 and Text 1 (0.89 and 0.84 respectively) show no difference in performance (p = 0.55, Z = -0.60); Video 2 and Text 2 accuracy (0.81 and 0.89 respectively) were also not statistically different (p = 0.23, Z = -1.2). However, comparisons between Trials 1 and 2 did show Video having a strong effect on performance. Video 2 indicated a statistically significant drop in performance relative to Video 1 (p = 0.04 Z = -2.4). Interestingly, this drop is only seen in the Video condition and not found in the Text condition - in fact, performance seemed to improve slightly in the Text condition although the improvement was not statistically significant (p = 0.34, Z = -0.95).

The drop in performance for the NNS in the Video condition indicates two possible difficulties for completing the Tangram task. First is that there may be less information (non-verbal gestures) provided by the leader through video channel in the second trial. Another possibility is NS Leaders may have assumed that NNS achieved common ground and removed descriptive phrases before grounding was achieved. This process will be visible through analysis of the commonality of the NS and NNSs' Survey answers. Our subsequent analysis will focus on investigating these possibilities.

Video Task Gestures

During the Video condition, the video channel was mainly used for providing supplemental non-verbal information. This information is normally provided by the Leader to clarify or provide further details on how a tangram looked or details that differentiate similar Tangrams. An example is the following excerpt:

> *Leader: Number 4, it's almost like he has a big martini glass (pretends to hold martini glass). He's relaxing, having a drink. Or maybe he's holding some 'ramen' (pretends to hold bowl of soup next to face)."*

For the NNS, such gestures seemed useful, resulting in high task performance in the first trial (Figure 4). Yet given the drop in task performance between the first and second Video conditions, a comparison was made between the number of hand and body gestures made during Video 1 and Video 2 for possible indications of its use effecting task performance.

As we speculated, a significant drop in the number of gestures was detected, from 33.4 gestures per trial in Video 1 to 19.4 gestures per trial in Video 2 (p <0.01, Z=2.8). These results show that native speakers used fewer gestures during repeated communication, affecting NNSs' task performance on subsequent trials.

Common References

We analyzed the Tangram Common Reference Surveys in order to identify how each supplemental media allowed NNS to detect the common reference after each trial. This measure indicates each participant's retention of the common reference for each tangram used during the trial. When some members during collaboration achieve a basis of communication via these references, yet others do not, it can imply misunderstandings or communication difficulties for some participants. From this data, we calculated the ratio of tangrams for which the NNS and NSs (NS leader and NS follower) did not achieve the same level of lexical entrainment. A lower number (0 being the lowest) indicates that the NNS missed fewer common references shared between the NS leader and the NS Follower, thus a lower ratio being a sign of better grounding.

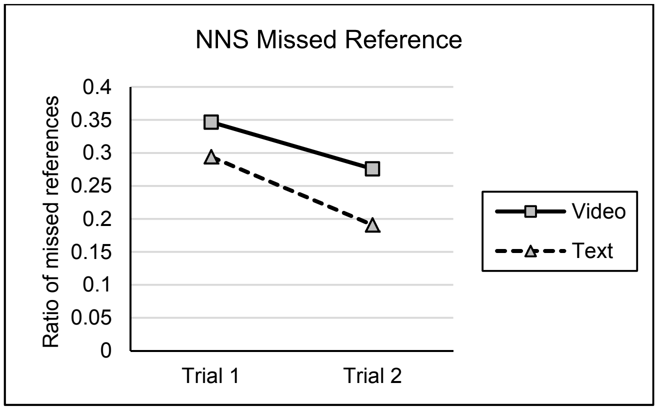

Figure 5. Non-native speaker Missed Reference (0 indicating no missed references)

While we found a consistent trend in NNS missing fewer references in the Text condition, the differences were not significant: the difference of average NNS Missed References scores between Video 1 (0.35) and Text 1 (0.29) were not significantly different (p = 0.50, Z=-0.67). We found the same between Video 2 (0.28) and Text 2 (0.19) (p = 0.20, Z = -1.3).

Trial order again had a strong effect on the survey results. While NNS Missed References in Video 1 and Video 2 showed no significant difference (p = 0.14, Z = -1.5), NNS Missed References significantly decreased from Text 1 to Text 2 (p < .05, Z = -2.7).

These findings indicate that as groups continue collaborating on a second trial, Text transcripts assist NNS in repairing misunderstood common references. For example, if both Leader and NS Follower labeled a tangram as "hawk" during Trial 1, yet NNS Follower labeled it as "owl", a misunderstanding is apparent on NNSs' part. In Trial 2, this misunderstanding is more likely repaired by the NNS when using supplemental real-time text transcripts. Thus as the groups collaborate, Text transcripts are more effective at repairing common ground misunderstandings.

DISCUSSION

Our results present three findings:

1. Non-native English speaker's (NNS) ability to match referents with common references (ground) diminished between Video 1 and Video 2.
2. Native speaker leader (NS Leader) reduced the amount of non-verbal gestures provided between Video 1 and Video 2.
3. NNSs' number of Missed References decreased between Text 1 and Text 2 (ability to detect common references increased).

These findings suggest two factors affecting communication for NNS. The first is the differences between NS and NNS in establishing and retaining common ground through repeated communication (trials 1 and 2) as seen through task performance. The second is how computer-mediation mitigates these differences in common ground establishment.

Differences in Common Ground Establishment between NS and NNS

Our findings highlight the differences between NS and NNS in establishing common ground, especially during repeated, multiparty collaboration. As demonstrated in Figure 3, NS had little trouble completing the tasks accurately with either supplemental communication medium throughout the experiment. Our results supports previous findings [19], demonstrating NSs' comprehension during audio conferencing as not changing significantly with the addition of supplemental communication media. Between the NS leader and NS Follower, common ground was established easily and confidently within the first trial.

The ease with which both NS established common ground likely influenced Leader's to reduce non-verbal gestures during repeated Video trials (Video 2), as seen in our results. Communication between NS became more efficient and less supporting information was provided. In some cases, NS Followers even notified Leaders that they don't need clarification, and dissuaded them from doing so:

> *Leader: John, you're fine?*
>
> *NS Follower: Yeah, I'm good. If I have a question, I'll let you know.*

This process was not specific to our study, as lexical entrainment and refining of established phrases or references during collaboration is a common behavior associated with establishing common ground [16,26]. Even with this reduction of non-verbal information, NS Followers' task performance was still consistent.

NNSs' results demonstrate that common ground is not as easily attained for them. The reduction in tangram-matching task performance seen in our study reflects a weak common ground understanding for NNS. The reduction of Leader's non-verbal gestures likely had a large effect on NNSs' Video 2 task performance results, with NS not acknowledging the status of the NNS Follower's understanding and the value of non-verbal gestures for NNS. NS Leaders may also have overestimated NNSs' level of understanding, and expected verbal feedback when common ground was not reached. Thus, just as the NS follower's silence indicated understanding, the same may not have been the same for NNS. The combined effect of NS Leader's lack of awareness of NNSs' imperfect understanding as well as the positive reinforcement for streamlined communication from the NS Follower demonstrate how multiparty settings further tax NNS in multilingual communications.

Communication Media and Common Ground Establishment for NNS

Communication media also had a strong effect on common ground establishment, as seen in our analysis of the Video and Text condition results. During repeated trials, continued use of Video as a supplemental communication media resulted in a reduction in common ground establishment. By comparison, the ratio of NNSs' missed references improved between Text 1 and Text 2. These results illustrates that even when common ground is not perfectly attained in the first trial (i.e. succeed in identifying the same tangram but having different references), Text transcripts may help NNS achieve common ground in subsequent trials by repeatedly and visually showing the key referring expressions. By doing so, NNS repair their references to the same reference shared among other NS members. Text may thus be a more robust supplemental communication media compared to Video during repeated collaboration.

Design Implications

The findings and discussion from our study support the implementation of instant messaging and text transcriptions for NNS collaborating with two or more NS. Especially if the collaboration is intended to extend beyond a single interaction among collaborators, Text is expected to provide continued benefits to the group. Given our use of a grounding task, our findings also demonstrate how negotiating common ground is affected by the use of these supplemental communication media. This understanding of how common ground is achieved, maintained, or deteriorated through Text or Video has significant implications for their efficiency and use in communication.

A technological implementation that would extend this work would be to automate text chat creation. During our study, in order to ensure accuracy, we asked NS to write down keywords. Automating this process may reduce the cognitive load on NS during the task. Allowing NS to edit and correct erroneously transcripts produced by ASR technology may also further the accuracy of this type of implementation.

Our research also hints at the beneficial aspects of combining Video and Text transcripts. During Video 1, task accuracy was similar to those of Text, indicating that the non-verbal gestures provided through video channel were useful for NNS. As the drop in NNS task performance in Video 2 coincided with a reduction of gestures, combining the additional cues Video provided in a persistent and reviewable manner would likely help NNS. This can be implemented by allowing the recording of gestures over video as they are performed. Recorded video gestures can then be tagged with a keyword (much like our Text condition). This allows for easier reference in the future and allows the supplemental visual information to be completely preserved.

CONCLUSION AND FUTURE WORK

Our research extends previous research in computer-mediated multilingual collaboration by providing a comparison of supplemental real-time transcripts or video conferencing when used alongside audio communication in a multiparty context. This study describes some of the methods in which each of these communication media mitigate common ground establishment for NNS. Video conferencing, though initially effective at negotiating common ground, is suspect to a decrease in task performance during repeated trials. Real-time text transcripts, conversely, are better at assisting NNS in repairing common references as displayed through common reference analysis during a second trial. Thus, Text and Video may affect NNS differently, yet Text seems to be a better supplemental communication media over repeated collaborations.

Our results thus pose two considerations for computer-mediation in multilingual, multiparty collaboration. The first is that NS and NNS do not build, repair, or attend to common ground in the same way. As seen in previous research, NNS are frequently faced with considerable difficulties in distance communication. When collaborating in larger groups, NNS' communication abilities are further taxed and their necessity for repair may go unnoticed.

The second consideration is how communication media mitigate communication difficulties, such as a lack of common ground. Repeated collaboration relies heavily on a firm foundation of common terms to ease communication. Video conferencing, due to an observed reduction in the necessary support information NNS need, significantly affects NNSs' communicative capabilities. By comparison, text transcripts are a more recognizable and comprehensible supplemental communication medium for NNS, and assist in keeping performance consistent and even promote common ground repair over time.

Future work should elaborate on how much detail Text should contain in order to be effective for the NNS during the collaboration. Given that the text chat in our study did not provide a detailed transcription of the audio, understanding how varying levels of detail in the text chat affects collaboration would give further insight to its use. In addition, text chat in our study was sourced from participants. Future work may compare text selected by the collaborators or by an automated process (such as transcripts and keywords automatically generated and extracted by the system). We believe that these results highlight non-native speakers' difficulties in multiparty CMC and how communication media can aid and advance collaboration.

ACKNOWLEDGMENTS

The authors would like to thank the support and assistance from all the members of the NTT Communication Science Laboratory, who have been tremendous partners over the course of the project. We would also like to thank the anonymous reviewers for their feedback and time.

REFERENCES

1. Brennan, S.E. Lexical entrainment in spontaneous dialogue. In *Proc. International Symposium on Spoken Dialogue* 1996, Acoustical Society of Japan (1996), 41-44.
2. Broadbent, D. Perception and communication. Oxford: Pergamon (1958).
3. Cho, H., Ishida, T., Yamashita, N., Koda, T. and Takasaki, T. Human detection of cultural differences in pictogram interpretations. In *Proc. IWIC 2009*, ACM (2009), 165–174.
4. Clark, H.H. and Marshall, C.E. Definite reference and mutual knowledge. In *Joshi, A.K., Webber, B.L and I. A. Sag, L.A. (eds.) Elements of discourse understanding*, Cambridge University Press (1981), 10-63.
5. Clark, H.H. and Wilkes-Gibbs, D. Referring as a collaborative process. *Cognition 22* (1986), 1-39.
6. Diamant, E.I., Fussell, S.R. and Lo,F.-L. Collaborating across cultural and technological boundaries: Team culture and information use in a map navigation task. In *Proc. IWIC 2009*, ACM (2009), 175–184.
7. Gao, G., Yamashita, N., Hautasaari, A., Echenique, A., and Fussell, S., Effects of public vs. private automated transcripts on multiparty communication between native and non-native English speakers. In *Proc. CHI '14*, ACM (2014).
8. Hirai, A. The relationship between listening and reading rates of Japanese EFL learners. *The Modern Language Journal 83, 3* (1999), 367-384.
9. Isaacs, E. and Clark, H.H. References in conversation between experts and novices. *Experimental Psychology, 16, 1* (1987), 26-37.
10. Jensen, C., Farnham, S.D., Drucker, S.M. and Kollock, P. The effect of communication modality on cooperation in online environments. In *Proc. CHI 2000*, ACM (2000), 470–477.
11. Koschmann, T., and LeBaron, C.D. Reconsidering common ground: Examining Clark's contribution theory in the OR. In *Proc. ECSCW 2003*, Kluwer Academic Publishing (2003), 81-98.
12. Krauss, R.M. and Weinheimer, S. Changes in reference phases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science 1* (1964), 113-114.
13. Li, Y., Li, H., Mädche, A. and Rau, P-L. P. Are you a trustworthy partner in a cross-cultural virtual environment?: Behavioral cultural intelligence and receptivity-based trust in virtual collaboration. In *Proc. ICIC 2012*, ACM (2012), 87–96.
14. Luisa, M., Lecumberri, G., Cooke, M. and Culter, A. Non-native speech perception in adverse conditions: A Review. *Speech Communication, 52* (2010), 864-886.

15. Nabelek, A.K. and Donahue, A.M. Perception of consonants in reverberation by native and non-native listeners. *Journal of Acoustical Society of America, 75* (1984), 632-634.
16. Olson, G.M. and Olson, J.S. Distance matters. *Human-Computer Interaction 15* (2000), 139-179.
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. In *Proc. CHI 2009*, ACM (2009), 2353–2356.
18. Rogers, C., Lister, J., Febo, D., Besing, J. and Abrams, H. Effects of bilingualism, noise and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics, 27* (2006), 465-485.
19. Setlock, L.D., Fussell, S.R. and Neuwirth, C. Taking it out of context: Collaborating within and across cultures in face-to-face settings and via instant messaging. In *Proc. CSCW 2004*, ACM (2004), 604-613.
20. Setlock, L.D., Fussell, S.R., Ji, E. and Culver, M. Sorry to interrupt: Asian media preferences in cross-cultural collaborations. In *Proc. IWIC 2009*, ACM (2009), 309–312.
21. Takano, Y. & Noda, A. A temporary decline of thinking ability during foreign language processing. *Journal of Cross-Cultural Psychology, 24* (1993), 445-462.
22. Veinott, E.S., Olson, J.S., Olson, G.M. and Fu, X. Video matters!: When communication ability is stressed, video helps. In *Proc. CHI EA 1997*, ACM (1997), 315–316.
23. Veinott, E.S., Olson, J., Olson, G.M. and Fu, X. Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proc. CHI 1999*, ACM (1999), 302–309.
24. Wang, H-C. and Fussell, S.R. Cultural adaptation of conversational style in intercultural computer-mediated group brainstorming. In *Proc. IWIC 2009*, ACM (2009), 317–320.
25. Yamashita, N., Echenique, A., Ishida, T. and Hautasaari, A. Lost in transmittance: How transmission lag enhances and deteriorates multilingual collaboration. In *Proc. CSCW 2013*, ACM (2013), 923-934.
26. Yamashita, N., Inaba, R., Kuzuoka, H. and Ishida, T. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proc. CHI 2009*, ACM (2009), 679–688.