Lost in Transmittance: How Transmission Lag Enhances and Deteriorates Multilingual Collaboration

Naomi Yamashita[1], Andy Echenique[2], Toru Ishida[3], Ari Hautasaari[3]

[1] NTT Communication Science Labs. Kyoto, Japan. naomiy@acm.org

[2] University of California, Irvine apexinandy@gmail.com

[3] Kyoto University ishida@i.kyoto-u.ac.jp arihau@ai.soc.i.kyoto-u.ac.jp

ABSTRACT

Previous research has shown that audio communication is particularly difficult for non-native speakers (NNS) during multilingual collaborations. Especially when audio signals become distorted, NNS are overburdened by not only having to communicate with imperfect language skills, but also compensating for the deteriorations. Under these faulty audio conditions, NNS need to pay extra time and effort to understand the conversation. In order to give NNS more time to process conversations, we tested the insertion of silent gaps (from 0.2 to 0.4 seconds) between conversational turns. First, gaps were inserted into a previously taped conversation, resulting in a significant improvement of NNS's understanding of the conversation. Second, gaps were inserted during a real-time audio conference by adding artificial delay between native speakers. The results show that the added delays have a combination of beneficial and detrimental effects for both native and non-native speakers. The findings have implications towards how audio conferencing can be improved for NNS.

Author Keywords

Audio conferencing, non-native speakers, mental resource, silent gaps, delay.

ACM Classification Keywords

H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work, Synchronous interaction.

INTRODUCTION

Audio conferencing is one of the most highly used communication tools in global enterprises and social organizations. It is very cost-effective, convenient, and allows rapid decision making among people in different locations.

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

Even though the use of audio conferencing continues to grow very rapidly, conference calls impose an excessive burden on non-native speakers (NNSs) [13, 15, 16]. In an audio conference, it is generally difficult to follow everyone's speech; audio quality is typically low, some terms or even some voices cannot be heard, and there tends to be extraneous noise. It is also sometimes difficult to identify who is speaking. Research has demonstrated that non-native speakers have particular difficulty perceiving speech in such imperfect conditions because they cannot instantly come up with a range of alternative possibilities once they miss a term [13, 16]. According to Rogers, "even true bilinguals cannot reach the ability of monolinguals in the presence of noise [17]." It is thus important for system designers to consider the plight of non-native speakers in imperfect audio conditions.

To date, however, virtually no research has focused on supporting non-native speakers in audio conferences. Although researchers have proposed novel solutions to improve the current audio conference (e.g., improving sound quality and adding visual cues [3, 4]), those approaches are not specifically designed for non-native speakers.

Our interest is in supporting non-native speakers in audio conferences. As a preliminary investigation to capture the difficulty they face when joining an audio conference, we conducted a small survey and a follow-up interview in the Japanese computer science research community. Seven researchers who play an active role in their respective communities (such as HCI, HRI, AI) participated in the investigation. All have served as international committee members more than ten times and have attended at least three audio conferences. The average length of their overseas experience was two years.

From the questionnaire results, we found that their perceived comprehension and production levels of speech during conference calls were much lower than in face-to-face meetings (Tables 1 and 2). Although the sample size is too small to derive any conclusions, astonishingly, 70% of the professionals who are considered to be Japanese representatives of various technical fields feel that they cannot contribute even half of what they want to say. The interesting point is that they seem to face particular

difficulty in audio conferences, while they feel comfortable using a second language in face-to-face meetings.

Table 1. NNS's perceived ability to follow conversation.

Comprehension level (%)	0-20	20-40	40-60	60-80	80-100
Face-to-face	0	0	1	2	4
Audio conf.	0	0	4	2	1

Table 2. NNS's perceived ability to make an utterance.

Utterance level (%)	0-20	20-40	40-60	60-80	80-100
Face-to-face	0	0	3	2	2
Audio conf.	3	2	0	1	1

Inspired by these questionnaire results, this study explores ways to support the verbal activities of such high-skilled, non-native speakers in audio conferences. Since most of these people could comprehend and contribute in face-to-face meetings, we believe their language skills are sufficiently high; if the same meeting were held face-to-face, their performance would have been better. Rather than a problem of their language skills, the problem is the lack of resources for concurrently processing multiple demanding tasks [1, 20]; when they are using their resources to compensate for the missed cues/information in an audio conference, their ability to think about the conversational content and to analyze the forthcoming input (such as phonetic analysis and parsing an ongoing conversation) is likely to decline [20]. Under such condition, non-native speakers will usually be left behind in a conversation.

Based on the previous works, we hypothesized that giving non-native speakers more time to understand the conversation would reduce their mental load. We developed the idea of inserting frequent silent periods for non-native speakers to compensate for the high demand on their mental resources. Ideally, those silent periods should allow non-native speakers to deal both with incomplete information and language processing in a timely manner (without being left behind in the conversation).

In this paper, we examine whether adding short silent gaps between native speakers' conversational turns improves non-native speaker's verbal activities. We explored this issue in two stages. First, we examined whether inserting small gaps in a pre-recorded meeting actually improved non-native listener's comprehension; second, we examined whether inserting small gaps in a real-time audio conference by adding artificial delays only among native speakers improved the non-native speakers' comprehension and production of speech. Note that the artificial delay allowed non-native speakers to listen to native speaker's speech earlier than other native speakers, which gave the non-native speakers a small break between utterances during which they could process the speech.

This study is important in two ways. First, we demonstrate that gaps as short as 0.2-0.4 seconds added between native speakers' conversational turns did indeed improve non-native listener's comprehension. Second, we report on how the small gaps were produced by controlling the amount of transmission delays between native and non-native speakers and how it affected native and non-native speakers' conversation. Although transmission delay is a common issue that has been studied for years [19], to our knowledge, no research has investigated its effects on conversation between native and non-native speakers.

In the remainder of this paper, we first draw on prior research that guided our work. Next, we describe our iterative studies and present the results for each study. Finally, we conclude with a discussion on the implications of our findings and some issues raised by them.

RELATED WORK

Audio conferencing is difficult not only for non-native speakers but also for native speakers. Indeed, previous research has pointed out various problems in audio conferencing. Many are related to the audio itself, such as poor audio quality, noise, and the inability to identify speakers. Others are related to the attendee's behavioral issues (e.g., speakers tend not to check others for their understanding) or technical issues (e.g., attendees are unable to check who is attending the meeting) [27].

Researchers have suggested many techniques for dealing with those problems. Many researchers have tried to improve the audio quality in audio conferences, starting with the reduction of reverberation. For example, previous works [9, 25] suggested that spatialized audio enhanced attendees' comprehension of speech by enabling them to distinguish between background noise and the focal speaker's voice. Researchers have also attempted to augment audio conferences with such visual cues as video images of attendees and/or expression buttons. Video images (spatialized video) have proven particularly useful in multiparty conversations: smoother turn-takings [24], and increased attendee involvement and ability to keep track of the conversation [7]. Besides adding a video channel, Yankelovich added a private chat channel so that attendees can consult with others without disturbing the main conversation [26]. In addition, for those who missed small parts of the conversation, Junuzovic et al. proposed an "accelerated instant replay" function [8], which allows the attendees to catch up on missed content.

Some of these techniques could be useful for non-native speakers as well. For example, spatialized audio also appears beneficial to non-native attendees by extracting speech information from background noise [5]. For augmenting audio conferencing with visual cues, Veinott studied non-native English speaking pairs and found that video images facilitated mutual understanding by helping them assess each other's state of understanding [23]. While

other techniques such as additional chat channels and accelerated instant replay might also be useful to non-native speakers (for asking others for clarification and catching up with the discussion), we must keep in mind that non-native speakers are already overwhelmed with multiple parallel processes such as linguistic processing (i.e. speech recognition, production of a foreign language, and recovering from the missed conversational context when necessary) and intensive thinking, which is typically accompanied by internal speech in their native language [20, 14], which could also be the target of linguistic processing. Thus, techniques that require further intensive work for non-native speakers are likely to be ineffective - it might even decrease performance. Instead, a technique that reduces the burden on non-native speakers would be desirable.

CURRENT STUDY

In this paper, we explore the ways for reducing the burden of non-native speaking attendees in an audio conference. Since the main problem of our target users (i.e., non-native attendees whose language skill is sufficiently high to participate in a face-to-face meeting in their second language) lies on the high demand on their resources to run multiple processing, we consider how to provide them with additional resources (e.g., processing power, time). As our first step, we attempt to supplement non-native attendees with additional processing time. We considered two ways for their support: either to slow down the conversation or to insert frequent silent gaps between speeches. Because the latter approach seemed more natural (i.e. people are usually exposed to silent gaps caused by transmission delay), we employed the latter approach for this study.

We carried out two studies to examine whether the short silent gaps improved the non-native speaker's verbal activities. The first study investigated the effects of short gaps on non-native speaker's perceived comprehension effort and actual comprehension level when they were manually inserted in a pre-recorded meeting. The second study explored their effects on non-native speakers' comprehension and production of speech when inserted in a real-time audio conference. The difference between the two studies is whether the short gaps were studied in an interactive situation.

STUDY 1: DO SHORT GAPS IMPROVE NON-NATIVE LISTENER'S COMPREHENSION?

The first experiment measured the effects of three different silent gaps (0, 0.2, or 0.4 seconds) on comprehension and perception of comprehension effort. To see whether the short gaps had similar effects on native speakers, the scores were compared between native and non-native speakers.

Method

Participants

Ten native English speakers and ten non-native English speakers were recruited for this study. The native speakers were born and raised in an English speaking country. The non-native speakers were all Japanese. Their English proficiency levels varied, but all had studied English for more than six years. Most were confident in their English skills; eight described themselves as proficient. Their average overseas experience in English speaking countries was 0.8 years.

Experimental Design

Each participant engaged in three listening tasks. The three tasks differed in their listening materials and in the length of silent gaps (0, 0.2, or 0.4 seconds). The order of the listening materials and the length of gaps were counterbalanced across participants.

Task Materials

As a pre-recorded meeting, we used the meeting corpus released by the Linguistic Data Consortium (LDC)[1]. The meeting consisted of four native English speakers discussing how "9-11" affected their lives and how innovative technologies might help prevent terrorism. The audio data was collected using two personal microphones attached to each meeting member (i.e. a close-talking noise-canceling boom microphone and an omni-directional lapel microphone) and several table microphones covering 360 degrees. Since each channel was recorded in a separate file (Figure 1, second and third top), we could independently handle each member's speech. This allowed us to change the timing of each member's speech by manually shifting each member's utterances and eventually mixing them all together (see bottom of Figure 1 and 2).

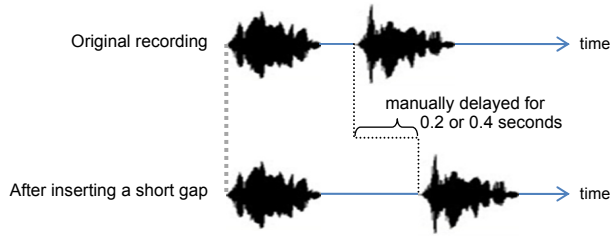

Figure 1. Insertion of short gaps

For the listening task material, we first randomly clipped three fragments from the meeting data. Each fragment lasted about 45 seconds with 10 to 11 speaker switches, 4 or 5 of them being overlapped with one another. We then prepared three versions for each fragment: no changes, a 0.2 second gap, or a 0.4 second gap inserted in between every speech. Note that the length of the latter two fragments

[1] http://www.ldc.upenn.edu/

became approximately 2 (or 4) seconds longer than the original version due to the insertion of gaps.

Note also that when there was an overlap between two speakers' utterances, the speech overlaps were resolved but the actual gap produced in the modified recording was less than 0.2 seconds (or 0.4 seconds) (Figure 2).

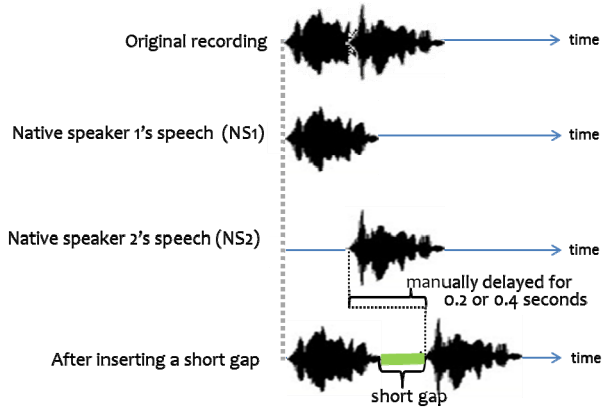

Figure 2. Resolving speech overlaps.

Tasks

The task was a simple listening task in which participants listened to each fragment (with different gaps) and rated their comprehension - after listening to each fragment, the participants were given a transcript, and were asked to mark all sentences or words they were able to follow. The native English speakers were provided with an English transcript while the Japanese participants were provided with transcripts translated into Japanese.

Procedure

Participants were first provided with a brief introduction to the study, including the topic of the recorded meeting and the experiment procedures. Note, however, that the differences between the three tasks were not explained to the participants to avoid any influence it may have on the outcomes.

After the introduction, the participants worked on a sample listening task to familiarize themselves with the actual experiment task. The fragment of the sample task was also extracted from the same meeting data. No changes were made (no silent gaps were added) to the sample task.

Next, the participants carried out a series of three trials. During the trials, they were allowed to take notes although they were only allowed to play each fragment once. Upon completion of each task, the participants filled out post-task questionnaires about their perception of their comprehension effort. After all three tasks, they were interviewed about the differences they have noticed among the trials.

Results

Below, we examine the effects of short gaps on non-native speaker's comprehension using two measures - in the first analysis, objective measures were used to compare actual comprehension level between the conditions; then, subjective measures were used to compare participant's comprehension effort. Throughout the paper, performance results (using objective measures) were analyzed in a repeated measures ANOVA, and survey results (using subjective measures) were analyzed in a non-parametric Friedman test, unless specified.

Each participant's comprehension level for each task was measured by calculating the rate of words he/she was able to follow by counting the number he/she marked in the trial and dividing it with the total number of words in the fragment.

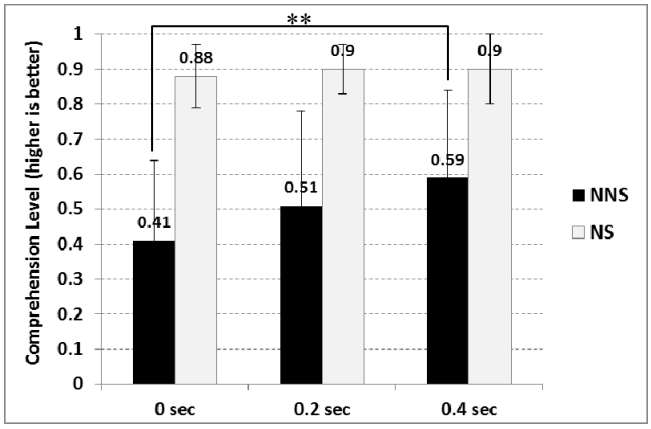

Figure 3. Mean comprehension scores by gap.

Overall, the comprehension level of the native English speakers ("NS") remained high throughout the conditions, while comprehension of the Japanese participants ("NNS") increased as silent gaps were added (Figure 3). A repeated measures ANOVA using gap lengths as repeated factors indicated a significant main effect of the gaps ($F[2, 18]= 5.91$, $p<.05$). A Bonferonni post hoc test indicated that the difference was significant between gaps of 0 and 0.4 seconds ($p<.01$). No significant difference was found for the native English speakers.

Next, we measured the participants' perception of comprehension effort using the questionnaire results on a five-point Likert scale. The results showed that the Japanese participants felt the tasks became easier as the length of the gap increased (Figure 4). A Friedman test indicated that this effect was significant ($\chi^2[2]=10.47$, $p<.01$). Post hoc tests (Wilcoxon Signed-Rank Test with Bonferroni correction) indicated that the difference was significant between gaps of 0 and 0.4 seconds ($p<.01$) and slightly different between gaps of 0.2 and 0.4 seconds ($p=.06$). Similar to the comprehension scores, no significant difference was found on comprehension effort for the native English speakers.

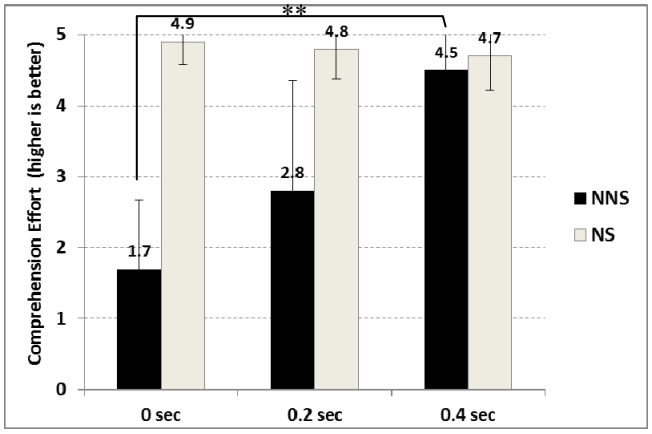

Figure 4. Mean ratings of perception for comprehension effort.

In the post-experimental interview, participants made some interesting comments about the differences between the trials. Many native English speakers commented that they felt the conversation (particularly with 0.4 gaps) somewhat awkward and unnatural, although it did not seem to affect their comprehension/effort. Meanwhile, many Japanese mentioned that conversation slowed down for some sessions and were easier to follow. They seemed surprised when notified that it was not the conversation speed but the silent periods inserted between conversational turns.

Additionally, some of the Japanese participants mentioned that they felt the conversation particularly fast and difficult to follow when there was a conversation overlap. Since the manually inserted gaps often eliminated the overlaps as in Figure 2, the elimination of overlaps might also have helped the Japanese participants follow the conversation with less effort.

In summary, inserting short gaps and resolving the overlaps between speeches helped the non-native speakers follow the conversation with less effort.

STUDY 2: HOW DO SHORT GAPS AFFECT REALTIME CONVERSATION?

Although silent gaps during conversation appear effective for non-native speakers, inserting frequent gaps is usually difficult during ordinary verbal activities - asking native speakers to temporarily halt before making an utterance is unrealistic. In our second study, we inserted the gaps by controlling the transmission delays between native and non-native speakers.

Figures 5 and 6 illustrate how the gaps were produced by adding artificial delays among native speakers. The direct sound waves of a speaker's voice are shown in black, and the sound waves that have passed through the network (i.e., sound waves heard at the listener's site) are shown in gray. Note that these examples show the case of zero network delay. Figure 5 shows that non-native speakers listen to native speaker NS1's voice ahead of other native speakers because the artificial delay is added among native speakers. This allows the non-native speaker to process NS1's speech ahead of the other native speakers, which eventually provides them with additional processing time. The gap is not only expected to increase non-native speakers' comprehension but also to allow them to prepare their own utterance.

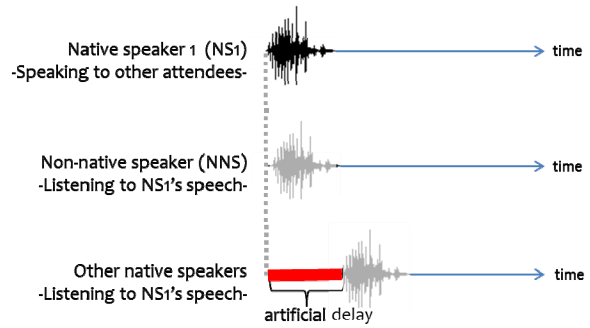

Figure 5. Providing NNS with additional processing time by inserting artificial delay among NSs.

The artificial delay inserted among native speakers is not only expected to supplement non-native speakers with additional processing time, but also to allow them to hear each utterance more thoroughly by reducing/resolving conversational overlaps (Figure 6).

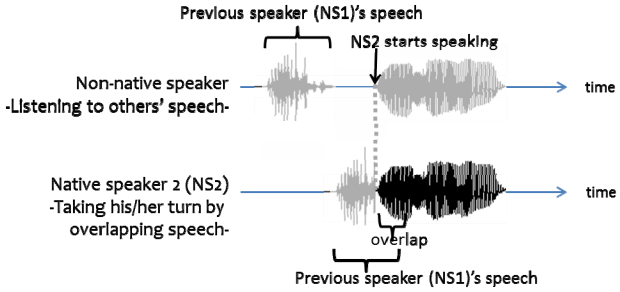

Figure 6. Resolving conversation overlaps between NSs.

While the short gaps are expected to be useful for non-native speakers, we must also pay attention to their effects on native speakers. In fact, previous research has significantly demonstrated the negative impacts of network delay on communication [6, 10, 22]. For example, Krauss et al. demonstrated that an audio delay of 0.3 seconds can have a detrimental effect on the communication process, and delays as large as 0.9 seconds can drastically impact a pair's ability to communicate [10]. Similarly, Tang et al. found that a delay of 0.57 seconds make turn-taking difficult to negotiate [22]. In summary, previous work on audio delay indicates that people hardly notice the delay if it is shorter than 0.2 seconds; delays between 0.2 and 0.4 seconds pose little if any problems; but delays longer than 0.45 seconds can severely impact communication and coordination processes.

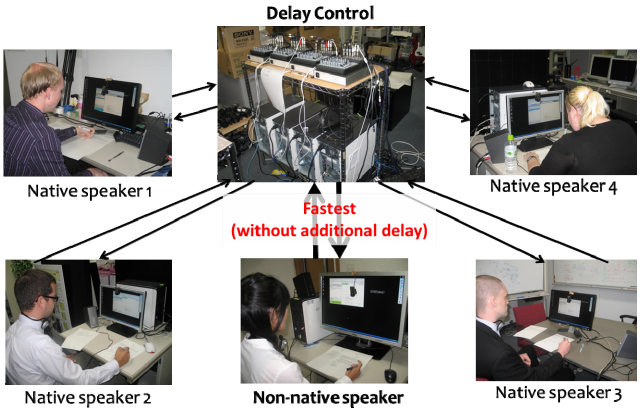

Figure 7. Five participants (4 NSs and 1 NNS) working on a decision making task over Skype with delay control.

In our second study, we examine the effects of adding artificial delays of 0.2 or 0.4 seconds among native speakers voice relay (Figure 7 and 8), which presumably have a small impact on native speakers' conversations. We were interested to know whether the small gaps would actually improve the non-native speakers' verbal activities without hindering the native speakers. Specifically, we sought to answer the following questions:

- Does our delay mechanism work as expected? – Does it actually produce small gaps and reduce conversation overlaps in a real-time audio conference.
- How does the mechanism affect communication among native speakers? – Does turn-taking get difficult, and do the native speakers feel more frustrated as the delay increase?
- Does the scheme improve non-native speakers' verbal activities? – Can we see an increase in their perceived comprehension and in their production of speech?
- How does our delay mechanism affect consensus building among native and non-native speakers?

Method

Participants

Fourteen groups of five adults (70 participants) were newly recruited for this study. Fifty-six were native English speakers, and fourteen were non-native speakers. The non-native speakers were Japanese participants whose English skills were sufficient to have daily conversations in English. Their TOEIC scores exceeded 860, and the average length of their overseas experience was 1.6 years.

Experimental Design

Five-person groups (four native speakers and one non-native speaker) participated in three decision making tasks with different silent periods: 0, 0.2, and 0.4 seconds. Groups of five were selected due to their ability to be small enough for participants to collaborate while still not being so small that participants will be chosen to take the floor by default. The order of the discussion topics and gap lengths were counterbalanced across participants.

Tasks

As a discussion topic, we chose a series of survival tasks (on a desert, at the arctic, and on the moon) [11] that are widely used for training group development. In these tasks, participants imagine that they have been stranded at one of these locations. Several items are presented, and each participant ranks the items for importance to survival. After ranking them individually, their discussion generates a group solution through an audio conference. Each task contained fifteen items, but for simplification, we excluded the items that are not used in daily conversations (such as a "large piece of insulating fabric") and randomly chose six items from the remaining pool.

Apparatus

Skype was used as the audio conferencing software. The Skype interface was displayed on each participant's screen. There was approximately 0.2 seconds of network plus Skype latency between the clients. Artificial delays were added using an audio delay device (Figure 7 and 8). In other words, a native speaker's speech arrived at other native speakers' site with 0.2, 0.4, or 0.6 seconds delay while it arrived at the non-native speakers' site with 0.2 seconds delay regardless of the conditions.

Figure 8 shows the wire diagram of the system used in the study. It illustrated how delays were inserted only among native speakers.

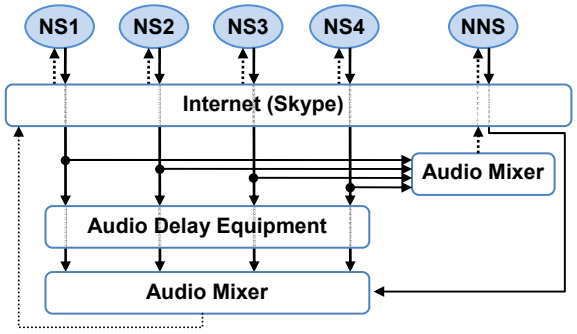

Figure 8. Wire diagram of the system.

During the experiment, conversations were recorded using Tapur, which is an exclusive Skype recorder. Tapur was installed on each client because it allowed us to store the exact voice data exchanged at each site.

Procedure

Procedure (1): On arrival, participants completed experimental consent forms and then moved to five separate rooms. After a short phonetic test, they were introduced to each other over Skype.

Procedure (2): The following procedure was repeated three times with different amounts of delay introduced among

native speakers: Participants were given five minutes to rank the six task items by themselves and to write down their solutions. Next, participants were given 15 minutes to generate a group solution. In the discussion, they were told to have a free-style conversation (i.e. with no chairman) because we were interested in investigating the effects of our delay mechanism on non-native speakers' unprompted turn-taking behaviors (e.g., whether it would allow them to produce more utterances without being prompted). After a group solution was determined, participants were separated and asked to rank their second individual rankings to determine the possible influence of the group discussion. Participants also completed post-task questionnaires about the conversations they had just experienced. Similarly to the first study, the participants were not notified about the differences between the trials.

Procedure (3): Following the completion of the three tasks, participants completed a final questionnaire and were interviewed about the differences between the three trials.

Results

Production of Small Gaps

First, we examined if our delay mechanism actually produced small silent gaps and reduced conversation overlaps at the non-native speaker's site[2]. To this end, speaker switches between native speakers were classified into two groups: the ones with no overlaps and the ones with overlaps. Speaker switches in the former group were used to examine if the mean length of gaps increased, and those in the latter group were used to see if the mean length of overlaps reduced.

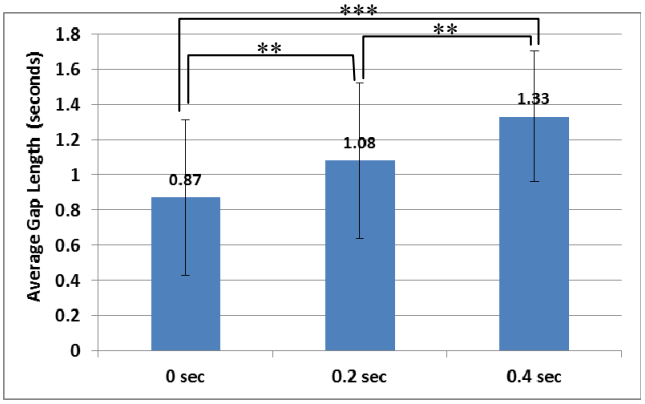

Figure 9. Mean length of gaps between native speakers' speech by delay scheme.

Figure 9 shows the average gap lengths of speaker transitions between native speakers. As expected, the results showed a steady increase in the average length of gaps between native speakers' speech as the delay increased. A repeated measures ANOVA indicated that the differences between the trials were significant ($F[2, 26]=24.09$, $p<.001$). A Bonferonni post hoc test indicated that the difference was significant between gaps of 0-0.2 seconds ($p<.01$), 0.2-0.4 seconds ($p<.01$), and 0-0.4 seconds ($p<.001$).

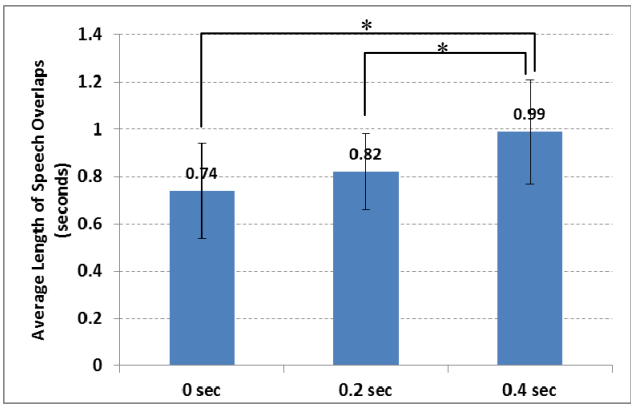

Figure 10. Mean length of overlaps between native speakers' speech by delay.

Similarly, we also measured the overlap lengths among native speakers' speech. Although we expected a constant decrease, the average length of speech overlaps increased as the delays increased (Figure 10). A repeated measures ANOVA indicated that the differences between the trials were significant ($F[2, 26]=9.70$, $p=.001$). A Bonferonni post hoc test indicated that the difference was significant between gaps of 0-0.4 seconds ($p<.05$) and 0.2-0.4 seconds ($p<.05$).

This was mainly caused by the increase in speech overlaps where multiple native speakers started talking simultaneously to take their next turn without noticing that another participant was taking their turn due to the longer delay (see Table 3 top row; further explanation provided below).

Effects on Native Speakers

Overall, we found a trend showing that native speakers had trouble taking their turns as delays increased. Table 3 shows evidences that support this tendency.

Table 3. Native speakers having difficulties in taking their turns as delays increase.

	0 sec	0.2 sec	0.4 sec
Percentage of conversations initiated simultaneously	22%	33%	41%
Frequency of speaker switches (# of speaker switches per min.)	7.8	7.3	6.6
Speaking effort (5pt Likert scale: higher is better)	4.2	4.2	4.0
Comprehension effort (5pt Likert scale: higher is better)	4.4	4.5	4.3
Frustration (5pt Likert scale: lower is better)	1.7	1.8	2.1

[2] All the analysis presented in this paper was conducted using the data stored at the NNS's client, unless specified.

(1) *Increase in simultaneous speech when multiple NSs start talking*: Native speakers in the delayed trials sometimes started talking without noticing that others had already started their speech because of the delayed arrival of other native speakers' speech. This caused an overlap in the beginning of their speech. The frequency of such overlapped speech was measured by first counting the number of incidences where multiple native speakers started talking to take their next turn in each trial, and then dividing it with the total number of speaker switches during each trial. Table 3 shows a steady increase in the average frequency of such incidences as the delays increase. A repeated measures ANOVA indicated a significant main effect for delay ($F[2,26]=11.58$, $p<.001$). A Bonferonni post hoc test indicated that the difference was significant between gaps of 0-0.2 seconds ($p<.05$) and 0-0.4 seconds ($p<.01$).

(2) *Decrease in the frequency of speaker switches among NSs*: Speaker switches were measured in order to identify the difficulty of turn taking. If speaker switching is low, it indicates a lower level of interactivity between speakers. Coding for this measure only counted instances when new speakers took the floor, excluding utterances that were not completed. The frequency of speaker switches between native speakers decreased as delay increased (Table 3). A repeated measures ANOVA indicated a significant main effect of delay ($F[2,26]=3.79$, $p<.05$). The difficulties in their turn-taking may be due to the increase in overlap and long pauses caused by delay. We found many instances in which native speakers continued talking because others could not take the turn at the right timing. The following excerpt[3] captures this tendency:

Excerpt 1 (0.4 second trial):

NS1: When you say fishing kit what do you mean↑ exactly↓ (0.8) uh:: well what do you have in mind when you say FISHING KIT↑(1.4)

NS2: Pro [bably ro::d and]

NS1: [Mosquito netting is] pretty thin it could rip (.) I mean you know (0.9) well it it it's pretty (0.2) I don't think it's uh good for uh::: for for catching fish.

(3) *Increase in NS's frustration*: From the post-experimental interview, it revealed that overlaps occurring in the early stage of speech occasionally sounded rude because they felt like someone broke into their speech without listening to the entire argument. This seemed to increase their frustration during the discussion. Table 3 shows the average scores of native speaker's perceived comprehension effort, speaking effort, and frustration. Results show that the scores of the 0.4 sec delay is consistent in being the worst of all trials. A Friedman test indicated a significant main effect of delay for frustration ($\chi2[2]=7.23$, $p<.05$). Post hoc tests (Wilcoxon Signed-Rank Test with Bonferroni correction) indicated that the difference was significant between gaps of 0-0.4 seconds ($p<.05$).

[3] [ ] indicates where overlaps take place; ↑ and ↓ indicates rising and falling intonation; : indicates an extension of sound; ( ) indicates a silent gap timed in tenths of a second.

Effects on Non-native Speakers

Overall, we found a trend that non-native speakers performed the best and preferred the 0.2 second trial. The 0.4 second trial was consistently the worst of all trials in terms of non-native speaker's performance and preference. Similarly to the previous section, we will show evidence that supports this tendency (Table 4).

Table 4. Non-native speakers performing best with 0.2 second delay trial.

	0 sec	0.2 sec	0.4 sec
Average rate of spontaneous speech	0.76	0.81	0.69
Speaking effort (5pt Likert scale: higher is better)	3.1	3.4	2.3
Comprehension effort (5pt Likert scale: higher is better)	4.1	4.1	3.5
Frustration (5pt Likert scale: lower is better)	2.9	2.7	3.1

(1) *Increase in NNS comprehension effort*: Based on our first study, we had initially expected that non-native speakers' comprehension effort would decrease as delay increase, because the artificial delays create gaps. However, it turned out that non-native speakers made slightly more effort to comprehend the discussion as delay increased (Table 4). A Friedman test indicated that this effect was borderline significant ($\chi2[2]=5.52$, $p<.06$).

(2) *Rise and drop in the rate of NNS's spontaneous speech*: The degree to which non-native speakers were eager to speak during the trials was measured by the rate of their spontaneous speech, which was calculated by first counting the number of non-native speakers' spontaneous speech (opposed to speech prompted by native speakers) and then dividing it with their total number of utterances. Table 4 shows that non-native speakers spontaneously spoke the most in the 0.2 sec trial and the least in the 0.4 sec trial. A repeated measures ANOVA indicated a borderline significant effect of delay ($F[2, 26]=3.23$, $p=.056$).

(3) *Drop and rise in perceived effort of producing an utterance*: Although the actual difference in the proportion of non-native speaker's spontaneous speech was subtle, it appeared that non-native speakers felt strongly that it was easier to produce an utterance in the 0.2 second trial and difficult in the 0.4 second trial (Table 4). A Friedman test indicated that this effect was

significant ($\chi^2[2]=6.59$, $p<.05$). Post hoc tests (Wilcoxon Signed-Rank Test with Bonferroni correction) indicated that the difference was significant between gaps of 0.2 and 0.4 seconds ($p<.05$). In the post-experimental interview, one non-native speaker explained her reluctance to speak "*(In the 0.4 second trial,) they started to jump onto others' conversation. At times, it sounded like they didn't want to listen to others' opinions. […] So, I felt reluctant to cut into the conversation. I just tried to concentrate on following their conversations.*"

In sum, even though our delay mechanism successfully created silent gaps between native speakers' speech, increased speech overlaps (caused by multiple native speakers trying to take their next turns) appeared to have a negative effect on non-native speakers. The negative effects grew apparent in the 0.4 second trials.

Effects on Consensus Building

Finally, we investigated whether our delay mechanism affected consensus building - how much the participants actually agreed with the group ranking, whether the agreement levels differed between native and non-native speakers, and whether the agreement levels varied across different delay conditions.

Table 5. NS and NNS's agreement levels of group ranking.

	0 sec	0.2 sec	0.4 sec
Average correlation between NS and Group rankings	0.90	0.89	0.91
Average correlation between NNS and Group rankings	0.68	0.88	0.77

We first calculated the correlation (Spearman's coefficient) between each participant's second individual ranking and the group ranking. Each correlation score indicates the participant's agreement level (i.e. degree of ranking similarity), ranging from -1 (complete opposite) to 1 (identical). We then compared the average correlation scores between native and non-native speakers using the Wilcoxon test (Table 5). Results indicated that non-native speakers were significantly less convinced by the group ranking than the native speakers when delay was not added ($p<.05$). The gaps between native and non-native speakers became less prominent in conditions with 0.2 or 0.4 sec delay (both n.s.). The gap found in 0 sec condition may becaused by the difference in their comprehension and speaking skills where non-native speakers have difficulties arguing back and defending their opinions at the right timing.

We further ran a Friedman test to see if the correlation scores varied across delay conditions. Results indicated that the effect of delay was not significant. Although not significant, non-native speakers seemed to be most convinced by the group ranking in 0.2 sec delay (Table 5). The result is consistent with the pattern found in the previous section where 0.2 sec delay was preferred and performed the best by the non-native speakers.

DISCUSSION

There are five main findings from our two studies:

- Reducing overlaps and inserting small gaps in a previously recorded meeting significantly improved the comprehension (Figure 3) and perceived effort (Figure 4) for non-native speakers when following the conversation. Similar effects were not found for native speakers.
- Our delay mechanism (i.e. adding artificial delays among native speakers) successfully inserted small gaps among native speakers' speech in a real-time audio conference (Figure 9). However, the speech overlaps did not decrease as expected, but increased along with delay (Figure 10).
- Native speakers had trouble taking their turns as delays increased (Table 3): Due to the delay, speakers were unaware when they would talk simultaneously or over the speech of others. This created higher instances of simultaneous speech as delay increased which usually ended with speakers ending their utterances abruptly. This resulted in lower frequency of turn-taking between native speakers. Native speakers' perceived efforts (both for comprehension and speaking) also grew substantially with 0.4 second delay.
- Non-native speakers spontaneously spoke the most in the 0.2 second delay condition (Table 4). Their perceived speaking effort was also the lowest with 0.2 second delay. Their perceived frustration also followed a similar pattern – 0.2 sec delay being the best and 0.4 sec delay being the worst. Regarding the non-native speakers' comprehension effort, although we expected it to decrease, it remained similar between the 0 and 0.2 second delays, and increased in the 0.4 second delay.
- Non-native speakers were significantly less convinced by the group solution than the native speakers when delay was not added. The agreement level of non-native speakers improved when delays were added - gaps between native and non-native speakers became less prominent in conditions with 0.2 or 0.4 sec delay.

Below, we explain our findings and discuss some design opportunities and directions for future work.

Explanations of the Findings

How did the gaps in a pre-recorded meeting affect NS and NNS's comprehension?

There are two possible reasons why the gaps worked well on non-native speakers when they were inserted in a pre-recorded meeting. First, the gaps allowed for more time to

process incoming information, placing less demands on the listener's mental resources. Second, inserting gaps in between utterances resulted in less overlaps in conversation. This improved the overall conversation clarity and reduced the effort needed to compensate for the missed parts of the conversation caused by speech overlaps. In short, the gaps between utterances allowed the non-native speakers to have a longer window to make out the meaning of each utterance before focusing on the following content.

Though the insertion of gaps improved comprehension for non-native speakers, the same benefits were not seen for native speakers. For native speakers, their comprehension scores were always high, perhaps indicating that they did not require much mental resource to understand the conversation. Thus, an improvement in comprehension (both in terms of score and effort) would not have been seen even when their mental resources were conserved.

How did the added delays affect NS and NNS's communication in a real-time meeting?

As previous literature indicated, our second study demonstrated that the added delays negatively affected the native speakers. The number of overlaps in the beginning of their speeches increased and there were also awkwardly long pauses as illustrated in Excerpt 1. We suspect that such difficulties in turn-taking led them to fewer speaker switches and increased frustration.

For the non-native speakers, it appeared that there were both beneficial and detrimental effects because of the added delays on non-native speakers. The beneficial aspects were that the delays were successful in creating gaps (Figure 9) where non-native speakers could utilize to process the language and think of what to speak. It also gave them an opportunity to participate in the conversation before other native speakers, resulting in a higher rate of spontaneous speech (Table 4). The detrimental effects were found for longer delays. This includes longer, more disruptive speech overlaps that occurred more frequently.

Why didn't the gaps improve NNS comprehension in a real-time meeting?

Non-native speakers were able to follow the conversations with less effort when the gaps were inserted in a pre-recorded meeting. However, those gaps appeared insufficient for improving their comprehension effort in a real-time audio meeting. We infer that the changes caused by the delay in the native speakers' conversation impacted the non-native speakers, requiring more attention and effort to follow the conversation.

Another consideration is the difference in the characteristics of overlapping speech, which might have affected the non-native speakers' comprehension effort - in the 0.2 second trial, the average number of words in an overlapping speech was approximately 1 or 2 words while that of the 0.4 second trial was around 2 to 3 words. Although this difference might look subtle, the participants' reaction to the overlaps was quite different. In the 0.2 second trial, when participants started talking together, they soon noticed the overlaps, halted, and then often repeated their utterance. In contrast, participants in the 0.4 second trial were more likely to be interrupted in the middle of their sentences. In such cases, they typically continued their speech or just halted without repeating their utterance. The excerpts below capture the tendency:

Excerpt 2 (0.2 second trial):

NS1: Ok=

NS2: =Yeah (.)

NS1: So [anyway]

NS3: [I:::] I need I need the brandy

Excerpt 3 (0.4 second trial):

NS1: my argument towards [the knife is tha:::t]

NS2: [the thing that kill you] the fastest is dehydration (0.8) so:::

NS1: yeah (.)

NS3: water or dead

The missed part (i.e. overlapped speech) in 0.4 second trial looks much more difficult to compensate than the 0.2 second trial, which had possibly led them to much higher comprehension effort and frustration.

Design Implications

Our findings and the above discussions suggest recommendations for the design of future audio conferencing systems to support non-native speakers. Since gaps were beneficial but overlaps overshadowed those effects for non-native speakers, a function that inserts gaps without increasing overlaps might better support their verbal communication in an audio conference.

For example, one solution for inserting gaps without increasing overlaps during audio conferencing would be to combine push-to-talk (PTT) functionality with our delay mechanism. The PTT function is commonly used to block out any possible background noise when there are a large number of participants, but moreover, works as a behavioral trigger to avoid overlaps in discussions [18]. An artificial delay interval combined with PTT functionality might reduce the negative effects of native speaker overlaps in longer delay settings. This would thus increase the positive effects of longer delays for non-native speakers in NS-NNS audio conferencing while still keeping overlaps to a minimum.

CONCLUSIONS AND FUTURE DIRECTIONS

Previous literature has shown the negative impacts of transmission delays on communication between collaborators with the same native language. Yet little is known about how these delays would affect multilingual collaborations. Our studies provide insight into the effects of delays on communication among native and non-native speakers. By

looking at how the delays have affected non-native speakers in isolation as well as native and non-native speakers in a group, we have been able to identify how comprehension and participation are affected. When listening to discussions with embedded delays (i.e. gaps), non-native speakers were capable of understanding the discussion content with greater accuracy. In a group, a short delay (0.2 second) also assisted non-native speakers by creating more opportunities to participate in the conversations. However, as the delay increased to 0.4 second the positive effects of the speech delay that assisted non-native speakers were overshadowed by the negative effects experienced by the native speakers, such as increased speech overlaps.

For future studies, we are interested in expanding our study to other language speaking participants. For example, the Japanese participants may be having difficulties collaborating in English due to the differences in grammatical structures [21]. Adjusting both the native language background of the participants and/or the language used to communicate may yield different results.

These studies illustrate how in a group interaction, higher delays deteriorate communication between native and non-native speakers. Especially when connections span oceans and continents, the combination of even seemingly minute delays along with imperfect language skills create a very taxing situation for non-native speakers. Yet even these imperfect short delay conditions, which have always been considered entirely detrimental, have given some insight into ways to enhance communication between native speakers and non-native speakers. We do not support extending delays, but neither can their benefits be overlooked. We hope that this work illustrates the impact these delays/gaps have on collaboration in the hopes that future work in this field will consider its effects.

ACKNOWLEDGMENTS

We are grateful to Hideaki Kuzuoka for his comments on an earlier draft of the paper. We would also like to thank Linsi Xia and Masanobu Ishimatsu for their assistance in running the experiment. Finally, we thank the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

1. Broadbent, D. *Perception and communication*. Oxford: Pergamon, 1958.
2. Clark, H. H. & Schaefer, E. F. Collaborating on contributions to conversations. *Language and Cognitive Processes, 2*, 1987, 19-41.
3. Ding, X., Erickson, T., Kellogg, W. A., Levy, S., Christensen J., Sussman, J., Wolf, T. V., Bennett, W. E. An Empirical Study of the Use of Visually Enhanced VoIP Audio Conferencing: The Case of IEAC. *Proceedings of CHI*, 2007, 1019-1028.
4. DiMicco, J. M, Hollenbach, K., and Bender, W. Using visualizations to review a group's interaction dynamics, *Proceedings of CHI*, 2006, 706-711.
5. Ezzatian, P., Avivi, M., Schneider, B. Do non-native listeners benefit as much as native listeners from spatial cues that release speech from masking? *Speech Communication*, *52*, 2010, 919-929.
6. Gutwin, C. The effects of network delay on group work in shared workspaces, *Proceedings of ECSCW*, 2001, 299-318.
7. Inkepen, K., Hegde, R., Czerwinski, M. and Zhang, Z. Exploring spatialized audio & video for distributed conversations, *Proceedings of CSCW*, 2010, 95-98.
8. Junuzovic, S., Inkepen, K., Hegde, R., Zhang, Z., Tang, J. and Brooks, C. What did I miss? In-Meeting Review using Multimodal Accelerated Instant Replay (AIR) Conferencing. *Proceedings of CHI*, 2011, 513-522.
9. Kilgore, R., Chignell, M. and Smith, P. Spatialized Audioconferencing: What are the Benefits? *Proceedings of CASCON*, 2003, 135-144.
10. Krauss, R. M. and P. Bricker. Effects of transmission delay and access delay on the efficiency of verbal communication. Journal of the Acoustical Society of America, 41 (2), 1967, 286-292.
11. Lafferty, J. C., Eady, P. M., & Elmers, J. The Desert Survival Problem. Plymouth, Michigan. Experimental Learning Methods, 1974.
12. Li, H. Z., Yum, Y., Yates, R., Aguilera, L., Mao, Y. and Zheng, Y. Interruption and Involvement in Discourse: Can Intercultural Interlocutors be Trained? *Journal of Intercultural Communication Research*, Vol. 34, No. 4, 233-254.
13. Luisa, M., Lecumberri, G., Cooke, M., Culter, A. Non-native speech perception in adverse conditions: A Review. *Speech Communication*, 52, 2010, 864-886.
14. Macnamara, J. and Kushnir, S. L. Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior, 10*, 1971, 480-487.
15. Mayo, L., Florentine, M., Buus, S. Age of Second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, Vol. 40, No. 3, 1997, 686-693.
16. Nabelek, A.K., Donahue, A.M., Perception of consonants in reverberation by native and non-native listeners. *Journal of Acoustical Society of America, 75*, 1984, 632-634.
17. Rogers, C., Lister, J., Febo, D., Besing, J., Abrams, H. Effects of bilingualism, noise and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, 27, 2006, 465-485.

18. Sawhney, N. and Schmandt, C. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *TOCHI*, 2000, 353-383.
19. Short, J., Williams, E. & Christie, B. The Social Psychology of Telecommunications. London: John Wiley, 1976.
20. Takano, Y. and Noda, A. A temporary decline of thinking ability during foreign language processing. *Journal of Cross-Cultural Psychology*, 24, 1993, 445-462.
21. Tanaka, H. Turn Projection in Japanese Talk-in-Interaction. *Research on Language and Social Interaction, 33,* (1), 2009, 1-38.
22. Tang, J.C., and Isaacs, E.A. Why do users like video? Studies on multimedia-supported collaboration. Sun Microsystems Laboratories, Inc., Mountain View, CA, 1992.
23. Veinott, E., Olson, J., Olson, G., and Fu, X. Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other. *Proceedings of CHI*, 1999, 302-309.
24. Vertgaal, R., Van der Veer, G. C. and Vons, H. Effects of Gaze on Multiparty Mediated Communication. *Proceedings of Graphic Interface 2000*, Morgan Kaufmann Publishers, (2000), 95-102.
25. Yankelovich, N., Kaplan, J., Provino, J., Wessler, M., DiMicco, J. M. Improving Audio Conferencing: Are Two Ears Better than One? *Proceedings of CSCW*, 2006, 333-342.
26. Yankelovich, N., McGinn J., Wessler, M., Kaplan, J. and Provino, J. and Fox, H. Private Communications in Public Meetings. *Proceedings of CHI*, 2005, 1873-1876.
27. Yankelovich, N., Walker, W. Roberts, P., Wessler, M., Kaplan, J. and Provino, J. Meeting Central: Making Distributed Meetings More Effective. *Proceedings of CSCW*, 2004.