Mitigating problems in video-mediated group discussions: Towards conversation aware video-conferencing systems

Marwin Schmitt
CWI: Centrum Wiskunde & Informatika
The Netherlands
schmitt@cwi.nl

Simon Gunkel
CWI: Centrum Wiskunde & Informatika
The Netherlands
gunkel@cwi.nl

Pablo Cesar
CWI: Centrum Wiskunde & Informatika
The Netherlands
p.s.cesar@cwi.nl

Dick Bulterman
CWI: Centrum Wiskunde & Informatika
The Netherlands
dcab@cwi.nl

ABSTRACT

Video has improved the experience of multi-party real-time communication, but users are still facing difficulties in the conversation. The varying delay introduced by the structure of the internet disrupts our natural timing of conversations. Users further face a lack or distortion of non-verbal cues like gaze awareness. Alterations in the temporal and spatial dimension of our media-streams are an inherent characteristic that can be reduced, but never completely solved by hardware advancements. Current video-communication systems only react to a minimal degree to the communication (e.g. by changing the focus to the participant with the loudest audio). In order to adequately support users in these difficult situations, we need systems that react conversationally aware and give the ability to alleviate communication problems.

In an experimental setup, we investigated how delay impacts the objective and subjective performance of a small group discussion. Ad-hoc groups of five people, with a moderator, solved a quiz question-select answer style task. This was done under different delay conditions, of up to 2000ms additional one-way delay. Even with a delay up to 2000ms, we could not observe any effect on the achieved quiz scores. In contrast, the subjective satisfaction was severely negatively affected. While we would have suspected a clear conversational breakdown with such a high delay, groups adapted their communication style and thus still managed to solve the task. This is, most groups decided to switch to a more explicit turn-taking scheme.

In this paper, we discuss what cues and problems a video-conferencing system needs to be able to detect and how recent advancements in computational social science can be leveraged. Further, we provide an analysis of the suitability of normal webcam data for such cue recognition. Based on our observations, we suggest strategies that can be implemented to alleviate the problems.

Categories and Subject Descriptors

H.4.3 [Communications Applications]: Computer conferencing, teleconferencing, and videoconferencing

General Terms

Human Factors; Design; Measurement.

Keywords

Multi-party; video-conferencing; performance; delay; conversation aware; cues

1. INTRODUCTION

Multiparty video-conferencing has become a reality enabling us to quickly have real-time conversations with a small group of people. The structure of the internet, coupled networks of different types and packet-based routing, introduces varying delays between participants. Delay has been shown to introduce conversational difficulties in video-conferencing [20][18][2]. Together with lack of social cues like gaze awareness the conversation is severely disturbed. But these conversational problems are not unique to computer-mediated communication, like disturbances in video and audio streams. Mishaps like double talk happens in everyday life collocated conversations as well and "conversational repair" actions often follow [15]. Thus it is hard for humans to identify whether a technical or interpersonal problem is causing the complication in the conversation [21]. The challenge that we are currently facing is not only to tune hardware and software to perfect the capturing, transmission and rendering, but to really provide support for the alterations in the temporal and spatial conversational realities. While turn-taking in telecommunication has been extensively studied, it is mainly used for offline processing of experimental data [19, 20, 22]. Current video-communication systems only provide a minimal adaptation to the conversation, e.g. Google Hangout highlights the participant with the loudest audio source. To build systems that provide an optimal experience also in difficult situations we need to understand the ongoing conversations and have mechanisms to ease the problems.

In this paper, we report about the objective and subjective performance and in video-mediated consensus small group discussion. We conducted a study in which groups of five people, with one randomly assigned moderator. Together they solved a quiz style question-select answer scenario. We introduced up 2000ms one-way delay and assessed the achieved quiz-score and various subjective ratings of the experience. Our analysis showed that the scores of our participants did not decrease with a higher delay. On the other hand participants were consistently less satisfied with the discussion, the result and their contribution. We observed that depending on the moderator groups employed an explicit turn-taking scheme in reaction to high delays. In the debriefing participants discussed various communication problems arising in high delay situations.

We started an investigation how we can make use of the available data to mitigate these effects. Our first investigation gives indication that the combination of audio- and video-streams from standard video-conferencing can be sufficient for advanced analysis. We discuss which cues can be detected with current state-of-the-art social signal processing. We further elaborate how

Table 1 System Configuration

System Setup	Desktop PCs (Core i7, 16GB Ram, SSD) Webcam (Logitech HD C920) Headset (Creative Soundblaster Xtreme) Video: 640x480px, 30fps, H264 Audio: Speex Network: Local Gigabit LAN, UDP, RTP
Conditions	•0ms-delay (avg = 75ms, sd = 31ms) •500ms-delay (avg = 564ms, sd = 34ms) •1000ms-delay (avg=1065ms, sd = 39ms) •2000ms-delay (avg = 2058ms, sd= 57ms).

we can raise awareness of social cues and provide support for explicit turn-taking schemes.

2. CONTROLLED EXPERIMENT

This section reports about the experimental investigation of the subjective and objective performance of small groups in video-mediated decision-making discussions under different delay conditions. We already reported about the developed testbed [17], the degradation of experience and comparison of symmetric, asymmetric and dyadic conditions [18] and an approach to use speech pattern for role distinction [19] to further qualify the experience.

2.1 Related Studies

When studying the user experience of remote real-time communication the effects of delay have been a long term field of interest [20, 22]. The effects of delay are hard to capture, as it is not a directly perceivable media degradation. Traditional subjective quality metrics fail to capture the important aspects in which delay influences us [21]. Known effects are in a slowdown of the conversation and higher perceived conversational difficulty. But these effects are not necessarily attributed to technical difficulties, but can easily be interpreted as characteristics of the communication partners [20]. In multi-party video-conferencing, a study investigating delay in high-end tele-immersive system showed that participants were still able to hold a conversation with 1000ms one-way-delay [5]. In a study with different devices (Desktop PCs, TVs) Berndtsson et al. [2] found that people would likely not continue a call with 800ms delay.

Our study focuses on standard desktop hardware and we selected up to 2000ms delay as we suspected a breakdown of the conversation with high delay.

2.2 Methodology

The study followed an experimental design with randomized conditions. In our study participated 39 (20 female, 19 males, Average age 36 years (min 20 years, max 60 years)). The experiment was conducted in English, in which all participants were fluent. Participants were in groups of 5 except one, with 4 participants, as one participant did not show up and it was not replaceable in time. All participants were seated in separate rooms, after an introduction round in which we explained our research and the experiment. The task of our participants was a quiz style question-select answer scenario. The participants had to discuss together the best answer to questions about surviving in the wilderness. Before the group discussion, each participant answered the question individually. The task is based on the team building exercise from [3]. One participant was asked to be the moderator, to submit the final group answers and move the discussion along to keep the 10 minutes time constraint per round. The order of the questions did not change in the experiment but the order of the delay. Each round of questions was in total 8 times discussed, twice under each condition.

Table 2 Assessed questions

Label	Question
satisfaction _discussion	I am satisfied about the course of discussions in our team.
satisfaction _outcome	I am satisfied with the quality of the outcome of our team.
contribution	To what extent do you feel that you have contributed to the team's final out-come?

We used the VMC-TB [17], the exact technical configuration and test conditions can be found in Table 1. For the objective data, we assessed the quiz group scores and individual scores. The score is the number of correct answers, thus in each round between 0 and 3 points can be achieved. We asked people about their satisfaction with the discussion, the satisfaction with the outcome of the discussion and their contribution to the discussion. The exact questions and used labels can found in Table 2.

2.3 Results

2.3.1 Subjective Performance

The average responses to the asked questions, see Table 2 , went down with more delay, as can be seen in the plot with 95% confidence intervals in .

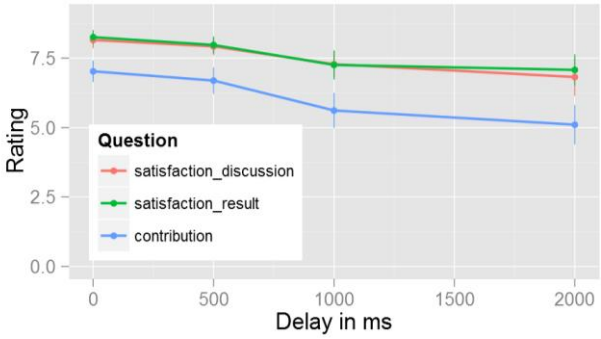

Figure 1 Responses to Questions

As can be seen the responses of satisfaction_discussion and satisfaction_outcome are very similar. They have a pearson correlation of 0.85 ($p < 0.05$) and it is quite likely that both questions measure the same underlying principle. We thus average satisfaction_discussion and satisfaction_outcome with the label satisfaction.

The responses to satisfaction_discussion are not normal distributed while our other two are (in respect to kurtosis and skewness below 2). The composite variable satisfaction does also not follow a normal distribution. They however follow a student's t-distribution. We hence use the paired two-sample student's T-Test to compare satisfaction between our conditions. This reveals the differences are not significantly perceptible between the 0ms and 500ms condition ($p = 0.137$) and between 1000ms and 2000ms ($p = 0.285$). For all other conditions we have $p < 0.001$. For the contribution we use ANOVA, comparing the fit of a linear function as within subject design and Group as a blocking factor to see if we have an influence of delay. This is the case ($p < 0.01$) and a pairwise comparison of the different conditions revealed that the conditions with 1000ms and 2000ms were significantly

worse rated that the conditions with 500ms or 0ms added delay. In other word the satisfaction people had with the discussion was better if they had a delay of 500ms or less compared to a delay of 1000ms and more.

2.3.2 Objective Performance

We investigated whether the added delay had an influence on the number of correct answers (survival_score_group). To check whether delay had an influence on the achieved group score we used ANOVA and modeled the group score as a linear function of delay with the individual score and question as error components. This revealed that there is no significant influence of delay ($p > 0.4$). In Figure 2, we plotted the group scores per round color coded for the different delay conditions. We further investigated the influence of the round (i.e. the different questions) and ANOVA shows that they are statistically significantly different ($p < 0.001$) in other words the questions were of different difficulty.

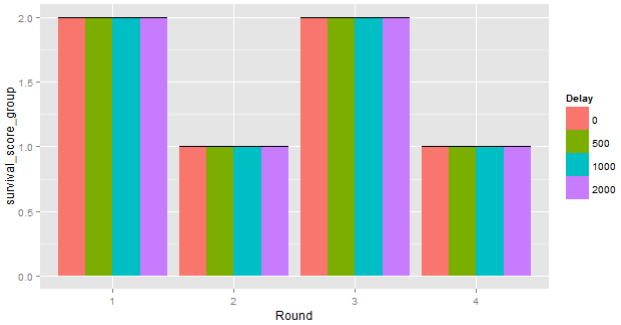

Figure 2 Group Survival Scores per Round

Even though we had chosen a task in which discussion is integral part in forming the solution we could not find a statistical significant influence in the task performance. On the other hand, the satisfaction with the discussion and feeling of having contributed is significantly less with higher delay.

2.4 Discussion

2.4.1 Communication Styles

Our moderator got the instruction not to make the final decision by himself but to make sure that everybody's opinion is heard, move the discussion a long if necessary and fill in the group answer in the form. As this description was intentionally very broad different styles of moderation emerged in interplay with the other participants. In six of the eight groups the moderator adapted at some point a systematic approach to hear the opinion of everybody. While in all groups the moderator inquired the opinion of particularly silent participants in some situations directly. Only in these groups the moderator started the discussion of a question in a structured way. The moderator either went by participant, inquiring from each participant what they had chosen as answer or by the answer, asking who had chosen it. In four of these groups an organic discussion without an explicitly enforced structure proceeded as the different options were debated. Two moderators maintained nearly throughout the whole session an explicit turn-taking scheme. These moderators showed an assertive behavior in situations when somebody spoke "out of turn" by interrupting them with a comment like "*I will come back to you in a moment.*". Three moderators started in the beginning of the session with an organization scheme while three others adopted it at a later stage. In three groups the moderator explicitly mentioned that the delay was so strong that he/she now adopts a strategy to handle it. Although in one of these groups the moderator had already employed this scheme in most occasions. Of the groups which adopted the communication style later, only one changed back to free discussion when there was less delay.

From the comments and behavior in the sessions and debriefing discussions we found indications that low and high moderated groups noticed the delay, but due to different interactions.

In three of our groups, the moderator switched (and said so) to an explicit turn-taking scheme with calling out names after particular many incidents of unintended interruptions preceded. In one group with many active participants, a lot of double talk occurred in lower delay conditions. A participant made the observation that she guesses there is now delay, since now the laugher in reaction to jokes came particularly late.

On the other hand, it was mentioned at least three times (one time during the debriefing) that participants noticed longer delay that participants needed to respond in strong moderated groups.

We did not find evidence that in the stronger moderated groups were more long pauses than in less strong moderated groups. This might be due to that in less moderated groups, often not directly addressed questions, like "*Who chose answer one?*" were asked, after which a long pause arises till a participant decides to answer.

2.4.2 Communication Problems

We present here two examples, from a more freely conversing group and a strongly moderated group which both occurred in the 2000ms condition.

Three participants in the more freely conversing group attempted to say something, a simultaneous talk start occurs, all participants turn silent, a long pause, all participants start again roughly at the same time, pause, all three start again and burst out in laughter after they realize it has happened again. After this the moderator decides to call out names.

In the stronger organized group:
P1 (Moderator): "*P2*"
P2: "*Can you hear me?*"
P1: "*Yes."*
As with a delay of 2000ms, the other participants do not hear the answer for another 2 seconds, two of them decide after roughly 3 seconds to answer.
P3: "*Yes*"
P4: "*We can hear you*."
After P2 hears P1, she/he starts to talk, but is interrupted shortly after that by P3 and P4. He/She is confused and annoyed and asks again: "*Can you hear me?*" Now the three people answer at the same time, after which P1 presents her/his reasoning (for the discussion).

Both incidents were mentioned in the discussion afterwards. People in the first group said they found it funny since this was an experiment. However, in real life, this would have probably been the point where they would have stopped the connection. P1 reported to be very annoyed by this incident: "*I was already at the top of my annoyance level, I was like "Hello can you hear me?*"

In the debriefing participants reported that, the delay was more problematic in situations where they wanted to discuss things in more detail.

P5: "*if you just vote it was okay - it was more problematic when we needed to brainstorm*"

They reported that the moderator had a more crucial role with higher delay.

P6: [discussing higher delay conditions] *for us it was easy … but you* [to moderator] *you needed to keep contro*l."

Besides calling out names of the participants, two moderators also employed the strategy of assessing who chose what by show of hands. In one session, this was attempted by two participants but was not adapted by the others. In the group where the moderator used this method particular often, he/she also stated that over the longer time till people rose their hand the delay was noticeable. In general the video was reported as a helpful addition in the group conferencing. It was mentioned as particularly better than audio-only conferencing.

P7: […] *in telephone conferences there is only noise* [everybody speaking over each other] *and silence… you really need a strong moderator.*

It was also an easy way to assess the opinion of particularly silent participants.

P5: […] *P8 didn't say much but it was always easy to see if she was agreeing and following along or had a different opinion.*

3. TOWARDS CONVERSATIONALLY AWARE COMMUNICATION SYSTEMS

3.1 Suitability analysis of recordings

Extending our previous study which used the audio data for investigating turn-taking[19], we started to investigate whether the video data is sufficient for cue extraction. Our first investigation was to check whether we can recognize faces and if the resolution would be high enough. Our participants had no instructions how to position themselves in front of the camera, so we were unsure if participants move away too often. We used OpenCV[1] based on the source code from the Attention Meter [11]. We analyzed 5 participants and a total of ca. 3 hours of video regarding the percentage of frames in which we could detect a (frontal) face, the size of face, eyes and mouth. We were able to track the face in average 96,4% (Stdev between participants 5.16) of the time. Our original data was recorded and transmitted in VGA (640x480px) and 30fps, although it would be possible to use FullHD resolution in the future. The average face area was a square with an average edge length of 152px (Stdev 28px). The area for an eye was estimated by the AttentionMeter to have an edge length of ca. 25px (Stdev 5px). For the mouth an area of 50px*90px was estimated (Stdev of 10px and 16px respectively). This resolution was sufficient for smile detection, which was adequately detected in a few manually checked samples, but we did not perform a systematic investigation yet. In theory the eye area should be sufficient for webcam based gaze detection [23].

3.2 Cues

If we want to build systems, which are able to appropriately act upon the conversation, we need to be able to identify meaningful situations. As previous real-time communication was mostly audio-only based this was used as the main source for investigating communicative problems [20, 22]. In the following we discuss which cues can be extracted and how the distortion of video-communication influences them.

3.2.1 Gaze

Several studies have found that gaze plays an important role in turn-taking [8]. Recently unmodified webcams have been identified towards their suitability for eye-gaze detection [23]. With a reported accuracy below 3.68° [23] this would be suitable, depending on screen size and layout, to detect at which participant the user is looking at. In a standard video-conferencing setup mutual gaze is not possible. However this gaze-awareness is crucial for several turn-prediction models [8]. With gaze detection we can assess at whom the user looks although this information is not available to the other participants, hence they are not aware of it.

3.2.2 Turn-Taking prediction

While some systems already react to user activity on a simple level (e.g. Goggle Hangout focusing the loudest participant), a finer detection and prediction of turn starts and ends can provide more support for the ongoing communication. Currently the detection in communication systems focusses solely on the timing of turns and thus is only able to react after the fact on a change. Current research agrees that humans in a conversation perform prediction of when the current speaker will end his/her turn. This is heavily researched in human-computer dialogue systems [4]. In the collocated setting, turn-prediction models movement, gaze [8], intonation and keywords A prediction of next speakers would also allow for a pre-preparation of streams which are likely to be shown in high-quality on a network level. However moving on to the video-mediated communication, besides that we need a high enough audio-video quality these patterns need to be investigated due to temporal distortions (i.e. delay, jitter, synchronization) and spatial distortion (size, layout and orchestration).

3.2.3 Backchannels

To signal the current speaker, whether we are still following, agree/disagree or want her/him to go on, we give small feedback often described as backchannels [1]. Short utterances like "yes" and "uhm" are used as a confirmation to the speaker that we are still interested in the conversation [22]. [25]. Hmm could be more from allwood about different kind of levels in there These cues are often in parallel to the main speaker and should not be confused with double talk indicating conversational problems. A common problem for automatic detection is that these short utterances are particularly hard to differentiate from small noises that can be picked up from the microphone by moving things, adjusting headset or computer etc. We can assess mouth movement to get a better differentiation, a similar approach (but not stated whether automatically or manually done) was used for the offline processing when investigating the Halo system [5]. Furthermore prosodic features can be used to infer some semantics and discriminate against noise [26][13]. Additionally head nods and shake are possible to detect [11], which are also used as backchannels [10]. Especially the non-verbal utterances agree or disagree "mhm" or "hmm" are accompanied by nodding and shaking and can be further qualified. Also facial expressions, like smiling, are known backchannels [9] and can be detected with OpenCV or higher level software like SHORE [14]. In light of delay in video-conferencing it is of particular interest how the distorted timing affects the experience, as it is known that wrong timing can be interpreted as inappropriate [9].

3.2.4 Double Talk

Even though we generally do not talk over each other in a conversation [15], there are instances where double talk is actually desired. We thus need to be able to discriminate normal communication flow from conversational problems.

[1] www.opencv.org

Shared laughter

Smile detection has become already a commodity in cameras but we are not making use of it in video-communication. The detection of shared laughter [16] helps to discriminate shared expression of emotion from accidentally overlapping speech.

Intended and unintended interruptions

The problems in video-communication make it on the one hand hard to interrupt other participants when desired and produces on the other hand unintended interruptions like simultaneous starts. To some degree the timing of turn-taking gives insight into discriminating simultaneous starts from the attempt to take over a turn [22]. As in video-conferencing every participant has a potentially different perception [20] it is important to know the synchronization between video-clients. Turn prediction models could help here to assess whether multiple participants were getting ready to speak. Interruptions are often connected to a different amplitude level [12] which is often distorted in video-conferencing due to speaker equalization.

3.3 Conversation Aware Adaptation

Our approach to mitigate conversation problems is two-fold: we can raise the awareness of cues and conversational problems. This can reduce conversational problems and provide better means for users to adapt their communication style. Further, we want to implement explicit turn-taking mechanisms that can be employed without the need for a human moderator.

3.3.1 Awareness

Speaker Discrimination

Especially in the case of small video-streams, the differentiation between participants can be difficult. In our experiments, this was only raised by a few participants and only in situations when the voice of participants was perceived very similar. A visual indication was proposed, like in Google Hangout, which shows an audio-level next to participants.

Gaze

Conveying virtual eye gaze has been demonstrated by artificial modifying the eyes in the video-stream which can lead to unrealistic looking eyes [6]. This approach is however only developed for dyadic scenarios. In multi-party scenarios, the correct warping of the eyes has additional challenges as also the head rotation is displaced for a group layout. The GAZE-2 system uses 3D rotated video streams which look in the direction of the target [25]. The system could also convey mutual gaze by employing multiple cameras, a translucent mirror and a professional eye-tracker.

For our setup with standard hardware, the rotating approach can convey some directional gaze to other participants. An increase in size or pop out effect could demonstrate gaze directed at the user.

Delay

Research suggests that if participants are informed about delay, they will be more sensitive [24]. Our study showed some users do not perceive the delay at all. On the other hand conversational problems will occur and there are known issues in the conversation that increase with delay [21]. Our study shows that participants can adapt to this situation and thus ease the problem. Thus it seems suitable to make users aware of delay (e.g. with a visual notification), if the delay is over a certain threshold and communication problems are detected. Skype and Google Hangout use both a visual feedback to indicate a bad connection, but not to particularly to indicate delay.

Backchannels

Backchannels are meant to be unobtrusive and in the background of the communication [7]. Due to the often unavailability of 3D audio, overlapping of speech is less intelligible, as we cannot focus easily on one audio-source (known as the "cocktail party effect") [34]. Depending on the layout and device, participants might be presented as small thumbnails, which hinder the recognition of non-verbal cues like nods and shakes. Thus, the challenge is to amplify the feedback, while rendering them in a non-disrupting manner.

Visual backchannels could be amplified by raising the presentation size or color saturation of attentive and agreeing participants, or performing nod or shake like movements on the whole stream. These methods are particular interesting in situations when the bandwidth is not sufficient for a video-stream, as they could also be performed on an image of the user.

For audio backchannels we suspect that they are more disruptive, depending on 3D audio availability and the delay. Thus, we can make the presentation of the backchannel depending on the synchronization. Here it is important whether the back channel would be still presented in the turn (in which it was meant). If it will collide with the next speaker, it might be better to omit the rendering in favor of a visual representation.

3.3.2 Turn-taking

In computer-mediated communication, it is relatively easy to enforce a turn taking scheme. The floor can be assigned to one participant and the audio of others oppressed. As this is an intrusive approach, such a mode should be specifically enabled by the users (but could be suggested by the system). Further, we observed that an explicit turn taking approach can have benefits in high delay situations. This is also a common approach in formal meetings and larger groups. In our scenario, the moderator often first assessed the opinion of each participant and then more dynamically reacted upon conflicting positions. While the latter (content sensitive) approach would not be feasible, the turn management would be possible. It is not uncommon, that discussion participants first state their opinion or introduce themselves. Therefore, a simple round-robin like approach could be used, although it has to be investigated whether it is acceptable if the system proposes turns on the users. A common more dynamic approach is the employment of a speaker queue. Users could signal (e.g. via hand gesture) that they want to speak. The users could be prioritized by their past speaking time (making sure that everybody has an equal chance of giving his/her opinion), whether they are direct addressees of the speaker or have direct answer to the last speaker. Of especial importance for a smooth turn transition is how the active speaker signals the end of his/her turn. An accurate end-of-turn prediction would allow the most natural transition compared to explicit signaling or long pauses. Our scenario was a consensus task, thus participants were interested in letting others speak. For scenarios that are, more competitive, users could "hog" their turn if no possibility of interruption is given. Besides setting a time limit (if other users indicate they want their turn), also a majority vote to end the current speakers turn could be feasible solution.

4. CONCLUSION AND FUTURE WORK

In this paper, we report current problems in video-conferencing and present approaches on how conversation aware systems can help participants in these challenging situations. The objective results of our study suggest that participants can still communicate all necessary information. The subjective ratings and feedback reveal that the communication gets severely disturbed and the satisfaction is lowered. Users adapt to video-communication

situation by employing explicit turn-taking schemes. Based on this, we investigated if our data would be sufficient for further analysis and which detectable cues could be leveraged. To ease the communication problems we proposed a set of awareness raising mechanisms and to support the explicit turn-taking we suggested an automatically managed speaking queue. In our future work, we will investigate how robust the discussed cues can be practically extracted from our data and implement based on this proposed conversation aware adaptions. We want to investigate how the synchronicity of non-verbal cues affects the experience and how much conversation aware adaptions can alleviate the problems.

5. ACKNOWLEDGEMENTS

This research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760 (Vconect project).

6. REFERENCES

[1] Allwood, J. et al. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*. 9, 1 (1992), 1–26.

[2] Berndtsson, G. et al. 2012. Subjective quality assessment of video conferences and telemeetings. *Packet Video Workshop (PV), 2012 19th International* (2012), 25–30.

[3] Biech, E. 2007. *The Pfeiffer book of successful team-building tools: Best of the annuals*. Pfeiffer.

[4] Bohus, D. and Horvitz, E. 2011. Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions. *Proceedings of the SIGDIAL 2011 Conference* (Stroudsburg, PA, USA, 2011), 98–109.

[5] Geelhoed, E. et al. 2009. *Effects of Latency on Telepresence*. HP labs technical report: HPL-2009-120 http://www. hpl. hp. com/techreports/2009/HPL-2009-120. html.

[6] Gemmell, J. et al. 2000. Gaze awareness for video-conferencing: A software approach. *IEEE Multimedia*. 7, 4 (2000), 26–35.

[7] Heldner, M. et al. 2010. Pitch Similarity in the Vicinity of Backchannels. (2010).

[8] Ishii, R. et al. 2013. Predicting Next Speaker and Timing from Gaze Transition Patterns in Multi-party Meetings. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (New York, NY, USA, 2013), 79–86.

[9] Kok, I. de and Heylen, D. 2011. When Do We Smile? Analysis and Modeling of the Nonverbal Context of Listener Smiles in Conversation. *Affective Computing and Intelligent Interaction*. S. D'Mello et al., eds. Springer Berlin Heidelberg. 477–486.

[10] Koutsombogera, M. and Papageorgiou, H. 2010. Linguistic and non-verbal cues for the induction of silent feedback. *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer. 327–336.

[11] Lee, C.-H.J. et al. 2006. Attention Meter: A Vision-based Input Toolkit for Interaction Designers. *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2006), 1007–1012.

[12] Meltzer, L. et al. 1971. Interruption outcomes and vocal amplitude: Explorations in social psychophysics. *Journal of Personality and Social Psychology*. 18, 3 (1971), 392–402.

[13] Noguchi, H. and Den, Y. 1998. Prosody-based detection of the context of backchannel responses. *ICSLP* (1998).

[14] Ruf, T. et al. 2011. Face Detection with the Sophisticated High-speed Object Recognition Engine (SHORE). *Microelectronic Systems*. A. Heuberger et al., eds. Springer Berlin Heidelberg. 243–252.

[15] Sacks, H. et al. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*. 50, 4 (1974), 696–735.

[16] Scherer, S. et al. 2012. Spotting Laughter in Natural Multiparty Conversations: A Comparison of Automatic Online and Offline Approaches Using Audiovisual Data. *ACM Trans. Interact. Intell. Syst.* 2, 1 (Mar. 2012), 4:1–4:31.

[17] Schmitt, M. et al. 2013. A QoE Testbed for Socially-aware Video-mediated Group Communication. *Proc. of the 2nd International Workshop on Socially-aware Multimedia* (New York, NY, USA, 2013), 37–42.

[18] Schmitt, M. et al. 2014. Asymmetric Delay in Video-Mediated Group Discussions. *To appear in 6th International Workshop on Quality of Multimedia Experience (QoMEX), 2014* (2014).

[19] Schmitt, M. et al. 2014. The influence of interactivity patterns on the Quality of Experience in multi-party video-mediated conversations under symmetric delay conditions. *Submitted to Proc. of the 3rd International Workshop on Socially-aware Multimedia* (New York, NY, USA, 2014).

[20] Schoenenberg, K. et al. 2014. On interaction behaviour in telephone conversations under transmission delay. *Speech Communication*. 63–64, (Sep. 2014), 1–14.

[21] Schoenenberg, K. et al. 2014. Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end. *International Journal of Human-Computer Studies*. 72, 5 (May 2014), 477–487.

[22] Sellen, A.J. 1995. Remote conversations: the effects of mediating talk with technology. *Hum.-Comput. Interact.* 10, 4 (Dec. 1995), 401–444.

[23] Sewell, W. and Komogortsev, O. 2010. Real-time Eye Gaze Tracking with an Unmodified Commodity Webcam Employing a Neural Network. *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2010), 3739–3744.

[24] Tam, J. et al. 2012. Video increases the perception of naturalness during remote interactions with latency. *Proc. of CHI'12* (New York, NY, USA, 2012), 2045–2050.

[25] Vertegaal, R. et al. 2003. GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2003), 521–528.

[26] Ward, N. and Tsukahara, W. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*. 32, 8 (Jul. 2000), 1177–1207.

Columns on Last Page Should Be Made As Close As Possible to Equal Length