Experimental Enquiry into Automatically Orchestrated Live Video Communication in Social Settings

Marian F Ursu
University of York
United Kingdom
marian.ursu@york.ac.uk

Manolis Falelakis
Goldsmiths University London
United Kingdom
m.falelakis@gold.ac.uk

Martin Groen
University of Utrecht
Netherlands
m.g.m.groen@uu.nl

Rene Kaiser
JOANNEUM RESEARCH
Austria
rene.kaiser@joanneum.at

Michael Frantzis
Goldsmiths University London
United Kingdom
m.frantzis@gold.ac.uk

ABSTRACT

’Orchestration’ refers to the ability of a live video communication system to adapt in real-time to the communication context with a view to enhance the quality of mediation and subsequently the quality of interaction between participants. For example, this can be done by reframing the cameras and changing the way in which the video content is mixed on each screen. To be a feasible solution, orchestration has to be an automatic process. This paper reports a study of orchestration carried out in the social setting of a group of friends playing social games from two separate living rooms. The quality of the communication was assessed via two measures: one objective, in the form of task efficiency, and one subjective, in the form of a questionnaire. The objective measure indicated that mediated communication can be improved through orchestration, but the subjective measure was inconclusive. The paper also uncovers some of the complexities of the experimental space associated with orchestrated mediated communication and aims to provide motivation for further research into this new communication paradigm.

Author Keywords

Videoconferencing; Telepresence; Mediated communication; Group communication; Orchestration; Virtual Director; Video; Live; Communication; Group; Interaction; Television.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

TVX'15, June 03 - 05, 2015, Brussels, Belgium.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3526-3/15/06:.$15.00.
DOI: http://dx.doi.org/10.1145/2745197.2745211

INTRODUCTION

The use of live video mediation in social communication is growing at a fast pace. Products such as Microsoft Skype and Google+ Hangouts are increasingly becoming part of our daily lives. Group communication is being addressed, but the current systems do not go beyond ’talking heads’ [14]. The communication setups currently dealt with by live video communication systems are still quite restrictive.

Imagine, for example, a conversation across a live video link among a group of friends located in two separate living rooms, carried out over the rooms’ TV screens. One fixed camera per location providing one fixed shot could suffice to mediate the conversation, particularly as people are able to adapt their behaviour [22] and find artful ways to overcome the limitations of the communication system [10]. However, there are also many reasons why it may not suffice [10, 22]. For example, the one shot may lack the necessary levels of detail to appropriately convey particular actions, facial expressions and messages expressed through body language, and so it may lead to ’unnatural behaviour’, for example by constraining people to be aware of the camera and ensure that they are always in the frame. An alternative would be to employ more than one shot per location, through multiple cameras, with the corresponding communication system able to ’follow the communication as it progresses’, choosing the right shots and mixing them in a manner similar to film and TV [21]. Figure 1 illustrates sample shots that could be captured from one of the locations and mixed on the TV screen of the other.

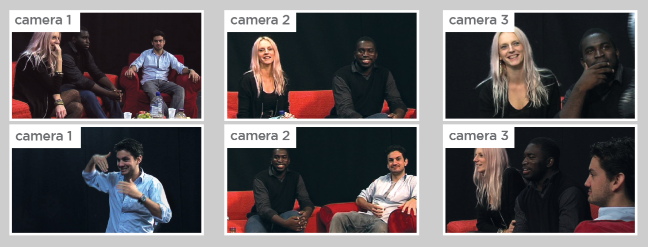


Figure 1. Example of shots that could be used in a multi-camera live video communication system.

Such an approach appears to allow people to have more natural interactions with each other, potentially increasing their sense of empathy [16], connectedness and presence, as well as counteracting spatial distortion [12], perceptual invariance [15] and increasing the 'transparency' of the communication medium itself. Furthermore, such an approach appears to be necessarily required in more comprehensive communication setups intended to be mediated through live audio and video, such as connecting more than two locations and providing for larger groups, involved in more complex social activities.

The ability of a communication system to dynamically reframe the cameras, control the microphones and mix the available live audio-video content on the available screens and speakers, to ensure, for each participant, an optimum in the perspectives and levels of detail perceived from the other spaces, has been denoted *communication orchestration* or, simply, *orchestration* [21]. Orchestration is similar to the compilation of live TV programmes, but it is fundamentally different, as its underlying principle is *communication*, not storytelling. Furthermore, in live orchestrated communication, the way in which the cameras and microphones are controlled and audio video content is mixed has a direct effect upon the conversation itself, whereas in TV, events are being merely depicted, the lens being a mere observer with no agency. Lastly, TV considers one viewpoint, namely that of the aggregated audience, whereas orchestration must take into account the perspective of each of the participating members. The success of such adaptable context-aware communication systems depends upon the 'recipes' according to which they perform orchestration – i.e., their orchestration *logics*, *grammars* or *knowledge*. Any body of orchestration knowledge subsumes two main reasoning processes:

- *understanding* the continuously changing context of the communication taking place, and
- *adapting* the communication system to appropriately provide for the identified communication context

The former process takes input from sensors and feature extraction procedures (e.g., voice activity extraction and face detection) and 'lifts' such primitive information into aspects of the communication context (e.g., the person holding the conversation turn). The latter process applies screen-language conventions to determine how to control the cameras and how to mix the available live content in order to best mediate the identified communication context (through the former process) among all the participants.

Orchestrated communication is a rather new area of work and, consequently, there is very little orchestration knowledge available. There is significant work regarding the understanding of the context of communication (see the related work section), but this has not been linked, yet, to directing cameras and screen 'vision mixing' for live video-mediated communication. Conversely, there is significant expertise and knowledge regarding directing cameras and mixing content, but this is geared towards film and TV, not to adaptive, context-aware communication systems. Orchestration logic represents the *grammar of the language of live video communication*. Orchestration could (and should) take video-mediated communication to where film and TV are today. Orchestration grammars should be refined incrementally, through experimental enquiry and ought to be possible to be automated.

Refining and validating orchestration knowledge is not straightforward. On one hand, orchestration knowledge is made of a combination of rules with potentially strong interdependencies. Its study, therefore, is radically different from that of the effect of a single variable (such as the size of the head) upon the mediated communication. On the other hand, orchestration is not a *general* recipe to any video-mediated communication setup. Each particular setup requires its own specific orchestration logic. Yet, more generic bodies of orchestration knowledge, applicable to wider communication setups, need to be identified, if the whole approach of orchestrated communication is to be feasible and effective. There is tight coupling between choosing the communication setup and refining the corresponding orchestration knowledge, each influencing the other. These aspects create a highly non-deterministic space, very dynamic and quite challenging for experimental enquiry.

This paper presents an experimental investigation into *automatically orchestrated* video-mediated communication within a particular setting of social group interaction: friends playing social games, chatting and having fun, from two separate living rooms, using the TV as the communication conduit to the other room. It is motivated by [21], which showed, by carrying out orchestration *entirely by human operators*, that orchestration can have a positive impact in such communication setups. The study presented here addresses the questions of whether a *completely automatic* system could have a similar impact. In the process of answering this question, the study also uncovered aspects regarding the complexity of the experimental enquiry into this novel communication paradigm.

RELATED WORK

As already mentioned, there is a significant body of research related to the automatic understanding of communication contexts. The conceptual underpinning is provided by 'conversation analysis' [20], with 'turn-taking' [19] being probably the most investigated concept. More recently, this line of research has been extended to 'social signal processing' [6, 23]. Its ultimate aim is the development of procedures for the automatic extraction of features of social interaction, taking verbal as well as non-verbal behaviour (such as prosody, posture, gaze, gestures, etc.) as the object of analysis. Various conceptual models have been designed and validated. The review provided in [6] groups them into four categories, namely those pertaining to *interaction management* (e.g., addressing and turn-taking), *internal states* (e.g., interest and emotional engagement), *personality traits* (e.g., dominance and extroversion), and *social relationships* (e.g., formal roles, such as interviewee, and social roles, for example determined by the social status). The more recent review provided in [24] uses a similar taxonomy, namely *social actions* and *interactions*, *social emotions*, *social attitudes*, and *social relations*. The state of the art is impressive when it comes to implementations able to automatically extract such features of social interaction. They are directly relevant to orchestration. However, orchestration places two very demanding requirements on such implementations: they need to work in real-time and operate in uncontrolled, complex and ’noisy’ environments. Many of the existing implementations do not meet these requirements. Nevertheless, they can certainly inform of what is possible and might become available in the (near) future. It is also rewarding to notice that some of these implementations are made available as open source, such as OpenSMILE [2], which provides a rich set of low level features that can be extracted from audio (e.g., waveform properties, signal energy, loudness) [3] and from video (e.g., face detection) [2]. They can be processed jointly in a single framework allowing for time synchronization of parameters, on-line incremental processing as well as off-line and batch processing, and the extraction of statistical data. In conclusion, there already are insightful research results with regards to the understanding of communication contexts, and the momentum behind this work is increasing. However, when it comes to orchestration, this is only half of the picture: the part that maps such identified traits into camera and mixing decisions still needs to be developed.

The other half of the orchestration reasoning process – i.e., controlling cameras and vision mixers – is also the subject of related research, but most of this work is carried out in the context of games, film and TV production. The main problem dealt with by games research is the selection, at each point in the game interaction, of the most appropriate viewpoint – referring to position, orientation, focal distance, etc. of the virtual camera – to represent the 3D world. The selection is constrained by editing principles, adapted from film and TV, related to concepts such as 2D/3D continuity, rhythm, story and emotion. Such editing principles, together with the techniques employed in their computational expression, are surveyed in [18]. The description is made from the viewpoint of the approach taken for knowledge modelling: procedural, declarative and optimization. Narrative techniques from film and TV have already made their way into game screen language to provide for more dramatic and engaging interactions. This is strong motivation for orchestration research, which is also hypothesizing that film and TV techniques could be used to enhance the means of social interaction. However, orchestration addresses communication within social groups which is mediated by video, whereas games are about interaction within virtual worlds.

On a parallel track, in the area of film and TV, the most relevant research to orchestration concerns the formalization of shot description, directing, and other aspects of the film production workflow, with a view to providing for more rigorous descriptions and to automate parts thereof. [25], for example, presents a symbolic language to express the content of film, which is accompanied by a number of software tools intended for use during film planning and visualization of ideas. This is relevant to orchestration, as such work, too, is concerned with deciding upon the right shots for a specific situation, according to some embedded logic. However, this is about narrating events to a passive audience, not about live communication, as orchestration is.

Related work that connects the two main reasoning components of orchestration – context understanding and directing video expression – is scarce. [1] investigated the identification of ’floor control’ in group meetings, using a multi-modal approach that combines patterns of speech (e.g., the use of discourse markers) with visual cues (e.g., gaze exchanges). Identifying who has control of the floor provides an important focus for information retrieval and summarization from audio-video recordings. This work focuses on the definition of a model, using manual annotations of audio and video recordings, and not on its implementation. Also, its application is summarization and retrieval, not communication. [17] reports a study that employed principles from television production to capture meetings of small groups. Events such as ’speaker change’, ’posture change’ and ’head orientation’ are mapped, according to an internal logic, onto different types of shots, such as ’close-up’, ’two-person’ and ’overview’. Automatically compiled representations are compared with representations compiled by a film crew. This work is closely related to orchestration, but its overall aim is to better recount meetings (to passive viewers), not to provide for real-time communication between participants. [1] and [17] consider face-to-face communication and aim to develop better ways for their recording and recounting, whereas orchestration considers mediated communication. Also, [1] and [17] are set in the more structured space of meetings, whereas orchestration, here, is studied within the less structured space of social interaction.

The work reported in [7] is about ’orchestration’ (our terminology). Two groups of people have a meeting across a video link. One direction of communication was orchestrated, employing one mobile camera, whereas the other one was ’conventional’, using a static wide shot. The orchestration logic was refined by analysing TV debate programmes. Eight types of shots were refined, including ’speaker’, ’listener’ and ’speaker and listener’. The duration of each shot was clocked and so a frequency distribution was compiled from the TV content. The orchestration logic was subsequently expressed as two-shot transition tables, each specifying the probability of transition from any shot to any other shot. One table was used when the speaker changed, the other was used in conjunction with the frequency distribution. The ’change in speaker’ cue was inserted manually, by a human operator. This is inspiring work, but the setup was limited: context understanding was not automatic, only one direction of communication was orchestrated and the interaction context was that of a structured meeting. Benefits of the orchestrated communication have been observed, such as better conveyance of the feelings and intentions of the active speaker, but also benefits of the conventional link have been noticed, such as making the situation easier to grasp. The conclusion was that orchestration can be a way of improving the quality of communication in structured meetings. However, interestingly, the study hypothesized that this paradigm is not portable to the less structured conversations of the social space, as speaker identification is not portable to situations when people take short turns or speak over each other.

The work reported in [21] attempts to invalidate this hypothesis and takes orchestration into a social setting. A group of friends play social games and have idle chats from two separate living rooms. Both communication directions are orchestrated, the system employing a number of fixed cameras per location, each able to provide a particular functional shot. The orchestration logic is expressed as a set of rules, but the corresponding reasoning process is being entirely carried out by human operators. The logic is based not only on 'conversation turns' (similar to active speaker in the previously cited work), but also on other features of conversation analysis, such as 'crosstalk' and 'quick turn taking'. The study concluded that orchestration can improve the quality of the communication and interaction experiences in a social setting, and motivated enquiries into automatically performed orchestration. Our paper reports on such an enquiry, in a setting similar to [21] and using a similar orchestration logic, but automatically executed.

In conclusion, there are various bodies of research related to the two main reasoning processes subsumed by orchestration, but there is no study yet of an automatically orchestrated video communication system employed in a social setting – i.e. a system *aware* of the communication context and able to *dynamically configure* itself in real-time to best address the communication needs. Such a study is presented in this paper.

EXPERIMENT DESCRIPTION

Communication setup

As stated in the introduction, orchestration is not a general recipe to any communication setup. Rather, particular setups require particular bodies of orchestration logic. However, such 'particular setups' ought to have a level of generality to ensure reuse of the orchestration logic to the extent that the proposition is feasible and cost effective.

The general setup chosen for this study is that of friends playing social games, chatting and having fun, from two separate living rooms (see Figure 2).

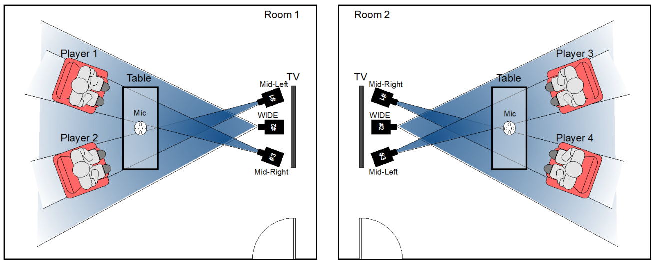


Figure 2. Communication setup.

Each room had two armchairs, a coffee table, a lamp shade hanging over it, and a TV table. Each room used one HD 50 inch TV screen and 3 HD cameras (Sony EVI-HD1), positioned next to each other on the same table as the TV. Each camera provided a specific functional shot which was framed at the beginning of each interaction session, but not dynamically reframed afterwards – orchestration was carried out through dynamic mixing only. There was stereo sound with echo cancellation between the two rooms. A star-4 microphone array was concealed in the lampshade, some 1.5m equidistant from the two people, and the speakers were aligned with the TV. Each group consisted of 4 people, 2 per room. They played a variant of the game of Articulate, where players have to describe words to their team as quickly as possible, in a window of 30 seconds. If the word is not guessed, the other team can join and the description continues for another 30 seconds.

Orchestration logic

Any corpus of orchestration knowledge should be founded on a combination of key interaction principles, such as: following the most active speaker(s); providing a balanced view of all the people involved in the conversation; following the conversation flow by focusing on the active speakers as well as the listeners who actively react to what is being said (e.g., by interjecting or nodding); or others. The principle behind the orchestration logic used here was that of following and giving visual prominence to the active speakers, but ensuring that the visual continuity of the representation is preserved.

The former requirement was based on the concepts of *turn shift* and *short turn taking*, which were automatically inferred from a more primitive cue, namely 'start voice activity by person P'. The definition of the concepts is given below:

> *Only one person can have the turn at any one time.*
>
> *Person P takes the turn if there is voice activity detected from P and P is not currently holding the turn and P is not cross talking over someone else.*
>
> *Person Q is cross talking over person P if P has the turn and there is voice activity detected from Q.*
>
> *There is a pattern of short turn taking if 3 or more turn shifts occur within 5 seconds.*

For the dynamic construction of the visual representation, a small set of types of shots was considered to be sufficient: a *wide shot*, framed to include both people in the room, allowing them also a bit of space for movement, and two tighter shots, *mid-shots*, one for each person in the room (Figure 3). Orchestration was done solely through their mixing (i.e., there was no camera reframing), using only clean cut transitions.

The mixing logic, expressed as rules in natural language, is described below (the rules are stated from the point of view of the room/screen *for* which the mixing is taking place, therefore, the persons referred to in the rules are from the other room):

1. *If person P starts a conversation turn (i.e., a turn-shift to P occurs), show the medium shot of P.*
2. *If there is a pattern of short turn-takings, then show the wide.*
3. *If there is no turn-shift for 5 seconds, then show the wide.*
4. *No change of shot is allowed within 2 seconds of the previous cut (highest priority, when there are conflicts).*

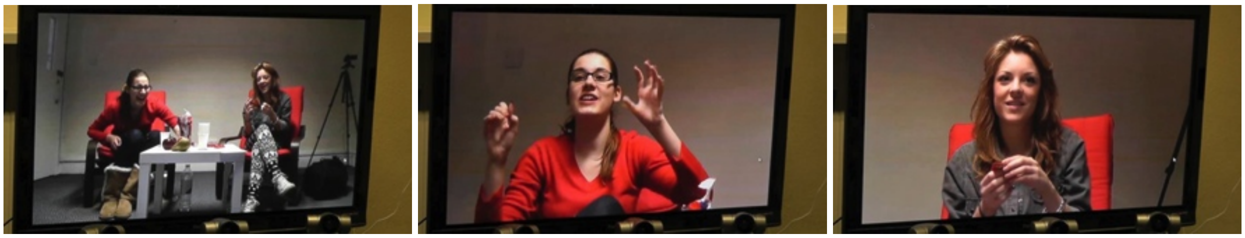


Figure 3. The shots for orchestration: a wide and 2 mid-shots.

The last rule, inspired by TV production, aimed to avoid too often cuts, which would break the visual continuity.

Conditions

Four experimental conditions were found necessary for this investigation, which were all mediated through the same communication infrastructure and differed only with regards to orchestration: *context aware automatic orchestration (CA-A)*, *context aware manual orchestration (CA-M)*, *static* and *context unaware mixing (CU)*. All the conditions were of equal length (15 minutes) and their order was counterbalanced as follows: CA-A was experienced by the participants 2 times as first condition, 2 as second, 3 as third and 2 as fourth; CA-M was 3 times the first one to take place first, 2 times second, 1 time third and 3 times fourth; the static condition was 2 times first, 3 second, 2 third and 2 fourth; while the CU condition was experienced 2 times as first, 2 as second, 3 as third and 2 as fourth.

In the CA-A condition, the whole reasoning process was entirely automatic, from the extraction of primitive cues to the issuing of mixing decisions. In the CA-M condition, the entire reasoning process was performed by two human operators (one per screen/room). They were instructed to follow rigorously the rule-set aforementioned and rely as little as possible on their mixing experience and knowledge (tacit or explicit). As orchestrated communication experiences are determined by the bodies of rules on which they are based, any rule change could have a significant impact upon the group communication experience. This was the reason why the operators were instructed to follow rigorously the given set of rules. Otherwise, there would have not been any comparable behaviour between manual and automatic orchestration, nor continuity across experiments (if the operators are not constrained by a given set of rules, in each experiment they would have applied their tacit knowledge differently). However, the experience of the manual operators was somewhat reflected in their interpretation of the rules, which were expressed in natural language.

In the static condition there is no reasoning process involved at all, a wide shot of the room being always shown in the other room. The CU condition was included as a reference for measuring the effect of context-awareness (both of the CA conditions). In the CU condition, a simple time-based algorithm was used to drive the mixing process: a change of shot, from the available three, was continuously made at a random interval of time varying between 2 and 5 seconds (uniform distribution). This simulated the pace with which human operators cut between shots for this communication setup (inferred through prior experimentation) and was aligned with the thresholds used in the orchestration rules.

A number of criteria informed the choice of these conditions. To make manual orchestration possible, the size and expression of the logic had to be sufficiently small and clear to allow the human operators to digest it and apply it rigorously. To ensure that the static condition was not disadvantaged by design, the setup was chosen such that the wide shot captured all the necessary information from the other room. In order to not distract the participants on reverse engineering the system or 'playing' with its capabilities, they were not briefed beforehand on the conditions they were going to experience. Although the CU condition involves random choices, it is still an informed way of mixing, as: (i) it mixes the available shots with, more or less, the same pace as the CA conditions; and (ii) out of the 3 shots available for mixing, 2 are always relevant, as they always capture the active speaker.

Sessions and participants

A total of 9 interaction sessions were carried out with a total of 36 participants – a session being defined by a group of 4 people undertaking all 4 conditions. There were 2 teams of 2 participants in each session who experienced the 4 different conditions in succession, giving a 4 factorial within-participants design. The participants were chosen such that all the team members knew each other from before, but they did not necessarily know the members of the opposite team. Out of the 36 participants, 13 were male and 23 female, they were aged between 18 and 51 (mean 23.1, variance 4.8) and 28 of them were native English speakers. 21 of them knew their teammate for longer than 6 months.

There were 2 orchestration operators for the manual condition, both with prior film editing experience and they were in charge of all 9 experimental sessions.

Communication system

The architecture of the bespoke communication system used in the experiment is depicted in Figure 4. Video and audio were processed on parallel channels. The round-trip video delay was about 600ms and audio was slowed down to ensure synchronicity. The audio chain was supported by the audio communication engine clients (ACEs) and the server component, the ACE Multipoint Control Unit (MCU) (described in [5]). They provided for a high quality stereo connection between the two rooms with very good echo cancellation. The video chain was supported by the video acquisition clients (VACQs), the video router (VR) and the video composition engines (VCEs). Of key importance here are the VCEs [8], which were able to execute the commands of the orchestration engine. The VCEs decoded in parallel all the streams received from the other room, to ensure that they could cut instantaneously between them.

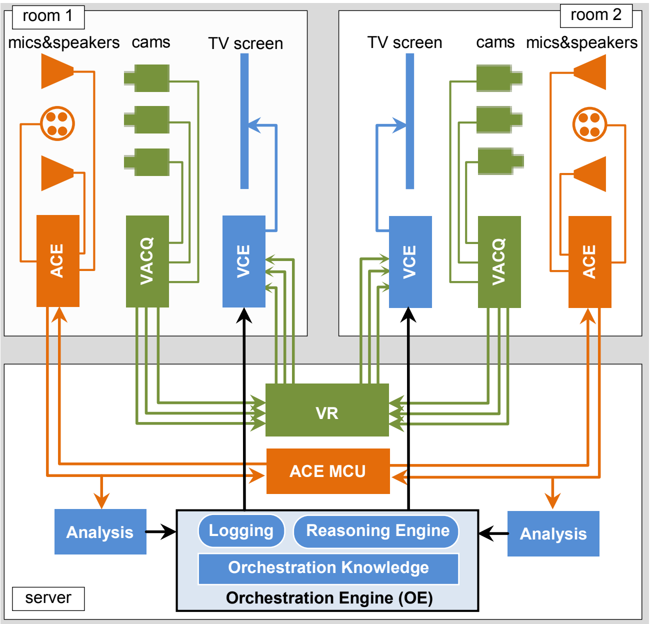


Figure 4. Orchestrated communication system for 2 rooms.

The network connectivity was chosen such as to impose no constraints on the system described above, i.e. there was sufficient bandwidth for each room to upload 3 HD streams and download 3 HD streams.

The output of the Orchestration Engine (OE) consisted of cut commands sent to the VCE. The input into the OE consisted of the primitive cues received from the Analysis components [13]. Two types of primitive cues – 'directional audio activity' and 'face detection' – extracted by the Analysis engines were further fused by it into the 'start/stop voice activity by face X' cue, which was the primitive cue used by the OE to infer the required conversation cues – turn-shift and short-turn-taking.

The Orchestration Engine (OE) was built in a declarative model, with the knowledge base separated from the reasoning engine [4]. The logic was implemented using rules in an event-based framework [9], using the JBoss Drools inference engine[1]. The main issues that needed to be considered in the implementation included: (i) conflict resolution (ii) processing of asynchronous events, (iii) management of imprecision in measurements and definitions, (iv) real-time event processing (v) ability for temporal reasoning.

All the inferences of the OE were time-stamped logged, including detection of conversation turn-shifts, short turn taking, and decisions to cut.

In the manually orchestrated condition, the OE was replaced by two human operators, one per room, each operating via a live-mixing interface, in-house implemented, plugged into the communication framework. Each operator had a view of the room *from* which they were mixing and the screen of the room *for* which they were mixing. Three buttons, corresponding to the three possible shots, were mapped onto three adjacent keys for shot selection. In order to assist them with the application of the rules, rule 4 was enforced in the live-mixing interface by disabling the keys and buttons for 2 seconds from the previous action.

For the context-unaware mixing condition, the OE was replaced by a simple piece of code implementing the time-based mixing algorithm.

RESULTS

The objective measure used was that of 'task efficiency', defined as the average game points won per condition (Figure 5). Significant differences were found here, as participants performed considerably poorer in the context unaware mixing condition (CU), achieving their best results in the automatically orchestrated one (CA-A). More specifically, the results show that different conditions affected the gameplay of the participants, $F(3, 30) = 4.34$, $p = .01$, $\eta\ p^2 = .30$. The mean number of game points won in the automatic condition (CA-A) was 25.91 (SD = 8.07), in the manual condition (CA-M) was 24.00 (SD = 9.74), in the static condition was 23.09 (SD = 11.42) and in the context-unaware condition (CU) was 17.27 (SD = 9.03). Planned pairwise comparisons, Bonferroni corrected, showed that only the automatic orchestration condition contributed to the observed difference in effect of orchestration, $p < .05$.

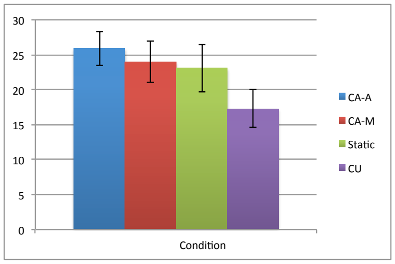
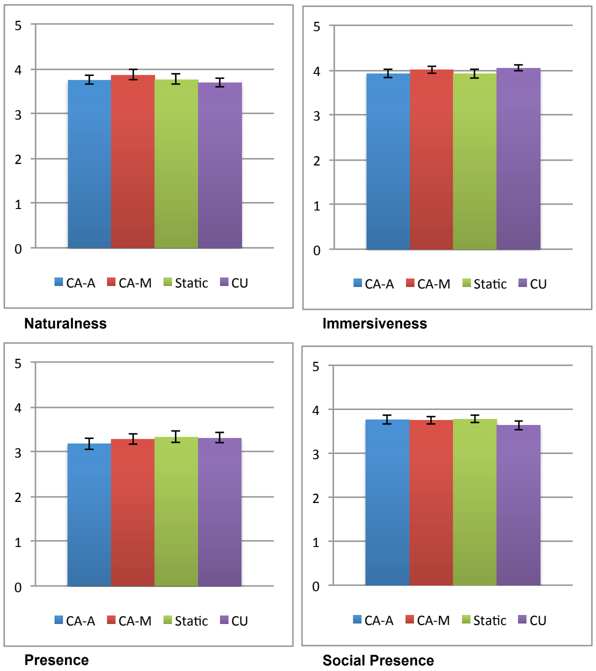


Figure 5. Objective evaluation: Average game points won per condition: Context Aware Automatic (CA-A) orchestration, Context Aware Manual (CA-M) orchestration, Static and Context Unaware (CU) mixing. The corresponding standard error is represented as error bars.

The questionnaire proposed in [11] was used for the subjective evaluation of the mediated communication across the four selected conditions. It targeted four factors describing mediated interaction: *naturalness*, *immersiveness*, *presence* and *social presence*. The average assessment for each factor is shown in Figure 6 by experimental condition. There were no significant differences measured in these four factors ($p > .05$), which means that all the four conditions were more or less equivalent with regards to the specific subjective evaluation.

The average number and duration of 'turn shifts' was also measured for each condition (see Figure 7), with a view to

[1] http://www.jboss.org/drools



Naturalness

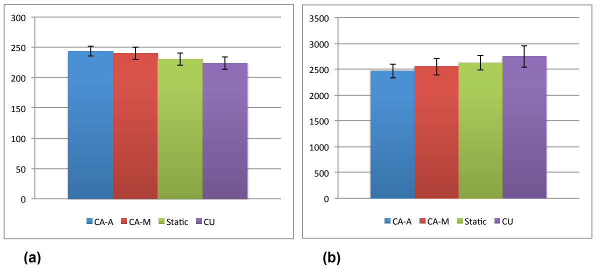
Immersiveness

Presence

Social Presence

Figure 6. Average participant experience per condition.

providing further data for the evaluation of the quality of the communication experience. The two are related, but not inferable from each other, as the state where no participant holds the turn is possible. These results were log-transformed and entered in a repeated analysis of variance. There were no differences in turn shifts across conditions (p >.05), see Figure 7a. Figure 7b shows that turn durations were also not affected by the different conditions (p >.05); no significant pairwise differences were observed in the duration of turns.



(a)

(b)

Figure 7. (a) Average number of turn shifts per condition. (b) Average duration of turns per condition.

DISCUSSION

Recall that the aim of the current experiment was to investigate the effect of *context-awareness and dynamic configurability*, defined as *orchestration*, on the quality of video-mediated communication, using an objective measure of task performance and a subjective assessment of the experience. The hypothesis was that orchestration can improve the ability of participants to communicate and collaborate effectively and that this would be reflected through both measures, objective and subjective.

The results did confirm, albeit marginally, the expectations with regards to the objective measure, as participants were indeed more efficient when collaborating in the two context-aware conditions. This can probably be attributed to the availability of more relevant information for each side of the group from the other, as provided through orchestrated mediation. The results overall, however, are more suggestive than conclusive. Yet a conclusive result which indicates the potential of automatic orchestration to improve live mediated communication did emerge from the fact that *statistically significant* (p <.05) differences were observed *only* between the automatically orchestrated condition (highest in terms of points won) and the context unaware mixing (CU) condition (lowest in terms of points won); neither the static condition nor, surprisingly, the manually orchestrated condition showed statistically significant differences with respect to the CU condition.

In order to further investigate the positive effect of automatic orchestration on task efficiency and to attribute this effect or correlate it with other factors, two other objective measures were considered: the average number of conversation turns and their average duration per condition. As shown in Figure 7, there is a decrease in number of turn shifts from automatic, through manual, static and down to context unaware and a converse increase in the average duration of the turns. They could be interpreted as indicating a decrease in the degree of how 'animated' the conversation was. Given the chosen communication setup – guessing within a short interval of time – there is a reasonable case to make that the best communication experience, from this point of view, is the most animated one, namely that had in the automatically orchestrated condition. Yet, this, too, is suggestive, not conclusive, but it backs up the same pattern observed with regards to game points won (Figure 5).

There is also an indication of automatic orchestration outperforming manual orchestration. When compared with each other, there are no significant differences between them, neither with regards to task efficiency nor to number of turns. However, the better experience in the automatic condition is suggested by the fact that the manual orchestration condition did not show statistically significant differences when compared with the CU condition (p >.05), whereas the automatic one did. On a first look, this appears counterintuitive, as manual operation in such a setup could be expected to establish some sort of 'ground truth' hard to reach through automatic operation; humans are expected to have a much better understanding of the context of communication than what could be implemented in automatic procedures. Yet, the opposite conclusion is suggested here. To further explore this aspect, we measured the average number of mixing decisions applied by the two manual operators, per communication session (Figure 8).

There are very significant differences in the numbers of decisions per operator, t(8) = 9.60, p <.001: in room 1 on average 93 decisions were applied (SD = 10.08), whereas in room 2 on average 147 decisions were applied (SD = 16.76). The results also show that one operator (for room 1) is more constant in

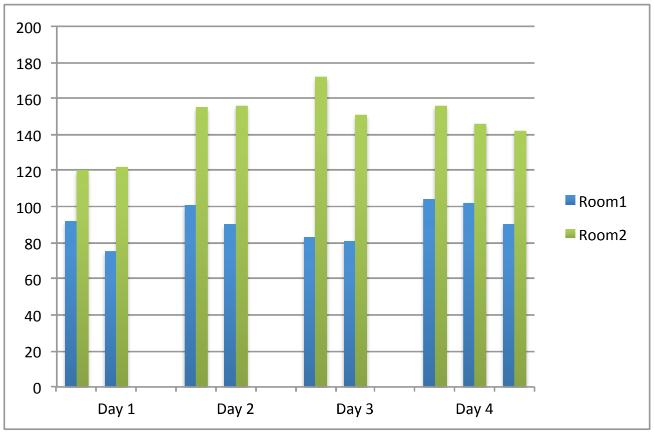


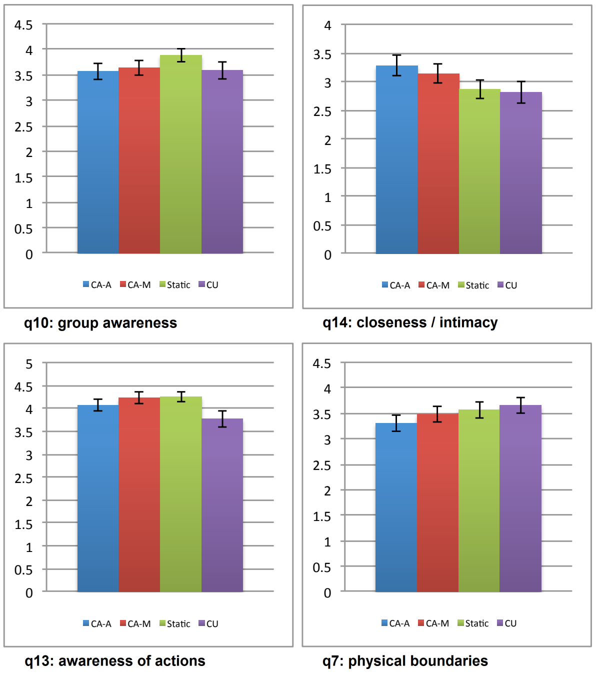
Figure 8. Number of mixing decisions per human editor per session grouped by experiment day.

number of decisions made across the four days, whereas the other one fluctuates quite considerably. Also, the number of decisions, per operator per day, decreases from one session to the other for all 4 days, for one of the operator, and for the last 2 days, for the other. This reduction in responsiveness may be attributed to fatigue.

These aspects clearly indicate that different human operators, and even the same operator at different moments in time, will interpret the same body of rules expressed in natural language differently, even when it is as simple as the one used here. Most probably, such differences would increase with the increase in complexity of the rule set. Any experimental evaluation of orchestration knowledge through manual operation must account for this aspect (e.g., train the operators in the rule set, average manual operation across a representative sample of operators, etc.) and cannot assume that manual operation provides the 'ground truth' of the experience. The implications are even more significant, as the whole process of orchestration knowledge elicitation and implementation is affected, giving rise to questions such as: *whose or which interpretation of a rule set expressed in natural language is being implemented*? and *how is a particular implementation validated*? This uncovers an added degree of complexity of an already rather complex experimental space.

The subjective evaluation of the communication experience was inconclusive (Figure 6).

The discrepancy between the objective evaluation and the subjective one provokes the thought that the questionnaire, despite being validated in a related application area [11], could somehow 'hide' more subtle but meaningful differences, possibly due to the aggregation of scores in different questions or participants going into narrative mode reflecting on their performance instead of actually recounting their experience. Trying to unpick this aspect, we considered each individual question separately. Four of the ones that showed significant differences between conditions are presented in Figure 9.



q10: group awareness

q14: closeness / intimacy

q13: awareness of actions

q7: physical boundaries

Figure 9. Responses to indicative questions.

For q10 – 'I felt equally aware of people in the other room as of people in my own room' – the static condition was evaluated higher than the other conditions. This can potentially indicate its appropriateness for better conveying *group awareness*.

For q14 – 'People in the other room felt more distant than people in my own room (transposed)' – automatic orchestration was evaluated higher than static and context unaware mixing, with manual orchestration closely following. This can potentially indicate its appropriateness for better conveying *closeness* and *intimacy*.

For q13 – 'I was aware of what the other participants were doing' – context unaware mixing performed worse than the other three (static and manual being top). This can potentially indicate that the *ability to keep track of a task* requires context awareness or an appropriate static condition.

For q7 – 'I felt like the boundaries between my and the other room disappeared' – automatic orchestration performed worse than the other three conditions. It is hard to find an explanation as to why *spatial awareness* was affected in this way.

These interpretations, save q7, are aligned with some of the conclusions presented in [7] and represent hypotheses for further research. They also indicate that other questionnaires than [11] might work better for the evaluation of orchestrated video mediated communication.

With regards to *manual* orchestration, the results of this study are not as strong as those reported in [21], as, here, manual orchestration does not show any statistically significant differences compared to the static and context unaware mixing conditions. There is no disagreement. This shows how relevant the details of the communication setup can be to mediated communication experiences. In all the three experiments (two reported in [21] and one here), the generic setting was that of a group of four friends, two in each living room, playing social games across a live video connection through the room's TV screens. However, in [21] the social game was Pictionary, and so the overall interaction had stronger visual requirements than that around Articulate, which was the game of the study here. The number of shots available for orchestration in the experiments reported in [21] was greater than those employed in this study and the orchestration logics employed in [21] were more complex that the logic used here. This illustrates the complexity of the orchestrated experience space, which presents various co-ordinates such as: (social) activity undertaken between people, number and type of cameras, their placement, the shots available for orchestration, the orchestration logic, the number of people and locations, and others. Any variation along these co-ordinates generates a new setup that needs to be investigated. Finding ways to evaluate orchestration in larger, more generic setups is an extra challenge.

In the communication setup considered here, the differences between the four experimental conditions were not substantial. One view may be that the measures employed were insufficiently sensitive to pick up existing differences. Another view may be that the communication setup itself (2 rooms, 2 people per room in fixed positions, relatively short communication session, high quality audio and video, wide shot in static condition captures all the necessary details, simple orchestration logic, etc.) was simple enough to be served well by all the four methods of mediation. In more complex setups, for example involving more locations, more people, and possibly different activities, it can be hypothesised that more sizeable differences might arise. The automatically orchestrated condition, conceptually, appears to scale better than both the static and the context-unaware orchestrated conditions. For a three room setup, for example, the level of detail provided by the static condition is halved compared to the two room setup – each room has to place two static shots, as opposed to one, on its screen. For the same setup, considering that each room provides three shots, similarly to the experiment reported here, the probability of showing a meaningful shot in the context unaware condition is more or less half of the same probability in the two room setup – a simplistic calculation indicates that 2 out of 6 shots will always include the active speaker in the three room setup, whereas 2 out of only 3 shots will include the active speaker in the two room setup. However, the *informed* choice of 1 shot out of a possible 6, in the three room setup, is hypothesized to work more or less similarly to the informed choice of 1 shot out of a possible 3, as in the two room setup, as 'following' the active speaker within a group of 4 is similar to 'following' the active speaker within a group of 6. Context awareness is thus hypothesized to provide more significant improvements to the communication experience in more complex setups compared to the two context unaware possibilities.

CONCLUSIONS

Live video mediation is penetrating the social sphere at a fast pace. Currently, all there is available for group social communication is offered within the paradigm of 'talking heads', which serves very well simple communication setups. This paper was motivated by richer, more complex setups of social communication, which were assumed to require multiple cameras to capture all the necessary details and, consequently, dynamic mixing of the available live content. The decision making process that directs the cameras and controls the mixing process was denoted as orchestration.

The paper presented an experimental enquiry into a completely automatic orchestrated live video communication system. The evaluation was carried out within a generic setup, namely that of a group of friends (4), located in 2 living rooms and playing social games (Articulate). The overall principle underlying the specific orchestration logic employed was to follow the conversation flow and to keep the focus of the representation on the person most actively involved in the conversation. This was implemented via a relatively simple rule set.

Automatic orchestration was found to enhance task efficiency and to better facilitate the communication style required by the social activity undertaken: animated conversation with short turns and quick exchanges of turns. This was attributed to the provision of more relevant information through automatically orchestrated mediation than through the other methods of mediation.

In an attempt to attribute the observed difference to and correlate it with other factors, additional objective and subjective instruments were employed. However, they led to no conclusive results. Further differences and trends were observed, but they lacked statistical significance. The objective measures of number of conversation turns and their average duration backed up the trend observed with regards to task efficiency, but not with statistical significance. The questionnaire used for the subjective evaluation of the communication experience through four factors (naturalness, immersiveness, presence and social presence) [11] did not lead to any significant differences either between the four experimental conditions. However, by considering individual questions, it showed that differences did exist, with orchestration appearing better at conveying closeness and intimacy, and the static condition appearing better at maintaining group awareness.

On one hand, these results could suggest that the measures employed were insufficiently sensitive to pick up existing differences in the communication experiences. On the other hand, they could also suggest that the investigated communication setup was rather simple and therefore well served by all the four methods of mediation and not conducive to more significant differences. More complex setups, it can therefore be hypothesised, could impose stronger requirements for context-aware mediation, i.e. orchestration. This hypothesis is backed up by the observation that, conceptually, context-unaware mediation (e.g. static and context unaware mixing) does not scale up for more complex setups as well as orchestrated mediation. We suggest both perspectives as avenues for further research required in the ever-growing landscape of social live video-mediated communication.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements no. ICT-2011-287760 and ICT-2007-214793.

REFERENCES

1. Chen, L., Harper, M., Franklin, A., Rose, T. R., Kimbara, I., Huang, Z., and Quek, F. A multimodal analysis of floor control in meetings. In *Proceedings of MLMI'06*, Springer (2006), 36–49.
2. Eyben, F., Weninger, F., Gross, F., and Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of ACM MM '13*, ACM (2013), 835–838.
3. Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of ACM MM '10*, ACM (2010), 1459–1462.
4. Falelakis, M., Kaiser, R., Weiss, W., and Ursu, M. Reasoning for video-mediated group communication. In *Proceedings of IEEE ICME'11* (2011), 1–4.
5. Fraunhofer. Fraunhofer IIS audio communication engine raising the bar in communication quality. In *White Paper* (2012).
6. Gatica-Perez, D. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Comput. 27*, 12 (2009), 1775–1787.
7. Inoue, T., Okada, K.-i., and Matsushita, Y. Learning from tv programs: Application of tv presentation to a videoconferencing system. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*, UIST '95, ACM (1995), 147–154.
8. Jansen, J., Cesar, P., Bulterman, D. C. A., Stevens, T., Kegel, I., and Issing, J. Enabling composition-based video-conferencing for the home. *Multimedia IEEE Transactions 13*, 5 (2011).
9. Kaiser, R., Weiss, W., Falelakis, M., Michalakopoulos, S., and Ursu, M. A rule-based virtual director enhancing group communication. In *Proceedings of IEEE ICME'12 Workshops* (2012), 187–192.
10. Kirk, D. S., Sellen, A., and Cao, X. Home video communication: Mediating 'closeness'. In *Proceedings of CSCW'10*, ACM (2010), 135–144.
11. Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. A cross-media presence questionnaire: The itc-sense of presence inventory. *Presence: Teleoper. Virtual Environ. 10*, 3 (2001), 282–297.
12. Liu, Z., Cohen, M., Bhatnagar, D., Cutler, R., and Zhang, Z. Head-size equalization for improved visual perception in video conferencing. *Multimedia, IEEE Transactions on 9*, 7 (2007), 1520–1527.
13. Motlicek, P., Duffner, S., Korchagin, D., Bourlard, H., Scheffler, C., Odobez, J.-M., Galdo, G. D., Kallinger, M., and Thiergart, O. Real-time audio-visual analysis for multiperson videoconferencing. *Adv. MultiMedia* (2013), 4:4–4:4.
14. Neustaedter, C., Oduor, E., Venolia, G., and Judge, T. K. Moving beyond talking heads to shared experiences: The future of personal video communication (workshop). In *Proceedings of ACM GROUP'12*, ACM (2012), 327–330.
15. Nguyen, D. T., and Canny, J. Multiview: Improving trust in group video conferencing through spatial faithfulness. In *Proceedings of CHI'07*, ACM (2007), 1465–1474.
16. Nguyen, D. T., and Canny, J. More than face-to-face: Empathy effects of video framing. In *Proceedings of CHI'09*, ACM (2009), 423–432.
17. Ranjan, A., Birnholtz, J., and Balakrishnan, R. Improving meeting capture by applying television production principles with audio and motion detection. In *Proceedings of CHI'08*, CHI'08, ACM (2008), 227–236.
18. Ronfard, R. A Review of Film Editing Techniques for Digital Games. In *Workshop on Intelligent Cinematography and Editing*, R. M. Y. Arnav Jhala, Ed., ACM (Raleigh, United States, 2012).
19. Sacks, H., Schegloff, E., and Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language 50*, 4, Part 1 (1974), 696–735.
20. Sidnell, J. *Conversation Analysis: An Introduction*. Language in Society. Wiley, 2009.
21. Ursu, M. F., Groen, M., Falelakis, M., Frantzis, M., Zsombori, V., and Kaiser, R. Orchestration: Tv-like mixing grammars applied to video-communication for social groups. In *Proceedings of ACM MM'13*, ACM (2013), 333–342.
22. van der Kleij, R., Schraagen, J. M., Werkhoven, P., and De Dreu, C. K. W. How conversations change over time in face-to-face and video-mediated communication. *Small Group Research 40*, 4 (2009), 355–381.
23. Vinciarelli, A., and further authors. Open challenges in modelling, analysis and synthesis of human behaviour in human-human and humanmachine interactions. *Cognitive Computation* (2015), 1–17.
24. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing, IEEE Transactions on 3*, 1 (2012), 69–87.
25. Yannopoulos, A. Directornotation: Artistic and technological system for professional film directing. *J. Comput. Cult. Herit. 6*, 1 (2013), 2:1–2:34.