Sonification for Supporting Joint Attention in Dyadic Augmented Reality-based Cooperation

Thomas Hermann
Ambient Intelligence Group
Bielefeld University
Bielefeld, Germany
thermann@techfak.uni-bielefeld.de

Alexander Neumann
Ambient Intelligence Group
Bielefeld University
Bielefeld, Germany
alneuman@techfak.uni-bielefeld.de

Christian Schnier
Interactional Linguistics & HRI
Bielefeld University
Bielefeld, Germany
cschnier@techfak.uni-bielefeld.de

Karola Pitsch
Interactional Linguistics & HRI
Bielefeld University
Bielefeld, Germany
karola.pitsch@uni-bielefeld.de

AM 2013, '13, September 18–20 2013, Pitea, Sweden

Abstract

This paper presents a short evaluation of auditory representations for object interactions as support for cooperating users of an Augmented Reality(AR) system. Particularly head-mounted AR displays limit the field of view and thus cause users to miss relevant activities of their interaction partner, such as object interactions or deictic references that normally would be effective to establish joint attention. We start from an analysis of the differences between face-to-face interaction and interaction via the AR system, using interaction linguistic conversation analysis. From that we derive a set of features that are relevant for interaction partners to co-ordinate their activities. We then present five different interactive sonifications which make object manipulations of interaction partners audible by sonification that convey information about the kind of activity.

Keywords

sonification, auditory display, mediated communication, assistive technology, social interaction

ACM Classification Keywords

H.5.2 User Interfaces Auditory (non-speech) feedback

Introduction

In natural human-human interaction, we have many communicative resources at our disposal to coordinate joint activity, such as speech, gaze, gestures or head movements. Their interplay allows us to establish and sustain joint attention when needed, such as in collaborative planning tasks. We deal with the latter in an interdisciplinary project between linguistics and computer science where we aim at better understanding the principles of successful communication[1]. As our method, we have introduced and developed an Augmented Reality (AR) system that enables us to '(de-)couple' two users engaging into co-present interaction for a collaborative planning task. The AR system allows us to precisely record what the interaction partners see at any moment in time – and thus to understand on basis of what information they select their next action. Besides this visual interception of visual cues, we extended the system to also enable an auditory interception by using microphones and in-ear headphones.

We have proposed and introduced various new sonic enhancement methods in [3] to increase the users' awareness of their interaction partner. In this paper, we take the next step and evaluate the approaches at hand of a user study with test listeners. One particular aim of this work is to better understand the principles of how sound can be successfully used, and what sounds are accepted.

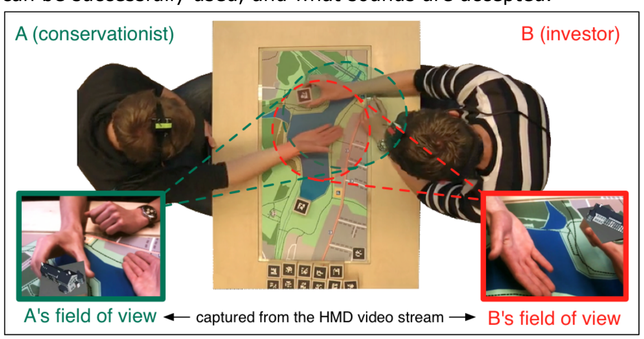

Figure 1: Participants argue about a fictional recreational area project. The markers on top of the wooden cubes are augmented with possible buildings.

[1] www.sfb673.org/projects/C5

Alignment in AR-based Cooperation

In the Collaborative Research Center 673 *Alignment in Communication* we combine proven communication research methods with new interdisciplinary approaches to get a better understanding of what makes communication successful and to gather insights into how to improve human-computer interaction. The project *Alignment in AR-based cooperation* uses emerging Augmented Reality technologies as a method to investigate communication patterns and phenomena. In experiments we ask users to solve tasks collaboratively, using an Augmented Reality based Interception Interface (*ARbInI*) which consists of several sensors and displays and allows us to record and alter the perceived audiovisual signals of a system's users in real-time. This feature allows us to monitor, control and manipulate the visual information available to both users separately during the negotiation process at every moment during the experiment [1].

The participants are seated at a table with a map on a fictional recreational are, equipped with wooden cubes with attached markers representing symbolic representations of possible attractions or construction projects as shown in Figure 1. The system detects the marker and augments a virtual representation into the participants' video stream.

For data analysis we combine the benefits of machine-driven quantitative data mining approaches with qualitative conversation analysis in a mutual hypothesis generation- and validation loop.

Mutual Monitoring in face-to-face and Augmented Reality-based interaction

In natural face-to-face interaction, participants rely on the possibility of mutual monitoring and on-line analysis of the co-participant's actions (speech, bodily conduct, gesture etc.) which enables them to adjust their ongoing actions on a fine-grained level to each other and to micro-coordination. By mutually monitoring each other's behavior they are able to interpret interactional goal-directed actions in situ and make use of the underlying projections of each other's conduct. This process enables interlocutors to anticipate certain relevant next actions. By using in-depth conversation analytical methods our interest focused on one particular aspect of the interactional organization in face-to-face (f2f) and AR-based cooperation:

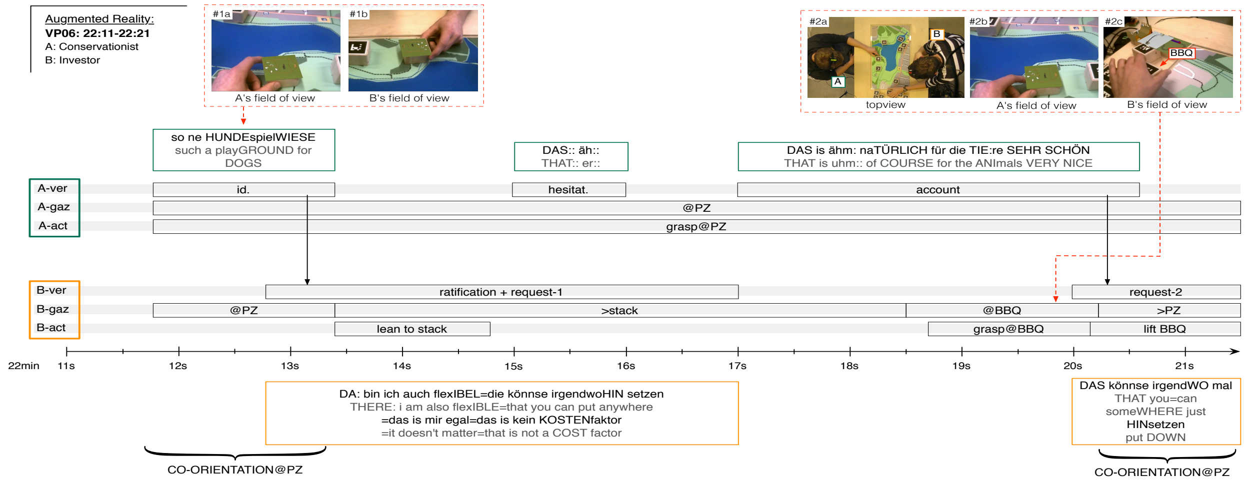

Figure 2: Lack of Mutual Monitoring in AR-based interaction

How do mutual monitoring or a lack of it influences the interactional organization in f2f and AR? While our analytical results in our f2f condition could reveal that interlocutors reciprocally adapt their behavior to each other in order to prevent simultaneous action and ensure the sequential organization of their activities, our AR-based dyads reveal a contrasting organization in cases where simultaneous activities emerge.

The lack of Mutual Monitoring in AR-based interaction

Let's consider a fragment from our AR-based dyads. The fragment's annotation and translation of the german text can be found in Fig. 2. At the fragment's beginning, A suggests the object Petting Zoo (PZ; here defined as "playground for dogs"). He grasps the object "PZ", identifies it as "so ne HUNDEspielWIESE" and orients to it (cf. 1a). Meanwhile, B follows A's action (cf. 1b). Comparing both participants' field of view (1a 1b), it is recognizable that they have a common focus of attention. This common focus of attention is different from joint attention sequences of our natural f2f condition: Both interlocutors haven's a profound knowledge about the co-participant's orientation. They assume joint attention, but due to the lack of mutual monitoring they can't be sure that each other's co-participant attends to the same location. For this reason, we want to term those sequences in AR as "co-orientation" in order to distinguish it from "joint attention". After co-orientation at the object "PZ" is established, B reacts to A's suggested object by a direct ratification, which includes a request to place the object ("DA: bin ich auch flexIBEL=die könnse..."). As he simultaneously shifts his gaze to the stack and transforms his posture by leaning forward to it, it is recognizable that the current interactional task "PZ" is finished for him at this point in time.

Due to the lack of mutual monitoring, B's shifting orientation (body + gaze) can't be used as a relevant semiotic signal by A. He continues the task "PZ" (cf. 2b) by giving the account "DAS is ähm: naTÜRLICH für die TIE:re SEHR SCHÖN", while B starts preparing a new interactional task: He orients to the object "Barbecue" (BBQ) and grasps it out of the stack (cf. 2c). Considering 2a we can recognize that both participants are working on different tasks during this time. In contrast to our observations in the f2f condition, participant A has no possibility to react to the emerging simultaneous task-preparation, introduced by participant B, as he is not aware of it. Shortly afterwards B lifts the object, carries it over the map, re-orients to A's grasped "PZ" and formulates the second request "DAS könnse irgendWO mal HINsetzen". Here, co-orientation is established again. But accordingly to the fragment's beginning, they have no profound knowledge about the co-participant's attention.

Comparative results

Mutual Monitoring-based procedures enable interlocutors to prevent emerging parallel activities. This ensures the sequential organization of their activities. However, the lack of Mutual Monitoring in AR leads in cases where simultaneous activities emerge to the impossibility to instantly solve parallel activities in situ. A time window to repair emerging parallel activities is short: In fact, seconds after the end of fragment 2, B's prepared object BBQ appears in A's field of view. A reacts to it by shifting his gaze to the object, but continues in his current task – the placement and account of PZ.

Non-Visual Guidance of Attention

In everyday interaction sound is an important cue to catch and orient our focus of attention, as for instance exemplified by situations where we hear our name being called from somewhere, or a sudden explosion or a car approaching on the street. However, there are also many situations where not a sudden event, but (even only a subtle) change of sound draws our attention, as for instance when driving a car and suddenly hear a change of the engine sound.

Sonification enables to profit from our auditory information processing – which operates largely in parallel and independent of our primary task – for interactional situations. An earlier system of this project made use of head gesture sonifications such as nodding and shaking the head: as the head-mounted displays allow either to look on the desk or to look to the interaction partner, but not simultaneously, the sonification of head gestures conveys analogic and subtle information to support interaction [2]. Furthermore, enhancing and augmenting object sounds with informative or aesthetic acoustic additions is a well established approach in Sonic Interaction Design [4], yet so far rarely considered for collaborative applications. More details about the sonification of object interactions for supporting dyadic interaction have been presented in [3].

Based on this, we developed a set of sonification methods, that not only imitate (and exaggerate) natural physical interactions, but allow also to associate sounds to normally silent actions such as carrying objects through air. From these methods we selected five for the following study, and they will be explained in the following section.

Sonification Designs

We are mainly interested in the object interactions (a) to move (shift/rotate) it on the desk, (b) to pick/lift an object, (c) to carry it to a different location through air, and finally (d) to place it on the desk. Such interactions are ubiquitous in our scenario and are partly accompanied naturally with interaction sounds (in our scenario: of wooden objects touching our glass table), specifically only (a), (b) and (d). Some actual interactions are silent (e.g. c), and many interaction go unnoticed as they can and are often rather silently executed. So the artificial sonification of all the interaction types will more reliably make the interaction partners aware of activities.

As for the data to practically implement our sonifications we use ARtoolkit tracking data captured from a camera mounted and looking downwards from the ceiling. The derivation of 'high-level' features that correspond to our interaction classes (a–d) is a complex computational process which is beyond the scope of this paper, but works reliably enough to provide the basis for the sonifications. The feature extraction results in either continuous features such as the current velocity, position or rotation of an object, or discrete events such as lifting or putting objects. With these tracking data, we implemented five sonifications.

For **Direct Parameter-Mapping** we turn the multivariate times series of features into sound. We use time-variant oscillators with frequency and amplitude parameters and map the vertical height of an object above the table to frequency, following the dominant polarity association [5]. The frequency range is 100Hz to 300Hz using sine tones without higher harmonics, so that the resulting sound is both rather quiet and has limited interference with the concurrent verbal engagement of the users.

The focus for the **Abstract signals** design was on clear and distinguishable abstract sounds. Lifting an object is represented by a short up-chirped tone, putting it down by its counterpart down-chirped tone. Pushing an object on the desk surface is sonified by pink noise that decays smoothly after the action stops, similar to pushing it through sand. Carrying an object above the surface leads to low-pass filtered white noise, again with smoothly decaying level as the action stops, representing wind sounds done by fast movement.

To examine how obtrusiveness sounds cause problems or disturb ongoing interaction, we created a design based on **Exaggerated Samples**: A high pitched blings for lift, crashing windows for put, creaking for pushing an object and a helicopter for carrying, in order to render the actions very salient.

Assuming that **Naturalistic Imitations** will be most easily understood, we created a sonification that uses the familiar sound bindings as true as possible. However, our sonification is different from what would be obtained by attaching a contact microphone to the table and amplifying the real sound signals in (a) that even silently executed actions (such as putting an object on the table) here leads to a clearly audible put-sound, and (b) that we here gain the conceptual ability to refine the sounds (as parameterized auditory icons) dependent on actions and circumstances we regard as important. The samples used have been recorded using a microphone and the same wooden objects that are used in the AR scenario.

Finally, we selected **Object-specific sonic symbols** corresponding to the model being shown on top of our objects. For instance while manipulating the ‘playground’ placeholder object, a sample recorded on a playground is played. Likewise for the petting zoo, animal sounds evoke the correct association. Technically, sample playback is activated whenever (but only if) an object is moved around, ignoring the object’s height above the desk. The sound is furthermore enriched by mapping movement speed to amplitude and azimuthal position to stereo panning, creating a coarse sense of directional cues.

Evaluation

To examine how the sonifications are understood by listeners and how they might affect interaction, we first conducted a preliminary study, asking subjects to rate the different sonifications at hand of a given interaction example according to a number of given statements.

Study Design

We prepared a short video clip of an interaction and augmented it with the sonification approaches explained before. The resulting five audio-visual stimuli are randomized for each participant in this within-subject design and were presented as often as wanted by the participants. Participants filled out a questionnaire containing statements and questions, and a 7-point Likert scale ranging from 1 (’false’) to 7 (’true’) (resp. ’no’ to ’yes’) [2]. We also collected basic data such as age, sex and profession as well as information about experience with computers and musical instruments and possible issues related to sound awareness.

Results

10 participants (6 female + 4 male), all right-handed, average age 26.3, age range 20–29, mostly students (expect one teacher and one therapist) participated to the study which lasted typically 20–35 minutes. Since no significant findings could be derived from the data we summarize observed tendencies.

In result, all sonifications allow to follow the dialogue. The naturalistic sounds cause the least incompatibility – we assume that is because we are used to such sounds in natural interaction which are also most easily subconsciously accepted. In contrast both object-specific and exaggerated sounds demand more attention. Additionally, naturalistic and abstracts sounds were rated to cover the conversation the least.

As expected, the naturalistic sounds are the least obtrusive, least disturbing, least irritating and least distracting. This may also be for the reason that in this sonification, there are less sounds played in total: carrying an object in air is silent and thus not represented by sound. An unexpected counterpoint is the very obvious bad evaluation of the OS method: this is most distracting, irritating, disturbing and obtrusive. The other methods are rated in between these extremes and particularly we find that the AS receives rather good ratings, often nearby NI, yet superior in terms of information, comprehensibility and ’well-soundingness’.

Certainly, participants can only vaguely extrapolate from their short experience. Results show that AS is best to get used to – but only a little better than NI. Particularly ES and OS are weaker concern-

[2]videos and the questionnaire can be found at http://www.techfak.uni-bielefeld.de/ags/ami/publications/HNSP2013-SFS/

ing long-term compatibility. It seems that AS were best understood in terms of what the meaning of the sound is, and thus the sound rather explain themselves instead of requiring a learning-by heart to interpret the meaning.

Discussion

The results of our study show tendencies on the basis of 10 subjects rating statements. Obviously, there is a rather high variance in the scores, and with only 10 subjects unfortunately t-test p values are not low enough. Yet the purpose of our study is to get guidance for our next design cycle iteration towards sonification candidates to be deployed into the running dyadic AR system.

From what we see we infer that comprehension, i.e. to understand what the sounds mean, is affecting acceptability and other judgements such as perceived obtrusiveness, pleasantness, irritation, distraction, etc. Furthermore, the subjects saw only a very simple situation where only a single object is manipulated. Characteristics and user acceptance of the chosen sounds could be evaluated without overlap. However, a usability study requires at least two 'active' objects where also object identification is required since interaction-critical situations might involve both subjects manipulating an object at the same time as seen in Section .

Generally, we were a bit surprised to see the AS sonification to work so well – having expected that the NI would perform best in most questions. This is a relevant guidance for us to experiment in future designs with a blend between abstract and naturalistic sonifications, in search of a sweet spot. We believe that parameterized auditory icons, starting from naturalistic sounds are the ideal starting point for that.

We are careful to not over-generalize the results towards how the sonifications would be perceived by users in the AR-setting. However, by using conversation analysis, we have a solid method to investigate this and to detect even subtle effects in sound-enhanced interaction – and this is our next step, once the sonification has been optimized and implemented for the running AR-system.

Conclusion

We have presented a sonification system to support joint attention in dyadic augmented reality-based cooperation. We derived the need for enhancing mutual monitoring between interacting users by a comparison of face-to-face vs. augmented-reality-mediated interaction using conversation analysis. From that we identified the problems that arise from lack of mutual monitoring. Five selected sonifications were compared in various characteristics in a within-subject experiment with 10 persons. The aim was to check how the sonifications would generally be accepted by users, and to extract from the feedback some guidance on how to proceed in our sound design.

In summary, the abstract sonification was unexpectedly well perceived and rated, and we conclude that a blend between naturalistic and abstract sonification, using parameterized auditory icons will be a good next design step. In our ongoing work we will implement several sonifications into the AR-system for testing in interaction.

Acknowledgments. This work has partially been supported by the Collaborative Research Center (SFB) 673 Alignment in Communication and the Center of Excellence for Cognitive Interaction Technology (CITEC). Both are funded by the German Research Foundation (DFG). Karola Pitsch also acknowledges the financial support from the Volkswagen Stiftung.

References

[1] A. Dierker, C. Mertes, T. Hermann, M. Hanheide, and G. Sagerer. Mediated attention with multimodal augmented reality. *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*, page 245, 2009.

[2] T. Hermann, A. Neumann, and S. Zehe. *Head gesture sonification for supporting social interaction*, pages 82–89. ACM Press, 2012.

[3] A. Neumann and T. Hermann. Interactive sonification of collaborative ar-based planning tasks for enhancing joint attention. Manuscript submitted for publication, 2013.

[4] S. Serafin, K. Franinović, T. Hermann, G. Lemaitre, M. Rinott, and D. Rocchesso. Sonic interaction design. In T. Hermann, A. Hunt, and J. G. Neuhoff, editors, *The Sonification Handbook*, chapter 5, pages 87–110. Logos Publishing House, Berlin, Germany, 2011.

[5] B. N. Walker and G. Kramer. Mappings and metaphors in auditory displays. *ACM Transactions on Applied Perception*, 2(4):407–412, Oct. 2005.