INTERACTIVE SONIFICATION OF COLLABORATIVE AR-BASED PLANNING TASKS FOR ENHANCING JOINT ATTENTION

Alexander Neumann

Ambient Intelligence Group
CITEC, Bielefeld University
Universitätsstraße 25, D-33615 Germany
alneuman@cit-ec.uni-bielefeld.de

Thomas Hermann

Ambient Intelligence Group
CITEC, Bielefeld University
Universitätsstraße 25, D-33615 Germany
thermann@techfak.uni-bielefeld.de

ABSTRACT

This paper introduces a novel sonification-based interaction support for cooperating users in an Augmented Reality setting. When using head-mounted AR displays, the field of view is limited which causes users to miss important activities such as object interactions or deictic references of their interaction partner to (re-)establish joint attention. We introduce an interactive sonification which makes object manipulations of both interaction partners mutually transparent by sounds that convey information about the kind of activity, and can optionally even identify the object itself. In this paper we focus on the sonification method, interaction design and sound design, and we furthermore render the sonification both from sensor data (e.g. object tracking) and manual annotations. As a spin-off of our approach we propose this method further for the enhancement of interaction observation, data analysis, and multimodal annotation in interactional linguistics and conversation analysis.

1. INTRODUCTION

In co-present human-human interaction, interaction partners have many communicative resources at their disposal to coordinate their joint activity, such as gaze, deictic gestures, speech or head gestures. In collaborative planning tasks these are accessed to establish and sustain joint attention. In the context of an interdisciplinary project between linguistics and computer science we strive at better understanding the underlying factors for these processes to organize[1]. For that we have developed over the past years an experimental setup that uses Augmented Reality (AR) to decouple two users interacting co-presently at a table in a cooperative task of planning a recreational area. The users' task is to jointly position physical objects (e.g. representing a hotel, waterskiing area, or playground) on a map, with given roles to play (investor vs. conservationist). AR allows us to precisely record what the interaction partners see at any moment in time – and thus to understand on the basis of what information they understand the ongoing interaction, and why eventually they behave as they do to coordinate the activity. So AR is utilized to visually intercept the authentic visual cues. We extended this idea towards the auditory domain and can likewise intercept the *audible signals* by using microphones and in-ear headphones – which allows us to record and analyze exactly on basis of what sounds users coordinate their actions.

However, with our given setup, we can not only *intercept*, we can also *manipulate* the media in manifold ways, both by introducing *disturbances*, e.g. to see how they affect the coordination processes, and *enhancements*, to find out how future technical assistance systems can better support their users during cooperation.

This paper introduces a new enhancement method that we expect to influence strongly how the interaction partners become (and remain) aware of the interlocutor's actions. First, we motivate the idea at hand of a qualitative examination of how we monitor activities which occur out-of-sight by listening in real-world scenarios. We then transfer the findings to the specific cooperation setup where two users equipped with AR-gears jointly plan the positioning of items/objects on a map. For that we provide more information about the setting and the measured data in Sec. 5. Finally, we discuss different sound designs and their benefits and drawbacks on the basis of an interaction example that is augmented with the sonification.

Currently we have both manual annotations and camera/tracking-based sensor data for object manipulations. In this paper we discuss how they differ and what differences are actually relevant for establishing an awareness of object interactions. We are at the stage of optimizing the sonification for online use in preparation of a study, and plan to report first insights on practical use, and feedback from users at the ICAD conference.

As an interesting side-line we can already use our auditory display to solve a data review problem that interaction researchers frequently encounter in data sessions of multimodal data: in today's annotation tools the information is mostly presented visually, so for instance, the altitude of an object over the table would be visualized as a function plot in a timeline. To follow an interaction, users would typically look at the running video, and in order to connect this with the given sensor information they need to look back and forth between the two visual displays. Our sonifications allow to solve the problem differently by providing the information via an auditory channel, allowing an undivided visual focus on the video. This data inspection support for complex multimodal interaction data corpora is a secondary yet also profitable outcome. Let us start with a motivation for using sound in this application area.

2. SOUND-INDUCED INTERACTION AWARENESS

When we manipulate our physical environment, this often produces sounds which give us feedback about the detailed manipulations, the objects involved and eventually even about subtle variables such as our own emotional state [1].

Sound is a valuable medium to provide information that is not constrained to a single location [2] which makes interaction sounds

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

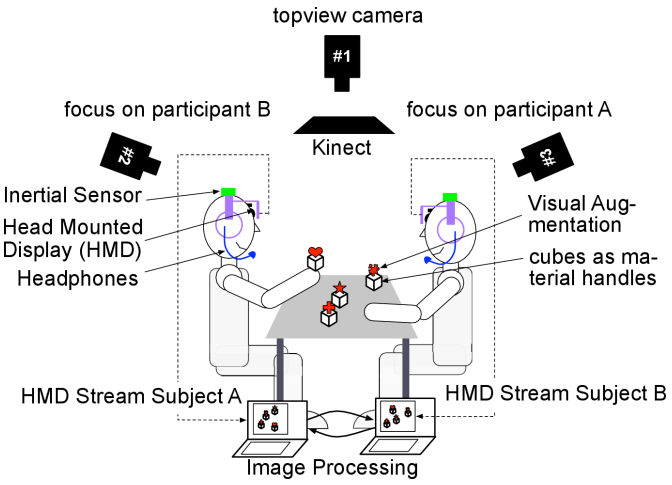

Figure 1: The schematic shows the setup and the hardware components that ARbInI aggregates. These include static components such as three DV cameras, a Microsoft Kinect and two to three workstations for data processing. Each participant wears a head-mounted display, a microphone headset and a BRIX motion sensor to measure head movement at high temporal resolution.

normally also accessible to others, and for them the sounds are an important information source to become and stay aware of activities in their environment. For example, a worker in an office could tell without even looking if her colleague is writing or not – thanks to the sound the keyboard emits while being typed on. Many details can be extracted from the sound signal, such as the writing speed, the error rate and perhaps even the urgency of the writing.

Sound also provides us information about the state of some delegated tasks which cannot be monitored visually. For instance when we start a water boiler we can hear when the water is boiling while we set the table next room. Or if we ask a person to set the table we expect to hear a certain sequence of impact sounds such as cupboard doors opening and closing, tinkling glasses, dishes and cutlery put on the table. The duration between cupboard opening and closing sounds could tell us how long it took the person to find the things he was looking for, the duration between to opening sounds gives us information about the average working speed, etc. Unexpected sounds like shattering implies that something went wrong and continued silence that the work is completed – or that it has been interrupted. Parents often use sound as a display for their children's activities out of their sight. Actually, it is the absence of steady noises that is used as an indicator that something might not be right and their offspring needs attention.

Since we normally process context sounds rather subconsciously and without effort, we tend to neglect their importance for staying connected – until they lack or other problems occur. Apparently sound is very effective in drawing our attention towards events outside our field of view, e.g. to become aware of somebody approaching from behind (e.g. from their footstep sounds), or of an alarm clock or mobile phone beeping on the table [3]. This is exactly the capacity which is useful in the case of AR-based cooperation, where typically the limited view angle of head-mounted displays shift most of our surrounding outside the visual field of view for most of the time. We argue that normal listeners are (both evolutionary and by learning) extremely tuned to understand how physical interactions manifest in sounds, and thus draw subconsciously conclusions about the source of a heard sound. Thus it makes sense – when aiming at augmenting object interactions with sonifications – to make use of these bindings.

However, beyond our acoustic reality we can also associate sounds to activities that are normally silent or inaudible such as moving an object through air or to embodied features such as body balance [4]. For these interactions we need to be more creative with sound designs and metaphors, designs need to be validated by empirical studies and listening tests.

3. ALIGNMENT IN AR-BASED COOPERATION

The Collaborative Research Center 673 *Alignment in Communication*[2] investigates the role of alignment and other communication patterns for successful communication. The goal is to gain new insights into how people communicate but also to find ways to improve human-computer interaction. In the subproject C5 *Alignment in AR-based collaboration* we use Augmented Reality (AR) as a technology for communication research which provides new features and methods for this discipline.

Within this context the *Augmented Reality based Interception Interface* (ARbInI) was developed and tested as a monitoring and assistance system in everyday dialogue scenarios [5]. The system allows a direct access to the audiovisual communication channels to monitor and alter information perceived by the users. Combined with other non-verbal communication cues such as gestures, posture and gaze direction these data form a complex multimodal data corpus.

3.1. ARbInI

Our system consists of several components which are either positioned around two chairs and a table, or worn by the users. All components are shown in Figure 1. The sensors attached to the users contain motion sensing devices from the BRIX toolkit which was developed in our working group [6] and headset microphones to record audio signals. The core component is a video-see-through head-mounted display (HMD) equipped with two Firewire cameras and a display for each eye. Monocular images captured by one of the Firewire camera are transferred to a computer and fed back without noticeable delay to the user via the displays. The HMDs also feature stereo vision but due to higher hardware demands and only little gain for the users in our current studies we decided to use a single video stream only. Three HD digital video cameras surround the participants, two of them are placed diagonally behind each participant and the third right above the table where also a Microsoft Kinect[3] is located. All data streams can be accessed, stored and manipulated in real-time except for the HD videos which we only record for later analysis.

3.2. Obersee II Scenario

For the study we have designed a recreation planning scenario which takes place in the surroundings of a lake called Obersee in the city of Bielefeld. The participants have to choose from two roles to elicit negotiation and a slight amount of competition: a financial investor who should focus on revenue by attracting tourists and a conservationist who wants to prevent serious damage to the

[2] www.sfb673.org
[3] www.xbox.com/en-US/kinect

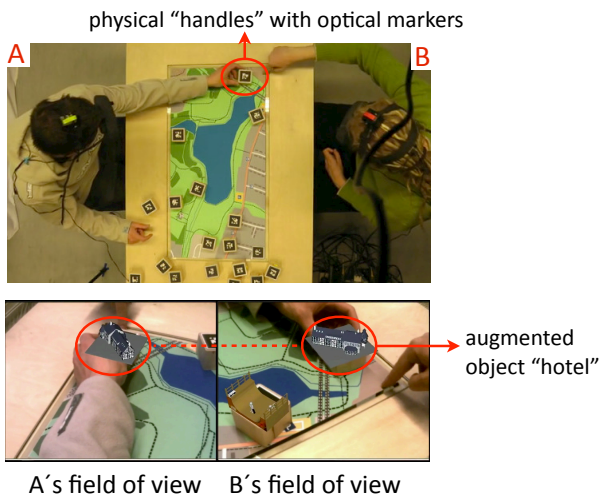

Figure 2: Obersee recreation scenario. In the ongoing study our participants collaborate to recreate a local lake and its surroundings. *ARbInI* monitors their actions. The markers on top of the wooden cubes are augmented with models representing concepts for possible projects (e.g. *hotel* or *skater park*)

surrounding nature. Both parties have to overcome their opposing goals and agree on a final result which should be presented after 20 minutes of negotiation. Figure 2 shows the setup from the top with the sketch of the Obersee area in the middle of the table. An important part for our AR approach is the introduction of mediating objects which represent constructions for the participants to use for their planning. They are wooden cubes which are used as “physical handles” with *ARToolkitPlus* [7] markers attached on top. When the system detects a marker it augments the corresponding visual representation of a building or concept on top of the cube as depicted in Figure 2. 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 [8].

4. JOINT ATTENTION AND COLLABORATION

In dyadic collaborative tasks interaction partners need to coordinate activities and the focus of attention, for instance to make sure that both interaction partners talk about the same object or topic. A frequent procedure to create a common reference in dyadic cooperation is for one of the partners (A) to gaze at an object or to point at it together with a verbal utterance [9]. The partner (B) would then interpret this as an invitation or prompt to follow the interaction partner’s gaze or directive gestures (e.g. pointing). Usually A monitors whether B orients towards the object pointed at and when satisfied they both can assume to have established a common reference.

But what happens if the setting reduces the access to our natural undisturbed resources? Augmented Reality, currently hyped by developments such as Google glass[4] and similar systems can be expected to influence our focus of attention, or at least to cause some interference with the Non-AR environment. More dramatically, the use of head-mounted display for AR affects our mechanisms to establish joint attention strongly, as both eye-contact and gaze following are made impossible. Even though this limitation is induced here by the AR system, the general phenomenon of unavailable visual cues which induces compensating actions appears independently of this specific scenario in “natural” workflows. Reasons vary from unexpected focus shifts to the pure impossibility of sharing visual cues when working together remotely. In these situations, interactants need to invent and establish new routines to co-orient their partner.

We have already observed in interaction data that the limited field of view causes reference processes to shift towards the verbal level, and that some teams invent new routines such as lifting an object in front of their partner to prompt reorientation [10]. These effects occur since the peripheral awareness of the partners’ interaction with objects is cut away by the limited field of view. However, we have already advertised in Section 2 that sounds are quite effective in drawing attention and establishing an awareness. So basically we aim at investigating how far a mutual auditory display (A perceives B’s activities and vice-versa) can help interacting users to better establish joint attention. For this we first need appropriate sonifications to implement this idea and we will continue with concrete designs after an introduction of the available data.

5. SYSTEM DESIGN

Our goal is to convey in an auditory display the basic information about the current activities and work in progress. To operate on the periphery of conscious attention, the display shall rather provide implicit cues than explicit/symbolic messages on the activities. Interruptions and ends of moves will for instance be communicated implicitly via the absence of sound and/or sound changes.

Object manipulations such as touching, moving, shifting, rotating, (for malleable objects also squeezing), etc. are naturally accompanied by sound. However, some manipulations such as rotating an object in air or lifting an object while holding it in air are silent. Some interactions (e.g. shifting on a table) cause continuous interaction sounds that depend on continuous variables (velocity, pressure, texture of the surface), others are event-like, such as dropping an object. A detailed characterization of how real interactions manifest as sound can be helpful for developing sonification concepts that are more easily understood by users.

In our scenario relevant actions are limited to moving solid objects around on a table. Since these objects represent installations or specific areas in the recreational task every object movement changes the current state of the final solution and therefore is relevant for both participants.

5.1. Features from Marker-based Object Tracking

In our first approach we access the manipulation parameters directly such as object speed and height above ground. As described in Section 3.2, all wooden cubes have a marker attached to their top surface. These markers can be used to retrieve information such as the marker ID, its spatial position and orientation and the screen coordinates (see Figure 5.2). The position is used to determine whether an object is placed or lifted. Additionally, the position changes over time is used to calculate an object’s movement speed. Screen position gives us information about where ob-

[4]http://www.google.com/glass

```
...
B-lH-O-rel 800 1185 385 ~@WP
B-lH-O-rel 1185 2275 1090 ~^@WP
B-lH-O-rel 2275 6950 4675 ^@WP
B-lH-O-rel 6950 9595 2645 ~^@WP
B-lH-O-rel 9595 10180 585 @WP
...
```

Figure 3: Annotation Snippet. Annotations are exported from ELAN[6]and contain information about the manipulator (in this case: participant B's left hand), the start of the activity (800 milliseconds after sample start), the end of the manipulation (1185 ms) and the duration (385 ms). The activity and the manipulated object is encoded in the last string.

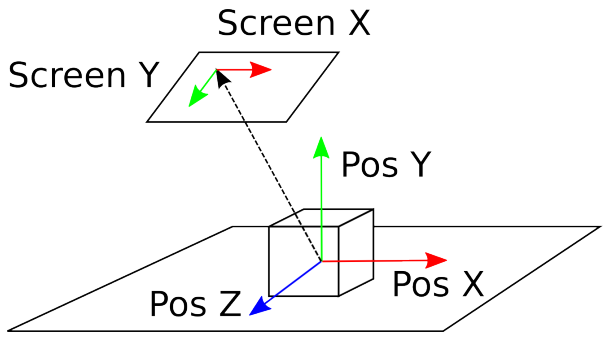

Figure 4: Marker Tracking allows us to retrieve the objects position (in cm) and screen position (in pixel) for every used symbol. Pos Y represents the height above ground and is used to identity whether an object is placed on the table or lifted. Object movement results in changes to Pos X and Pos Z. The screen is identical to the observers point of view which makes the screen position a good feature to adapt parameters according to that perspective.

jects are located from the observers point of view. Rotation and rotation speed are available as well but ignored for the time being.

5.2. Analysis-based Event Classifier

As an alternative to the the tracking-based features we also designed the system "top-down", i.e. with the help of data gathered from our experiments. This is motivated by the hypothesis that only meaningful manipulations need to be communicated. But how do we determine if an action is meaningful? The first step was to identify possible actions that can take place and find these actions in the annotation data. Annotations include the starting time of the manipulation, duration and end, the acting participant, and a string which encodes the object and the kind of manipulation. In this string the following characters represent certain activities:

Symbol	Meaning
@	Object grabbed
~	Object moved
^	Object above ground
mm	minor manipulation on Object
<	Hand moves away from Object
>	Hand moves toward Object
|	Hand rests close to Object

[6]http://tla.mpi.nl/tools/tla-tools/elan/

Starting from the analysis conventions, we designed a simple optimized state model (depicted in Figure 5) and compared that model to the annotated data. According to the model every object should fit in one of the following states at every time during the planning process:

- **Static**: The object rests on the table. A participant touching the object or performing smaller manipulations are ignored.
- **Pushed**: An object is moved to another position without leaving the table's surface.
- **Lifted**: A participant holds an object above the table nearly motionless.
- **Carried**: An object is moved to another position without touching the surface.

Since the state model is 'object-centered' in contrast to the annotations which are 'user-centered' object states had to be rearranged. Minor manipulations – so called micro manipulations – were treated as regular object movements and all activities that do not change the object position were merged into the static state. All theoretical state transitions were modeled according to Figure 5.

The next step was to evaluate this model with the annotation data and check if it contains every state transition observable in the data. When this approach failed we generated a comparable model from the data which included significantly more transitions, in fact transitions between almost all states. This implies that state changes can occur fluently, sometimes even indistinguishable for an analyst's eye which needs to be considered for automation attempts of this process.

Based on these findings we created two classifiers with varying accuracy and features. First, we used a parser-like classifier to convert the annotation results into a synchronized data stream which behaves like an automatic classifier. It features beginning and ending of an action and returns the responsible participant. This classifier is also used as a benchmark for further automatic approaches.

Second, we created a marker-based classifier which used the tracking data to retrieve position and velocity for each object. The basis for this classifier is the Marker-based object tracking & feature extraction code mentioned in Section 5.1 which was extended with a finite-state machine. In addition to beginning and end of an action it also contains location and velocity. With theses details on the available data, let us now introduce sonification designs.

6. SONIFICATION DESIGNS

For the designs presented in this paper we distinguish between continuous mappings and event-based elements. Both approaches can be fused into a hybrid approach. In this sense, the real-world sounds constitute already a hybrid acoustic representation. Our designs range from very data-oriented direct mappings via ecologically inspired real-world sound imitations to symbolic and artificial mappings that serve information functions beyond what a naturalistic sound is capable of. We motivate and describe them in turn.

6.1. Direct Parameter Mapping Sonification

From the standout of the available tracking data – which are (x, y, z)-positions of objects in space on a 20 millisecond grid as a sampling of the analogue and steady position function – we deal

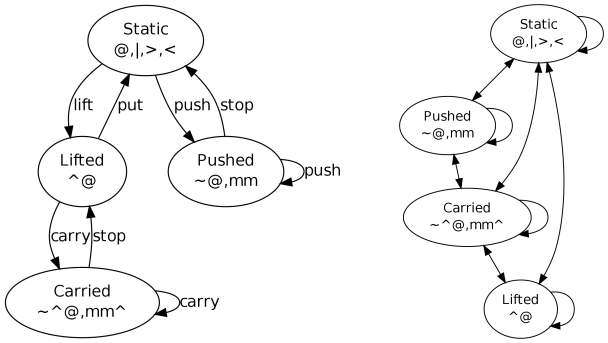

Figure 5: Theoretical transition model and annotation data model. The left graph shows the conceptual state model; the right one the model generated from the annotation data. Ideally every manipulation action ends in the 'static' state where 'carried' or 'lifted' objects are put down before they can be pushed. However, real world data show that this distinction cannot be made in every situation when actions transit seamlessly within a fraction of a second.

with continuous timeseries, so that any event must be computationally derived. From that perspective, a simple and direct mapping seems to deliver a good baseline.

The simplest approach is a direct mapping of the tracking data to amplitude and frequency parameters of a time-variant oscillator. We map the vertical height to frequency as this is a dominant association (supported also by default notation in musical scores) [11]. Since the sounds shall be concurrently used while engaged in verbal dialogue we chose a very narrow spectrum (of a single frequency) at low frequencies, where the perceived loudness is rather low. Specifically we map the object's attitude to the frequency range from 100 Hz to 300 Hz.

For the control of the amplitude there are various alternatives. Most intuitively, the absolute velocity $d(x^2 + y^2 + z^2)/dt$ can be mapped to amplitude, resulting in objects to remain silent without any movement. Frequency and amplitude interfere to some degree. If both data features (height / velocity) would be required to be perceived equally well, a different mapping would have been needed. However, the mapping to amplitude here serves solely as *excitatory mapping*, i.e. to let sound fade into silence without sustained interaction, and for this purpose the mapping works well.An additional extension for a more vivid mapping is that of controlling the frequency of a sinusoidal amplitude modulation by velocity, so that the faster an object moves the higher the audio rate oscillator pulses. Specifically we use a pulse range from 0.5 Hz to 10 Hz depending on the velocity input.

6.2. Abstract Signal Sonification

This design aims at signaling events with minimal dialogue interference. We use clear and distinguishable sounds inspired by the conceptual background: Lifting is represented by a short up-chirped tone. Consequently, putting an object down leads to the corresponding down-chirped tone. Pushing an object around on the table surface is represented by pink noise with configurable envelope release time. Carrying an object in air is modeled with a low-pass filtered white noise and a similar envelope. As guiding metaphor, the sounds are abstractions of sand and wind sounds for translation on ground or in air. Distinct events convey their directionality via the chirp direction.

6.3. Exaggerated Samples

We developed as design as contrast them to the Abstract Signal sonification, particularly concerning the degree of obtrusiveness. Just as in the Abstract Signal Sonification, the actions 'lift', 'put', 'pushing' and 'carrying' are sonified. However, instead of having unobtrusive sounds we have chosen very harsh signals: a high pitched blings (for lift), crashing windows (for put), creaking (pushing) and a helicopter (carrying) to render the actions very salient. Our observations are further described in Section 7.

6.4. Naturalistic Imitation

We can assume that sounds will be most easily understood if they fit perfectly to the performed actions, thus are naturalistic. The question arises why this needs sonification at all since the natural sound occur anyway. A sonification could be as annoying as the familiar MS Windows artificial 'click' sound that followed the real physical Mouse click sound. However, in our sonification, lift and put actions are identified by the object's z-coordinate, and thus an action becomes audible even if the physical manipulation is performed completely silent. Furthermore, by adding the sound we are in full control of the sound level, and can for instance set the sound level dependent on whether the interaction is within or out of the interaction partner's field of view. As sound samples we manually performed and recorded the actions using our wooden cubes on a desk. As additional (new) degree of freedom, we can for instance select samples dependent of the type of object: objects that matter to the investor could sound differently than objects that are critical for the conversationalist. Alike object sound redefinitions allow to add (subtle) task-dependent information layers beyond those that are prevalent in real interaction sounds.

6.5. Object-specific sonic symbols

As described in Section 3.2 all our objects represent an installation or building for the recreational area 'Obersee'. As the sounds aim at making the interlocutors aware of activity of an object outside their own field of view, the natural interaction sounds lack any information about the particular object/model being manipulated by the interaction partner. The sonification can disambiguate the sounds by associating samples to objects so that they are typical for the object. For instance while manipulating the playground placeholder, a sample recorded on a playground is played. For the petting zoo, animal sounds are a sample that causes a fitting association. The sample is activated whenever (but only if) an object is moved around – the height of the object above the desk is currently ignored, although it could be interwoven into the sound by a cascaded broadband bandpass filter.

We also mapped object movement speed to sound amplitude and screen position to panning. These parameters however, were only available in the case of using the Tracking Classifier.

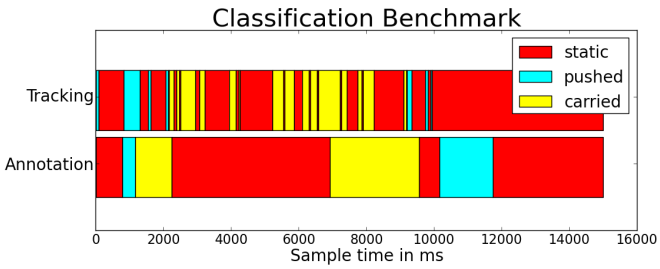

Figure 6: Model States Benchmark. The tracking classifier produces scattered results in contrast to the annotations. Rotations (e.g. from about 10000 to 12000) were annotated but were not detected by the tracker.

7. OBSERVATIONS

To evaluate the classifiers and the above sonification designs we took an approx. 15 seconds sample interaction from a trial. The sample video illustrates the mapping at hand of a typical scene from our data corpus where one of the subjects places an object and rotates it. It includes the top camera's audio and video recordings, an annotation snippet and preprocessed tracking data. All data were synchronized and differing formats and data representations unified.

We tested our five sonification concepts with two classifiers and two feature mapping approaches:

- **Classifiers**
  - A Annotation Classifier
  - T Tracking Classifier
- **Sonification Designs**
  - 1 Direct Parameter Mapping
    - a absolute velocity
    - b sinusoidal modulation
  - 2 Abstract Signal Sonification
    - a short envelope release time
    - b long envelope release time
  - 3 Exaggerated Samples
  - 4 Naturalistic Imitation
  - 5 Object-specific sonic symbols
- **Feature Mapping**
  - I Triggering (on/off)
  - II Velocity to amplitude, Position to panning

Sound examples are provided on our website [7] and are named according to the scheme:

`S*Concept**Classifier*_*Feature*`

For instance, `S2aT_II` refers to the Abstract Signal Sonification with a short envelope release time, input from the tracking classifier and extended feature usage. All sound samples and mappings were selected and iteratively optimized within the bounds of the defining metaphor using the authors' subjective evaluation and estimation of the sounds' information quality, information quantity and obtrusiveness. An evaluation of the designs with test listeners is in preparation.

[7] `http://www.techfak.uni-bielefeld.de/ags/ami/publications/NH2013-ISO/`

7.1. Classifier Benchmark

When comparing the classifiers' results we see and hear significant differences (see Figure 6). States are not congruent which results in different audio output. Additionally, the Tracking Classifier causes fractioned states. This is audible for every concept except for the abstract signals with a long release time (`S2aT`) where the longer fade out fills the gaps. However, this does not necessarily mean the fractioned results are less accurate.

7.2. Information Quantity

The characteristics of the used data sources had the highest impact on information quantity. Velocity and position were not annotated and therefore not available from this data source. The available information however, is more reliable. Direct parameter mapping involves all available parameters and therefore provides the richest experience (`S1a/b`). All event-based sonification designs feature nearly the same information except for the object-specific sonic symbols (`S5`) which do not distinguish between manipulations but provide details about the object kind instead.

7.3. Information Quality

Information quality varied depending on the frequency spectrum which was used. The choice of pink noise as well as white noise for the abstract signals results in significant masking of verbal utterances (`S2a/b`). This effect can be observed for the exaggerated samples as well with crash, scratching and explosion sounds (`S3`).

Adding velocity and position did not improve all sonifications. Samples and amplitude-velocity mapping clashed and resulted in a interrupted experience (`S3T_II` & `S5T_II`). Obviously, the amplitude modulation version (`S1bT_I`) causes sounds even without any ongoing manipulation which is rather annoying.

7.4. Obtrusiveness

Object-specific sonic symbols were considered to be the most unobtrusive design (`S5`). In combination with the annotation classifier and no direct parameter mapping the sound blends into the environment quite well (`S5A_I`). In general one can say, less features and therefore sound changes decrease obtrusiveness, a wide frequency spectrum increases it. However, we observed that the abstract signals were less obtrusive with the shorter activation periods from the tracking data in contrast to the longer annotated periods.

8. DISCUSSION

Our sonifications appear to be functional in conveying object interactions, both for interaction partners and interaction researchers. The five different methods only scratch the surface of possible designs.

Even though the different data sources deliver data of the same phenomena. It's not just the quality but also the way certain events are treated. The annotation data as a reference implementation which should be used to rate varying automated classifiers since it is the most reliable one.

We expected less accurate results for the tracking classifier since a computervision based tracking approach depends on many aspects such as image resolution, viewing angle, frames per seconds and marker quality. However, even if the data produced by

the tracking classifier are less robust and show different patterns it appears to us the sonification results are not worse. In our opinion the combination of the abstract signals and the velocity mapping produces one of the favorable designs we experienced in our observation. The fractioned nature is not compatible with the chosen sample-based sonifications though. Variations appear very rapidly and do not fit the characteristics of the chosen samples and made the impression there is something wrong with the sample playback rather than adding information.

Sonification approaches that allocate a wide frequency spectrum are not suited for scenarios in which dialogue occur. Pink and white noise as well as wide-banded crash and explosion sounds interfere with spoken words and are very likely to overlay the conversation. Low frequency solutions as used in the direct parameter mapping approach seem more acceptable yet they depend on high quality loudspeakers as small and cheap loudspeakers often fail to project the sound audibly.

But our system is not limited to online usage. The sonifications can also be used to support data analysis of recorded sensor data. For conversation and interaction research on multimodal corpora we expect analysts to benefit from such auditory data representations, especially if certain communication patterns are spread across several modalities and are otherwise difficult to detect. In these situations frequency overlap of sonifications and verbal utterances might not be such a severe issue if and only if language is not part of the data to be analyzed.

9. CONCLUSION

During collaboration an auditory display which communicates manipulation conducted in the context of interest is a valuable asset if these information are otherwise not available. In scenarios where visual attention has to cover an area wider than the (limited) view, such an auditory display can provide helpful cues about the current state of the environment and enable the user to perceive actions of collaborators which would otherwise have been unrecognized. Besides that, sonification of physical manipulations also provides a data representation which allows to *rapidly* process data and search for interaction patterns and distributions.

Features provided by our system include object identity, object state such as 'resting' or 'lifted', manipulation speed or intensity, location of the manipulation and the responsible manipulator. The question is which information is necessary in the context in which the display should be used. Every feature may increase the cognitive load necessary to process the sound and also increases obtrusiveness due to often changing patterns. Additionally, it is more likely to cover other verbal information which hinders collaboration more than it helps.

The obvious next step is to test the presented sonifications in an interaction study based on the Obersee II scenario with a qualitative evaluation using questionnaires and interviews. In addition conversation analysis should be used to analyze the impact of different designs onto the users' interaction and joint attention with the help of the already recorded data. Evaluating effectiveness however, requires a performance measure and therefore possibly a more specific task.

10. ACKNOWLEDGMENT

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).

11. REFERENCES

[1] S. Serafin, K. Franinovic, T. Hermann, G. Lemaitre, M. Rinott, and D. Rocchesso, "Sonic Interaction Design," in *The Sonification Handbook*, 1st ed., T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin: Logos Publishing House, 2011, pp. 87–110.

[2] W. W. Gaver, "Sound support for collaboration," in *Proceedings of the second conference on European ...*, 1991, pp. 293–308.

[3] C. Spence, J. Ranson, and J. Driver, "Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention." *Perception & psychophysics*, vol. 62, no. 2, pp. 410–24, Feb. 2000.

[4] M. Droumeva, A. Antle, G. Corness, and A. Bevans, "Springboard: exploring embodied metaphor in the design of sound feedback for physical responsive environments," in *Proceedings of the 15th International Conference on Auditory Display (ICAD2009)*, Copenhagen, Denmark, 2009.

[5] A. Dierker, K. Pitsch, and T. Hermann, "An augmented-reality-based scenario for the collaborative construction of an interactive museum," Bielefeld University, Tech. Rep., 2011.

[6] S. Zehe, "BRIX - An Easy-to-Use Modular Sensor and Actuator Prototyping Toolkit," in *The 4th International Workshop on Sensor Networks and Ambient Intelligence*, Lugano, Switzerland, 2012, pp. 823–828.

[7] D. Wagner and D. Schmalstieg, "Artoolkitplus for pose tracking on mobile devices," in *Proceedings of 12th Computer Vision Winter Workshop (CVWW'07)*, 2007.

[8] 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*, p. 245, 2009.

[9] G. Klein and P. Feltovich, "Common ground and coordination in joint activity," *Organizational ...*, vol. 53, pp. 1–42, 2005. [Online]. Available: http://csel.eng.ohio-state.edu/woods/distributed/CGfinal.pdf

[10] C. Schnier, K. Pitsch, A. Dierker, and T. Hermann, "Collaboration in Augmented Reality: How to establish coordination and joint attention?" in *ECSCW 2011: Proceedings of ...*, no. September, 2011, pp. 24–28.

[11] B. N. Walker and G. Kramer, "Mappings and metaphors in auditory displays," *ACM Transactions on Applied Perception*, vol. 2, no. 4, pp. 407–412, Oct. 2005.