Head Gesture Sonification for Supporting Social Interaction

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

Sebastian Zehe
Cognitronics Group
Bielefeld University
Bielefeld, Germany
szehe@techfak.uni-bielefeld.de

ABSTRACT

In this paper we introduce two new methods for real-time sonification of head movements and head gestures. Head gestures such as nodding or shaking the head are important non-verbal back-channelling signals which facilitate coordination and alignment of communicating interaction partners. Visually impaired persons cannot interpret such non-verbal signals, same as people in mediated communication (e.g. on the phone), or cooperating users whose visual attention is focused elsewhere. We introduce our approach to tackle these issues, our sensing setup and two different sonification methods. A first preliminary study on the recognition of signals shows that subjects understand the gesture type even without prior explanation and can estimate gesture intensity and frequency with no or little training.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Auditory (non-speech) feedback

General Terms

Sonification, Auditory Display, Interaction Technology

Keywords

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

1. INTRODUCTION

In co-present interaction, we use head gestures as a natural channel to signal our agreement or disagreement. Often they accompany or precede verbal feedback signals such as 'uhu', 'mhmm' or the words 'yes' or 'no'. Same as a spoken word is more than the mere symbolic information but contains rich information in intonation, head gestures are a complex sub-symbolic information carrier. For instance a nodding may vary in frequency, amplitude, number of repetitions or even in details such as how the gesture is synchronized with verbal signals or how exactly it builds up or decays over time. We normally do not pay much *conscious* attention to these signals, but nonetheless we use them, as apparently compensation mechanisms are required in mediated communication (e.g. on the phone) to maintain the interaction. For instance, if your conversation partner would not react verbally on the phone for a longer time, you might ask if he/she is still there. In co-present communication, however, we would tolerate much longer times without verbal feedback if head gestures are used for back-channelling.

Visually impaired persons, however, cannot access such non-verbal means. Since they do not experience head gestures directly, they lack important information to build up normal competences how to use their own head movements as communicative signals, e.g. for back-channelling. The needs and requirements have been analyzed for instance by Krishna et al. (2008) and Winberg et al. (2004) [15, 11]. Fortunately, the current trend towards ever more wearable sensing and actuation systems offers new opportunities to develop *sensory substitution systems* that enable users to perceive such signals on another channel. In this paper, we focus primarily on the auditory channel, using *interactive sonification* to represent head movements as sound so that the user can learn to understand and interpret correctly arbitrary activity.

There are manifold applications for such a sonification system, and we will only touch few aspects in this paper. First, our system (consisting of sensors, signal processing, sonification rendering and audio projection) can be used by a visually impaired user alone: thereby the user can directly experience how their own head movements sound, and how the sound correlates with proprioception. After some learning, blind users can give the sensor to their interaction partners and perceive their head gestures in real-time. This might facilitate communication and even create a higher sense of presence or connection. As a side effect, the blind user will also understand how sighted people usually use head gestures in natural communication, and may in turn use the own's head movements more alike for back-channelling. If visually impaired users could successfully replace their own verbal back-channeling by head gestures, this would – since this does not occupy the auditory channel – even keep their most important sense of listening free. In summary we aim at the plasticity of the human brain to learn and explore correlations between one's own movements and feedback sound to create a new link between them, with the question in mind whether and how far external sounds from head gestures of others can

activate an intuitive understanding and strong coupling between the interlocutors. If successful this could contribute a new *alignment* channel which otherwise wouldn't be at hand for visually impaired conversation partners. Further applications for our approach (such as mediated communication, Augmented Reality) will be considered in the discussion.

There is a body of research in sonifying movements for skill learning and for aiding movement (see [8] and references therein, and ICAD proceedings), and some sensory substitution systems which mainly focus on image sonification (see [3] and references therein), yet we are not aware of any sensor-based sonification for the purpose of supporting communication.

In this paper we focus primarily on the single user case and on methods to represent head movements as sounds so that – without any intermediate signal classification – listeners can correctly interpret them as head gestures. We start with a brief introduction to interactive sonification and a discussion of the role of head gestures in communication. The subsequent sections describe the hardware setup and introduce two different sonification methods, excitatory continuous and event-based sonification, which are also demonstrated in an example video. Afterwards, we present a first study where we ask users to classify and describe a set of different typical head gestures before and after they have experienced the system themselves. We discuss the results and draw some conclusions for future work.

2. INTERACTIVE SONIFICATION AS BIO-FEEDBACK

Sonification is the data-dependent generation of sound, if the transformation is systematic, objective and reproducible, so that it can be used as scientific method [6]. Interactive Sonification puts a focus on how users can interact with a sonification system, e.g. by selecting data, tuning parameters or even creating the data to be sonified in real-time [7]. The presented application is a special case of interactive sonification which can be called auditory bio-feedback since the users (listeners) themselves cause the sensor data by their own movements and the bio-feedback closes the loop into a 'closed-loop auditory interaction'. There are two interaction types in this closed-loop auditory display: the first is just head movements that cause sensor data to change as described before, the second interaction is the tuning of parameters (e.g. adjusting thresholds, frequencies, etc) while performing movements for personalization/optimization of the auditory display. Both interaction types play a different role: the first affects closed-loop learning of how head-gestures relate to sound, or for learning movements themselves according to a given sonic pattern, while the second will be important for users to adjust the sonification according to the own's preferences or other situational factors, yet we do not support any interfaces for that interaction yet during prototype development.

3. COMMUNICATIVE SIGNALS

Humans use a wide range of non-verbal communication signals in their everyday life. In conversations, head movements are intuitively used by speakers and listeners in many different ways from which we will introduce three exemplary.

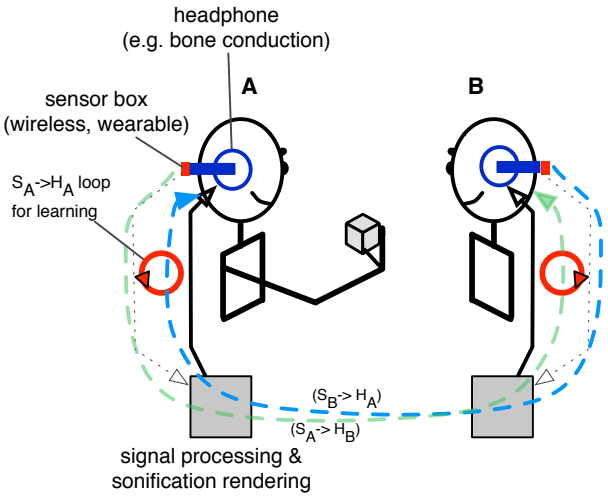

Figure 1: System sketch showing the components and the information flow between two interacting users. Depending on the application, either only the self-learning path (circles), the partner perception (Sensor B → Headphone A, marked blue) or the mutual coupling (both dashed lines) will be active. For most cases, a single sonification computer suffices.

First, they support dialogue structuring. Repeated movements such as nodding and shaking are commonly used by listeners to show agreement, disagreement or to emphasize that they still follow the statements of the speaker. Additionally, changes in head position can often be observed right before a current listener attempts to take the next turn in speaking [4]. This form of back-channelling creates a continuous information channel from the listeners to the speaker which is easy to access and process without any disturbance on the acoustic/verbal level. Second, speakers use their gaze and head direction with an indexical function, to reference to points or objects of interest during the conversation. The change of direction is often accompanied by rather less emphasized nod movements. And third, the speakers' head movements – like many other gestures used in communication – do not necessarily reference the listener but also provide stimulation for the speaker and assist cognitive structuring mechanisms [1] as well as they provide internal reinforcement through self-validation [2]. However, the influence of nodding and head shaking on confidence is not just limited to the speaker. There is evidence that a conversation listener nodding and shaking the head can influence the perception of spoken statements by other listeners in an unobtrusive manner as well [12]. To investigate the role of vision in early gesturing, studies have been conducted with blind and sighted infants that show that gestures evolve even in the absence of vision but differ in intensity and usage [9]. If no visual model is needed to understand gestures in general, making head movements perceivable for blind people might lead to an instant benefit for understanding non-verbal communication signals.

Besides the better grip on the ongoing conversations, the sonification-based perception of head gestures can be used *to*

train the usage of head gestures which may lead to faster and more robust conversation with sighted people. The active use of non-verbal communication channels also supports the spoken statements and could reduce the effort needed to convince the co-participant in a conversation. In addition to the interactive use of such information, *observation assistance* is another interesting application which is not limited to blind users: for instance, listeners tend to give positive back-channeling signals such as agreeing noddings more frequently when listening to a person with a higher status within a group than they do when listening to an equal-ranked speaker[5]. For interaction researchers that investigate alike research questions on the coherence between back-channelling and gender or group hierarchy, it might become easier to detect and identify promising hypotheses with the help of head gesture sonification instead of visual tracking.

4. SYSTEM DESIGN

Figure 1 depicts the modular system setup for the case of two interacting users A and B. Our setup incorporates motion sensors and headphones, either for one or for both communication partners as required by the application. A PC is used for handling the signal processing and rendering the sonification. This way, we are able to establish a closed feedback loop. In order to track the head motion, we used a BRIX module, a wireless sensor node attached to the head of each subject, see Figure 2. These small and lightweight wearable modules contain a battery, a wireless Bluetooth module as well a triple-axis gyroscope and a triple-axis accelerometer [16]. The BRIX system was generally designed to support rapid prototyping of sensing applications and it here suits our experiment well.

For the current prototype we can use either loudspeakers or transparent headphones, but we plan to replace these by bone conduction headphones as these permit to convey the information private to the listener and with minimal interference with the normal sense of listening. For a later iteration of the sensing setup, we even plan the integration of the sensors into the bone conduction headphone itself. The integration of sensing and actuation into a single unit will make the system much easier to attach and less obtrusive to use.

In the data processing step, we used only the gyroscope sensor readings because they contain already sufficient information for the auditory representation of head gestures. Figure 4 depicts the angular velocities (ω_x, ω_y). Apparently, due to the mounting of the sensor, nodding and shaking the head lead to rather decorrelated oscillations in these sensors. The gyroscope contained in the BRIX system has a maximum measurement range of $\pm 2000\ ^\circ/\mathrm{s}$ and a maximum resolution of 16 LSB per $^\circ/\mathrm{s}$. We operate the BRIX system for this application at a frame rate of approx. 100 Hz.

The sensor readings are received via a Bluetooth-to-serial interface. They are processed in a BRIX-server written in python, which processes all incoming sensor data and sends in turn OSC[1] messages to the programming language SuperCollider used for online sonification. The sonification module processes incoming OSC messages in an OSCresponder, stores the sensor data for later analysis or data replay, and executes on each step the update function of the selected sonification method. A simple GUI allows us to switch between sonification types, to save recorded data, or to replay previously stored data files.

[1]Open Sound Control

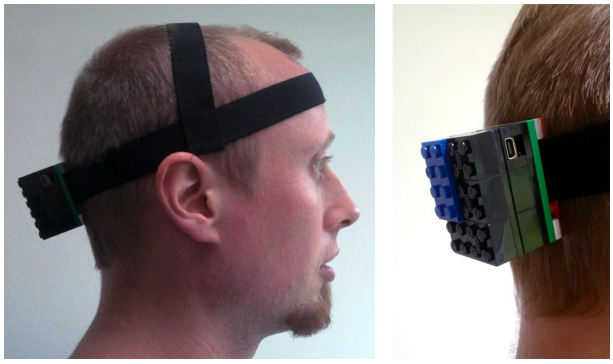

Figure 2: Test subject wearing the wireless motion sensors (left) and a close-up of the sensor system (right). The headphone is not shown.

5. HEAD GESTURE SONIFICATION METHODS

There are manifold ways how to represent the sensor data as sound. With the selected designs in this section, we present two new methods that aim at fulfilling selected requirements and goals as described next: Firstly, and importantly we here aim at sonifications that provide a relatively *direct* or *immediate* transformation of sensor data to sound. This means that the temporal evolution of the signal should rather directly influence the detailed sound pattern. The opposite approach would be to use symbolic representations, for instance generated by machine learning algorithms, to classify the head gesture type, to estimate the relevant properties and then to play a selected sound as acoustic signs to convey the information. Instead, we here aim to profit from the highly skilled auditory system of the listener to interpret the sound stream. This not only reduces the risk of false positives (which could dramatically reduce acceptance of the system), but it also enables trained listeners to pick up details which are not modeled in the machine-learning approach. Furthermore, we assume that users would prefer and appreciate the higher sound variability and richness (compared to a playback of acoustic signs) which results from the more direct and data-driven sonification. Additionally, direct approaches allow a lower latency than any approach that classifies gestures before displaying the result as sound. Secondly, we require the head gesture sonification to exhibit a significant contrast between relevant communicative signals such as nodding and shaking the head, etc. The requirement is that these gestures are particularly easy to distinguish, even if they occur at same frequency and intensity of the gesture. Thirdly, the sound should demand no or only little cognitive load for its interpretation. Though we believe that – after some training – many very different mappings would be understood, discerned and profitably used, we focus our design on mappings that are as natural or intuitive as possible.

As described in Section 4, we aim at a practical sound projection setup using bone conduction headphones so that the ear and thereby natural listening which is the most important far-range sensory channel for visually impaired users, is disturbed as little as possible. Furthermore, the interference with other environmental sound sources should be low and verbal utterances of dialogue partners should not be derogated. Therefore, we select sound signals that are compatible with this sound projection method and compatible with the sound ecological context. The sonifications should be heard only a little above the threshold of conscious perception, so that they are naturally in the perceptual periphery, yet listeners may attend to the information by active listening. Although the sound examples in the following sections are rendered at 0 dB, we suggest to listen to them at very low sound level, as low as possible yet still high enough to pick up the information.

The two methods presented in the following sections demonstrate different conceptual approaches for the problem that can be distinguished by their position on the analogic/symbolic continuum in auditory display theory [10, 14]: The continuous excitatory sonification is more on the analogue side, creating a rather direct continuous mapping, whereas the event-based contact sonification uses some mild signal- and task-driven criteria to condense the raw sensor readings into few 'key events' that encode relevant features of the gestures. Thus the event-based approach is more in the middle of the analogic/symbolic-continuum.

5.1 Excitatory Continuous Sonification

As explained in section 4, the angular velocities around the vertical and inter-ear axis are useful channels that contain the information about the head-gestures nodding and head shake. From the visual inspection of the signals (see Fig. 3) we can see that the intensity, frequency, duration and type are manifested in the temporal evolution (as amplitude, oscillation frequency, duration and dominant axis). Even subtle features such as whether a subject starts the nodding with a head rise or a head lowering can be seen from the signal. Since the temporal evolution of these continuous signals matters, and since the time scale is in the appropriate order of magnitude to be understood as rhythm, we selected a direct mapping of signal values to sound features of a single complex synthesizer.

Basically, we mapped the angular velocity around the between-ears axis – which is active during nodding – to a pitch parameter. This choice is motivated by the experience that height is usually intuitively characterized by higher pitch [13]. We mapped the angular velocity around the vertical axis – active during head shaking – to a stereo panning between the left and right audio channel. This choice is motivated from the indexical function of sound, and represents a movement on the horizontal sound source position.

As an important ingredient we introduce an *excitatory component*: we compute the activity (as the absolute value $a(t)$ of the angular velocities' first derivative) and feed these into a leaky integrator $A_\lambda(a(t))$ with adjustable leak rate λ. A leaky integrator collects data and decays without input to zero. A nonlinear mapping of A to the sound amplitude ensures a sonification which becomes quickly audible with the onset of head gesture activity and fades into silent as the activity stops.

The following SuperCollider code represents the key parts of the mapping:

```
   // gyro-x: v[1] head-shaking
pan=(v[1].sign*v[1].abs.linlin(0.02, 0.05, 0, 1)).neg;
   // gyro-y: v[2] nodding
freq=((v[2].sign*v[2].abs.linlin(0.02, 0.12, 0, 16)).neg
        +90).midicps;
   // pv is previous signal vector
activity = (v[1]-pv[1]).abs + (v[2]-pv[2]).abs;
A = (A * lambda) + (activity*(1 - lambda));
   // t0 is a threshold for silence
if(A>t0){
   level = A.linlin(t0, 0.18, 10, 50);
}{ // else
   level = A.linlin(0, t0, -40, 10); };
amp = (level + q.level).dbamp;
syns.set(\pan, pan, \amp, amp, \freq, freq);
```

The synthesizer is a simple bandpass-filtered pink noise with controllable center frequency, panning and amplitude. The filter's Q has been manually adjusted to deliver sufficiently pitched sound while maintaining some soft noise 'windlike' structure. Interaction examples are shown in Section 6.

5.2 Event-based Contact sonification

During the design phase of the previous sonification we realized that the sound delivers more information than actually required to understand most gesture details by listening. Same as complex body movement sequences can be understood by a low number of 'key frames' of a video, we assume that similarly key time points in the sonification are sufficient to reconstruct (or understand) the detailed gesture. The excitatory continuous sonification presented above created some sound during the whole gesture. In contrast, a reduction of information to 'key events' leads to a sparser occupation of the sound space which in turn may reduce the derogation of regular listening.

We identified the turning points of the movement as relevant key frames. For nodding and shaking, this would be where the head reaches the maximal and minimal angle. Since our sensor measures the first derivative of the angle, this corresponds to zero-crossings in the sensor data. Thus, we create (spawn) short transient parameterized sound events at these time points, mapping as above data features to sound synthesis parameters.

Please note that gestural features such as the duration or frequency of a nodding are inferred from the duration and tempo of these events: a perceptual skill which we exploit unconsciously when we estimate the walking speed from footstep sounds. Different from the previous excitatory continuous sonification approach, the focus on precisely placed transient sound events makes the perception of variations in rhythm (such as stumbling) more salient.

What transient sound events would be suitable for the upper/lower resp. left/right turning point of the head? Most of the everyday interaction sounds are contact sounds, e.g. when we touch an object, press a key, put an object on

the table, etc. Such contact sounds can be extremely short, they are sonically complex and can be added to the auditory scene at little interference with the perception of speech. Furthermore, our natural listening expertise for environmental sounds enables us to extract complex features (such as the estimation of source properties) from an impact sound at low cognitive cost.

As a starting point for the design, we recorded real sounds when a pencil strucks a surface. We then modify these samples according to features of the sensor data. For the up/down sound we used a sound of a pencil touching a glass of water. The reason for this choice is that (a) the material is easily recognized, and (b) the sound is clearly pitched, allowing us to stick to the already well-motivated pitch-mapping to discern between head-up and head-down events. For the left/right sound we used the sound of a pencil touching a plastic object[2]. This sound is extremely short, spectrally very broadband without much pitch structure, and also very different from the glass sound. However, the similar cause as 'pencil strikes object' may make it easier for listeners to accept both events as part of a single information stream. For the plastic sound, we used again stereo panning to discern between the left and right turning points.

The technical implementation demands some tricks: the first pitfall is that the sensor readings at zero crossings are zero and thus cannot be used directly to drive panning or pitch. This problem is solved by stepwise updating to new variables $v_{\max}$ and $v_{\min}$ to store the maximum/minimum value for each sensor channel. At a zero crossing from + to − the actual $v_{\max}$ value is used for the mapping and afterwards reset to zero and likewise for the opposite polarity. Same as above the activity is computed and used to suppress random jitter zero crossings without significant gestural activity. The following code snippet sketches the mapping for the zero crossings. The synthesizer is a simple sample player with amplitude, panning and rate control. Interaction examples are provided in section 6

```
2.do{ |i| // update min/max counter
   if(v[i] < vmin[i]) { vmin[i] = v[i] };
   if(v[i] > vmax[i]) { vmax[i] = v[i] };
};

// t0 is a threshold
if(A[0]>t0){
   if(v[1].sign*pv[1].sign<0){ // head shake
      if(v[1]>0){ // -/+ transition
         pan = vmin[1].neg.linlin(0.02, 0.2, 0, -1);
         vmin[1] = 0;
      }{ // else +/- transition
         pan = vmax[1].linlin(0.02, 0.2, 0, 1);
         vmax[1] = 0;
      };
      Synth.new(\syn, [\bufnum, b1, \amp, a, \pan, pan]);
}};
```

6. INTERACTION EXAMPLES

Figure 3 depicts sensor data during a recording of head gestures where the user created deliberately head gestures of given type (nodding, shaking), intensity (low, strong) and frequency (slow, medium, fast). It can be seen that the nodding is generally a bit faster than the headshaking. Secondly, the intensity manifests as higher amplitude of the signal. The frequency varies as expected, and in consequence, since the duration of the head gesture remained rather constant, leads to more oscillations per gesture.

[2]specifically we touched the computer keyboard

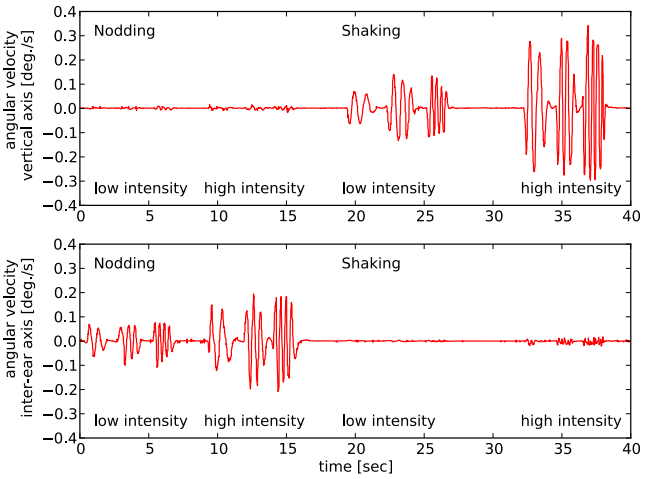

Figure 3: Plot of the gyroscope data during some head gestures.

Sonification example **S1** is the auditory representation of the data (rendered in real-time) using the excitatory continuous sonification. Sound examples are provided on our website[3]. It can be heard that the sound varies continuously and shows the expected frequency and spatial contours..

Sonification example **S2** is the auditory representation of the same data using the Event-based Sonification. Apparently, the concentration on few events causes a more sparse soundscape. Nonetheless detailed analogous information remains audible from the pitch, panning and level of events. We assume that users of this mapping might decrease the sound level much lower without losing awareness of the information, since the very transient events stand much stronger out.

Finally, example **S3** is a demonstration video of a user showing some head gestures. The video is the next best thing to trying-yourself to get a feeling or understanding how actions and sonification interrelate in practical use.

7. EVALUATION AND DISCUSSION

7.1 Preliminary Study

How well are the sonifications understood by listeners? How effective are the approaches to convey correct interpretations of type (nod / shake) or gestural details such as intensity or velocity? We conducted a first preliminary study to collect basic feedback from a few subjects, and to compare the two approaches. For this we asked subjects to listen to sonifications and to estimate head-gestural features. Specif-

[3] http://www.techfak.uni-bielefeld.de/ags/ami/publications/HNZ2012-HGS

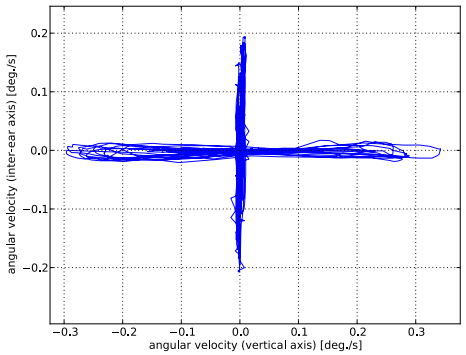

Figure 4: Plot of the gyroscope data (angular velocities). The plot shows that the mounting of the sensor is suitable to separate gestures with the two sensor channels.

ically, we gave as introduction only the information that the sounds represent head gestures and we left it open what sound would represent what type or characteristics. This allows us to access the immediate association from sound to head gesture.

The stimuli were pre-recorded sonifications, created from one of the authors deliberately performing head gestures at specific characteristics, i.e. of given type, strength and frequencies. With 2 types (nodding/shaking), 2 intensities (weak, strong), and 3 frequencies (slow, medium, fast), we obtain in total 12 combinations. A random permutation of these 12 examples forms a block. Such a random block (B_0) is played to the subject to become familiar with the range of sounds to follow, without any explanation. After that, a block of stimuli (B_1) is played, each sound one by one while the subject provides the ratings. The operator starts the next example on a verbal signal from the subject. Thus we also measure roughly how long subjects take to label each stimuli.

After B_1, the subjects are equipped with the sensor and can explore from own experience how their own head gestures sound using that sonification technique. With this experience we then ask the subject to rate another block (B_2) of 12 stimuli. This procedure (listening-only (B_0), rating a block (B_3), live test, rating another block (B_4)) is then repeated with the other sonification method. We balanced the initial sonification types. Subjects were finally asked to fill a questionnaire, including questions on their preference and expected performance.

7.2 Results

6 subjects (3 male, 3 female), age 23–37 participated to the preliminary study. Four subjects reported to play a musical instrument, 3 of them for less than 3 years.

7.2.1 *Overall performance*

First, we were interested to see what assignment of sounds to gesture type the subjects had chosen in the blocks B_1 and B_3, where they had not received any previous explanation or gained any prior experience of how the sonification represents head gestures. The first main result of this study is that all subjects (correctly) associated the sounds with changing pitch structure to nodding and the sound with changing panning to shaking the head. Together with phases B_2 and B_4, the overall recognition rate for gesture type is rather perfect, with 287 out of 288 correct classifications of the whole dataset.

Secondly we looked at how accurate the other features are rated. The overall accuracy for the intensity rating was 64.2%, and for the frequency rating 63.8%. Since however, frequency had 3 different values (slow/medium/fast), random guessing would yield only 33% accuracy, so obviously listeners are well capable to pick up some information. Full details on the performance can be seen in the class confusion matrices in Figure 5.

type	a.\p.	shake	nod
	shake	143	1
	nod	0	144

intens.	a.\p.	low	high
	low	78	66
	high	37	107

velocity	a.\p.	slow	medium	fast
	slow	52	40	4
	medium	17	56	23
	fast	2	18	76

Figure 5: Class confusion matrices for the pooled stimuli: the columns list the number of predictions for the actual labels shown in rows.

7.2.2 *Differences between sonification types*

Are there differences between the sonification types? Figure 6 shows the class confusion matrices of Figure 5 decomposed by sonification type, first summand indicating the number of cases under the excitatory continuous sonification condition. The number of correct intensity ratings differs only marginally (65% vs. 63%, being slightly better for the continuous sonification. The number of correctly rated gesture frequencies (61% vs. 65 %) is a bit better for the event-based sonification. From the class confusion matrix it seems that the main difference is in the perception of head gestures of medium velocity: under the excitatory continuous sonification, the velocity is rated much more often as fast whereas the event-based sonification is skewed to the slow rating. This is an interesting point, and we do not have a convincing explanation for that yet.

7.2.3 *Questionnaire*

Figure 7 depicts the subjects' replies on selected statements. We asked subjects to indicate their degree of agreement on a 4-point scale (strongly disagree, rather disagree, rather agree, strongly agree). Concerning the gesture type, subjects indicate their highest agreement to the statement Q1 ("It was easy to distinguish between nodding and shaking the head"). The figure differentiates the replies between the 2 sonification types (EBS, CS) and the phases, i.e. whether the block was the initial block (iEBS, iCS) or after live experience (eEBS, eCS).

We see that all subjects find it easy to distinguish between gesture types. Only with the iCS block there is some weak disagreement. Rating the intensity (Q2) and velocity (Q3) has been rated as much more difficult. Here, it is interesting

type	a.\p.	shake	nod
	shake	71+72	1+ 0
	nod	0+ 0	72+72

intens.	a.\p.	low	high
	low	38+40	34+32
	high	16+21	56+51

velocity	a.\p.	slow	medium	fast
	slow	25+27	21+19	2+ 2
	medium	4+13	24+32	20+ 3
	high	1+ 1	8+10	39+37

Figure 6: Class confusion matrices decomposed by sonification type: the columns list the number of predictions for the actual labels shown left. The cell entries are listed as a sum where the first term corresponds to the number of cases under the excitatory continuous sonification, the second term respectively for the event-based sonification.

to see that the ratings for the blocks after live experience are more positive. This may account either to learning effects or to enhanced identification due to the first-hand experience. We see no strong disagreement to the 'it was easy...' statements in any of the questions, indicating that the task is quite feasible for the subjects.

Question Q4 shows that subjects regarded the sound choice in EBS as slightly more intuitive than in CS. Q5 asks whether the sonification would disturb when accompanying a conversion. No one strongly agrees and no one strongly disagrees, but most subjects tend to find that the sonification would disturb. This has three reasons: (a) we did not explain that they would hear the *conversion partner's* head gestures (some subjects actively stated afterwards that they thought they would hear their own movements, and (b) we did not explain that this may be useful since the real user would be visually impaired or could for other reasons not see the partner. Some subjects expressed that under such knowledge the answer would have been different. Finally, we did not take enough care to adjust the sound to be very quiet and low in level, as intended for the real application. We agree that sonifications at such high level would be indeed disturbing.

Finally, subjects rated that they profit from training in rating the features of head gestures (Q6). This is very interesting as the data show that overall performance (accuracy) between the initial blocks (B_1 and B_2) and the blocks after live experience (B_2 and B_4) is pretty equal: less than 1% difference for type and intensity. Surprisingly the accuracy is even lower for the velocity rating for the 'trained' blocks (72.2% vs. 55%). However, the number of subjects and repetitions is probably too low to derive significant statements on the basis of these initial preliminary study.

In summary, the study gives us good confidence that the sonification designs can be basically understood even without any explanation of how head movements are encoded by sound. We received valuable new feedback from subjects via a free comment field in the questionnaire. For instance, two subjects commented that the strong spatial sound changes in head shake sonification is very displeasing, for one subject even to the level of causing balance problems. We did not

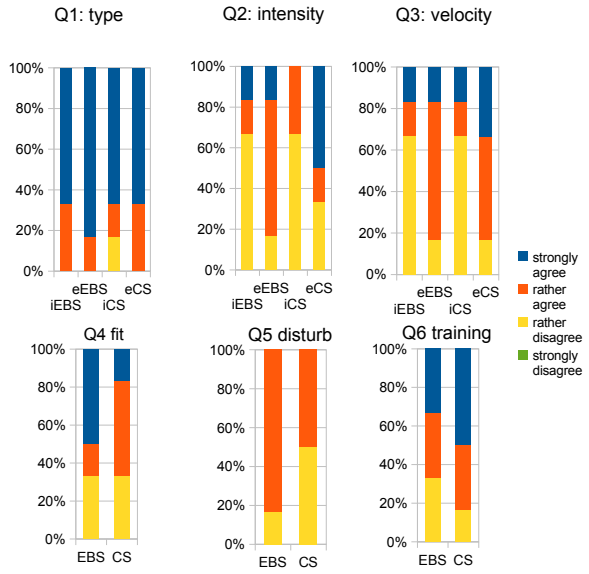

Figure 7: Histogram results of the agreement to 6 selected statements Q1–Q6 in the questionnaire: Q1: "It was easy to distinguish between nodding and shaking the head", Q2: "It was easy to distinguish between the intensities", Q3: "The differences in head movements velocity was easy to hear", Q4: "The chosen sound fits the movement", Q5: "During conversations the sounds would disturb me", Q6: "It was easier to categorize head gestures after training".

expect that! Explicit preference expressions occurred both for the excitatory continuous and event-based sonification by different subjects. Subjects commented that the sounds were too loud for use in communication settings. Apparently we can improve our test by adjusting the level more carefully at the beginning. One subject expressed that the excitatory continuous sonification works faster and more precise. Two subjects explicitly mentioned that they had a better subjective confidence after experiencing the sonification themselves.

8. CONCLUSION

We have introduced our approach and two methods for head gesture sonification. We have described our wearable sensor and sound synthesis system to generate real-time sonifications at low latency. The main contributions are two new methods for representing head gestures by sound, (a) continuous excitatory sonification and (b) event-based sonification, which we designed to allow users to understand in particular nodding and shaking the head. To test how users can extract information from the sound, we conducted a preliminary study with 6 subjects. The first result is, that all subjects associated correctly the gesture class (nodding / head shaking) without having received any explanation how head gestures would be represented as sound. Second, the study shows that subjects can pick up gesture details such as velocity/frequency and intensity at a level well above chance. This preliminary study helped us to collect some feedback on problems and preferences. We were surprised that the strong left/right panning has been experienced as irritating

and even interfered with the sense of balance for one person. Some subjects commented on the aesthetic quality. After the experiment, most subjects responded generally positive after they were informed how we think the system may be useful to support visually impaired users.

We consider various other application areas for head-gesture sonification. We see the opportunity for sighted users that head-gesture sonification can be an additional communication channel in mediated communication, e.g. to couple interlocutors more tightly while speaking on the phone. In co-present interaction sighted users may benefit from our system, if the setup or situation demands a narrow visual focus on an object under examination (e.g. blackboard, planning, cooperative repair tasks). The tight focus on this object would most probably demand head movements to attend to the other's head gesture signals, and thereby disrupt the primary focus of attention.

Concerning the primary application as sensory substitution for visually impaired people to experience their interlocutors' nonverbal actions, we are curious to apply and test our system with visually impaired subjects in the near future. We also plan to test how a bidirectional, mutual coupling between cooperating users (each hearing the partners' head gestures) will affect their cooperation and particularly their use of gaze and orientation in cooperation settings. Finally, an untapped application area is to provide interaction researchers with better methods to understand head gestures when analyzing social interaction of three or more interacting users: while we believe that in such situations the simultaneous *visual observation* of the groups' head gestures will be difficult, we expect that sonification will allow to attend to overall 'collective' gesture patterns in a way that would otherwise be difficult to achieve. The sounds certainly need some optimization of sound quality, sound level, timbre, mapping and aesthetic qualities, but this depends on the selected application (analysis, monitoring, skill learning, etc.), user group, sound projection system (e.g. bone conduction headphones) and use context, and may also be a matter of personal preferences or taste. We hope that our approach of head gesture sonification will offer a useful and helpful contribution in some of the proposed areas.

9. 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).

10. REFERENCES

[1] M. Boholm and J. Allwood. Repeated head movements, their function and relation to speech. In *Proceedings of the Workshop on Multimodal Corpora Advances in Capturing Coding and Analyzing multimodality LREC 2010*, pages 6–10, 2010.

[2] P. Briñol and R. E. Petty. Overt head movements and persuasion: A self-validation analysis. *Journal of Personality and Social Psychology*, 84(6):1123–1139, 2003.

[3] A. D. N. Edwards. Auditory display in assistive technology. In T. Hermann, A. Hunt, and J. G. Neuhoff, editors, *The Sonification Handbook*, chapter 17, pages 431–453. Logos Publishing House, Berlin, Germany, 2011.

[4] U. Hadar, T. Steiner, and E. Grant. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983.

[5] M. Helweg-Larsen, S. J. Cunningham, A. Carrico, and A. M. Pergram. To Nod or not to Nod: An Observational Study of Nonverbal Communication and Status in Female and Male College Students. *Psychology of Women Quarterly*, 28(4):358–361, Dec. 2004.

[6] T. Hermann. Taxonomy and definitions for sonification and auditory display. In *Proceedings of the 14th International Conference on Auditory Display (ICAD 2008)*, pages 1–8, 2008.

[7] T. Hermann and A. Hunt. An introduction to interactive sonification (guest editors' introduction). *IEEE MultiMedia*, 12(2):20–24, 04 2005.

[8] O. Höner, A. Hunt, S. Pauletto, N. Röber, T. Hermann, and A. O. Effenberg. Aiding movement with sonification in "exercise, play and sport". In T. Hermann, A. Hunt, and J. G. Neuhoff, editors, *The Sonification Handbook*, chapter 21, pages 525–553. Logos Publishing House, Berlin, Germany, 2011. Höner, O. (chapter ed.).

[9] J. Iverson, H. Tencer, and J. Lany. The relation between gesture and speech in congenitally blind and sighted language-learners. *Journal of Nonverbal Behavior*, 24(2):105–130, 2000.

[10] G. Kramer. An introduction to auditory display. In G. Kramer, editor, *Auditory Display*. Addison-Wesley, 1994.

[11] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan. A Systematic Requirements Analysis and Development of an Assistive Device to Enhance the Social Interaction of People Who are Blind or Visually Impaired. In *Workshop on Computer Vision Applications for the Visually Impaired*, Marseille, France, 2008. James Coughlan and Roberto Manduchi.

[12] J. E. Simpson. *Does Nodding Cause Contagious Agreement? – The Influence Of Juror Nodding on Perceptions of Expert Witness Testimony*. Dissertation, University of Alabama, 2009.

[13] B. N. Walker. Magnitude estimation of conceptual data dimensions for use in sonification. *Journal of Experimental Psychology: Applied*, 8:211–221, 2002.

[14] B. N. Walker and M. A. Nees. Theory of sonification. In T. Hermann, A. Hunt, and J. G. Neuhoff, editors, *The Sonification Handbook*, chapter 2, pages 9–39. Logos Publishing House, Berlin, Germany, 2011.

[15] F. Winberg and J. Bowers. Assembling the senses: towards the design of cooperative interfaces for visually impaired users. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, pages 332–341, New York, NY, USA, 2004. ACM.

[16] S. Zehe, T. Grosshauser, and T. Hermann. BRIX – An Easy-to-Use Modular Sensor and Actuator Prototyping Toolkit. In *Proceedings of the 4th International Workshop on Sensor Networks and Ambient Intelligence*, 2012.