EMPIRICAL AESTHETIC EVALUATION OF SONIFICATIONS

Katharina Vogt, Visda Goudarzi
Institute of Electronic Music and Acoustics
University of Music and Performing Arts, Graz, Austria
vogt@iem.at, goudarzi@iem.at

Richard Parncutt
Centre for Systematic Musicology
University of Graz, Austria
richard.parncutt@uni-graz.at

ABSTRACT

This paper discusses three experiments on the aesthetic evaluation of different sonifications. The effects of training and understanding of the auditory display on its aesthetic appealing were tested. Results showed no significant effect, but a trend towards less acceptance due to longer exposure to the sounds in general. Furthermore, there might be effects of musical ability and gender that should be further explored.

1. INTRODUCTION

Evaluation of auditory displays has gained importance in sonification research. Standardized methods have been successfully transferred from psycho-acoustical experiments, cf. to [1]: data can be collected in tasks where stimuli are identified, their attributes are rated, they are discriminated among each other, their dissimilarity is rated or they are sorted. When working with sounds, special requirements have to be taken into account, e.g., initial hearing tests might be included, and limitations of auditory memory have to be regarded in the test design. Analysis of such data employs classical statistical methods.

In the study of hearing, the classical psycho-acoustic approach proved to be insufficient [2]. Ecological psycho-acoustics are necessary, where not only basic results of static stimuli are taken into account. Listening behavior has to be regarded as well, being influenced by higher level of cognitive factors. Following the same argument, sonification research needs ecologically valid evaluation. This approach is challenging. Especially in the sonification of scientific phenomena, experts who both understand the data/ task and aspects of sound are very rare. The direct application of standard procedures is not possible because the extraneous variables cannot be controlled in the real-world as it is possible in a laboratory setting. Due to such difficulties, the pushing of researchers of auditory display towards legitimating their work by extensive tests has been criticized [3].

One possibility of ecologically valid testing in the context of sonifications is implementing methods from human-computer interaction (HCI). Usability testing includes, e.g., time factors in learning an application or finishing a task, the counting of the number of errors in completing a task, or the subjective satisfaction of the user, [1]. Methods for data collection include surveys, verbal protocols, focus groups, and expert appraisals.

1.1. Specific research context

In the research project SysSon – *A systematic procedure to develop sonifications* [4] – ecologically valid evaluation is conducted throughout the development of the sonification design and a sonification interface. As part of this evaluation we conducted some preliminary empirical studies on the aesthetic evaluation of sonifications. The results are discussed in this paper.

1.2. Why aesthetics?

Sonification research has been conducted systematically for over 20 years now. Still, sonifications have not yet found an accepted place in scientific research practice. It seems that the distribution of sonifications among a larger user group, e.g., in science, is lacking. It is known from other fields, that, e.g., the sound of a product affects the perception of its quality, and attractive products are easier to use [5]. Sonifications were, until recently, mostly developed from a technical perspective, while it has been shown that aesthetics is at least as important as functionality for the long term experience, as discussed by Barrass and Vickers [6]. For developing an aesthetic sonification design, these authors suggest to include persons in the process, who have skills in sound design or an ”aesthetic thinking and practice” (p. 164).

In order to evaluate the aesthetics of sonifications, we developed a test design in cooperation with musicologists. The design was employed in three similar experiments of three different sonifications. These sonifications have been developed by thirds, and were selected by the authors as good and diverse examples of recent research work – all presented at ICAD 2012. Student groups at masters level (see acknowledgements) chose them according to their preferences from a larger set of sonifications.

2. TEST DESIGN

2.1. Research question

Our research question was whether prior knowledge about the meaning of a sonification (e.g., type of mapping, origin of data) influences its aesthetic evaluation, and if so how.

Test subjects were randomly assigned to an experimental group (EG) and a control group (CG). While all subjects underwent a first and second round of aesthetic assessment, the groups were treated differently in between (see Tab. 1): the EG got to know what the sounds mean and how they were produced; the CG listened but did not get an explanation. Instead, they had to draw what came to their mind when they heard the same sounds as the EG. This should balance the effect of habituation to sounds that we hear more often, which might have an effect on the second aesthetic rating. In final interviews demographic data on the participants was gathered.

Round	Experimental Group	Control Group
First Round	Evaluation 1	Evaluation 1
Treatment	Training & Testing	Control Task
Second Round	Evaluation 2	Evaluation 2

Table 1: Test procedure

During the aesthetic evaluation the participants rated how much they liked a group of randomly presented sound samples on a scale from 1 (”not at all”) to 7 (”very much”). Furthermore, words were collected that participants associated freely with each sound for qualitative analysis. Quantitative analysis was conducted comparing CG and EG with each other, male and female participants, and (in some experiments) musicians vs. non-musicians. Significance of results were checked with an ANOVA analysis. Aggregated statistical results are collected in the Appendix. For the qualitative analysis, the associated words were assigned as positive, neutral, or negative by the students. Furthermore, the drawings of the CG were taken into account for the discussion of results.

2.2. Chosen sonifications

Three sonifications – as part of a larger set – have been suggested by the authors of the paper and picked by student pairs, see Tab. 2. The sonifications are completely different in their goals, implementations, and designs, but were regarded as interesting by the authors. All are recent developments by third parties and have been presented at ICAD 2012.

Name	Literature	Data
Tremor	[7]	3D movement data
Tweetscapes	[8]	real-time online data
VOSIS	[9]	b&w images (pixels)

Table 2: Sonifications chosen by the students.

Tremor is a project in cooperation with neurologists who analyze the rhythmic, involuntary movement of a part of the body (a tremor). Depending on the oscillation type, different neurological diseases can be differentiated, one of which is Parkinson disease. Different sound designs were developed and are at the moment tested with neurologists, based on frequency and amplitude modulation. In pilot tests, the sounds proved to be very helpful to determine the kind of disease. For examples see/ hear [10].

Tweetscapes is a project of sonification experts, media artists, and a radio broadcaster. Online data of German Twitter streams is sonified and visualized in real-time. The project received the Award of Distinction in the Digital Musics and Sound Art category of Prix Ars Electronica. The sounds are based on a large sound database and randomly – but reproducibly fixed – assigned to different semantic terms (hashtags). These sounds are then modified according to meta-information, e.g., from which location in Germany the tweet was sent. For examples see/ hear [11].

Voice of Sisyphus - VOSIS is a free online program to transfer black and white images into sound, based on a multi-media installation. The graphical synthesis technique is based on raster scanning of pixel data, and the sounds are refined by filtering and spatialization. For examples see/ hear [12].

2.3. Test subjects

The students recruited 10 to 20 participants for each of the 3 experiments. Details on their demographics (age range, gender, musicians or not, number of subjects in EG an CG) can be found in Tab. 3.

Exp.	Age (Av.)	M/F	Musicians/Non	#EG/#CG
Tremor	19-44 (26)	10/10	–	10/10
Tweetscapes	19-40 (25)	5/5	6/4	10/–
VOSIS	20-59 (28)	7/11	9/9	10/8

Table 3: Details on Participants.

3. EXPERIMENTS AND RESULTS

3.1. Tremor experiment

In the Tremor experiment, 27 sound samples were used that have been generated from data of diagnosed patients of three different diseases (Parkinson, essential, and psychogenic tremor). Out of these samples, 18 samples for the aesthetic assessment were cut out. For the training of the EG and the drawing task of the CG sound samples of a length of one minute were cut out. The stimuli were presented over headphones using PsychoPy software.

The training phase for the EG took ten minutes, which was followed by a test of how many tremors could be identified. Participants of the CG were confronted with the same sounds but only instructed to draw something along. The total duration of the experiment took about 35-45 minutes.

No significant differences have been found between the ratings of evaluation 1 and 2 (details for all statistical results can be found in the Appendix). The results also did not show a significantly different rating by female and male participants from evaluation 1 to 2, but male participants rated generally higher. The test results after the training phase in the EG showed that only 50% of the essential tremor, 57% of Parkinson, and 33% of psychogenic tremor were detected correctly.

The qualitative assessment of words showed that mostly neutral words (61%) and only a few positive ones (3%) were used to describe sounds. A trend could be observed that more negative words were used in the second evaluation. This coincides with the report of the interviewing students that the majority of test subjects became more annoyed by the sounds during the experiment. The words have been grouped by the students into seven categories:

- outer space
- ocean
- musical instruments and musical terms
- movement
- machines
- (illness)
- (negative) adjectives

Words related to illness have only been mentioned by participants of the EG after they were explained the origin of the sounds, which is not surprising. 14 out of 20 participants used negative attributes, such as annoying, boring, alarming, sad, or intrusive.

Some examples from the drawings of the CG during the control task are shown in Fig. 1. Seven out of ten participants drew

different kind of waveforms. The second biggest group were drawings of flying objects/ UFOs.

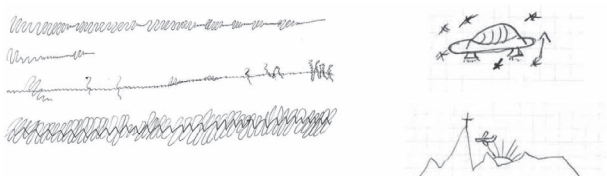

Figure 1: Examples of drawings of the CG in the tremor experiment. Waveforms are depicted at the left hand side. The right hand side shows a picture of a UFO and a plane flying over the mountains.

3.2. Tweetscapes experiment

The Tweetscapes experiment didn't have any control group due to personal reasons of the students conducting the experiment. The sound stimuli consisted of 18 sounds of 10 seconds each. They were gathered by the students at three days and different day times to showcase the diversity of tweets over the span of the time. The sounds were recorded from the web using audio-hijacking, i.e., recording the live-stream from the internet. In each evaluation round 10 sounds out of the 18 stimuli were played and rated on a scale (1-7) how much the participants like/dislike the sound. In the training phase 3 exemplary sounds from the Tweetscapes website were presented and explained to the participants. Testing of their understanding was difficult because the exemplary sounds showcased the basic features of the Tweetscapes sound mapping, while the real recorded sounds were much more complex.

In this experiment, also the difference between musicians and non-musicians was assessed. A person who has been active in singing or playing an instrument at least during the last five years was regarded a musician.

In the first evaluation round the mean rating was 5.20 and musicians rated the sounds higher than non-musicians. The qualitative interviews gave a hint on different listening behaviors. It seems that non-musicians concentrated more on the sound in general, whereas musicians tried to filter parameters such as harmony, dynamics, and other musical characteristics of the sound.

In the second evaluation round, the rating dropped down to a mean of 4.00. The average ratings of musicians and non-musicians changed comparing to the first evaluation. While non-musicians rated the sounds similarly in both evaluations, the musicians, apparently, didn't like the sounds the more they listened to them.

In the qualitative analysis of the freely associated words the students sorted them into the following categories, ordered by their frequency:

- environment
- technology
- condition (i.e., qualitative descriptions)
- fantasy
- (Twitter-related)

The category environment is used the most (40 times), including descriptions such as nature, environment, animals, jungle, and birds. This might be an effect of the random recording of sounds, where by chance many sound samples came up that have an environmental background. Twitter related words were only mentioned by the EG after they had been explained the mapping.

The descriptions were also classified in their positive, negative, or neutral nature. The majority were neutral ratings (66%), with a little trend towards negative terms (21%) as compared to positive ones (13%).

3.3. VOSIS experiment

The sounds for the VOSIS experiment were recorded by the students using ten different basic geometric objects, such as a square, a triangle, and a circle, see an example in Fig. 2. More complex figures, e.g., faces, proved to be too difficult to differentiate in an initial testing phase of the students. Again, participants were grouped into an EG, who were explained the sound mapping in a training phase, and a CG, who had to draw along the presented sounds.

Figure 2: Examples of the basic objects in black and white used to produce the sounds for the initial evaluations.

Participants were classified as musicians when they have had a regular practice of active music making for more than 3 years. Only the ten participants of the EG were analyzed according to this criterion.

20 sound samples were selected for the evaluations, and 8 sounds from slightly more complex objects were used for the training phase. The testing was conducted using nine different sounds from the basic objects. The participants saw the object and heard the sound and then they had to guess which region of the object was played out of a given choice of 6 possible regions. If they associated the correct section of the image, the task was rated as successful.

Generally, ratings were lower in the second round of tests, but results were not significant. Musicians rated the sound samples higher than non-musicians. While the ratings of non-musicians did not change from evaluation 1 to 2, the musicians rated lower. These trends are not significant due to the small number in each of the compared groups (5 musicians and 5 non-musicians). Around half of test conditions were successfully identified by the EG, therefore training can be regarded as successful compared to a chance of guessing the right region of 1/6th.

The qualitative assessment of the phrases showed that both EG and CG described the sounds as mostly neutral (57%) or negative (40%). They rarely described a sound using a positive word (2%). These relations did not change significantly between the two evaluation rounds. The participants described many of the sounds with science-fiction and industrial terms (but the students in this experiment did not categorize the terms completely).

An example of drawings of a participant of the CG is given in Fig. 3. The associations in this example comprise different concepts, from sound making objects, e.g., the didgeridoo, to behavior resulting from sound, e.g., dancing. Students, following their comments in the final report, were disappointed that never the "right" objects, which were the basis for the sonification, have been redrawn in the blind test. This is not surprising to the authors and probably not the aim of the VOSIS project.

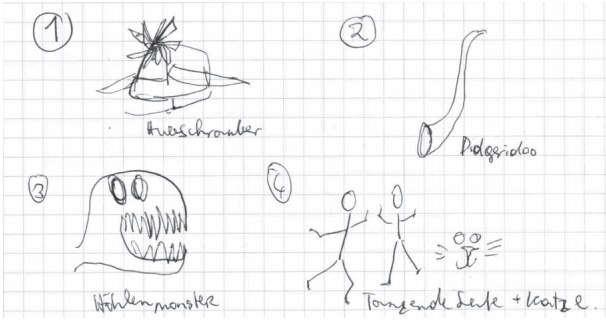

Figure 3: Examples of drawings of one participant of the CG in the VOSIS experiment. The participant in this case marked an associative word as well for each sound (the captions are in German: "Hubschrauber" - helicopter, "Didgeridoo", "Höhlenmonster" - cave monster, "Tanzende Leute + Katze" - dancing people plus cat).

4. AGGREGATED RESULTS

The nature of the experiments being from different projects with slightly differing test design makes it difficult to generalize the results. As one reviewer remarked, an important factor might be that the three chosen sonifications differ in their *function*: Tremor is purely functional, and has been developed for a specialized user group; VOSIS is functional as well, but for a broader public and based on the experiences of a multi-media project; finally, Tweetscapes is mostly a public installation/work between science and art, disseminating the idea of sonification. The authors chose the diverse set of sonifications on purpose, in order to be able to derive general results on sonifications. But the functionality of sonifications might actually influence the results in so far, as sounds that are useful are likely more accepted as listeners can identify or understand the need.

Concluding from all three experiments, several trends can be observed, even if the amount of test participants in each single comparison group did not allow significant results. The general aim of the experiments was to find a difference in the aesthetic evaluation with and without the knowledge of the underlying sonification. Two other interesting comparisons turned up during the experiments: first, there might be a difference between the rating of musicians and non-musicians (however they are defined), second, men rated generally higher throughout the experiments than women.

Answering the research question, we did not find a significant effect over all experiments. There is a trend that a longer exposure to the sounds worsens the rating in general. The knowledge of the sonification seems to have a slightly positive effect, but it does not outweigh the worsening completely.

There might be different effects that are reflected in these results: the students who conducted the tests reported that the participants became more and more frustrated with the sounds. This is partly reflected in the qualitative results, e.g., where the positively annotated associations in all experiments were clearly outnumbered by the number of negative ones. The general effect of sustained exposure can be caused by fatigue or increased boredom. The test design tried to filter the effect of knowledge gain from these general ones, but this was not successful or did not show significant effects in all experiments. The most serious problem with the test designs of our experiments is that training of the understanding of the sonification was not or not sufficiently tested or not sufficiently achieved (a success rate of up to 100% would be necessary, or participants who fail the test would have to be excluded from the analysis.) This proved to be difficult, because the students partly could not generate their own didactic examples, and a real training would probably make more effort than could be demanded in a university course.

Musicians seem to rate higher than non-musicians, as was tested in two out of three experiments. Again, the results are not significant, mainly due to the small number of participants. It seems that musicians could generally get more interested in the sounds of sonifications for different reasons such as curiosity about the sound parameters or dynamics and texture of the sounds whereas the non musicians had no point of reference in sounds to start with. There have been studies revealing inconsistent results for differences according to musical ability, as reported in [1]. How *musical ability* is measured has not been standardized in the literature. Even in our two experiments, we found two different concepts, both related to singing/ instrument playing. The concept of openearness within musicology research might be an interesting factor to regard in this context.

Another effect that turned up in the results is the difference between men an women: it seems to be a good idea to evaluate an auditory display with a group of men if one wants to achieve a good ranking. Again, the effect is not significant but a trend observable in all experiments.

5. DISCUSSION

For the authors the general outcome of the study was surprising. We would have expected a clear positive effect, e.g., because the sounds become more interesting to listen to, and richer, when their origin has been understood. Furthermore, the three chosen sonification designs are good examples of sonifications from our point of view, one even winning a prestigious media arts award. Still, the results showed and the students reported that the participants disliked the sounds more after some time of exposure. An explanation might be that they expected something more "musical", as they were interviewed by musicology students. Another interpretation could be found in the final report of the VOSIS group: *"Although training enables to explain the sounds, we believe that after the training the participants understand less why the sounds sound so uncomfortable."* In VOSIS the students found the sounds very harsh and unbearable from the beginning. Maybe when the interviewer doesn't find the sounds appealing, it doesn't help to encourage the participants to listen to them either and it can influence their bias. In general, students gave harsh comments for the sonifications. For instance, the Tremor group stated that the sonifications are *"definitely not suitable for everyday use"*. Compared to other tests that the authors conducted in the past, where the test

participants mostly came from a scientific or sound background, this general attitude is surprising to us.

Another concern is that when the concepts that are sonified are not in the domain knowledge of the participants, it gets more difficult for them to distinguish the differences in sound. As we know from personal communication with the authors of the Tremor sonification, the experiments run in a workshop for neurologists showed that they found the learning curve in training relatively easy and the sounds made more sense to them. Both, testing a sonification with the specialists in the field and with a general public (as in our experiments) seems to be useful in a complementary way. Definitely, even a broad public should associate more than 2 or 3 % of positive terms with a sound, if a sonification should be successful.

6. CONCLUSIONS AND OUTLOOK

In this paper we discussed three experiments that should test the effect of understanding an auditory display on its aesthetic appealing. In order to prevent an effect of habituation to the sounds while being exposed to them for a longer period of time we defined a control task. Due to the experimental test design and the small number of participants in each experiment and each testing condition (experimental vs. control group; men vs. women; musicians vs. non-musicians) no significant results could be found. Still, we argue that the focus on aesthetic testing in sonification and the choice of a general pool of test subjects instead of specialists reveals interesting effects, that might explain the difficulties of the dissemination of sonification in our society.

In our research project SysSon we will continue to evaluate our sonification designs with various groups (domain experts, sound experts, general), implementing the experience we learned from the experiments discussed in this paper.

7. ACKNOWLEDGMENTS

We would like to thank the students of musicology who completed the seminar ”Aesthetics in sonification” in winter term 2012/13 at University of Graz: Lukas Auer, Florian Eckl, Andreas Juwan, Eva Matlschweiger, Sigrun Mogel, and Jaqueline Wilfer, and the research assistant Sabrina Sattmann.

APPENDIX

Experiment	1/EG	1/CG	2/EG	2/CG
Tremor	3.68 (1.14)	3.48 (0.93)	3.61 (0.93)	3.18 (1.03)
Tweetscapes	4.18 (0.52)	–	3.66 (0.33)	–
VOSIS	2.78 (1.14)	2.53 (0.67)	2.58 (1.14)	2.10 (0.82)

Table 4: Ratings of Evaluation 1 and 2 for EG and CG (Mean (Std.Dev.)) on a scale from 1 (”not at all”) to 7 (”very much”).

Experiment	1/men	1/women	2/men	2/women
Tremor	4.03 (1.01)	3.14 (0.68)	3.66 (0.98)	3.13 (0.95)
Tweetscapes	4.28 (1.80)	4.08 (0.83)	3.82 (1.46)	3.50 (0.87)
VOSIS	–	–	–	–

Table 5: Ratings of Evaluation 1 and 2 for men and women (Mean (Std.Dev.)) on a scale from 1 (”not at all”) to 7 (”very much”).

Experiment	1/Musician	1/Non	2/Musician	2/Non
Tremor	–	–	–	–
Tweetscapes	4.08 (1.40)	3.46 (2.19)	3.76 (1.26)	2.80 (1.81)
VOSIS	3.77 (0.68)	2.38 (0.17)	2.79 (1.35)	2.38 (0.99)

Table 6: Ratings of Evaluation 1 and 2 for musicians and non-musicians (Mean (Std.Dev.)) on a scale from 1 (”not at all”) to 7 (”very much”).

Experiment	Neutral	Negative	Positive
Tremor	61%	36%	3%
Tweetscapes	66%	21%	13%
VOSIS	57%	40%	2%

Table 7: Subjective valuation of associated words (%).

8. REFERENCES

[1] T. L. Bonebright and J. H. Flowers, *The Sonification Handbook*. Logos Publishing House, Berlin, 2011, ch. 6: Evaluation of Auditory Display.

[2] J. G. Neuhoff, Ed., *Ecological Psychoacoustics*. Elsevier Academic Press, 2004.

[3] A. Schoon and A. Volmar, Eds., *Das geschulte Ohr*. Bielefeld: Transcript, 2012.

[4] [Online]. Available: http://sysson.kug.ac.at/

[5] S. Serafin, K. Frankovic, T. Hermann, G. Lemaitre, M. Rinott, and D. Rocchesso, *The Sonification Handbook*. Logos Publishing House, Berlin, 2011, ch. 5: Sonic Interaction Design.

[6] S. Barrass and P. Vickers, *The Sonification Handbook*. Logos Publishing House, Berlin, 2011, ch. 7: Sonification Design and Aesthetics.

[7] D. Pirrò, A. Wankhammer, P. Schwingenschuh, A. Sontacchi, and R. Höldrich, “Acoustic interface for tremor analysis,” in *Proc. of the International Conference on Auditory Display*, 2012.

[8] T. Hermann, A. V. Nehls, F. Eitel, T. Barri, and M. Gammel, “Tweetscapes - real-time sonification of twitter data streams for radio broadcasting,” in *Proc. of the International Conference on Auditory Display*, 2012.

[9] R. McGee, J. Dickinson, and G. Legrady, “Voice of sisyphus: An image sonification multimedia installation,” in *Proc. of the International Conference on Auditory Display*, 2012.

[10] [Online]. Available: http://iem.kug.ac.at/en/projects/workspace/2011/akustisches-interface-zur-tremor-analyse.html

[11] [Online]. Available: http://iem.kug.ac.at/en/projects/workspace/2011/akustisches-interface-zur-tremor-analyse.html

[12] [Online]. Available: http://www.imagesonification.com/