EVALUATING THE MAGNITUDE ESTIMATION APPROACH FOR DESIGNING SONIFICATION MAPPING TOPOLOGIES

Jamie Ferguson

Glasgow Interactive Systems Sections
University of Glasgow
Glasgow, G12 8RZ, Scotland
j.ferguson.4@research.gla.ac.uk

Stephen Brewster

Glasgow Interactive Systems Sections
University of Glasgow
Glasgow, G12 8RZ, Scotland
stephen.brewster@glasgow.ac.uk

ABSTRACT

A challenge in sonification design is mapping data parameters onto acoustic parameters in a way that aligns with a listener's mental model of how a given data parameter should sound. Studies have used the psychophysical scaling method of magnitude estimation to systematically evaluate how participants perceive mappings between data and sound parameters - giving data on perceived polarity and scale of the relationship between the data and sound parameters. As of yet, there has been little research investigating whether data-to-sound mappings that are designed based on results from these magnitude estimation experiments have any effect on users' performance in an applied auditory display task. This paper presents an experiment that compares data-to-sound mappings in which the mapping's polarity is based on results from a previous magnitude estimation experiment against mappings whose polarities are inverted. The experiment is based around a simple task in which participants need to rank WiFi networks based on how secure they are, where security is represented using an auditory display. Results suggest that for a simple auditory display like the one used here, whether or not the polarities of the data-to-sound mappings are based on magnitude estimation does not have a substantial effect on any objective performance measures gathered during the experiment. Finally, potential areas for future work are discussed that may continue to investigate the problems addressed by this paper.

1. INTRODUCTION

Parameter mapping is a technique for data sonification: "the use of non-speech audio to convey information" [1]. In a parameter mapping sonification system (commonly shortened to *PMSon* for parameter mapping sonification) data values are used to manipulate acoustic parameters which facilitates the communication of the data. One of the most fundamental design challenges during the development of a PMSon system is the mapping topology - the relationship between the data parameters and acoustic parameters. In their chapter on PMSon in *The Sonfication Handbook* [2], Grond & Berger posit that "effective PMson often involves some compromise between intuitive, pleasant and precise display characteristics". However, there is little theory or evidence to guide designers toward what is the most effective acoustic parameter to convey a particular data value. Negative consequences caused by a deficit in Grond & Berger's trio of necessary characteristics can be grave in high-stakes contexts. This has been seen in noted instances of nuclear control room operators, locomotive drivers and aircraft pilots turning off auditory displays due to sounding unpleasant, or the information that they intend to convey being misleading or false [3]. Therefore, the imperative to move towards designing parameter mappings with a balance of these three characteristics is clear, yet it remains an under investigated problem.

Walker proposed the use of *magnitude estimation* as a tool which could be used by sonification designers to aid in establishing what the most effective acoustic parameter would be to represent a particular value of data [4]. Magnitude estimation maps the relationship between a sensory stimulus and its associated perceived intensity [5]. Walker's method provides two psychophysical measurements: polarity and scale. Polarity is the directional aspect of the mapping (e.g. increasing or decreasing pitch mapped to increasing temperature). Scale defines the amount of change in the acoustic parameter for a given change in the data parameter (e.g. for an increase from 10 °C to 20 °C, increase pitch by 50 Hz).

Walker used polarity as a measure of the "naturalness" of a mapping - the more unanimously participants perceived the polarity of a given data-to-sound mapping, the more "natural" this mapping was and therefore more effective as a parameter mapping topology. This methodology has been used in more recent studies for a variety of types of data and acoustic parameters [6, 7] and even beyond data-to-sound mappings into data-to-vibration mappings [8]. These studies indicate that magnitude estimation is a useful predictor of the effectiveness of a data-to-sound mapping in that it tells the researcher how unanimous the mapping polarity is amongst participants. However, there has been no research investigating the extent to which using results from these experiments to influence the design of a data-to-sound mapping has in an actual sonification task. This paper describes an experiment which investigates the effect that using parameter mapping polarities based [7] has on the performance of a simple sonification task, when compared with using parameter mappings with arbitrary polarities - in this case, the inverse polarity. The goal of this study is to establish to what extent the data from magnitude estimation experiments are generalisable when used in actual sonification tasks and therefore, further understand how to use experimental methods

This work was funded by the EPSRC

and techniques to design parameter mappings for sonification.

2. RELATED WORK

2.1. A Brief Introduction to Parameter Mapping Sonification

There are three main aspects that must be considered when designing the mapping between a data parameter and an acoustic parameter. Firstly, there are psychophysical aspects: polarity and scale [2]. In addition to the psychophysical aspects, there are contextual factors. For example, the semiotics of the parameter mapping - what is the nature of the acoustic representation of the data? Kramer described a continuum for sound representation which ranges from analogic to symbolic, where analogic representations are more directly connected with the object (such as a Geiger counter) and symbolic representations are more abstract and indirect such as using pitch to represent temperature.

Walker & Kramer conducted the first study to investigate the effect that the choice of acoustic parameter(s) had on participant performance in a PMson system (originally presented in 1996 [9], published in 2005 [10]). They used of a number of acoustic parameters commonly used in sonification systems (pitch, onset, loudness and tempo) to convey simple data variables (temperature, pressure, size and rate) in a process-monitoring task. These parameters were split into four ensembles: *Intuitive*, *Okay*, *Bad* and *Random* based on how "natural" the designers believed a mapping to be. Results showed that the mappings which the sound designers believed to be optimal, e.g. temperature:pitch, did not result in either the most accurate or the fastest responses. Contrarily, mappings in the *Bad* ensemble yielded the fastest response time and *Random* led to the best performance.

Walker continued this line of research [4, 6], which investigated the use of the psychophysical method of magnitude estimation for systematically evaluating the perceived relationship between a data concept and an acoustic parameter in the context of sonification, with the goal of determining how "natural" a mapping is. Walker's studies obtained polarity and scale data for a number of data-to-sound mappings for basic data concepts and acoustic parameters such as temperature:pitch or pressure:tempo. Based on this data, Walker used the following criteria for evaluating a data-to-sound mapping: if a given polarity obtained a majority of all responses by participants in a block it was predicted to be a "good" polarity choice and it could therefore be predicted that the mapping itself was effective.

Present in the discussion section of these studies [10, 4, 6] is the importance of the listener's *mental model*. The "representation of some domain or situation that supports understanding, reasoning, and prediction" [11], or how they expect a data value to sound when it is sonified. A general assumption may be that an increase in a data value should be represented by an increase in an acoustic parameter, however findings from all of these previous works show that in many cases this is false. An example of this can be seen in the results from Walker's 2007 study [6] in which they found that when frequency was used to represent a value of size, more participants responded in a negative polarity than a positive. This suggests that these participants felt that "bigger" things are better represented by lower acoustic frequencies - aligning with a more physically-based mental model of sonification mapping, as larger things in the world often produce lower sounds.

Ferguson & Brewster conducted an experiment using the same magnitude estimation paradigm in which they investigated a number of additional data-to-sound mappings [7]. This study focused particularly on mappings in which both the data concept and the acoustic parameter are both generally deemed to be "undesirable" - attempting to further investigate the role of listeners' mental models in their determination of polarity and scale. The data concepts explored in this study were all semantically negative (*danger*, *error* and *stress*) and the acoustic parameters used to represent them (*roughness* and *noise*) would generally be considered undesirable from the standpoint of music or sound quality. Findings from this study suggested that for all of these mappings, the majority of participants perceived the mappings in a positive polarity - suggesting that for the data-to-sound mappings presented, an increase in a musically "undesirable" acoustic parameter like noise was perceived as conveying an increase in a semantically negative data variable such as stress. This again supports Walker & Kramer's assertion of the importance of the listener's *mental model* of the data being sonified being an important factor in how they expect a given data value to sound when it is sonified.

In a section entitled *Continuing Research Needs* in Walker's paper detailing initial investigations into using magnitude estimation for parameter mapping design [4], they posit that "the final test would always be instantiating these and other findings in more and varied sonification applications and systematically evaluating their effectiveness". However, this step has not yet been taken and the work detailed here aims to provide a starting point for this next stage of investigation into this method.

3. EXPERIMENT

An experiment was conducted to investigate if using mapping polarities based on results from a prior magnitude estimation experiment has any effect on performance during a simple auditory display task. In this experiment, a task consisting of ranking three WiFi networks based on their security level was used, in which the level of security for each network (low, medium, high) was conveyed using an auditory cue. The data-to-sound mappings and the polarities were based on results from Ferguson & Brewster [7], specifically here using the mappings of *danger:roughness* and *danger:noise* - danger in this context being how insecure or "dangerous" a WiFi network may be. In this magnitude estimation study, they found that when noise was used to represent danger, 13 of 15 participants perceived this mapping in a positive polarity (increasing noise = increasing danger). Similarly, when roughness was used, 12 of 13 participants responded in a positive polarity.

3.1. Participants

Twenty four participants took part in the study. Participants were: 12 female, 11 male, 1 non-binary, mean age = 28.2 years, SD = 6 years, 23 right-handed, 1 left-handed. All participants reported no uncorrected vision impairment and no hearing impairments.

3.2. Design

Eight conditions were investigated in which the independent variables were the acoustic parameter used and the polarity in which it was mapped to the security of the network. The polarity of each data-to-sound mapping was either based on results from [7], or were the inverse from the prior study, meaning the polarity was inverted such as increasing roughness = *increasing* danger would thus be inverted to increasing roughness = *decreasing* danger. For the sake of brevity, in this paper we will refer to all polarities based results from [7] as *aligned* and the others as *inverted*. The main dependent variable collected during the experiment include completion time, correctness of responses and NASA Task Load Index (TLX) [12]. The experiment used a within-subjects design. For this experiment, it was important to consider the order in which participants were presented with each polarity, as a participant may favour whichever polarity they were exposed to first, thus reducing the quality of the data gathered. Therefore counterbalancing was used to ensure that equal numbers of participants received each polarity first.

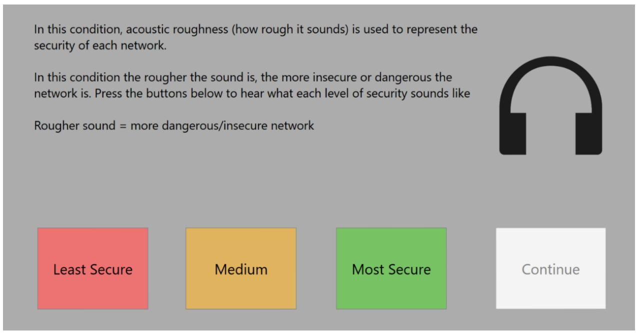
3.3. Stimuli

Roughness and noise were used as acoustic parameters in this study, based on the stimuli used in Ferguson & Brewster's magnitude estimation study [7]. These parameters were chosen in this prior work due to the effect of roughness on the perception of danger [13] and noise's effect on the perception of image focus [14] (with lack of focus or "bluriness" also being a semantically negative or "undesirable" data concept). This prior study used ten levels for each acoustic parameter, an example being the noise condition in that study containing ten sound cues ranging from a clean tone to total white noise. Results from another study that Ferguson & Brewster carried out using both acoustic roughness and noise to convey information suggested that participants found it difficult to interpret ten levels of these stimuli [14]. Therefore for the experiment described in this paper we reduced the number of levels to three to ensure that the task would be simple and the sound cues could easily interpreted. As in [7], each stimulus was 2 seconds in length. Each stimulus had an amplitude envelope with a 0.2 second linear ramp onset (attack) and offset (release). An amplitude envelope was included in the sound design, as an abrupt start or stop of a sound can be perceived as unpleasant [15]. All stimuli were created in the Supercollider programming language [1]. The acoustic design of each stimuli is described below.

- **Roughness**
  100 % sinusoidally amplitude modulated 1000 Hz pure-tone with modulation frequencies of 0, 11 and 70 Hz.
- **Noise**
  This condition consisted of a 1000 Hz pure tone for the first level, and equal blend of a pure tone and broadband white noise for the second level and the final level was solely broadband noise.

3.4. Procedure

The experiment consisted of four blocks - each acoustic parameter (roughness, noise) mapped in each polarity (*aligned*, *inverted*). Each block consisted of three trials. At the beginning of each condition, participants were presented with a screen which explained the acoustic parameter being used in that block and how it was mapped to each level of network security (Figure 1). In this screen, participants could use the three coloured buttons to hear the sound cues for each level of security for the given condition's data-to-sound mapping. Participants could not press the continue button until the button for each level of security was pressed at least once, however they could listen to each sound cue as many times as they needed. In each trial, participants were presented with a screen showing three WiFi networks, each with a button to play the sound cue to convey their level of security (Figure 2).


In this condition, acoustic roughness (how rough it sounds) is used to represent the security of each network.

In this condition the rougher the sound is, the more insecure or dangerous the network is. Press the buttons below to hear what each level of security sounds like

Rougher sound = more dangerous/insecure network

Figure 1: Condition introduction page where the data-to-sound mapping is explained, including the polarity. Three coloured buttons allow the participant to hear the sound for each level of security.

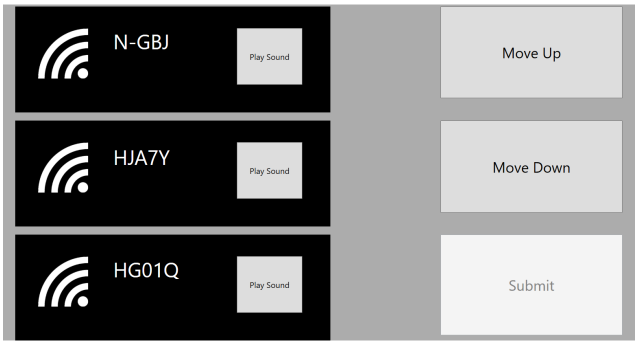


Figure 2: Screen for each trial showing three networks, each with a button to play their associated sound stimuli.

Sounds were presented using a pair of Beyerdynamic DT100 headphones. The participants were tasked with rearranging the networks based on their security (top = most secure, bottom = least secure) as conveyed by the sound cues. Similarly to the first screen (Figure 1) participants had to listen to each stimuli at least once before being able to submit their ranking, however there was no upper limit to how many times they could repeat each sound. For each condition participants completed three rankings, each time the ordering and name of the networks were randomised. The randomisation was implemented such that it was ensured that the network ordering was always mixed - ensuring the ordering wasn't correct at the beginning of each trial - thus ensuring the participant

[1] http://supercollider.github.io

had to rearrange the ordering. At the beginning of the experiment, participants completed a practice condition to allow them to familiarise themselves with the experiment. This practice condition was identical in procedure to all the subsequent conditions, except participants were required to rank the networks based on colour rather than using sound cues. Sound was omitted from the practice to ensure that participants were not influenced in any way which may have an effect their future responses. At the end of each condition, participants completed the NASA Task Load Index.

4. RESULTS

Firstly, since each ranking was completed three times, the mean completion time was calculated for each participant, for each sound/polarity combination. No statistically significant difference was found between the completion times for both polarities in the roughness ($F_{1,46}$ = 1.6, *p* = 0.2) and noise ($F_{1,46}$ = 0.02, *p* = 0.9) conditions. Of all 288 ranking trials completed, only 10 were ranked in an incorrect order. In this case, it is of interest to attempt to gain a more thorough insight into how little polarity choice impacted the participants' completion time. For example, it may be useful for an auditory interface designer to understand if carrying out a magnitude estimation experiment to evaluate a particular data-to-sound mapping is necessary for their particular use-case. Therefore, we calculated effect sizes (using the recommendations set out by the Transparent Statistics in Human-Computer Interaction Group [16]). Looking at the results for the roughness conditions, a Welch's t-test shows that the estimated difference in the means between the *aligned* and *inverted* polarities is -1774ms (95% CI: [-4601, 1053]). Cohen's *d* = 0.36. For the results for the noise conditions, a Welch's t-test shows that the estimated differences in the means between the *aligned* and *inverted* polarities is -149ms (95% CI: [-2505, 2206]). Cohen's *d* = 0.03. A Wilcoxon signed rank test found no statistically significant differences between NASA Task Load Index workloads for both polarities in the roughness and noise conditions (Table 1 shows TLX results).

Condition	Polarity	Mean (Workload)	SD (Workload)
Roughness	*Aligned*	31.1	13.6
Roughness	*Inverted*	32.2	12
Noise	*Aligned*	25.8	9.7
Noise	*Inverted*	30.1	15.7

Table 1: Summary of NASA Task Load Index results including mean workload and standard deviations

5. ANALYSIS

The small effect sizes and estimated differences between polarities for both acoustic parameters is surprising, as it would be a reasonable expectation *a priori* to assume that the mapping design in which the polarity is based on results from a previous magnitude experiment would result in a faster completion time. The estimated differences in the means for the roughness conditions is less than two seconds and less than a quarter of a second for the noise conditions. For many applications this very small difference may be acceptable, meaning that carrying out a magnitude estimation experiment to gather polarity data may not be necessary in some cases. In order to further explore this notion, we utilised the *Akaike Information Criterion* to investigate whether a model in which the completion times for both polarities are *equal* is more representative of the data gathered from this study, than a model in which the completion times for each polarity is assumed to be different. The following section provides an introduction to this method and describes its application to the results from this experiment.

5.1. A Brief Overview of the Akaike Information Criterion

As this method is uncommon in auditory display literature (with the notable exception of Frohmann et al. [17]) a brief overview of the process of using the method is given in this section. Hirotugu Akaike's information criterion (*AIC*) [18] is a method to estimate the relative quality of statistical models for a given data set. AIC estimates the amount of information lost when data is fitted to a given model, thus when comparing two potential models, the model with less information lost is more representative of the data in question. Null hypothesis testing cannot allow acceptance of the null hypothesis (in this case being "the completion times for mappings that are based either on polarity data from a prior experiment or inverted polarities are equal") however AIC can be used to reframe the question to be "is there more support for a model in which the completion times for both polarities are equivalent than one in which they are not". The following equation is used to estimate the AIC of a model [18, 19]:

$$AIC = -2log(\hat{L}) + 2k \tag{1}$$

where k is the degrees of freedom and $\hat{L}$ is the maximum value of the likelihood function of the model. To quantify the quality of each model, the raw AIC score must be converted to weighted scores. The first step is to calculate the differences in AIC for each model with the respect to the AIC of the best candidate model [19, 20]:

$$\Delta_i(AIC) = AIC_i - AIC_{min} \tag{2}$$

Where AIC_{min} is the minimum of the AIC values. This transformation causes the best model to have $\Delta_i(AIC) = 0$, while the rest of the models have positive values. The next step is to establish the relative likelihood L for each model i given the data

$$L(M_i|data) \propto exp\{-\frac{1}{2}\Delta_i(AIC)\} \tag{3}$$

where $\propto$ denotes "is proportional to". Finally, the relative likelihoods for each model are normalised to obtain weighted AIC scores for each model (w_i). Here each model's relative likelihood is divided by the sum of the likelihoods of all other models being compared, like so:

$$w_i(AIC) = \frac{exp\{-\frac{1}{2}\Delta_i(AIC)\}}{\sum\limits_{k=1}^{k} exp\{-\frac{1}{2}\Delta_k(AIC)\}} \tag{4}$$

Finally, the weighted AIC scores can infer the best fitting model. For example, if two models: A and B are being compared, with weighted AIC scores of $w_a(AIC) = 0.6094$ and $w_b(AIC) = 0.2242$, their weighted AIC scores would be used to show that model A is around 2.7 times more likely to be a better fit for the data than model B [2]:

$$\frac{w_a(AIC)}{w_b(AIC)} = \frac{0.6094}{0.2242} \approx 2.7$$

[2] example taken from [21].

5.2. Applying AIC to the Current Experiment

As discussed in the prior sections, we use the Akaike Information Criterion to answer the question:

> *"Is there more support for a model in which the completion times for both polarities are equivalent than one in which they are not?"*

Firstly, we fit two models: a linear model in which the completion times for both polarities is forced to be equivalent (effectively treating the data as if there were only one polarity category) - henceforth written as *Equivalent Model*, and a linear model in which they are assumed to be not equal - here named *Unequal Model*. By using the process described in the previous section, the quality of these models can be compared. Table 2 shows the results of the AIC analysis.

6. DISCUSSION

From the AIC analysis results in Table 2 we can see that for both roughness and noise, the *equivalent model* i.e. the one in which the completion times for each polarity are assumed to be equivalent is the best model for the given data (1.2 and 2.7 times more so for roughness and noise respectively). The AIC results in combination with the small effect sizes reported earlier support the argument that for this task, the polarity of the data-to-sound mapping did not have a substantial effect on the time it took participants to complete the task, with the estimated effect being $\sim$1.8s for roughness and $\sim$0.15s for noise. Furthermore, the NASA TLX results also suggest similar levels of workload in both polarities.

These results are surprising as the previous magnitude estimation study [7] showed that nearly all the tested participants perceived *increasing* noise or roughness as *increasing* danger, therefore it was expected that the mappings used in this study that were based on this would result in faster completion times, however this was not found to be true. This suggests that for *simple* auditory displays using roughness or noise such as the application used in the experiment here, the polarity in which the data is mapped to the acoustic parameter does not have a substantial effect. This means that for designers working in a similar space, the expenditure of resources to carry out a magnitude estimation experiment to establish polarities may not be necessary if the design can afford the potential discrepancies in completion time as discussed earlier.

7. LIMITATIONS AND FUTURE WORK

The generalisability of data-to-sound mapping polarities obtained from magnitude estimation studies like [4, 6, 7] has yet be fully determined. This study provided insight into the how generalisable polarities obtained for two data-to-sound mappings: *danger:roughness* and *danger:noise* [7], but it is only a first step toward understanding how generalisable data from these magnitude estimation experiments are in practice. The following sections discusses some limitations of the current study and puts forward potential future work that may address them.

7.1. Difficulty of The Task

The primary limitation of this study is that participants could potentially work through a "bad" data-to-sound mapping, because the task was relatively simple - 96.5 % of rankings were completed correctly. For example, even if a participant thinks that a more natural representation of increasing danger for them is using increasing roughness to convey this increase, they may still be able to complete the task in a fairly quick amount of time using a conflicting representation (i.e. increasing danger conveyed by *decreasing* roughness) due to the relative easiness of the task. The task was intentionally designed to be easy to carry out - both to account for participants who may be new to the notion of an auditory display and so that we could begin investigating this area with a simple auditory display. Many situations where auditory displays are commonly confronted by most people are quite simple such as mobile phone notifications or in-car displays etc. so we wanted to reflect that in this study before moving onto more complex sonifications. We intend to carry out a similar study with a more complex task by using a similar auditory display but in a more cognitively demanding and potentially more ecologically valid situation - again, to attempt to reflect the fact that many auditory displays are specific in context and complex, such as used by aircraft pilots or process-monitors.

7.2. Specificity of Context

This experiment focused solely data-to-sound mappings conveying danger - specifically the danger posed by an insecure WiFi network. The previous magnitude estimation study [7] presented danger in general and contextually agnostic terms, therefore it may be useful for future works attempting to investigate the generalisability of polarities gathered from magnitude estimations to evaluate multiple contexts for a given data-to-sound mapping. For example, a sonification of a value of danger in terms of WiFi security may be perceived vastly differently than a much more severe context such as a process-monitoring sonfication system in a nuclear power station. Therefore evaluating a broader range of contexts may afford a more well-rounded view of how generally polarities and scales from magnitude estimation experiments may be applied.

8. CONCLUSIONS

The research presented in this paper presents a first attempt to investigate the effect of designing data-to-sound parameter mapping polarities based on data from a magnitude estimation experiment. We presented a study in which we compared the time it took participants' to complete an auditory display based ranking task using two data-to-sound mappings in a simple auditory display task: one mapping in which the the polarity was based on results from a previous magnitude estimation experiment and one mapping in which the polarity was arbitrarily designed - in this case inverted. Based on results from this experiment we used the Akaike Information Criterion to discuss statistically that the polarity of the data-to-sound mappings did not have a substantial effect on the time it took participants to complete a ranking. Finally, we discussed some limitations of this study and suggest some future work which may address them. This work represents a first step toward researching how data obtained from magnitude estimation experiments can be appropriately applied in real-world sonification tasks and results from this study underline the need for further research in this area.

Condition	Model	DoF	$\log(\hat{L})$	AIC	ΔAIC	$w_i(AIC)$	$\frac{w_a(AIC)}{w_b(AIC)}$
Roughness	*Equivalent*	2	-475	954.8	0	0.5449	1.2
Roughness	*Unequal*	3	-474	955.1	0.3601	0.4551	
Noise	*Equivalent*	2	-466	935.6	0	0.7294	2.7
Noise	*Unequal*	3	-466	937.6	1.983	0.2706	

Table 2: Summary of results from AIC analysis.

9. ACKNOWLEDGMENTS

The authors would like to thank David Borchers for their suggestions regarding statistics during the development of this work, in addition to the reviewers for their insightful and useful comments.

10. REFERENCES

[1] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, *et al.*, “The sonification report: Status of the field and research agenda. report prepared for the national science foundation by members of the international community for auditory display,” *International Community for Auditory Display (ICAD), Santa Fe, NM*, 1999.

[2] F. Grond and J. Berger, “Parameter mapping sonification,” *The sonification handbook*, pp. 363–397, 2011.

[3] R. D. Sorkin, “Why are people turning off our alarms?” *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 1107–1108, 1988.

[4] B. N. Walker, “Magnitude estimation of conceptual data dimensions for use in sonification,” *Journal of Experimental Psychology: Applied*, vol. 8, no. 4, pp. 211–221, 2002.

[5] R. Teghtsoonian, S. Stevens, and G. Stevens, “Psychophysics: Introduction to its perceptual, neural, and social prospects,” *The American Journal of Psychology*, vol. 88, no. 4, p. 677, 1975.

[6] B. N. Walker, “Consistency of magnitude estimations with conceptual data dimensions used for sonification,” *Applied Cognitive Psychology*, vol. 21, no. 5, pp. 579–599, 2007.

[7] J. Ferguson and S. A. Brewster, “Investigating perceptual congruence between data and display dimensions in sonification,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 611.

[8] J. Ferguson, J. Williamson, and S. Brewster, “Evaluating mapping designs for conveying data through tactons,” in *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. ACM, 2018, pp. 215–223.

[9] B. N. Walker and G. Kramer, “Mappings and metaphors in auditory displays: An experimental assessment,” in *ICAD 1996, Proceedings of the International Conference on Auditory Display*. Georgia Institute of Technology, 1996.

[10] ——, “Mappings and metaphors in auditory displays: An experimental assessment,” *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 4, pp. 407–412, 2005.

[11] D. Gentner, “Mental models, psychology of,” in *International Encyclopedia of the Social & Behavioral Sciences*. Pergamon, 2001, pp. 9683 – 9687.

[12] S. G. Hart and L. E. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[13] L. H. Arnal, A. Flinker, A. Kleinschmidt, A.-L. Giraud, and D. Poeppel, “Human screams occupy a privileged niche in the communication soundscape,” *Current Biology*, vol. 25, no. 15, pp. 2051–2056, 2015.

[14] J. Ferguson and S. A. Brewster, “Evaluation of psychoacoustic sound parameters for sonification,” in *Proceedings of the 19th ACM International Conference on Multimodal Interaction-ICMI 2017*. ACM Press, 2017, pp. 120–127.

[15] P. Bergman, A. Sköld, D. Västfjäll, and N. Fransson, “Perceptual and emotional categorization of sound,” *The Journal of the Acoustical Society of America*, vol. 126, no. 6, pp. 3156–3167, 2009.

[16] T. S. in HumanComputer Interaction Working Group, “Transparent Statistics Guidelines,” Feb 2019, (Available at https://transparentstats.github.io/guidelines).

[17] L. Frohmann, M. Weger, and R. Höldrich, “Recognizability and perceived urgency of bicycle bells.” Georgia Institute of Technology, 2018.

[18] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[19] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag, 2002.

[20] ——, “Multimodel inference: understanding aic and bic in model selection,” *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.

[21] E.-J. Wagenmakers and S. Farrell, “Aic model selection using akaike weights,” *Psychonomic bulletin & review*, vol. 11, no. 1, pp. 192–196, 2004.