Psychophysical Scaling of Sonification Mappings

Bruce N. Walker
Psychology Dept.
Rice University
6100 Main St.
Houston, TX 77005 USA
walkerb@rice.edu

Gregory Kramer
Clarity
310 NW Brynwood Lane
Portland, OR 97229
greg@metta.org

David M. Lane
Psychology Dept.
Rice University
6100 Main St.
Houston, TX 77005 USA
lane@rice.edu

ABSTRACT
We determined preferred data-to-display mappings by asking experiment participants directly and then examined the psychophysical scaling functions relating perceived data values to underlying acoustic parameters. Presently, we are extending and validating the scaled mappings in practical data interpretation tasks. The resulting scaling functions, in conjunction with the experimental paradigm developed here, should spark further research in this area and have implications for the design of future sonifications.

Keywords
Psychophysical scaling, auditory display, sonification, data-display mappings

INTRODUCTION
Our investigations focus on data sonification, or the representation of scientific data with sound [1]. In particular, this research begins to determine the best way to vary auditory display dimensions (i.e., sound parameters such as frequency or tempo) in order to communicate a data set to a listener for interpretation and analysis. Little research has been done to determine exactly how to map data to sounds, despite the finding [5] that this mapping procedure is not necessarily straightforward or intuitive. An effective and practical approach to sonification can only result from the determination of actual listener preferences and performance. The most rigorous method for this is experimentation and validation of sonification schemes through representative listening tasks.

IMPORTANT QUESTIONS FOR PRACTICAL AUDITORY DISPLAYS
To begin to create auditory displays with widespread utility, three basic questions must be answered. First, we need to know what auditory display dimensions best represent a given data dimension. Is it best to use frequency, tempo, or some other auditory parameter to represent the data dimension of, say, velocity? Second, what is the direction or polarity of the preferred data-to-display mappings? If frequency is used to represent velocity, one might postulate that increasing frequency would represent increasing velocity. However, when representing the dimension of size, increasing frequency might best represent *decreasing* size, since high-frequency sounds tend to come from small objects. And third, once we establish a mapping and polarity, what is the scaling factor for that data and display pair? That is, exactly how much change in frequency must we use to represent a given change in velocity?

Certainly there are many more questions, but these three are primary and critical, and as yet have not been systematically investigated. We begin by answering these questions, and then we begin to validate the resulting mapping solutions in a representative practical listening task.

PREVIOUS WORK
Mapping Choices
Very few of the sonification projects to date have systematically addressed the three questions we raise. Most systems rely on mappings, polarities, and scaling functions determined either by system constraints (e.g., only frequency may be varied) or by design decisions made by the system programmers (e.g., increasing frequency is always mapped to increasing data values). Unfortunately, the resulting solutions may not match listeners preferences or expectations. This may lead to errors or slower performance by users of the system [5]. Further, most sonifications have not taken into account the actual type of data being displayed, or the specific listening experience of the listeners.

Walker & Kramer, 1996
In the sonification of a fictitious Crystal Factory, Walker and Kramer [5] showed that the specific choices of both mapping and polarity can affect reaction time and accuracy in monitoring tasks. Three different sound designers agreed about what ought to be the best mappings, however the data showed that such intuitive design decisions may not result in the best actual performance.

Psychophysical Scaling

There is a vast literature of auditory perception research, much of it important for sonification. However, one particularly relevant line of research is the work of S. S. Stevens and others [3,4] in determining the psychophysical scaling functions between acoustic parameters and subjective perceptions of a stimulus. For example, if the physical amplitude is doubled, what is the resulting change in the perceived loudness?

Here we adapt the psychophysical scaling paradigm known as magnitude estimation [4] to examine how changes in the physical sound attributes (e.g., frequency) result in different estimations of the data that the sound is supposed to represent. That is, if temperature is mapped to the frequency of the sound, then what change in *temperature* will the listener report when the frequency is doubled? Is it simply the same as the change they report for *pitch* when frequency is doubled? Is it the same change they would report for pressure or velocity or the value of the dollar when the frequency is doubled? The question thus simplifies to whether it matters what data dimension the sound is supposed to represent, and if so, how.

EXPERIMENT 1: MAPPING PREFERENCES

To address the question of which display dimension best represents a given data dimension, we sought a straightforward approach that would be very easy to apply by designers of new auditory displays. We simply presented pairs of sounds that differed only along a single auditory dimension at a time, and asked listeners to indicate which sound best represented a particular data dimension.

Apparatus and Procedure

Listeners sat in front of a 17-in. Macintosh computer monitor in a sound-attenuated room, listening to sounds via Sony MDR-V200 headphones, and responded using the mouse.

Sounds

The sounds were presynthesized at 16-bit, 44.1 kHz using Csound. They were composed of a one-beat long pure sine wave tone, followed by a half-beat of silence. These elements were looped to create a continuous pattern. The sounds were synthesized with all nine combinations of the auditory dimensions of frequency (400, 1000, and 2400 Hz) and tempo (60, 210, and 420 beats per minute or bpm; at 60 bpm, one beat lasts one second). The amplitude envelope of the tones included a 0.1-second linear ramp onset and offset. We normalized the stimuli for loudness, starting with relative amplitude values from equal-loudness contours [2] then making minor adjustments based on pretesting.

Word Cues

We chose four data dimensions that were both widely known and commonly used in a variety of scientific fields (temperature, pressure, velocity, and size). For each dimension, matching cue questions were created to cover each end of the cue dimension. An example pair of cues would be: Which sound best represents something with a hotter temperature? and Which sound best represents something with a colder temperature?

Trial Structure and Task

For maximum transportability the experiments were written in HTML and Javascript, and run with Netscape 4.6. On each trial a cue question was centered near the top of the screen, with two 1-inch squares centered just below, separated by two inches horizontally. While the cursor was located over square A, sound A would play. While the cursor was over square B, sound B would play. On a given trial, the two sounds would differ only along only one of the dimensions (e.g., frequency). The participant would listen to each of the sounds, then simply click on the square whose sound they felt best answered the cue question. During the half-hour session, all sounds differing only in frequency (nine pairs) and all sounds differing only in tempo (nine pairs), were combined with all eight word cues, for 144 total trials presented in random order.

Participants

Eleven Rice undergraduate psychology students (2 male, 9 female) participated for partial course credit.

Results

For each cue (data) and sound (display) dimension pair, a signed preference score was determined for each subject. If, for example, the higher-frequency sounds of a pair were always judged better for representing hot, and the lower-frequency sounds were always judged better for representing cold, then the individual preference score would be +18. If, on the other hand, the higher frequencies always matched cold, and the lower frequencies always matched hot, then there would be perfect agreement, but with the opposite polarity (i.e., a score of −18). Scores in between +/− 18 would indicate a less-consistent or weaker preference. Average preference scores across subjects provided a measure of how effective a mapping was, and in what polarity it was preferred. Figure 1 shows the individual preference scores for each data and display dimension for the 11 subjects.

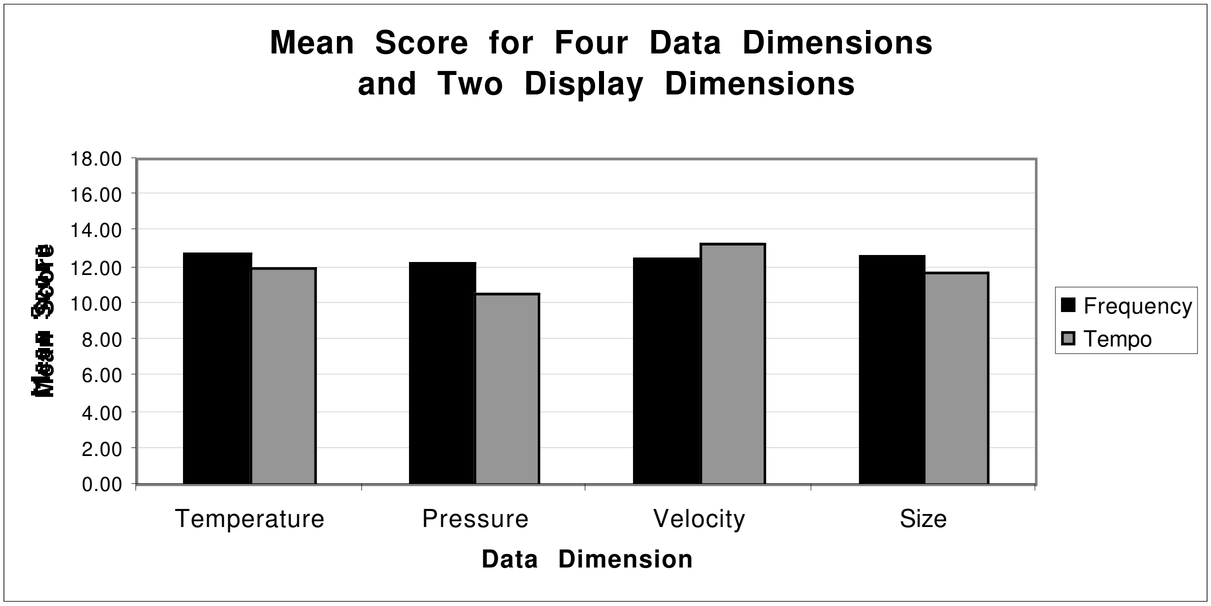

Figure 1. Mean preference scores from Experiment 1, across subjects. Black bars represent sounds differing only in frequency; gray bars represent sounds differing only in tempo. None of the differences reached statistical significance.

Discussion of Experiment 1

The preference scores in this experiment were meant to indicate which display dimensions are preferred for representing the data dimensions. However, it seems that most subjects just determined a mapping polarity and applied it consistently across all data dimensions. In debriefing the listeners, many reported simply trying to be consistent and get a perfect score. Thus this simple paradigm does not seem to be as effective as we had hoped in determining preferences. For example, increasing pitch should map to decreasing size for at least some listeners (which would make the mean preference score lower, or even negative). It was clear that we required a more sophisticated, and perhaps less transparent, experimental paradigm.

EXPERIMENT 2: SCALING THE DIMENSIONS

In addition to discovering mapping polarity preferences, we also need a way to determine what the scaling function is between the physical sound attributes and the data dimensions they are meant to represent. We turned to the psychophysical scaling paradigm of magnitude estimation [4] to assess directly both the polarity and scaling function issues. We hoped that our cognitive twist on this well-established paradigm would be less transparent than Experiment 1, so it would be less likely that listeners would respond in a particular pattern throughout the study (i.e., attempting to attain some perfect score).

Procedure

Participants and Stimuli

From the same subject pool, 132 different listeners participated. Two sets of sound stimuli were created as in Experiment 1. One set consisted of 10 stimuli that were all at a frequency of 1000 Hz, but varied along the tempo dimension (45, 60, 105, 150, 210, 270, 420, 500, 550, and 600 bpm). The second set of 10 stimuli were all at the same tempo (60 bpm), but varied along the frequency dimension (100, 200, 300, 400, 800, 1000, 1400, 1800, 2400, and 3200 Hz).

Each participant completed two blocks of trials, separated by a brief rest. In one block they heard only the sounds from stimulus set 1, and in the other block they heard only the sounds from set 2. Block order was counterbalanced across subjects.

Trial Structure and Task

Participants read instructions like the following:

> You will hear a series of sounds, one at a time in random order. Your task is to indicate what pressure they represent, by assigning numbers to them. For the first stimulus, assign it any number of your choosing that represents a pressure. Then, for each of the remaining sounds, indicate its pressure, relative to the first sound. For example, if the second sound represents a pressure that is 10 times the first pressure, assign it a number that is 10 times the number you assigned to the first sound. If the second sound represents a pressure that is one fifth that of the first sound, assign it a number that is one fifth the first number; and so on. You may use any range of numbers, decimals, or fractions that seem appropriate, as long as they are greater than zero.

On each trial the participant heard one sound, and provided a number for that sound's pressure in a text-entry box on the screen. Listeners simply filled in their responses and clicked a Next button to proceed to the next trial.

In a block of 20 trials, each sound was randomly presented twice, with the constraint that the highest or lowest frequency or tempo could not occur first, and subjects only responded to one data dimension per block. Following a brief rest, the experimenter introduced the second block with new instructions that indicated a different data dimension (e.g., size), but were otherwise identical. Thus, for example, a particular participant might first listen to all the stimuli that differed in frequency and make judgments about temperature, and then in the second block listen to all the stimuli that differed only in tempo and make judgments about the size that the sounds would represent.

Results

Individual Data

For each subject, we kept the responses from each block (i.e., each data/display dimension pair) separate. We observed that within a block most of the listeners responded with a particular mapping polarity, and despite the random order of trials, produced a set of responses that either increased or decreased monotonically with the stimulus dimension. When plotted on log-log axes, the typical results fall along a straight line. This is entirely in accordance with the findings generally reported for psychophysical judgments [4], with the important difference that in our experiment the listeners were making *conceptual* judgments.

We analyzed each participant's data as follows. First we calculated the geometric mean of the two responses to each stimulus (the use of the geometric mean is justified in [4]). We then log-transformed the responses and the stimulus values (e.g., the frequencies), and calculated the Pearson correlation coefficient between the stimulus values and the geometric mean of the responses. A large absolute value of the correlation coefficient meant that the participant had responded in a consistent and generally linear fashion. We decided to remove the data of those few whose correlation coefficients did not reach conventional levels of statistical significance (i.e., for 10 data points and $\alpha = .05$, $r_{\text{critical}} = 0.44$).

Some listeners obtained highly significant, but negative correlation coefficients for some blocks. Positive correlations indicated that the listener had assigned increasing data values to increasing display values (i.e., increasing temperature mapped onto increasing frequency). Negative correlations, on the other hand, indicated that increasing data values were mapped to *decreasing* display values (e.g., increasing size was represented by decreasing frequency). Some listeners obtained a highly significant positive correlation in one block, and obtained a highly significant negative correlation in their other block. In debriefing some subjects, it was clear that this was intentional. For example, one reported, It seemed obvious to me that in the first part [of the experiment] hotter things were faster [in tempo], but in the second part bigger things make deeper [lower frequency] sounds. It appears, then, that this paradigm provided a better way to investigate mapping polarities. In terms of our data, we considered the actual number of participants using a positive versus a negative polarity for a given data-display dimension pair as a good and simple indication of the popularity or general preference of a mapping direction. These values are included in Table 1.

Demographic Variables

In addition to the magnitude estimation data, we recorded age, sex, handedness, and number of years of musical training for each participant. We calculated correlation coefficients between these four variables and the slopes for each listener. None of the demographic correlations reached statistical significance for this group of listeners.

Dimension Scaling: Aggregate Slopes

The next measure of interest was the actual scaling of the data dimensions. For this we sorted the total data set by display dimension (frequency or tempo), data dimension (temperature, pressure, velocity, and size), and mapping polarity (positive or negative correlations). For each subset of data, the geometric mean of all responses was calculated for each stimulus (across subjects). These ratings of pressure, velocity, etc., were plotted against the actual values of the physical parameter that was varied (e.g., frequency) on log-log axes (see Figure 2 for an example). These graphs were all very linear, with high R^2 values. The slope of each of these graphs, as indicated by the exponent of the x-term in the power-fit equation, provides the psychophysical scaling function between the data and display dimensions. For example, if the slope of the temperature-frequency graph is +0.7, then a ten-fold increase in frequency represents only a seven-fold increase in temperature. If the slope were -0.8, then a ten-fold increase in frequency would represent an eight-fold *decrease* in temperature.

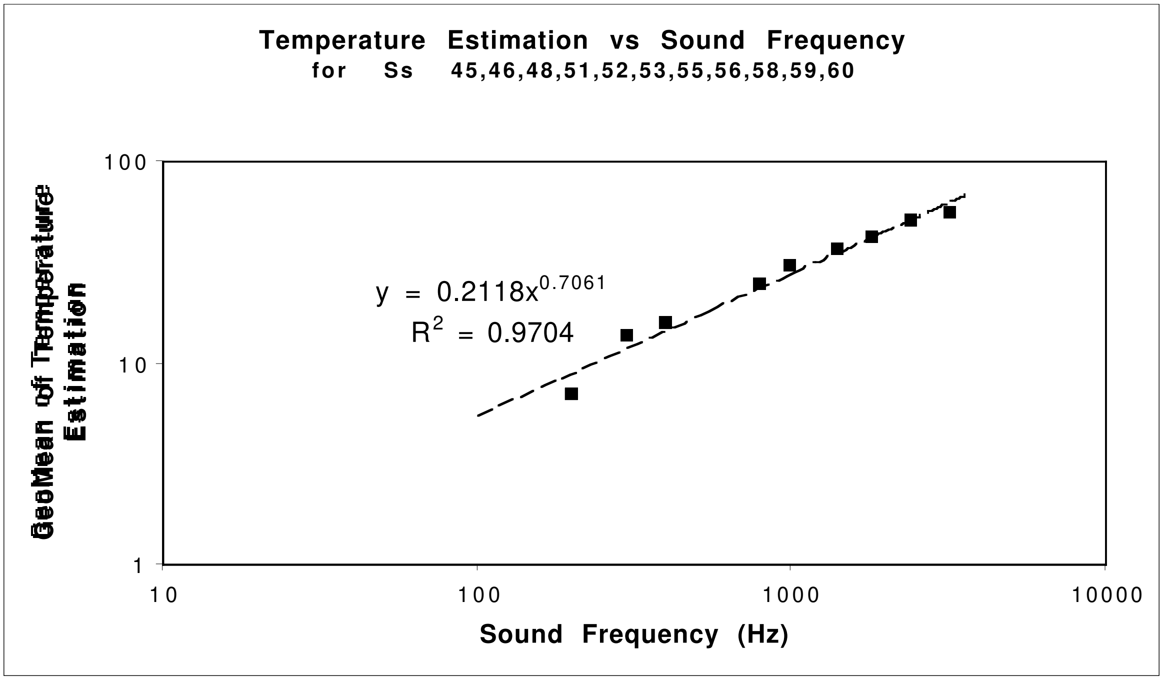

Figure 2. Geometric mean temperature estimations made for sounds of different frequencies in Experiment 2. The line through the data is a power fit; the slope is represented by the exponent of x in the equation, in this case, 0.71 (s.d.= 0.02). If the frequency is increased by a factor of 10, the perceived temperature will only increase by a factor of 7.1.

Table 1. Summary of slopes of graphs of data-display pairs. Slopes represent the psychophysical scaling function between the data and display dimensions. Positive slopes (listed first) indicate that increasing data values map to increasing display values. Negative slopes indicate increasing data values map to decreasing display values. Slopes in bold indicate the most popular polarity. The number of subjects contributing to each slope is indicated in parentheses.

Display Dimension: Frequency				**Display Dimension: Tempo**			
Data Dimension	**(N)**	**Slope (b)**	**SE(b)**	**Data Dimension**	**(N)**	**Slope (b)**	**SE(b)**
Temperature	(11)	**0.95**	0.12	Temperature	(11)	**0.43**	0.04
	(2)	-0.69	0.04		(6)	-0.48	0.05
Pressure	(8)	**0.78**	0.05	Pressure	(10)	**0.68**	0.06
	(4)	-0.49	0.08		(5)	-0.72	0.06
Velocity	(14)	**1.06**	0.03	Velocity	(11)	**1.04**	0.06
	(2)	-0.17	0.03		none		
Size	(7)	0.90	0.05	Size	none		
	(12)	**-0.76**	0.08		(16)	**-0.94**	0.07

Discussion of Experiment 2

Table 1 summarizes the data-display pairs and the obtained slopes, plus the numbers of listeners whose data contributed to the slope. The fact that these slopes are different for each of the data dimensions, and that they are different from 1.0, is an extremely important and interesting result. Thus it really does matter how the participant expects to interpret the stimuli, and this paradigm is sophisticated enough to distinguish user preferences. It also provides the transfer function to be used to convert changes in a data dimension (e.g., pressure or velocity) into the appropriate changes in the display dimension (e.g., tempo or frequency). This is the key to effective data representation. These scaling functions provide the first experimentally determined, scaled sonification mappings.

With the results of Experiment 2, we now have the beginnings of a set of transfer functions to provide the basis for an effective sonification tool. Data can be translated into the appropriate sound dimensions for optimal comprehension. Of course, these findings are for a relatively small sample of listeners from a homogeneous cultural group. Using this paradigm, these results will be replicated with new listeners with widely different musical and cultural experience.

The fact that significant numbers of listeners responded with different polarities for the same data-display pair indicates that it does matter what the sounds are supposed to represent. Further, while some polarities could be guessed (e.g., increasing frequency mapped to increasing temperature), there are clearly others that must be determined experimentally (e.g., increasing size to decreasing tempo). And of course, the actual slopes of the lines cannot be guessed.

Many other interesting questions are raised by these findings. It remains to be seen how listeners would respond if they were instructed to use a certain polarity for a mapping, but otherwise free to respond. Also, one could ask whether participants would respond as systematically if they were asked to switch the polarity they used when responding the first time. Another whole area awaiting investigation is that of the range of stimuli used, as well as the choice of modulus, or starting response. Stevens [4] recommends against introducing a starting value (i.e., the listener may choose any number whatsoever for the first sound). However, it may matter if the listener is seeded with temperatures in different ranges. That is, if the instructions mention outside daily temperatures and ask the participant to call the first sound 70, a different scaling function might arise compared to temperatures in a nuclear reaction starting at 3000.

Finally, many other display dimensions are yet to be tested, as well as an unlimited number of data dimensions. We are continuing with work in this area.

EXPERIMENT 3: VALIDATION TASK

Experimentally derived mappings and scaling functions are presumably better than simply using what sounds good to design a sonification. However, the utility of the scaling functions still needs to be experimentally validated within a practical task. A sonification that takes advantage of the preferred mappings and scaling functions should result in better performance than a sonification that does not.

Procedure

Rice undergraduate participants are performing data analysis and interpretation tasks with weather data. This domain allows for instantaneous measurement (What temperature is it?), trend analysis (Is the wind speed increasing?), and integration of multiple data dimensions (What is the wind chill factor?). Accuracy and task completion speed can also be measured in this way. The use of weather data provides a common task that is of interest to all listeners. It also requires little or no training of subjects – virtually everyone is familiar enough with rain, wind and heat. In addition, the data dimensions used in meteorology (temperature, pressure, velocity, etc.) are the same dimensions commonly used in other sciences. This choice should again help to make the results easily transferable to different scientific domains. This work is currently underway.

We predict that performance on all tasks will be faster and more accurate when participants use the optimal mapping polarities and appropriate scaling functions. This will validate the experimental determination of sonification schemes.

GENERAL DISCUSSION

With the determination of optimal mappings and scaling functions, design of data sonifications need no longer rely on the intuitions and educated guesses of sound designers. More importantly, the development of a paradigm for quantitatively evaluating the countless possible mappings between data and display dimensions advances the scientific foundations of the field of auditory displays.

ACKNOWLEDGMENTS

NSF Grant IIS-9906818 supported this research.

REFERENCES

1. Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N.; Neuhoff, J., Bargar, R., Barrass, S., Berger, J., Evreinov, G., Fitch, W., Gr hn, M., Handel, S., Kaper, H., Levkowitz, H., Lodha, S., Shinn-Cunningham, B., Simoni, M., Tipei, S. *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. ICAD, Santa Fe, NM, 1999.

2. Robinson, D. W., & Dadson, R. S. A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics, 7*, May, 1956, 166-181.

3. Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.*), Handbook of Experimental Psychology*. Wiley, New York, 1951, 1-49.

4. Stevens, S. S. *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Wiley, New York, 1975.

5. Walker, B. N., & Kramer, G. Mappings and metaphors in auditory displays: An experimental assessment. In S. Frysinger & G. Kramer (Eds.), *Proceedings of the Third International Conference on Auditory Display, ICAD 96*, (Palo Alto, CA, November 1996), ICAD, Palo Alto, CA, 71-74.