RICE UNIVERSITY

Magnitude Estimation of Conceptual Data Dimensions for Use in Sonification

by

Bruce N. Walker

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

David M. Lane, Associate Professor, Chair
Psychology and Statistics

James R. Pomerantz, Professor
Psychology

Michael Byrne, Assistant Professor
Psychology

Arthur Gottschalk, Professor
Music

HOUSTON, TEXAS, USA
SEPTEMBER, 2000

ABSTRACT

Magnitude Estimation of Conceptual Data Dimensions
for Use in Sonification

by

Bruce N. Walker

Most data exploration tools are exclusively visual, failing to exploit the advantages of the human auditory system, and excluding students and researchers with visual disabilities. Sonification uses non-speech audio to create auditory graphs, which may address some limitations of visual graphs. However, almost no research has addressed how to create optimal sonifications.

Three key research questions are: (1) What is the best sound parameter to use to represent a given data type? (2) Should an increase in the sound dimension (e.g., rising frequency) represent an increase or a decrease in the data dimension? (3) How much change in the sound dimension will represent a given change in the data dimension?

Experiment 1 simply asked listeners which of two sounds represented something that was hotter, faster, etc. However, participants seemed not to make cognitive assessments of the sounds. I therefore proposed magnitude estimation (ME) as an alternative, less transparent, paradigm.

Experiment 2 used ME with visual stimuli (lines and filled circles), replicating previous findings for perceptual judgments (length of lines, size of circles). However, judgments of conceptual data dimensions (i.e., the temperature, pressure, or velocity a given stimulus would represent) yielded slopes different from the perceptual judgments, indicating that the type of data being represented influences value estimation.

Experiment 3 found similar results with auditory stimuli differing in frequency or tempo. Estimations of what temperature, pressure, velocity, size, or number of dollars a sound represented differed, indicating that both visual and auditory displays should be scaled according to the type of data being displayed.

Experiment 4 presented auditory graphs and asked which of two data descriptions the sounds represented. Data sets based on the equations determined in Experiment 3 were preferred, providing validation of those slope values. Results also supported the use of the unanimity of mapping polarities as a measure of a mapping's effectiveness.

Replication with different users and sounds is required to assess the reliability of the slopes. However, ME provides an excellent way to obtain a function relating conceptual data dimensions to display dimensions, which can be used to create more effective, appropriately scaled sonifications.

Acknowledgments

First and foremost, I would like to acknowledge the contribution of my advisor, David M. Lane, who granted me the freedom to venture into uncharted waters, while providing support and timely advice throughout the voyage, even in its uncertain moments. Also, I would like to thank Greg Kramer for his patient encouragement over all these years, as I stumbled around in the labyrinthine foundations of a still-nascent scientific field that owes much of its present form to his tireless efforts.

I am grateful for the considered thoughts of my committee members all throughout the process, and for their cooperative spirit and careful reading of my final document. Their comments and suggestions have sparked many ideas regarding both the current work and future projects.

My family has forever been a constant source of encouragement and moral support. I thank them all, and in particular my parents, for the countless practical favors and emotional boosts over the years. As I have lived so many miles from home for so many years, my close friends here in Houston (and many around the world) have been my second family, and have shared greatly in the practical and emotional efforts involved in this dissertation.

And I would like to acknowledge the support of the National Science Foundation, and in particular the efforts of Ephraim Glinnert, who by funding project IIS-9906818 also lent this young field yet more credibility at a critical time in its development.

Thank you all, in so many ways!

Bruce Walker

Table of Contents

List of Tables

List of Figures

Introduction

Project Background

Motivation

In virtually every science classroom and research laboratory, researchers gather, analyze, and attempt to interpret data. Fundamental pedagogical and investigative practices rely on determining patterns in data, and reaching conclusions based on those data patterns. Projects such as the Sloan Digital Sky Survey[1] and the Human Genome Project[2] are generating vast amounts of new data, pointing out the need for more powerful and easier-to-use tools for making sense of all this information. In many cases, the data sets are not only huge; they are also multidimensional and rapidly changing. Researchers must use all of the resources available to us, both technical and perceptual, to display and interpret our scientific results. Computers in the laboratory and classroom have made this data analysis task easier. There are many software tools available for data exploration and analysis, and often a great deal of student learning and many very promising scientific hypotheses result from interacting with and manipulating the data. However, most data exploration tools in widespread use are exclusively visual in nature, including graphing and plotting software, modeling programs, and 2 or 3D visualization

[1] http://www.sdss.org

software. These tools fail to exploit the excellent pattern recognition capabilities of the human auditory system, and they also continue to exclude students and researchers with visual disabilities.

What is Sonification?

Sonification is the use of non-speech audio to convey information such as that used in the interpretation of scientific results. Specifically, data sonification is “the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation” (Kramer, et al., 1999). That is, scientific data, of any sort, is used to change the parameters of a synthesized tone. In the simplest case, it is often helpful to think of sonification as the process of creating sophisticated “auditory graphs,” analogous to the visualizations produced by modern graphing applications. However, sonification can involve many more than the 2 or 3 dimensions of data typically found in visual graphs.

Auditory display is a somewhat broader term referring to the use of any type of sound to present or display information to a listener. This may include, but is certainly not limited to, data sonification per se. The focus of the present research is on data sonification in specific, but the findings should be of interest to the field of auditory display in general.

[2] http://www.ornl.gov/hgmis/

Why Use Sound?

There are three main reasons why the use of sound to display data is important. First, sonification provides scientists with a new and powerful display medium with multiple distinct advantages over existing visual displays. Findings from the rich history of research on auditory perception point to distinct advantages of the auditory system over the visual system when dealing with temporal patterns and changes in data (e.g., Hartmann, 1997; Moore, 1997; Sanders & McCormick, 1993), and when the visual system is busy with another task (e.g., Brooks, 1968; Cohen, 1994; Wickens, 1992; Wickens & Liu, 1988). The growing body of research on sonification clearly indicates that auditory representation of data can, indeed, capitalize on these benefits of sound: in the representation of temporal and high-dimensional data (Kramer, 1993; McCabe & Rangwalla, 1994); in data monitoring tasks where the eyes are busy (Fitch & Kramer, 1994; Walker & Kramer, 1996); and in high-stress or critical conditions where cross-modal correlations would be of value (Begault, Wenzel, Shrum, & Miller, 1996). Sonification used for radiation monitoring (e.g., the Geiger counter) is better than either visual or combined audio-visual displays (Tzelgov, Srebro, Henik, & Kushelevsky, 1987). Some recent major discoveries, such as the “quantum whistle” (Pereverzev, Loshak, Backhaus, Davis, & Packard, 1997), have demonstrated sonification can help researchers make discoveries even when all manner of visual displays have failed.

Second, sonification provides students and teachers with new pedagogical tools that are effective and engaging. Science students need to learn to take advantage of the benefits that sonification tools will provide to them as researchers. Using sound in the classroom or teaching laboratory can provide a richer, more dynamic and effective learning environment, while preparing students to use cutting-edge investigative technologies. Woolf (1992) has shown that students are more eager to use multimedia training systems, and as a result spend more time engaged in learning (see also, Atlas, Cornett, Lane, & Napier, 1997). Flowers has shown that students judge statistical properties equally well with auditory histograms, compared to either visual-only or multimodal histograms (Flowers, Buhman, & Turnage, 1997; Flowers & Hauer, 1992, 1993, 1995). Flowers recommends the development of sonification software to help students and teachers make full use of the auditory system in the classroom (Flowers, Buhman, & Turnage, 1996). Thus, not only is sonification a powerful research tool, but it is also a valuable pedagogical tool.

Third, for students or researchers with visual disabilities, simple and standardized sonification methods will provide immediate inroads into the data analysis and interpretation tasks so prevalent in modern science. While all researchers and students stand to gain from the effective use of sonification, even more significant gains may be realized when the display users are students or researchers with visual disabilities. Science is still largely a visual endeavor, certainly not because it needs to be, but because the tools for learning and conducting science are more available for the visual modality only. Some researchers have begun to look into the ways auditory displays can help blind students. For example, Michael Kress at the City University of New York has investigated the use of sound to teach calculus; John Gardner at Oregon State University has worked on providing mathematics and physics materials to blind students; Alastair Edwards at the University of York has used sounds to help blind algebra students; and researchers at Purdue University are investigating audiotactile diagrams for blind chemistry students. A solid theory of sonification design, and simple yet powerful methods for creating them, will give both sighted and visually-impaired colleagues a common reference point and provide the shared understanding necessary for collaborative learning and discovery.

Brief History of Sonification: A Story of Trial and Error

While it is clear that sonification may have many benefits, it is not at all clear how to design effective sonifications. Sound, in general, has been used for many years to display warnings and simple status messages (see Sorkin, 1987; Wickens, Gordon, & Liu, 1998). However, only fairly recently has there been any investigation of even this most basic use of auditory displays (Edworthy, Loxley, & Dennis, 1991; Patterson, 1982; Sorkin, 1988). Unfortunately, the auditory system, an excellent pattern recognition device (Bregman, 1990; McAdams & Bigand, 1993), has still been underutilized in the display of scientific data. There has been almost no scientific investigation into how best to design such displays.

Until recently, most of the extant auditory displays, other than caution and warning tones, have been used to monitor dynamic events or processes, capitalizing on “the listener's ability to detect small changes in auditory events or the user's need to have their eyes free for other tasks” (Kramer, et al., 1999). Forbes (1946) mapped an airplane's flight data, such as altitude, compass heading, and fuel level, onto attributes of a sound, such as pitch, loudness, and left-right stereo location. This early auditory display allowed pilots to perform well without visual instruments after only a brief training period. Quite sophisticated auditory displays have been developed more recently for monitoring models of a cola bottling factory (Gaver, Smith, & O'Shea, 1991) and a crystal factory (Walker & Kramer, 1996), and for monitoring multiple streams of patient data in an anesthesiologist's workstation (Fitch & Kramer, 1994).

In the education realm, Gaver and Smith (1990) developed a collaborative multimedia learning environment called SoundShark, which allowed students to learn about physics. The sounds, or auditory icons, indicated user interactions, ongoing modes and processes, and physical attributes of the “objects” in the virtual learningscape. As an example, the activation of “gravity” in the world was indicated by smooth, low-volume sounds. Distance between a user’s hand and an object was indicated by the amplitude and low-pass filtering effects on the sound that represented that object.

A few researchers have begun to use sonification to display static data for the purpose of exploration and analysis. Early work in auditory presentation of seismic data showed that subjects could successfully discriminate between earthquakes and bomb blasts (Speeth, 1961). Hayward (1994) has followed up on this work by including oil exploration and earthquake sonifications. By shifting the seismic data into the audible frequency range, and simply "listening to the earth sing," researchers have been afforded another powerful tool for efficiently exploring and presenting huge volumes of data.

Data sonification is becoming more accessible, with the ubiquity of multimedia computers and powerful data-manipulation software. It is also being used in a much broader range of sciences. Part of sonification's growing appeal is that it can be used to display highly complex and multidimensional data. For example, participants in Fitch and Kramer's (1994) auditory anesthesiologist's workstation monitored eight simultaneously changing patient variables, and performed significantly better with the auditory version than the visual display. Kramer has also described the sonification of five-dimensional financial data (Kramer, 1994a) and nine dimensional chaotic data (Kramer & Ellison, 1991). The list of data dimensions that have been successfully sonified is growing, and it seems that many more data types, in a wide range of scientific fields, are amenable to sonification. Some very recent research has used sonification to detect tumors in a medical application (Martins & Rangayyan, 1997), make discoveries in physics (Pereverzev, et al., 1997), and analyze structural data from concrete highway bridges

(Valenzuela, Sansalone, Krumhansl, & Streett, 1997). It is quickly becoming clear that we are only scratching the surface of possibilities for using sound to display and help interpret complex scientific data.

Problems with Past Sonification

While it certainly holds great promise, sonification is still a nascent field and almost no research has been done to determine how to design auditory data displays for maximum effectiveness. There is little theory and virtually no experimental evidence to guide sonification researchers and designers (though see Barrass, 1998, for a start). The result has been that designers have used whatever “sounded good” or “made sense” to them. Not infrequently researchers simply had to use whatever display dimensions they had available (in many cases, only pitch) to represent whatever data dimension their application needed. For example, McCabe and Rangwalla (1994) mapped blood pressure in a heart valve onto pitch, but mapped turbine pressure onto loudness. Fitch and Kramer (1994) mapped blood pressure onto pitch, and heart rate onto the tempo of a repeating sound. Kramer (1996) mapped heartwall pressure simultaneously onto pitch, loudness, and two timbre dimensions. However, Papp and Blattner (1994) mapped heart rate to either pitch or loudness, but *not* tempo. These are just a few examples that show how the same data concept, such as pressure or rate, can be mapped to different auditory dimensions not only by different designers, but even by the same researchers in different experiments!

Emerging Questions in Sonification

There are basic research issues that need to be addressed in advance of any successful applications. First, the effective development of sonification depends on there being at least some agreement among users about what sound attribute most clearly represents a given data dimension. The consensus in Western culture that high frequency sounds are associated with physically higher locations (i.e., high pitch maps to "up;" Mudd, 1963; Walker & Ehrenstein, 2000) is an example that indeed there is agreement about certain mappings. However, it remains to be determined if such agreement exists for other data and display dimensions. Hence, sonification researchers should not consider it entirely axiomatic that there be agreement (at all) about what constitutes a "good" mapping. Indeed, Kubik (1975) shows that there are cultural groups with very different conceptions of sounds (see, also, R. Walker, 1987).

Research Question 1

Nevertheless, assuming there is some agreement about mapping preferences, what are they? What is the best sound parameter to use to represent some data, say, temperature? I consider this to be Research Question 1. Follow-on questions would include whether there are gradations of "goodness." That is, there may be some mappings that are considered excellent, or very obvious; some that are considered to be acceptable; and some mappings that are agreed to be poor mappings. It is not at all clear if agreement of mapping acceptability would be binary (only a good or bad mapping), or if there might be "shades of gray," as it were.

Research Question 2

Next, Research Question 2 is what the best polarities are for those mappings. For example, listeners might agree that pitch should increase in order to represent increasing temperature (a "positive" mapping polarity), whereas pitch should *decrease* in order to represent increasing size (a "negative" polarity).

Research Question 3

Once a designer does decide on what sound dimension to use to represent the data, she will still need to know how much change is required in, say, the pitch of a sound, in order to convey a given change in, for example, temperature. This issue of appropriate psychophysical scaling of the dimensions, which I consider Research Question 3, is critical if sonifications are to be used to make accurate comparisons and absolute judgments.

Later Questions

There are many other interesting and important research questions that will have to be asked in order to bring sonification science up to the point of practical utility, but which are unfortunately beyond the scope of this dissertation. However, since consideration of these impending questions has helped to direct my research, I will mention some of them here.

One critical question that must be considered in the near future is whether population preferences or stereotypes translate into performance. That is, do listeners actually perform sonification-based tasks better with the mappings that are preferred (i.e., considered intuitively “better”) by the population group? It should be stressed here that the key measure of utility is in performance, and not necessarily in stated preferences. It is sometimes the case that users claim to prefer something that is actually not best for their performance (Andre & Wickens, 1995).

Then, assuming that there is some agreement within a group of people as to what constitutes a good or bad way to map data onto display dimensions, it remains to be determined whether agreement about “good” and “bad” mappings would be stable across different populations of listeners.

A final over-arching question in all of this is whether sonification is a practical solution for the needs of researchers and students. This question needs to be held in mind, but cannot be fairly evaluated until a well-designed, principled sonification software application can be created.

It should be added at this point that another basic and pressing need is for an understanding of sound metaphors in general. When a listener hears a sound, what does that sound automatically connote? What could it possibly be used to represent? How is it that some sounds immediately evoke a certain thought or feeling? How can we capitalize on this metaphorical use of sound? Answers to these questions, which are certainly tied into the questions already raised, would help sonification researchers design display systems that are easier to learn and use because they capitalize on listener expectancies. They may also shed light on fundamental cognitive processes, providing a concrete forum for linking perceptual input and meaning making.

Mappings and Metaphors: Walker and Kramer, 1996

Perhaps the only study intended specifically to address the issue of data-to-display mapping choices (Walker & Kramer, 1996; but see, also, Barrass, 1998) showed that it *does* matter which auditory dimension is used to display a given data dimension. In research conducted at Rice University, Walker and Kramer created a simulated crystal factory, where undergraduates monitored the data dimensions of temperature, pressure, size, and rate. Listeners used a sonification to simultaneously track all of the data values, and if, for example, the sound indicated that the temperature was increasing then the operator would respond by rapidly pressing a button to turn on a cooling fan. In this way, both reaction time and response accuracy could be measured.

The four data dimensions were represented by the perceived sound dimensions of loudness, pitch, tempo, or onset time (i.e., attack time, how quickly the sound reaches maximum amplitude), in mapping arrangements that differed for each experimental group. Clearly, even with only four data dimensions and four display dimensions there are many possible mapping ensembles. Therefore, based on their experience with sound and sonification design, the researchers chose one mapping that seemed to them (and to others around them) to be the “best” or most “intuitive.” As shown in Table 1, an increase in temperature was represented by an increase in pitch, pressure by onset time, size by loudness, and rate by tempo. Each of these mappings, forming the “Intuitive” ensemble as a whole, seemed like it was most natural, and therefore should produce the best performance. The researchers then picked a second ensemble that they felt would be an “okay” mapping, but probably not optimal. A third mapping ensemble was chosen that seemed like it would actually be “bad” or “counterintuitive.” Finally, a fourth mapping arrangement, denoted the “Random” ensemble, was chosen that happened to map the data dimensions to the display dimensions in such a way that over the set of four ensembles, each data dimension was mapped exactly once to each display dimension.

Table 1. Data-to-display dimension mappings used by Walker and Kramer, 1996.

Display dimension	**Data dimension**			
	Temperature	Pressure	Rate	Size
“Intuitive” ensemble	Pitch	Onset	Tempo	Loudness
“Okay” ensemble	Loudness	Pitch	Onset	Tempo
“Bad” ensemble	Onset	Tempo	Loudness	Pitch
“Random” ensemble	Tempo	Loudness	Pitch	Onset

The first question for Walker and Kramer was whether or not there was a “best” way to map these four common data dimensions onto this set of four auditory display dimensions, and whether or not the researchers’ design decisions would capture that best mapping arrangement. A summary of Walker and Kramer’s results for reaction time (RT) is presented in Figure 1 (the accuracy data yielded comparable results). To the surprise of the researchers, the mapping ensemble that resulted in the best performance was not the “Intuitive” ensemble, but rather the “Bad” ensemble! Even the “Random” ensemble outperformed the supposed best choice. It is clear from these results that what sounds “intuitive” to several experienced sound designers may not produce the best performance with average listeners. Of course, all mapping combinations were not tested, but from this research it seems quite clear that even experienced sonification designers who must rely on their own intuition, in the absence of any guiding theory, may not be making the best sonification design.

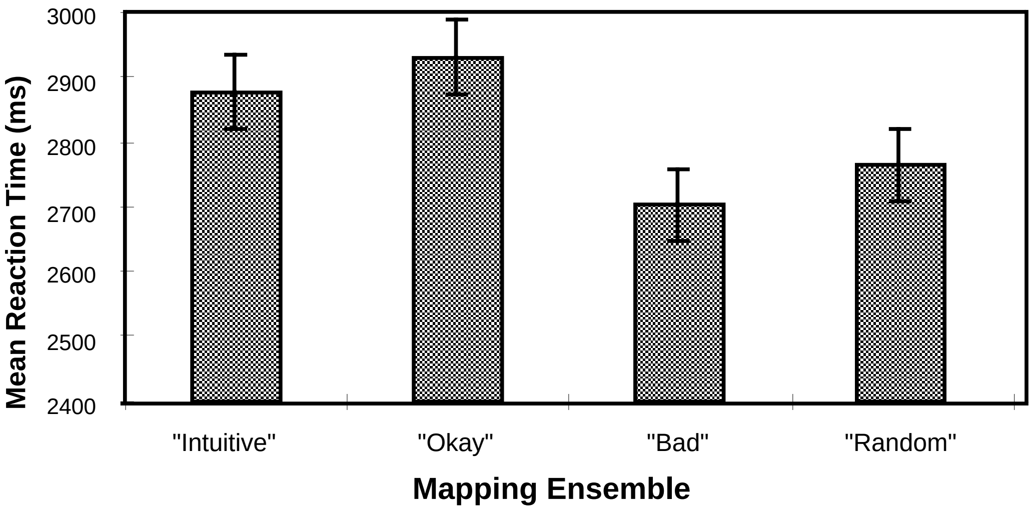


Figure 1. Reaction time for each of four mapping ensembles from Walker and Kramer, 1996. The mapping ensemble that was expected to be most intuitive led to poor performance, while the mapping that was specifically intended to be “bad” resulted in the best performance. This demonstrates the need for experimentally validated sonification design recommendations.

The second question for Walker and Kramer was whether there was a particular display dimension that was best for representing a given data type. That is, if one has to represent temperature, which of these four dimensions is best for the job? Taking into account both RT and accuracy, they looked at the performance with each mapping pair, across ensembles. Walker and Kramer summarized their results in the chart shown in Figure 2.

	Temperature	Pressure	Rate	Size
Pitch	~	~	√	~
Tempo	~	*x*	~	*x*
Loudness	√	~	~	~
Onset Time	*x*	*x*	*x*	√

Figure 2. Summary of best individual data-to-display dimension mappings, redrawn after Walker and Kramer, 1996. Checks indicate a strong mapping based on both RT and accuracy; x's indicate a poor mapping; tildes indicate a mapping that was okay, but neither the best nor worst.

Walker and Kramer initially considered that either pitch or loudness might simply be the best auditory dimension to use, regardless of the type of data, since listeners are likely to be most familiar with these auditory dimensions. That is, perhaps what really matters is how good a job the auditory display dimension will do, regardless of what it is supposed to represent. Indeed, both pitch and loudness were acceptable for representing all of the data types studied. Tempo was not as effective a display dimension, overall, and onset time was a fairly poor dimension. However, as seen in the final row of Figure 2, onset time was actually the best choice of display dimensions for representing size data, despite its poor performance with the other data types. Further, from Figure 2 one can see that pitch was not as effective as loudness for representing temperature, despite the common experience that hot things (e.g., a tea kettle) tend to make higher-pitched sounds as they become hotter. Finally, tempo, which would seem naturally suited to represent "rate" information, was only moderately successful in that role.

Thus, it is clear that the specific data-to-display mapping has a large impact on performance in a rapid-response sonification-monitoring task. Recall that the only difference between experimental groups was what the listeners were told a particular sound meant. They all heard the same sounds and made the same manual responses. Certainly, there are many open questions, here. It may be that tempo is, in fact, good for representing rate information, but just not in the range of tempos used in this experiment. Perhaps there is a particular reason why onset time is effective in representing size. For example, one colleague reasoned that a small water droplet falling into a bucket makes a "blip" sound with a fairly sudden onset, whereas a larger drop makes a "bloop" sound with a slower onset. Perhaps listeners were subconsciously capturing this metaphoric connection between onset time and size. Further investigation is needed to expand on these findings, in order to build a theoretical basis for sonification design decisions.

Apparently, while there are at least some preferred data-to-sound mappings about which most listeners will agree, these results demonstrate that even a team of sound designers might not know how the wider listening audience will respond—and certainly not without more research. It is simply not evident how common data values like "acidic," "salty," "smooth," or "high voltage" should sound. Less evident is how more subjective or affective information, such as "good" or "successful" should be sonified (see Kramer, 1994b).

Dimensions Used in the Present Research

In order to implement a successful sonification application, designers will need to know a great deal about the underlying psychophysical functions between the data they are representing and the sound dimensions being used to display them. However, I could not study all the possible sound dimensions, and certainly not all the possible data dimensions. It was critical, then, to make wise choices for the data and display dimensions to use here, in order to learn about dimensions that will have practical applications. And on the other side of the same coin, I will eventually want to test the mappings and scalings that are obtained, so for this project I chose dimensions for which an appropriate “proof-of-concept” task could be constructed in the future.

Display Dimensions

There are a great many possible parameters of a sound that can be varied, some gradually and systematically, and other not. My goal in this research was not to investigate auditory perception, per se, nor to redo the 100 years worth of psychophysics experiments that have been amassed (see, e.g., Falmagne, 1985; Marks, 1974; Stevens, 1975). Rather, I am treading into the largely uncharted waters of using sound to represent a *data-based* conceptual dimension.

To start with, the auditory display dimensions that I used, at least at the outset, needed to be both systematically variable and easily reproducible. That is, if the results of this research are to be eventually implemented, I had to study sound attributes that can actually be manipulated by applications programmers in a straightforward and standardized manner, without the need for a composer or sound designer at every step of the way (of course, this does not guarantee the sounds will be aesthetically pleasing, but more on that later).

In addition, the sound dimensions should be able to produce a large number of perceptually distinct values, or steps. These steps should lie in some stable order along the continuum. Consider the visual display dimension of saturation. It is a physical attribute of a visual stimulus, it can be varied either continuously or in many more-or-less-discrete steps, and the gradations in the resulting stimuli fall into an immediately recognizable and ordered series (see Figure 3). This allows viewers immediately to “see” the relationship between data elements, rather than having to “read” a legend in order to decode the stimuli (see Bertin, 1981).

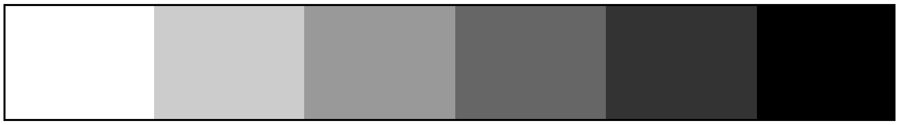


Figure 3. Illustration of the dimension of saturation with six distinct and immediately ordered levels.

On the other hand, the visual dimension of hue (i.e., different “colors”) is a poorer choice than saturation for the display of visual data. Colors can certainly be used to create (more or less) distinct or different perceptual categories. That is, for the most part we can tell red from blue. Thus, we could use colors to represent categorical data, like brands of toothpaste or different countries on a map. However, despite the common misperception resulting from memorizing the colors of the rainbow, colors do not lie along a perceptual continuum (although, wavelength and saturation do) (Bertin, 1981). That is, there is no reliable ordering of colors that immediately affords understanding. We can use the "rainbow sequence," but that requires people to look up the color order mentally before each comparison of the data.

With this in mind, I considered as candidates the auditory dimensions such as frequency, amplitude, tempo, onset time (how quickly the sound reaches maximum amplitude) and the ratio of high to low frequency components in a frequency-modulated (FM) sound. These are basically objective attributes of the sound wave itself, irrespective of the listener, and can be varied by a synthesis program in standard, systematic steps that can be objectively compared from one sound to another, and from one synthesizer to another. Further, these attributes can be considered as continuous variables. For example, we technically have countless many frequencies to use. These dimensions are also regarded as producing an ordered series of perceptual experiences. For example, increasing frequencies produce increases in the perception of pitch.[3] Of this short list I

[3] However, see, for example, Moore (1997) or Deutsch (1982) for more on the complex phenomenon of pitch perception. See, also, Kubik, 1975, for a culture with a different conception of pitch.

decided to begin with frequency and tempo, since they would be the easiest for other researchers to replicate.

Note that in discussing some of the display dimensions, some authors refer to the physical stimulus dimension (e.g., intensity or frequency), or the perceived value of these dimensions (loudness and pitch, respectively), in a somewhat interchangeable fashion, since quite well-known scaling functions exist between them. That is, authors might describe the temperature-to-pitch mapping or the temperature-to-frequency mapping, referring to the same relation. In particular, the auditory display literature has tended to include references to pitch and loudness, rather than to frequency and intensity, while the psychophysical literature has generally been more strict in using the physical dimensions where technically appropriate. At a functional level, they are essentially equivalent, if not so at a strict definitional level. For the sake of clarity and consistency, I generally use the physical (acoustic) parameter (e.g., frequency) when referring to the display dimension. The subjective impression of that acoustic parameter (e.g., pitch) is referred to as the "perceptual" dimension.

Data Dimensions

To obtain the maximum utility from this research, I decided to begin with data dimensions that are most commonly used in a variety of settings, across a broad range of investigative domains: temperature, pressure, velocity, and size. To this list I would add other common data types, in progressively more abstract steps, as time and resources allowed. The number of dollars, pleasantness, and urgency were other dimensions in line for investigation.

Experiment 1: Which Sound is Hotter?

Purpose

As described in the Introduction, the first task was to determine prevailing population stereotypes for the best data-to-display mappings and to determine their preferred polarities. That is, which display dimension do listeners feel best represents each of the data types, and how? The ultimate goal was a straightforward diagnostic tool that could be very easily used by designers of new auditory displays to poll their ultimate user population for mapping preferences.

Over the past years I have casually asked many people which of two hummed, whistled, tapped, or sung sounds seemed to represent something that is hotter, colder, faster, bigger, and so on. While there is not universal agreement, nearly everyone can make a quick response, then support their decision with some sort of post-hoc rationale. For example, “Well, bigger animals make slower steps, so slower tempo in a sound must represent a larger object.” Thus it seemed that I could simply ask listeners to do this task repeatedly with different sounds and different data dimensions, and thereby determine which mappings were most commonly preferred, and in which polarity. As a first step in this new research territory I decided to do just that: I 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 something hot, cold, fast, and so on.

Choice of Dimensions

As already described, the choice of display dimensions is bounded by the eventual practical applications of sonification. I decided to begin with frequency and tempo, since both are easily produced and controlled, well studied, and are relatively unaffected by the environment in which they are heard.[4] That is, a noisy lab, a chatty classroom, or a quiet office; with or without people talking; on a ship, a plane, or a college campus—these do not have a large effect on the perception of relative frequencies and tempos. As for the data dimensions, temperature, pressure, velocity, and size are all in use in a variety of fields, and do not require special training on the part of experimental participants.

Predictions

I predicted that for some data and sound pairs there would be general agreement on an increasing-to-increasing mapping, which I call a positive-polarity mapping. For example, if I were to present two sounds that differed only in frequency, and then asked which one represents something that is hotter (the temperature-to-frequency mapping), I would expect, based on both experience and intuition, that most listeners would say that

[4] Throughout the rest of this report I use “tempo” to describe what Fraisse (1982, p.151) calls “cadence.” That is, “the simple repetition of the same stimulus at a constant frequency.”

the sound with a higher frequency represents the hotter temperature. In other words, increasing temperature maps onto increasing frequency. As a kettle heats up, the sounds it makes are generally of higher frequencies. Despite the fact that this pseudo-physical explanation might not be 100% accurate from a physics perspective, it is still a very workable mental model for most people (see Smith & Casati, 1995, for more on the topic of naïve physics). For some other mappings, I predicted that there would be broad consensus about a negative-polarity mapping. For example, most listeners would likely agree that increasing frequency represents *decreasing* size. One might rationalize this with the fact that larger objects (e.g., drums, bell, or animals) tend to emit lower-frequency sounds.

These strong mapping preferences and polarities, both positive and negative, should be in keeping with our cultural and ecological experiences with sound. However, there are many more mappings for which it is not clear how sound should best represent data. For example, it is not clear whether increasing tempo should represent increasing or decreasing size. Further, even if one person decided on a set of “best” mappings, Walker and Kramer (1996) have shown that the intuitions of one person may not match the ideas of another. Mapping preferences can only be determined by obtaining responses from a number of listeners.

Method

Participants

Seventeen undergraduate students (3 male, 14 female; mean age 19.5 years) from Rice University participated. All reported normal or corrected-to-normal vision and hearing, signed informed consent forms, and provided demographic details about age, sex, handedness, and number of years of musical training. Each participant received partial credit in a psychology course.

Apparatus

Participants sat alone at a table with an Apple Macintosh G4 computer, located in a laboratory room at Rice University. The computer monitor was a 17-in. Apple Macintosh G4 display set to a resolution of 1024 x 768 pixels, with millions of colors (24-bit). The listeners wore Sony MDR-V200 headphones, adjusted for fit and comfort, and used the Apple G4 mouse and keyboard to respond. There were two identical stations in the laboratory room, facing opposite walls, so that up to two students could participate at the same time. The ambient sound level in the room averaged 47 dB,[5] with most of the noise coming from computer equipment fans.

[5] SPL-A measured at the head of the participant using a Lutron SL-4001 sound pressure level meter.

For maximum transportability the experiments were written in HTML and JavaScript, and run with Netscape Navigator 4.6 on Macintosh OS 8.6. The sounds were played by having JavaScript commands control the Beatnik audio plugin for Netscape.

Stimuli

Sounds

The nine sound stimuli were presynthesized at 16-bit, 44.1 kHz using Csound, and saved in AIFF format. The Csound orchestra and score files used in the synthesis are included in Appendix A. Each stimulus was composed of a one-beat long pure sine wave tone, followed by a half-beat of silence. These sound and silence elements were looped to create a continuous pattern. Note that the length of a beat when measured in milliseconds depends on the tempo at which the sound is repeated. At 60 beats per minute (bpm), one beat lasts one second. 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 bpm). Figure 4 depicts the relationship between the sounds in the frequency-tempo plane, and how the sound and silence were assembled.

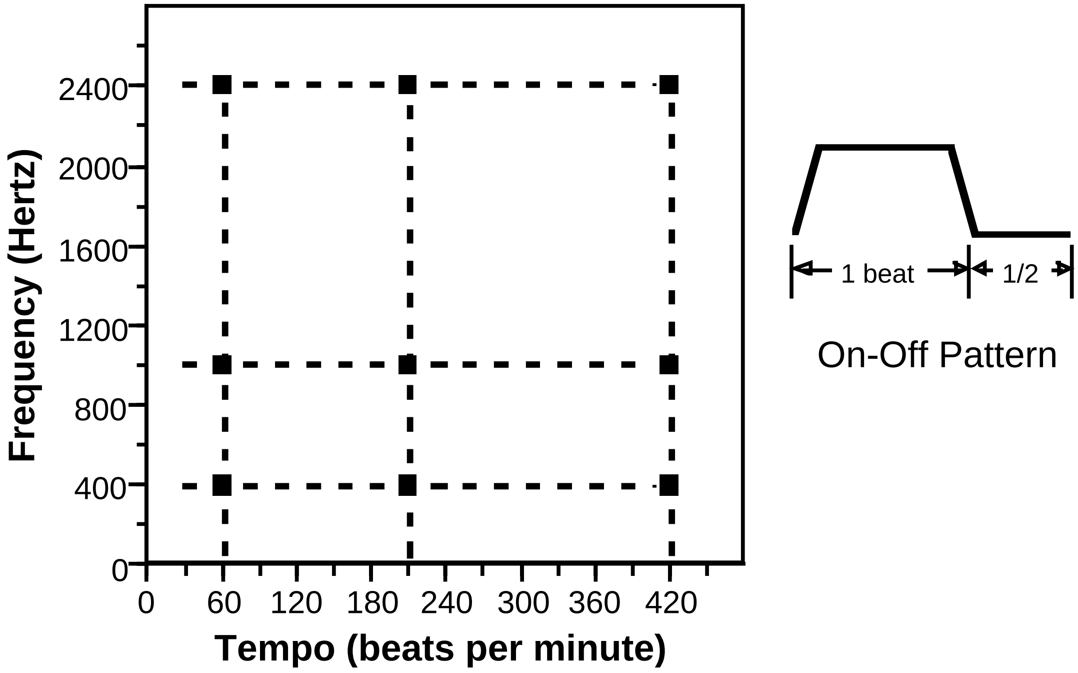


Figure 4. Schematic representation of the frequency-tempo space used to create the 9 sounds used in Experiment 1. Each sound was a repeating pattern of 1 beat of sound plus 0.5 beat of silence. The "on" portion had 1 of 3 frequencies (400, 1000, and 2400 Hz), and the patterns were looped at 1 of 3 tempos (60, 210, and 420 bpm). Note that a 1.5 beat pattern lasts 1.5 s when repeated at 60 bpm.

The amplitude envelope of the tones included a 0.1-beat linear ramp onset (attack) and offset (decay), reflected in the on-off pattern presented in the figure. I normalized the stimuli for perceived loudness by starting with relative amplitude values from equal-loudness contours (Robinson & Dadson, 1956), then making minor adjustments based on pretesting. The actual adjustments were made by changing the Csound amplitude parameters for each sound. To ensure the same relative levels, all of the sounds were synthesized into one long sound file and then separated into individual files using a sound-editing program. Table 2 contains the sound pressure level (SPL) values for the stimuli relative to the central reference tone of 1000 Hz at 60 dB.

Table 2. Relative sound pressure levels for stimuli that varied in frequency in Experiment 1.

Frequency (Hz)	400	1000	2400
SPL (dB)	62	60	58

Note: The sounds were adjusted relative to the reference tone of 1000 Hz at 60 dB. The approximate loudness values were determined from equal-loudness contours (Robinson & Dadson, 1956), and minor adjustments were made by ear.

Word Cues (Data Dimensions)

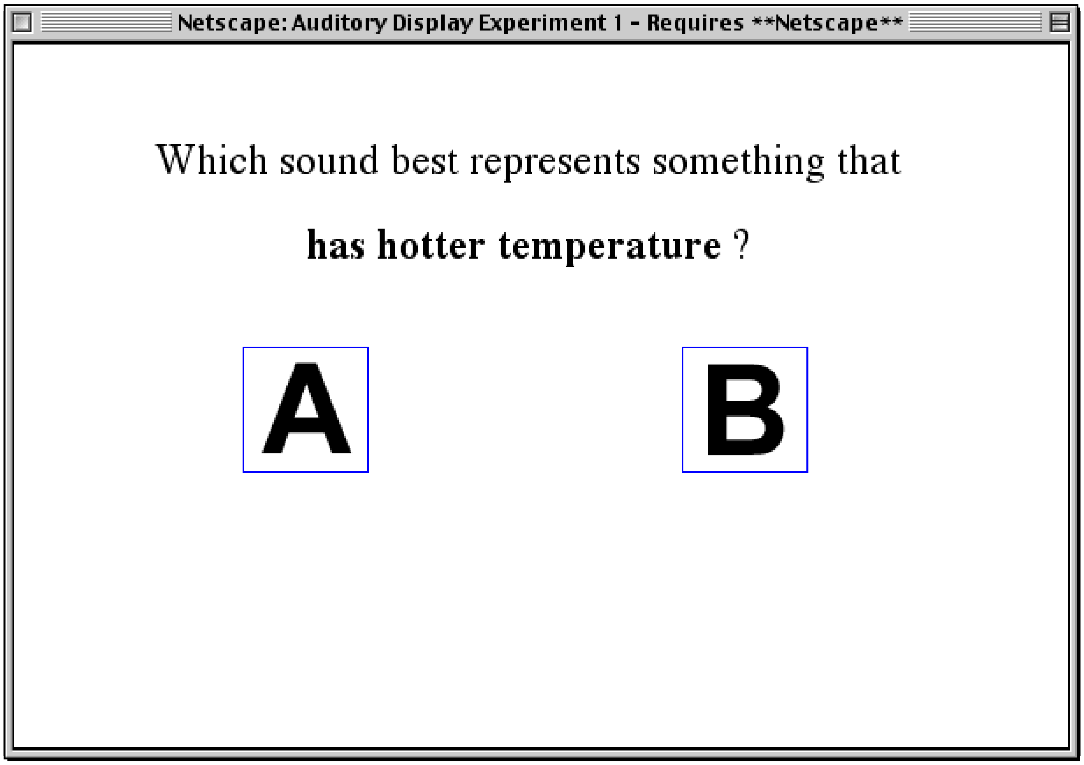
For this first experiment I chose four data dimensions that were both widely known and commonly used in a variety of scientific fields: temperature, pressure, velocity, and size. Matching cue questions were created to cover both ends of each cue dimension. For example, the pair of cues for the temperature dimension consisted of: “Which sound best represents something that has hotter temperature?” and “Which sound best represents something that has colder temperature?”

Design

The experimental design was fully factorial, including the within-subjects factors of data dimension (temperature, pressure, size, velocity) and display dimension (frequency, tempo). The dependent measure was a signed mapping preference score, described below.

Trial Structure and Task

Figure 5 shows an example of the screen layout in this experiment. On each trial a cue question was centered near the top of the screen (e.g., “Which sound represents something that has hotter temperature?”). In the middle of the screen there were two 2.5-cm squares labeled “A” and “B,” separated by 5 cm horizontally.


Netscape: Auditory Display Experiment 1 - Requires **Netscape**
A
B

Which sound best represents something that

has hotter temperature ?

Figure 5. Image of the screen layout seen by participants in Experiment 1. The experiment was written in HTML and JavaScript, and run in a Netscape 4.6 window with extraneous menus and scrollbars removed, and maximized to fill the screen.

The participant would move the cursor over square A, and would hear sound A (e.g., a 400-Hz sound repeating at 210 bpm). When the cursor moved off the square, the sound would stop. The listener would then move the cursor over square B and hear sound

B (e.g., a 1000-Hz sound repeating at the same 210-bpm tempo). The listener was able to move the mouse back and forth from one square to the other, listening to the sounds. After deciding which sound seemed to “answer” the cue question better (i.e., which represented something with a hotter temperature), the listener would click on the corresponding square to indicate a response and to go on to the next trial.

Each of the cue questions (e.g., which is hotter?; which is colder?), for each of the data dimensions (temperature, pressure, size, velocity), was presented factorially with each pair of sounds that differed only in frequency (nine pairs) and with each pair of sounds that differed only in tempo (nine pairs). With counterbalancing this yielded 144 total trial types. The trials were presented in a different randomized order for each participant.

Signed Preference Scores

For each data dimension and display dimension pair a signed preference score was determined for each subject as follows. For each combination of data and display dimensions I defined the increasing-data to increasing-display mapping as the “positive” polarity, regardless of any preconceived notions about “best” or “preferred” mappings. Then for each response I determined whether the listener had responded in accordance with the positive polarity or not. For example, if the listener responded that the higher-pitched sound of a pair better represented “hot,” then that response supported the positive polarity, and was scored as a +1. If, throughout the session, 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 between temperature and pressure would reach the maximum score of +18. It is also possible for a participant to have an equally consistent, but entirely opposite concept of how hot and cold should sound, and therefore score −18, with hot corresponding to low frequency and cold corresponding to high frequency. Scores in between ±18 would indicate a less-consistent or weaker preference. Thus, the magnitude of the mean preference scores across subjects should provide a measure of how effective a mapping was, and the sign should indicate in which polarity it was preferred.

In the example above, one might predict that nearly all listeners would respond that the high-frequency sound represents something “hotter.” Nevertheless, even the listener may not be able to explain why she responded in such a way. In fact, for many of the trials the listener will make what she feels is a guess or a random choice. However, even if a listener cannot “explain” why a particular sound answers the question, the population stereotypes should emerge over the set of trials for each participant, and moreover in the aggregate of all participants. For example, Patterson (1982) recommended certain onset times that affect the perceived urgency of a sound. A listener is unlikely to know why a sound seems “more urgent,” but the fact is that, overall, people do agree on the relative urgency of sounds that differ only in onset times (e.g., Edworthy, Loxley, & Dennis, 1991).

Results

Aggregate Means

The most important data patterns to examine were meant to be the aggregate, or “population” mean results. Figure 6 shows the mean preference scores for each data and display dimension collapsed across the 17 subjects. For the display dimension of frequency, there was no significant difference between the mean preference scores for the data dimensions of temperature, pressure, velocity, and size [*M*s = 12.53, 11.88, 12.29, 12.53, respectively; *F*(3, 64) = 0.017, *p* = .9967]. That is, across subjects, frequency was judged to represent all of the data dimensions equally well. For the display dimension of tempo, there was again no significant difference between the mean preference scores for the data dimensions of temperature, pressure, velocity, and size [*M*s = 13.06, 12.24, 14.06, 12.71, respectively; *F*(3, 64) = 0.2509, *p* = .8604]. Thus, tempo was also judged to represent all of the data dimensions equally well.

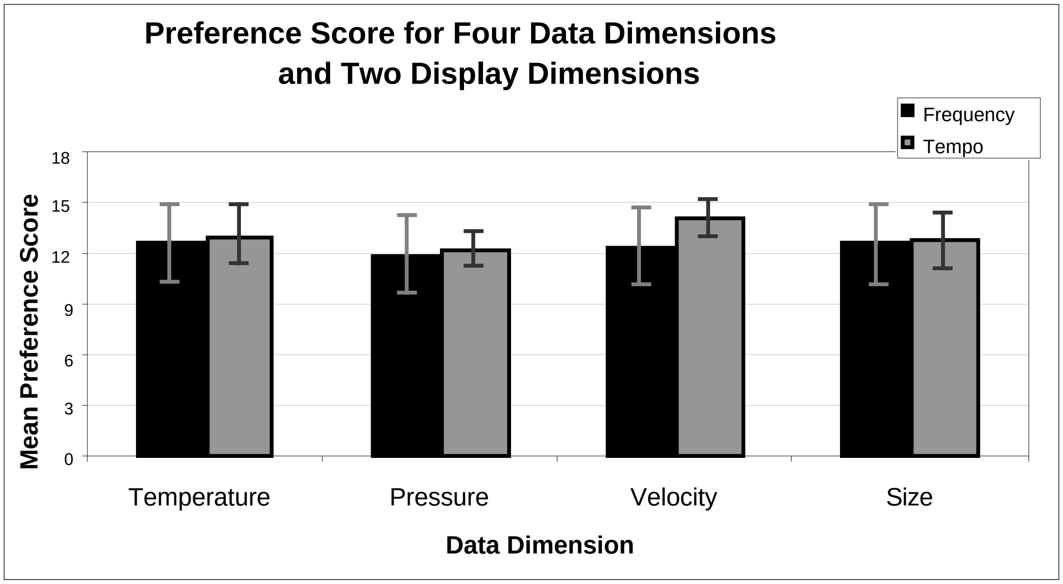


Figure 6. Mean preference scores from Experiment 1, across subjects. Black bars represent sounds differing only in frequency; gray bars represent sounds differing only in tempo. Note that a maximum score of 18 is possible in both the positive and negative directions, however none of the means was negative. None of the differences reached statistical significance. Error bars represent one standard error above and below the mean.

The goal of this experiment was simply to ask listeners to make judgments, and then use the easily-obtained preference scores to determine which data dimension a given display dimension should represent, and in which polarity. The means shown in Figure 6 would seem to indicate that it simply does not matter. However, this is a particularly curious result. In the case of temperature and frequency, a highly positive result (perhaps near +18) was expected, since increasing frequency is commonly associated with increasing temperature (e.g., a kettle boiling). On the other hand, when frequency represents size, a highly negative result was expected, if listeners reason, for example, that smaller objects usually make higher-pitched sounds.

Individual Means

Apparently the listeners in this experiment were not making distinctions between the different data dimensions. To investigate this possibility, I examined the individual responses. Figure 7 shows the individual patterns of responses for each of the 17 participants for the display dimension of frequency. If listeners were responding as predicted, then some of the bars should be highly positive (e.g., temperature and possibly velocity) while at least the bar representing size should be strongly negative. However, it seems that each listener simply decided on a mapping polarity (e.g., an increase in the frequency means an increase in every data dimension), then applied that polarity to all of the trials, more or less consistently (see, e.g., Participant 3 in Figure 7). Participant 16 shows a similar, but inverted response pattern. They are both quite consistent in their responses, but seem to have internalized a different mapping polarity scheme. The results for the tempo dimension are shown in Figure 8, with the same pattern emerging. It is interesting to note that Participant 16 seems to have taken a completely consistent but opposite polarity for the frequency and tempo trials. Recall that the trials were all randomly interspersed.

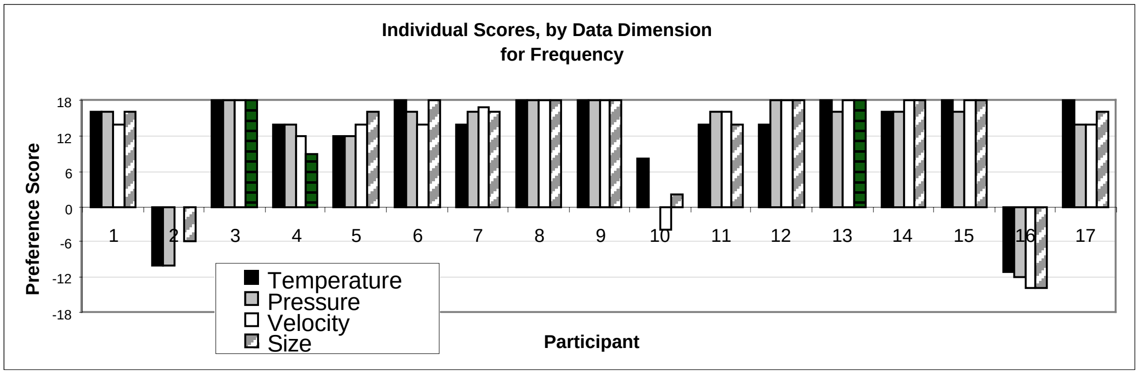


Figure 7. Preference scores from Experiment 1 for each participant, for each data dimension, with frequency as the display dimension.

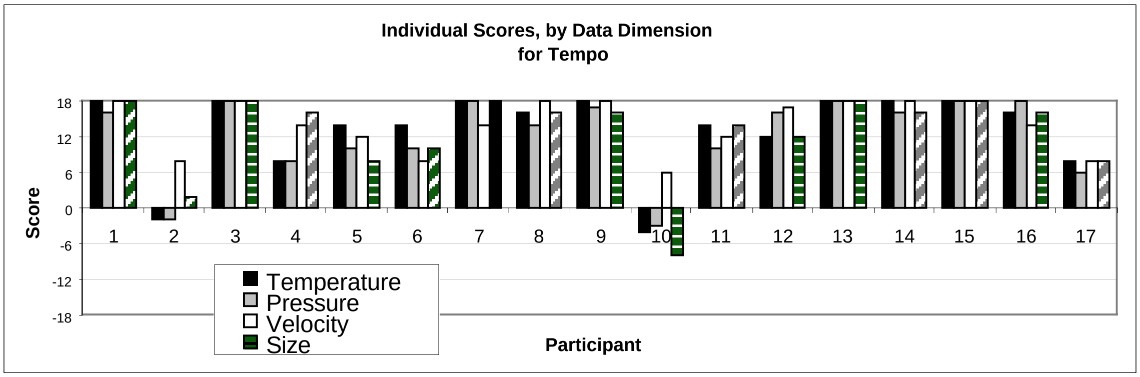


Figure 8. Preference scores from Experiment 1 for each participant, for each data dimension, with tempo as the display dimension.

Discussion of Experiment 1

This first experiment used a homogeneous population—young-adult undergraduate psychology students—in order to minimize the effects of any population variables. The focus here was on the differences between the data-to-display mappings, and not on the variability between populations.

The preference scores in this experiment were meant to indicate which display dimensions are preferred for representing the data dimensions. Pre-testing and plenty of informal and anecdotal evidence have given me reason to believe that there really are differences between preferred mappings. When I casually ask people to tell me which sound represents something that is hotter, then hum two different-pitched tones, I reliably get different responses for different data dimensions.

However, when I looked at the overall means in this experiment, there was considerably less variability in the mappings and polarities than I had predicted. In fact, none of the polarities was even negative, despite what I thought was a “natural” (or dare I say, “intuitive”) mapping between increasing frequency and decreasing size. In looking at the individual subject data, it is clear that most subjects just determined a mapping polarity and applied it consistently across all data dimensions. For example, they responded that increasing frequency indicated an increase in whichever data dimension it was mapped to, be it temperature, pressure, velocity, or size. The listeners in this experiment were not really making a cognitive assessment of the mapping, but rather listening for the changing sound parameter, and applying a simple rule to determine a “correct” response. 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 I had hoped in determining preferences. That was an unfortunate result, since this might have been a very simple paradigm that display designers could use to do some simple research on their particular display parameters. A variation of this paradigm might still be useful, basically as an automated version of my casual interrogations. In that case, though, perhaps listeners would have to pause between each trial and provide some kind of verbal justification for their decision, thereby forcing more perceptual-to-conceptual processing on the part of the participant.

On the other hand, even if this paradigm had been more effective, it could only have addressed the issues of mappings and polarities. I would still not know if the relationship between, say, frequency and temperature were linear. Further I would still need to determine the scaling function between the data and display dimensions. That is, how much change in temperature would a given change in frequency represent? It was clear that I required a more sophisticated, and perhaps less transparent experimental paradigm. For that, I turned to a procedure that I hoped would provide not only the mapping preferences and polarities, but also the actual numerical relationships that will be critical for the creation of practical sonifications.

Psychophysics and Dimension Scaling

Determining the preferred mapping and polarity is the initial step of any sonification design. Unfortunately, the simple and straightforward approach of asking for preferences (at least, as attempted in Experiment 1) did not turn out to be a successful way to accomplish that initial step. However, as I have already pointed out, even after an auditory display dimension has been chosen to represent a data dimension, a more difficult question arises. In representing data, it is crucial to know how much change in the display dimension is required to represent a given change in the data dimension. That is, how much louder should 400 ºC sound, compared to 200 ºC? This is known as “scaling” the dimensions, and, in a general sense, has been the topic of psychophysical investigation for more than 100 years. However, to date most attention has been paid to the way that one physical dimension is subjectively perceived. In the present context, though, psychophysical methods should provide a well-grounded approach to determining all three of the critical components in a sonification: the effective dimension mapping; the polarity; and the scaling of that mapping.

Physical versus Perceptual Dimensions

The fact that various sound attributes exist, and can be varied in a sound stimulus, does not mean that listeners perceive changes to the sound in a way that matches the actual variations in the underlying sound parameter. The study of psychophysics has been concerned with precisely this matter; namely, determining the function that describes how changes in a physical attribute of a stimulus are perceived and judged by an observer. In the traditional method of describing this psychophysical scaling function (see Falmagne, 1985; Stevens, 1951, 1975), the underlying physical dimension, Φ, is perceived along the psychological dimension, Ψ. The two are related by an arbitrary "psychophysical function" f (which is usually presumed to be monotonic):

$$\Psi = f(\Phi). \tag{1}$$

S. S. Stevens and others have amassed a large number of experiments that attempt to determine the actual psychophysical scaling functions between several physical dimensions and their resulting perceptual or psychological dimension. Across these many studies, it has emerged that the vast majority of perceptual dimension pairs are related by a "power law" (Stevens, 1975; Stevens & Galanter, 1957). That is,

$$\Psi = \Phi^{\beta} \tag{2}$$

where β is the exponent describing how much a given physical change in the stimulus appears to change, according to the observer. The actual exponent β is different for each pair of dimensions, but has been found to be stable across time, observer groups, and experimental conditions (Stevens, 1975, p.15). I should clarify, however, that while aggregate measures of β are relatively constant and stable, there is a fair amount of variability in individual responses. See Teghtsoonian and Teghtsoonian (1983) for more on this issue.

Magnitude Estimation

The most common experimental technique used in these psychophysical studies, particularly by Stevens and his colleagues, is the method of magnitude estimation. Consider, for example, the scaling of the perceptual dimension of loudness to its underlying physical dimension, intensity. The experimenter creates a set of several stimuli, across a large range of sound pressure levels (several orders of magnitude on the physical dimension, if possible). The middle stimulus is played first as the "standard," and the listener is told that the standard stimulus is to be called, say, "100." Subsequent stimuli are played in random order and for each one the listener estimates the ratio of that stimulus to the standard. Thus, if the stimulus seems to be twice as loud as the standard, then the listener reports "200." If it seems half as loud, it should be called "50," and so on.

The geometric mean of responses from all the subjects at each stimulus value yields a function relating the physical intensity to perceived loudness changes.[6] This,

[6] S.S. Stevens (1975, p.269) indicates: "After experiments with large groups of subjects had supplied sufficient data, J. C. Stevens (1957) determined in his thesis study that the distribution of the logarithms of the magnitude estimations is approximately normal. The efficient average for a log-normal distribution is the geometric mean. Consequently, it has become customary to use geometric averaging with magnitude estimation. Caution is still in order, however, for outlandish values can turn up to distort the picture. The experimenter may then want to resort to the less efficient but less vulnerable median."

indeed, turns out to be a power function, as in Equation 2, with an exponent β of about 0.67 for loudness (see Stevens, 1975, p. 15, for a variety of dimensions). Note that if a doubling of intensity were to correspond to a doubling in perceived loudness, the exponent would be 1.0, indicating a linear relationship. The fact that this relationship is not linear (i.e., the exponent is not 1.0) for intensity and loudness is the very reason that we have different terms to describe the physical parameter ("intensity") and the perceived quality ("loudness"). Stevens (1966) also provides a review of several "social" dimensions that have been scaled with magnitude estimation, such as the perceived value of money and the aesthetic value of handwriting.

Intramodal Matching

Determining the power function for a single physical attribute of a stimulus, as described above, is described as "dimension scaling" via magnitude estimation. However, there is another set of questions that can be asked about perceptions within a given sensory modality (e.g., within taste or within hearing). Stevens suggests a number of these "intramodal matching" questions, such as, "What amplitude of vibration applied to the fingertip feels as strong as a given vibration applied to the arm?" or "What concentration of fructose tastes as sweet as a given concentration of sucrose?" (Stevens, 1975, p. 63). Work in this area has led to families of curves called equal-perception contours, which can be used to predict, for example, how loud one sound will seem in comparison to another, even if the frequencies of the sounds are different. This has significant practical applications, most notably in the use of equal-loudness contours to determine noise exposure limits (see, e.g., Sanders & McCormick, 1993).

Cross-modal Matching

Once magnitude estimation has been used to determine how the perception of several stimulus attributes behaves (i.e., what the slope of the power function is for each dimension), yet another interesting question presents itself. As Stevens puts it, “Can the scales obtained by number matching be used to predict how subjects will perform when they are asked to make a direct comparison between two *different* sensory continua?” (Stevens, 1975, p.99, emphasis added). For example, if a subject feels vibrations of various intensities on the finger, and is asked to adjust the loudness of a tone to “match” each vibration event, what will the results show?

It turns out that the two dimensions often combine in a straightforward way. If the sensation of the first dimension is of the form

$$\Psi_1 = \Phi_1{}^m \qquad (3)$$

and the perception of the second stimulus dimension is

$$\Psi_2 = \Phi_2{}^n \qquad (4)$$

then we can equate sensations Ψ_1 and Ψ_2 (e.g., by adjusting a tone so it “matches” a vibration) to obtain

$$\log \Phi_1 = (n/m) \log \Phi_2. \qquad (5)$$

If the data are plotted on log-log axes, this last equation represents a straight line, where the slope of the line (n/m) is the ratio of the exponents obtained for each dimension separately.

Many experimental results confirm this relationship, between a variety of stimulus dimensions both within and across different sensation modalities (see Marks, 1987; Stevens, 1966, 1975). Thus, it is possible to determine the function that specifies how much of an increase in loudness corresponds to a given increase in brightness.

Non-Sensory ("Conceptual") Dimensions

Stevens (1975) claims that *any* two dimensions, in any modalities, can be compared, and therefore scaled, in a cross-modality matching paradigm. He points out that even the basic dimension scaling via magnitude estimation is essentially a special case of the cross-modality match between the perceptual dimension (say, loudness), and the dimension of positive natural numbers.[7]

Some researchers have tested this assertion, and have created cross-modal scaling functions for some "non-sensory" dimensions, such as matching the loudness of a tone to the level of racism attributed to certain acts, the pronounciability of trigrams, and the desirability of certain professions (Dawson & Brinker, 1971). I call these "conceptual" dimensions, since they are not within any particular sensory modality. In general, these comparisons for conceptual dimensions result in power functions, as do the perceptual dimensions, but with different exponents.

The mapping of conceptual dimensions onto sounds (e.g., Dawson & Brinker, 1971) is very close to what I have described as critical goals in sonification research, except that in sonification the goal is to scale “data-type” conceptual dimensions onto sounds. Even though the data dimension of “temperature” that is represented by pitch in an auditory thermometer can be a sensory dimension (i.e., one can feel temperature), in the case of scientific data analysis it is really the “concept” of temperature that is being used. With this in mind, I was interested in seeing how the basic scaling of loudness to numbers compares to scaling loudness to conceptual dimensions like pressure or temperature.

The range of conceptual dimensions that have been scaled to perceptual dimensions has been extended to sociology, criminology, and political science. While some studies have determined scaling functions between auditory dimensions and conceptual dimensions (e.g., Dawson & Brinker, 1971; Stevens, 1966), I am not aware of any case where these scaling functions have been used to predict subsequent performance, nor have they been used specifically to create auditory displays. It is this

[7] See Banks and Hill (1974) for more on the issue of scaling the set of integers themselves. It turns out that the perceived magnitude of numbers is, itself, not quite linear, with a slope less than 1.

crucial next step that would tie the psychophysical scaling literature to the field of auditory displays.

Some Additional Considerations

Research in psychophysical scaling has pointed out two additional issues that may be of particular relevance to auditory displays more in the future. I point them out here since I have considered them in the design of the experiments described later.

Choice of Modulus

Initial experiments using the paradigm of magnitude estimation used an experimenter-defined modulus, or number to which the first stimulus should be associated. The description of magnitude estimation provided above includes a modulus of “100” so that twice the standard stimulus would be called “200.” However, it turns out that allowing the subject to choose his or her own modulus provides for a less constrained assignment of numbers by the subject. Stevens (1975, p.252) specifically recommends a subject-defined modulus, where the listener assigns the standard (or first) stimulus “any number.”[8] In order to produce a general scaling function, for the purposes of extracting just the exponent of the power function, it seems perfectly sensible to afford the listener

[8] Actually, Stevens (1969, p.251) makes the point more vehemently: “As was noted in the study of loudness…the best thing to do with the standard is to omit it. Except in special circumstances, the presentation of a standard *and the designation of a modulus* have proved to be quite unnecessary. It is especially pointless and wasteful to repeat the standard before each variable stimulus” (emphasis added).

with this response freedom. My present research uses this so-called modulus-free paradigm.

However, in dealing with conceptual dimensions, especially those which may be either simple ordinal or interval dimensions (see Stevens, 1975), such as temperature in degrees Celsius, as opposed to ratio dimensions like temperature in Kelvin, practical applications may need listeners to be thinking in terms of a particular range. That is, it may fall into Stevens's category of "special circumstances." For example, when scaling loudness to temperature, most listeners will be thinking about temperature in the range from, say, 0 ºC (the freezing point of water) to 100 ºC (the boiling point of water) since lower or higher temperatures are less often encountered in everyday life. For the eventual needs of sonification, it will be important to know whether loudness scales to temperature equally across temperature ranges. The imposition of several specific values for the modulus may help in that investigation. That is, the standard sound could be "called" 50 ºC for one group of listeners, 100 ºC for another group, and 1000 ºC for a third group. However, detailed investigation of the modulus effect with conceptual dimensions will have to await the completion of the present project.

Magnitude Estimation vs. Magnitude Production

A second issue that may be relevant to auditory display is the difference between magnitude estimation and magnitude production, and the so-called "regression bias" that results. When the experimenter sets various levels of a stimulus parameter (e.g., intensity), and the subject adjusts the number scale to match, this is magnitude estimation, as already discussed. When the experimenter provides values along the number line, and the subject adjusts the stimulus level (e.g., intensity) to match the numbers, this is called magnitude production. The slopes of the scaling functions determined by each of these two methods are close, but systematically different. Magnitude estimation produces a flatter curve (a lower slope). As depicted in Figure 9, the geometric mean of the slopes of the two lines produced by estimation and by production should provide the “actual” (or at least “theoretical”) matching function between the dimensions (the dotted line in the figure). See Stevens and Greenbaum (1966) and Stevens (1975, p.31) for a more complete explanation of this effect.

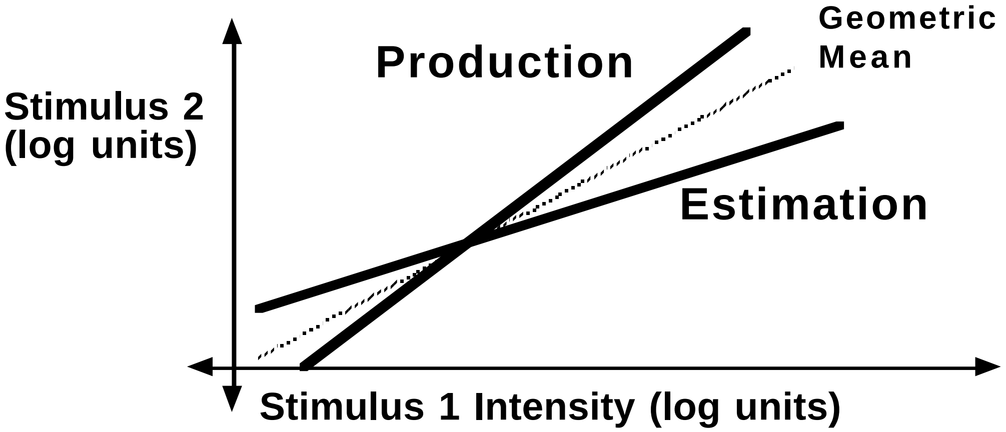


Figure 9. Representation of the different slopes obtained through magnitude estimation and magnitude production. The dotted line depicts the geometric mean of the two lines, which may represent the “actual” or “theoretical” scaling function. However, as described in the text, it may be more relevant to auditory displays to focus on the “estimation” function. After Stevens, 1975, Figure 12, p 31.

For our purposes, the important point is that there is a difference between the two functions (i.e., the slopes of the lines). Consider the task of an auditory display designer: Given a set of data values (numbers) he or she must produce sounds that represent those values. This is clearly magnitude production. However, the listener will hear those sounds, and associate them to numerical values in the data set (magnitude estimation). Thus, the designer is bound to create a display (a set of stimuli) that is not optimally matched to the listener's expected perceptions. However, if the designer used the same function to generate the sounds from the data as the listener will use to "decode" the sounds, then a better-understood display would likely result. This provides some psychophysical explanation for the observation by Walker and Kramer (1996) that what "sounds right" to the designer might very well not sound right to a listener. Hence, in order to combat this problem, the listener's scaling function should be computed through magnitude estimation, and the designer should rely on the *listener's* function to calculate the best mapping function between the data and the auditory display. Of course, this computational approach will never replace the human sound designer, but it will certainly provide a tool for a more accurate, effective, and more easily comprehended display.

Experiment 2: Magnitude Estimation with Lines and Circles

Purpose

In addition to discovering mapping polarity preferences, this line of research is also intended to find a way to determine what the scaling function is between the physical sound attributes and the data dimensions they are meant to represent. As described in the last chapter, I turned to the psychophysical scaling paradigm of magnitude estimation (Stevens, 1975) to assess directly both the polarity and scaling function issues. I hoped that my 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). In order to take a smaller leap, I decided to explore the magnitude estimation paradigm with visual stimuli first, using both perceptual and conceptual data dimensions. I would then move on to auditory stimuli in the next experiment if the approach proved successful here.

In this experiment the dimension of size (e.g., line length) was treated as a perceptual dimension and was included as a calibration of the experimental procedure, since there is some literature on the magnitude estimation of line length (e.g., Stevens, 1975; Stevens & Guirao, 1963; see also Teghtsoonian, 1965). Stevens and Guirao (1963) studied magnitude estimation of line lengths and, as Stevens later wrote, “demonstrated clearly that the exponent for line length is very close to 1.0. In other words, the relation is linear. Hence, from the subject's point of view (or, as we say, subjectively), number and line length are directly proportional" (Stevens, 1975, p.109).[9]

Subsequent examination of the perceived lengths of lines has shown exponent values to be systematically just less than 1.0. For example, Teghtsoonian and Teghtsoonian (1971) obtained a value of 0.93 for apparent length. They also obtained mean values of 0.98 in each of two later experiments (Teghtsoonian & Teghtsoonian, 1983, Experiments 1 and 2). However, as I described in the previous section, the magnitude estimation procedure tends to slightly underestimate the "actual" value for the exponent, which may be why Stevens (1975) asserted a linear relationship between perceived and actual line length (i.e., a "real" exponent value of 1.0).

However, this says nothing about what to expect for the matching of line length to the conceptual data dimensions. In this experiment, temperature, pressure, and velocity were meant to be conceptual data dimensions, and were described as the temperature, pressure, or velocity of "something." Few, if any, studies have actually matched perceived line length to a conceptual dimension, and certainly not to other conceptual data dimensions like temperature. Indow (1961; cited in Stevens, 1966, p.534) reportedly matched the desireability of (preference for) different wristwatches to the length of a line

[9] Stevens (1975, p.15) presents a table of exponents of the power functions relating subjective magnitude to stimulus magnitude. His book contains a summary of results for many different dimensions.

marked on a paper, though not in a magnitude estimation paradigm, and obtained an approximately logarithmic relationship.

There have been magnitude estimation studies of the perceived size of two-dimensional visual stimuli, as well. Stevens and Guirao (1963) used filled squares as stimuli, and obtained an exponent of 0.70. In other words, as Teghtsoonian (1965, p.392) put it, for two-dimensional stimuli “judged size increases somewhat more slowly than the stimulus.” It turns out that the word “somewhat” is appropriately vague. Teghtsoonian indicates that Ekman (1958) found an exponent of 0.86 for circles. Teghtsoonian (1965), in a more thorough examination of the factors affecting perceived size, obtained an exponent of 0.76 with circles. Teghtsoonian and Teghtsoonian (1971) later found an exponent of 0.69 using outline circles.

There has been very little work relating two-dimensional shapes to any sort of conceptual data dimensions. Map makers confront this issue all the time, but very little has been done experimentally to examine the relation between perceived size and perceived amounts of some data represented by a symbol. Williams (1956, Experiment 4) conducted an interesting study in the context of map symbols. His goal was to answer questions such as, “What size should a [symbol] be drawn to represent an army of 1,000,000 men in comparison to other figures representing armies of different sizes?” Essentially this is the scaling of the conceptual data dimension size (of an army) to the display dimension of area (of the circular map symbol). Unfortunately, though, in the actual experiment Williams did not include a specific conceptual data dimension. Rather, participants provided responses about circles that would have 2X, 3X, 5X, 10X the "value" of the first circle. Williams reported an overall exponent of 0.81 between the magnitude estimations of "value" and the actual area of the circles. However, the analysis methods Williams employed were somewhat unlike those used in typical magnitude estimation experiments. Using the data provided (medians, rather than geometric means, unfortunately) I re-plotted Williams' results for circles. Figure 10 shows the results in log-log coordinates, with a regression slope of m = 0.73, in the middle of the range of results found for simple size estimations for filled circles. Unfortunately, there is no way to know what, if any, particular data dimension the participants had in mind. The obtained value of 0.73 could conceivably be a sort of average slope across the various data dimensions in use in the minds of the participants. On the whole, it is not clear whether one should expect any difference between estimations of size and estimations of "value" at all, let alone differences between conceptual data types.

With all of these questions in mind, I set out to examine how perception of linear and circular stimuli would map onto a variety of conceptual data dimensions.

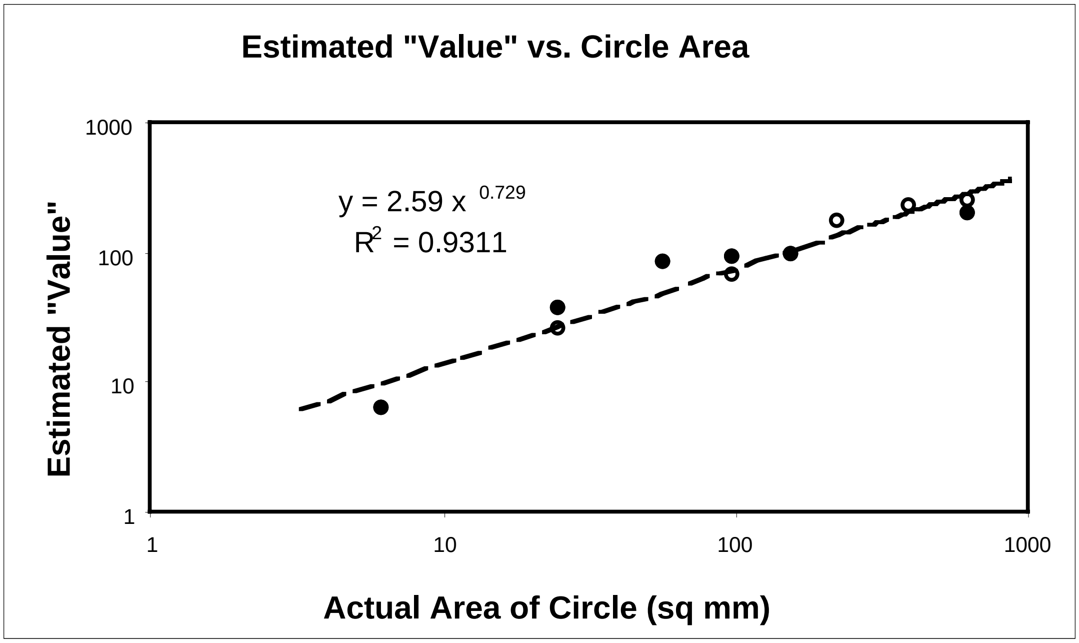


Figure 10. Estimated “value” of circles versus their actual areas. Plotted from median data provided in Williams, 1956, Experiment 4. Filled circles in the plot represent the data obtained relative to comparison circles with 5 mm diameter; open circles represent data relative to 10 mm comparison circles. Note the slope in this case m = 0.73; r^2 = 0.931.

Method

Participants

From the same subject pool as Experiment 1, 69 students began the experiment. Due to computer problems, the data from two of the participants were lost, leaving a final total of 67 participants (22 males, 45 females; mean age 19.3 years).

Apparatus

The apparatus was identical to that of Experiment 1. However in this experiment the headphones were not used, as the stimuli were presented visually on the computer screen.

Stimuli

Over many years and many different dimensions, researchers (e.g., Stevens, 1975; Teghtsoonian & Teghtsoonian, 1978) have found that with the magnitude estimation paradigm six to ten levels of the stimulus dimension, repeated no more than twice each, provide consistent results. I used nine stimuli, and presented them twice each. The line stimuli were nine black rectangular bars 5 pixels wide and 10, 20, 40, 60, 80, 100, 400, 500, or 600 pixels long. They were displayed one at a time in the center of the screen with a white background, and were oriented either horizontally or vertically.

The circle stimuli were nine black filled circles with diameters of 10, 30, 50, 70, 100, 300, 400, 500, and 600 pixels. They were also displayed one at a time in the center of the screen with a white background. All of the stimuli were created with Adobe Photoshop and exported as GIF images, suitable for use with the HTML-based experiment code.

Design

The experimental design included the between-subjects factor of data dimension (temperature, pressure, velocity, and size) and the within-subjects factor of display dimension (horizontal lines, vertical lines, and filled circles). Each data dimension (e.g., temperature) could be paired with one of the display dimensions (e.g., horizontal lines) for an entire block of trials (named, e.g., temperature+horizontal lines). Thus, there were 12 possible block types. Each participant completed two blocks of trials separated by a brief rest. The block types were assigned pseudo-randomly, with the constraint that each participant would see two different display dimensions and two different data dimensions. An example experiment might include a size+circles block followed by a pressure+horizontal lines block.

Initially, the plan was to have 12 participants in each block type. However, to conserve subject hours, that number was reduced to 10 per block type and the temperature+circles block type was not used. Table 3 summarizes the number of participants in each block type.

Table 3. Number of participants in each block type in Experiment 2.

Display dimension	**Data dimension**			
	Temperature	Pressure	Velocity	Size (of Stimulus)
Horizontal Lines	10	12	10	12
Vertical Lines	12	10	10	10
Circles	-	12	10	12

Trial Structure and Task

Before each block of trials, the experimenter read aloud instructions like the following, while the participants followed along on the computer screen:

> You will see a series of lines on the screen in random order. Your task is to indicate what temperature they would represent, by assigning numbers to them. For the first line, assign it any number of your choosing that represents a temperature. Then, for each of the remaining lines, estimate its "temperature", relative to your subjective impression. For example, if the second line seems to represent a temperature that is 10 times as hot as the first, then assign it a number that is 10 times bigger than the first number. If the line seems to represent a temperature that is one-fifth as hot, assign it a number that is one-fifth as large as the first number, and so on. You can use any range of numbers, fractions, or decimals that seem appropriate.

Participants in the “size” block saw slightly different instructions like the following:

> You will see a series of lines on the screen in random order. Your task is to indicate their lengths, by assigning numbers to them. For the first line, assign it any number of your choosing. Then, for each of the remaining lines, estimate its length, relative to your subjective impression. For example, if the second line seems to have a length that is 10 times as long as the first, then assign it a number that is 10 times bigger than the first number. If the line seems to have one-fifth the length, assign it a number that is one-fifth as large as the first number, and so on. You can use any range of numbers, fractions, or decimals that seem appropriate.

Figure 11 shows the simple screen layout for the trials in this experiment. On each trial the participant saw one stimulus from the set being used for that block (e.g., horizontal lines) and provided a number for that stimulus's "temperature" in a text-entry box on the screen. Participants simply filled in their response and then clicked the "Next" button on the screen or pressed the "*return*" key on the keyboard to proceed to the next trial.

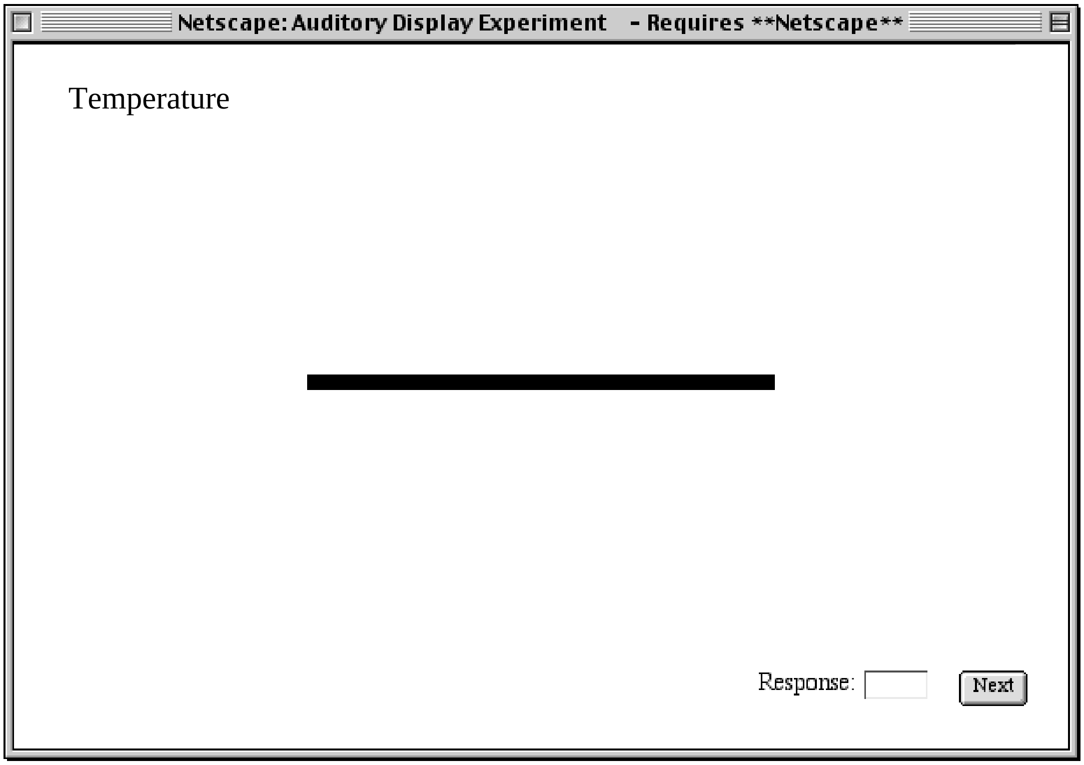

Netscape: Auditory Display Experiment - Requires **Netscape**
Temperature
Response:
Next

Figure 11. Image of the screen layout seen by participants in Experiment 2. The data dimension was displayed at the top left. The stimuli were either horizontal or vertical lines, or filled circles and were located in the center of the screen. Participants entered their responses in the text-entry box at the bottom right, and clicked the "Next" button to continue.

In a block of 18 trials, each of the nine stimuli was randomly presented twice, with the constraint that the largest or smallest stimulus in that set could not occur first (see Teghtsoonian & Teghtsoonian, 1978). Following a brief rest, the participant began the second block with new instructions that introduced different data and display dimensions.

Results

I combined the responses from all of the participants for each trial block type (i.e., for each combination of data and display dimensions) and then sorted the combined results by stimulus number. I also converted all responses to decimals, because fractional responses like 1/3 were allowed. Since participants could pick any starting number, some chose a small number like 1, and ended up with responses in the range of, for example, 0.01 to 120. Others started with a number like 100 and ended up with a range like 2 to 3,000. To obtain a measure of the average response for each stimulus, I calculated the geometric mean of all responses for that stimulus, across participants in a given block. As mentioned, the geometric mean is less susceptible to the effects of very large or small numbers.

Perceptual Dimensions: Size

I began by looking at the perceptual dimension of size. Specifically, I examined how the magnitude estimations of the size of a stimulus compared to the actual size of the stimulus. I plotted the geometric mean response for each stimulus against the actual size of the stimulus in pixels (recall that the subjects' responses are unitless). Figure 12 shows the plot of the geometric mean of estimated length against the actual length of horizontal lines. Note that the axes are in log units. The participants whose responses contributed to the means are indicated in the title area of the plot. The dashed line through the points represents the best fit using a power equation. The equation of that line is of the form $y = b\ x^m$, and is included inside the plot along with the r^2 coefficient, which is a measure of goodness of fit in the regression equation. Values of r^2 near 1.0 indicate an excellent fit. The slope of the line for a power equation is given by m, the exponent of x in the regression equation. In the case of length estimation vs. horizontal line length, the regression slope m = 0.96 (SE_m = 0.02), and r^2 = 0.995.

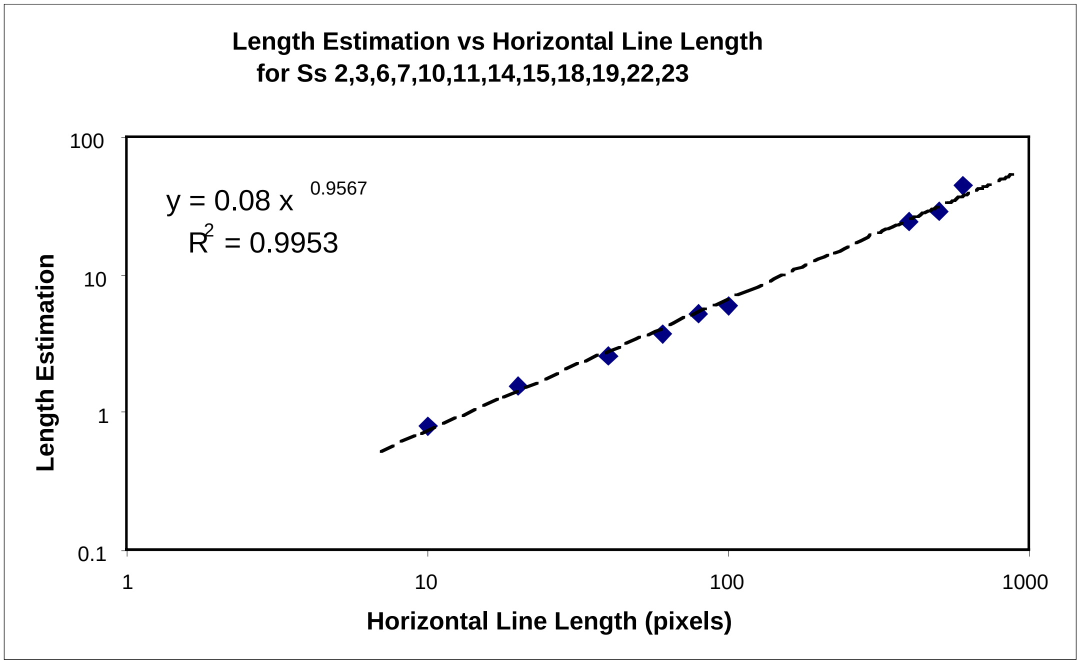


Figure 12. Geometric mean of estimated length against the actual length of horizontal lines in Experiment 2. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The dashed line through the points represents the best fit using the power equation that is included in the plot. The slope of the fit line is given by the exponent of x. In this case, the regression slope m = 0.96 (SE_m = 0.02; r^2 = 0.995).

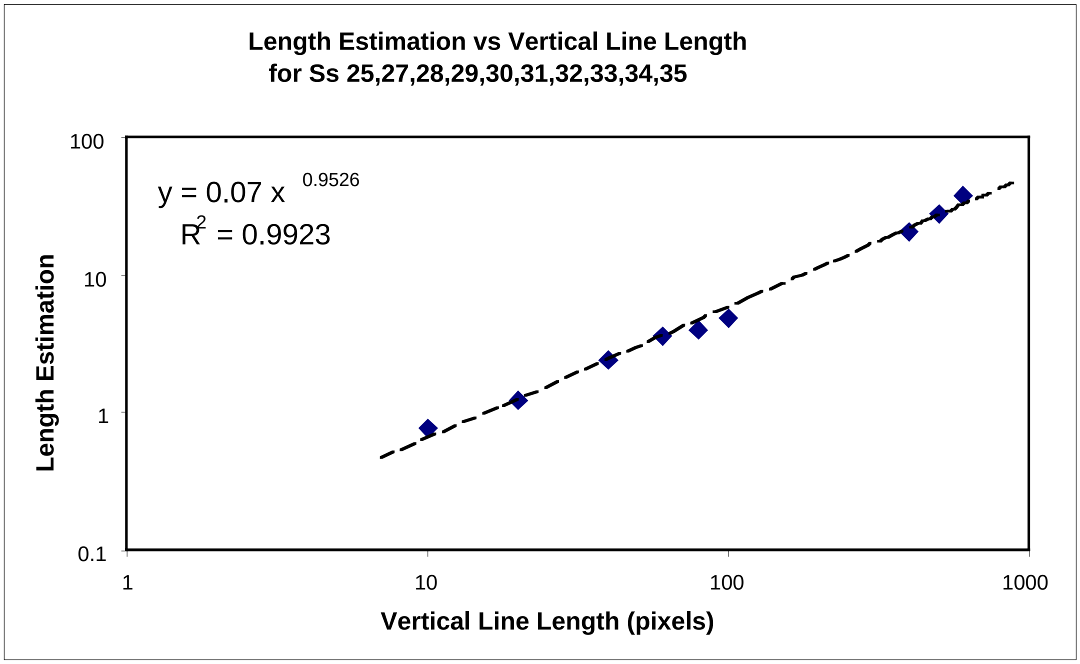


Figure 13. Geometric mean of estimated length against the actual length of vertical lines in Experiment 2. The slope of the fit line is given by the exponent of x. In this case, the regression slope m = 0.95 (SE_m = 0.03; r^2 = 0.992).

Figure 13 presents the estimated length against actual length of vertical lines. Again, the points are very collinear, with the regression slope m = 0.95 (SE_{m} = 0.03), and r^2 = 0.992.

The final perceptual dimension was the size of circles. Figure 14 presents the estimated size of circles against their actual areas (in square pixels). As with the horizontal and vertical lines, the points are very collinear. The regression slope m = 0.59 (SE_{m} = 0.01), and r^2 = 0.997. The fact that the points are spread across such a large range on the x-axis (four orders of magnitude range in circle area), in conjunction with the high r^2 value, indicates a good determination of the equation for the regression line.

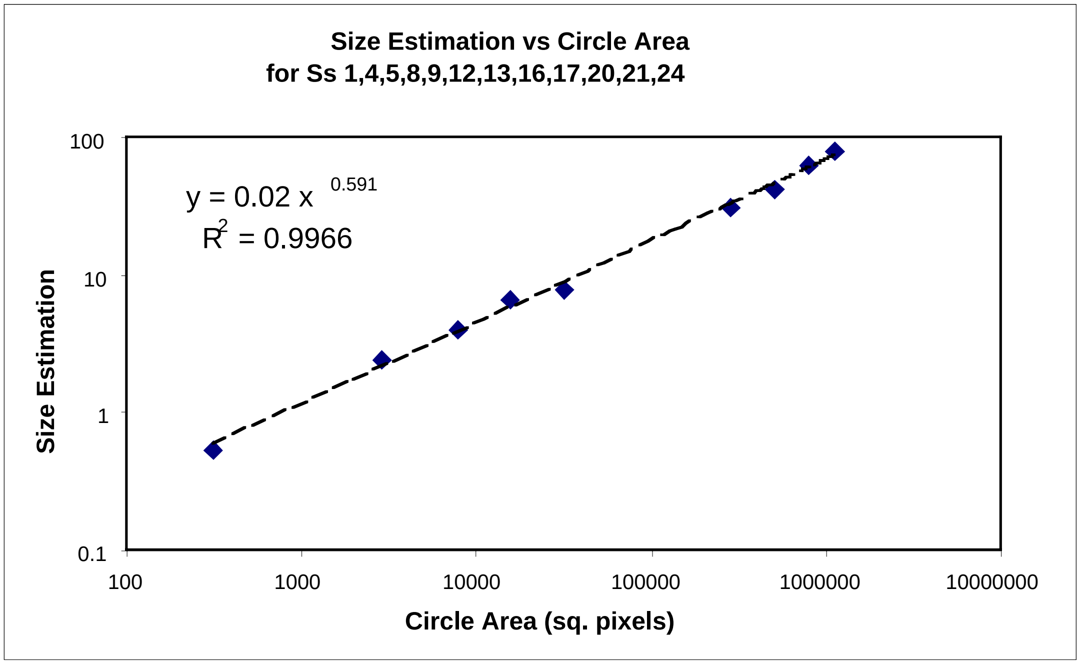


Figure 14. Geometric mean of estimated size against the actual area of filled circles (in square pixels), in Experiment 2. The slope of the fit line is given by the exponent of x. In this case, the regression slope m = 0.59 (SE_{m} = 0.01; r^2 = 0.997).

Conceptual Dimensions: Temperature, Pressure, Velocity

Horizontal and Vertical Line Stimuli

I analyzed the responses for the conceptual data dimensions mapped to the straight-line stimuli next. The analysis was identical to that described for the perceptual dimension of size. Figure 15 shows the plot of the estimated temperature against actual length of horizontal line stimuli. The data are, again, collinear, with high r^2 values. However, the slope of the regression line is m = 0.86 (SE_m = 0.01). This is certainly very different from the slopes just less than 1.0 that were obtained for the estimates of line length already reported here. Figure 16 and Figure 17 include all of the conceptual-to-line length mapping results. The slopes, plus 95% confidence intervals about the means, are included in the summary table, Table 4. Note that the slopes for the different dimensions are all very reliable, and in nearly every case are statistically different from 1.0.

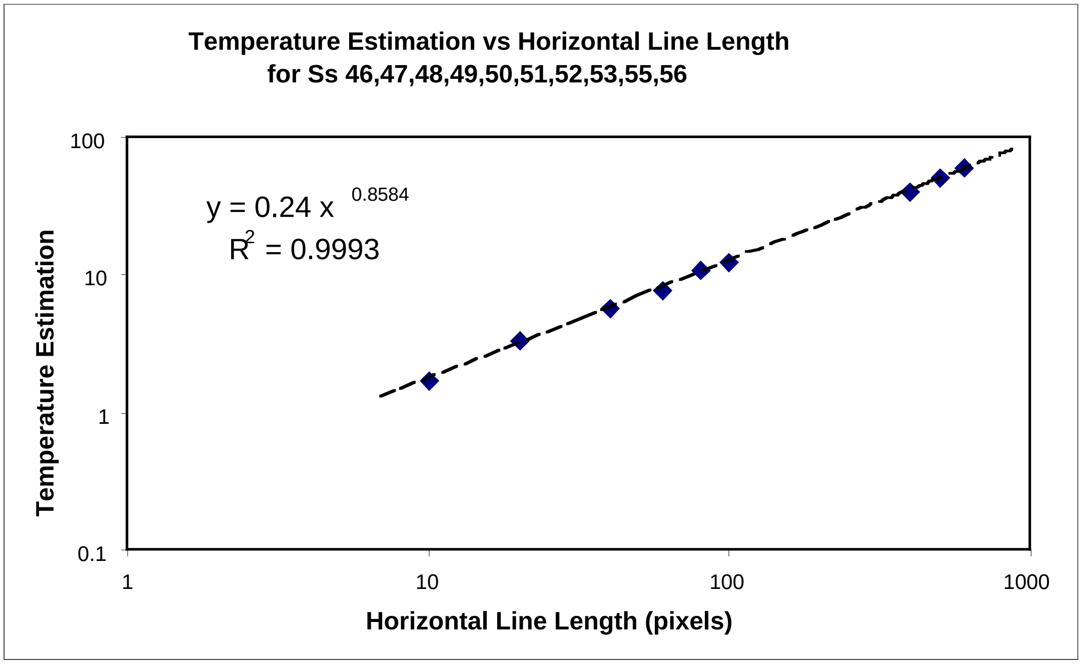


Figure 15. Geometric mean of estimated "temperature" against the actual length of horizontal line stimuli (in pixels), in Experiment 2. The slope of the fit line is given by the exponent of x. In this case, the regression slope m = 0.86 (SE_m = 0.01; r^2 = 0.999).

Circle Stimuli

As I was calculating the geometric means for the conceptual dimensions pressure and velocity compared to the area of circle stimuli, it became clear that while participants' responses were monotonic and quite consistent, for a few of the mappings some responded with polarities different from the majority. For that reason I separated the data that reflected different polarities within a mapping, and analyzed them separately.

The results for magnitude estimations of the data dimensions versus the area of circle stimuli are presented in Figure 18. Note the collinear data and high r^2 values. Also compare the slopes of the lines in this figure to the estimations of perceived size made for the same circle stimuli (refer back to Figure 14). When some subjects produced polarities opposite to the majority, I separated them and produced separate graphs. In those cases (the right side of Figure 18) the data in the graphs are typically not as collinear, likely because they are derived from only three participants.

Table 4. Summary of results from Experiment 2.

Display dimension	**Slope of regression line (center, bold), inside 95% confidence interval (number of participants in that cell)**			
	Temperature	Pressure	Velocity	Size (of Stimulus)
Horizontal Lines	0.84 ≤ **0.86** ≤ 0.88 (10)	0.67 ≤ **0.74** ≤ 0.81 (12)	0.45 ≤ **0.47** ≤ 0.49 (10)	0.92 ≤ **0.96** ≤ 1.00 (12)
Vertical Lines	0.83 ≤ **0.89** ≤ 0.95 (12)	0.49 ≤ **0.53** ≤ 0.57 (10)	0.96 ≤ **1.00** ≤ 1.04 (10)	0.88 ≤ **0.95** ≤ 1.02 (10)
Circles	-	0.51 ≤ **0.56** ≤ 0.61 (9)	0.51 ≤ **0.56** ≤ 0.61 (7)	0.57 ≤ **0.59** ≤ 0.61 (12)
Circles (negative polarities)	-	-0.48 ≤ **-0.95** ≤ -1.42 (3)	-0.45 ≤ **-0.62** ≤ -0.79 (3)	-

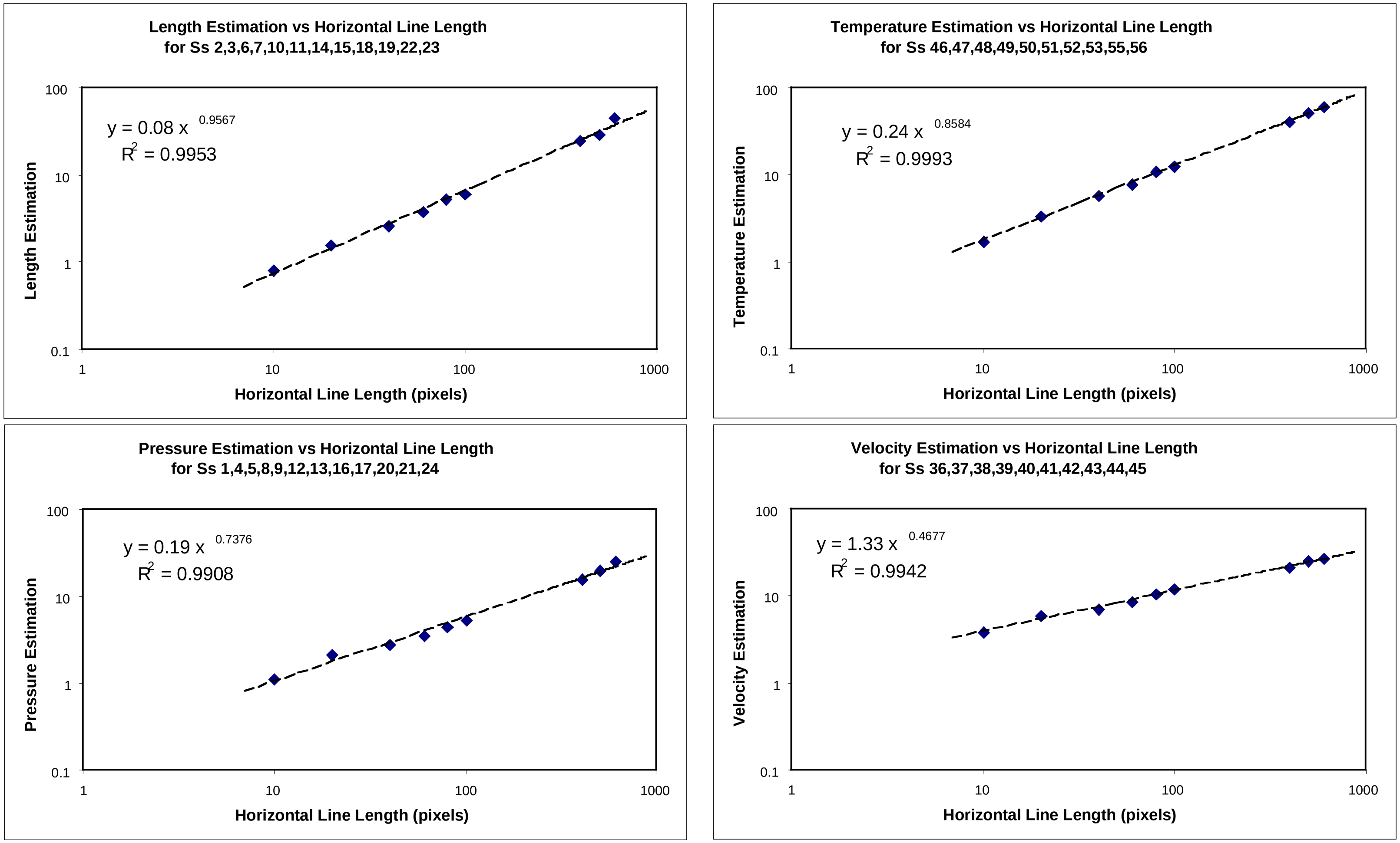


Figure 16. Plots of geometric mean of estimated values against the actual length of horizontal lines in Experiment 2. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The dashed line through the points represents the best fit using the power equation that is included in the plot. The exponent of x gives the slope of the fit line.

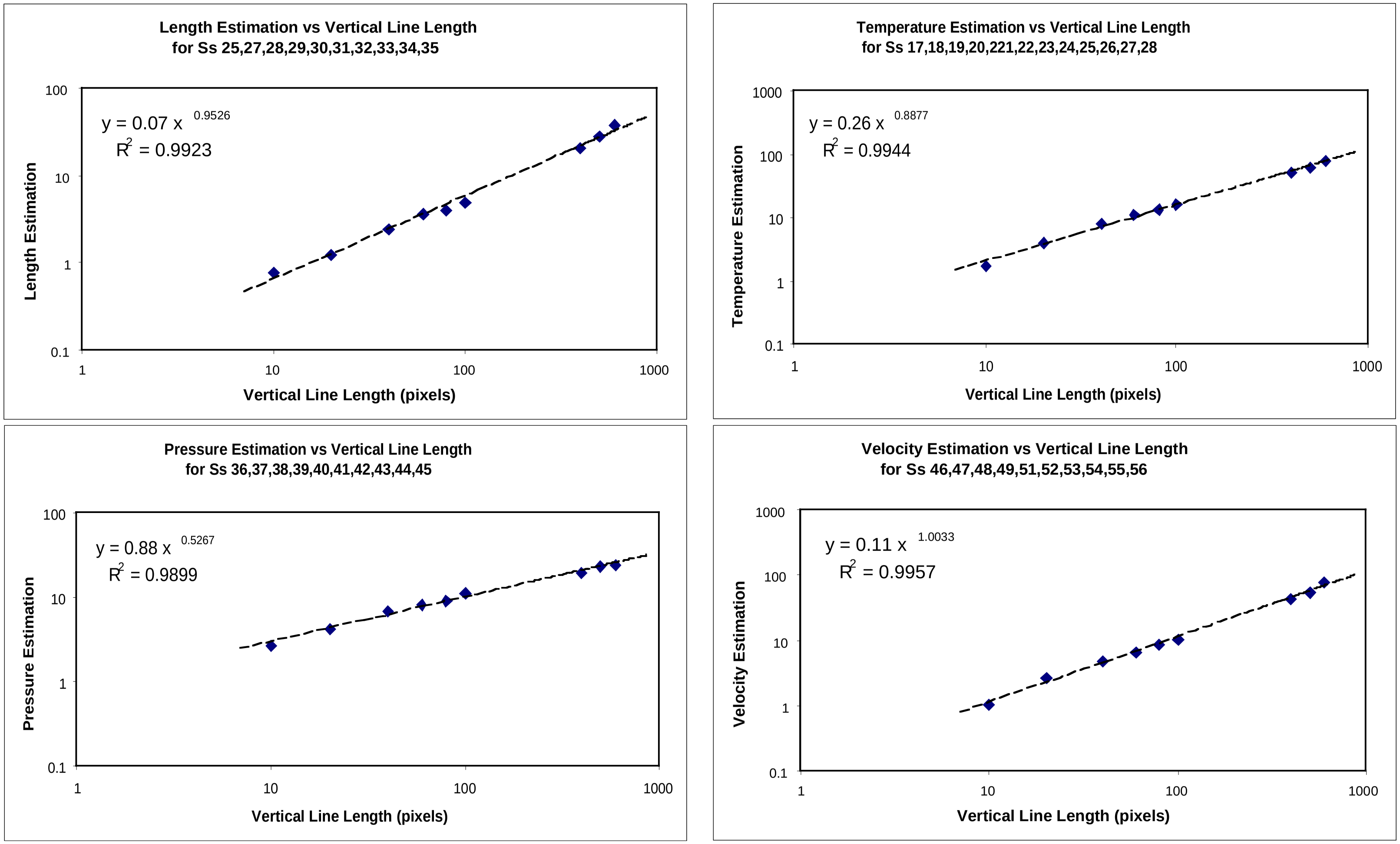


Figure 17. Plots of geometric mean of estimated values against the actual length of vertical lines in Experiment 2. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The dashed line through the points represents the best fit using the power equation that is included in the plot. The exponent of x gives the slope of the fit line.

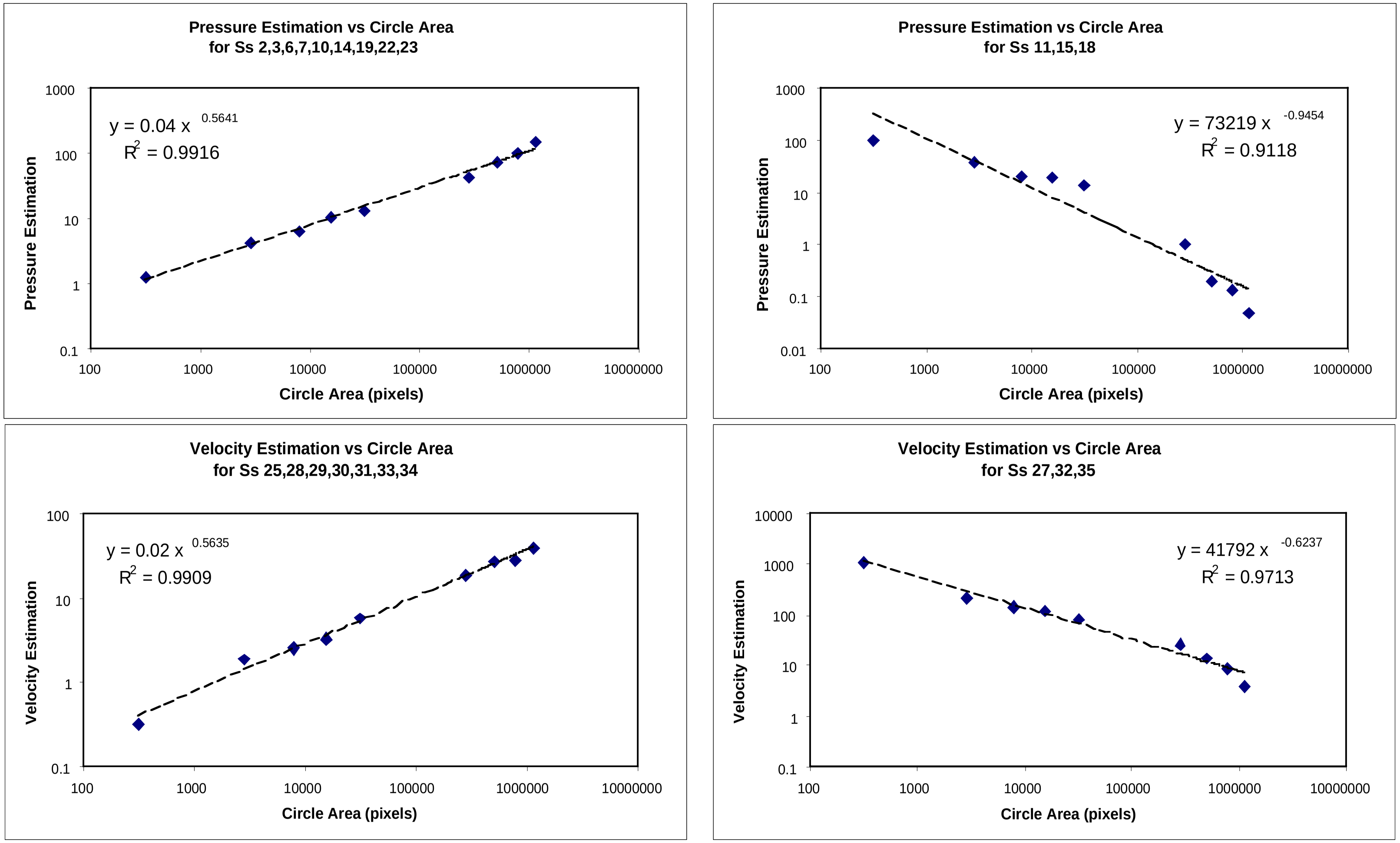


Figure 18. Plots of geometric mean of estimated values against the actual area of filled circles in Experiment 2. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The dashed line through the points represents the best fit using the power equation that is included in the plot. The exponent of x gives the slope of the fit line.

Discussion of Experiment 2

Perceptual Dimension: Size

When participants saw horizontal lines, vertical lines, or filled circles and were asked simply to estimate the size of the stimuli, the resulting plots of perceived size versus actual size were straight lines in log-log coordinates, with high values of r^2. For line stimuli the regression slope was 0.96, very near to the results with similar stimuli reported in the literature (e.g., Teghtsoonian & Teghtsoonian, 1978). This result supports the claim by Stevens (1975) that number and line length are very nearly proportional. For circle stimuli the slope in the present experiment was 0.59, which is slightly lower than, but in the range of previous results (e.g., Teghtsoonian & Teghtsoonian, 1971). The collinearity of the data, and the close agreement with previous results, validates the use of the Web-based experimental procedure to obtain magnitude estimation slopes.

Conceptual Dimensions: Temperature, Pressure, and Velocity

When participants made conceptual magnitude estimations the results were again very collinear, with high r^2 values for the regression lines. However, for line stimuli the slopes were nearly all different from the slopes obtained for the perceptual dimension. For circle stimuli, some of the slopes were quite different from the slope obtained for the perception of circle size. Further, some of the data-to-display pairings yielded different polarities with circle stimuli. While most of the participants responded to the circles with a positive polarity, 25% of the participants responded with a negative polarity. This resulted in two separate slopes, one positive and one negative, being computed for velocity and pressure estimations with circle stimuli.

When asked about her responses, one participant said that she had thought of the circle as a two-dimensional balloon, and figured that there would be lower pressure if the balloon got larger, hence the negative-polarity mapping. This implies that when the participants consider the conceptual data dimensions they at least sometimes use mental models to guide their interpretations, and this may lead to responses that are not what one might expect, given the positive polarities used in “traditional” graphs and figures. Other participants mentioned that they did not know why they had responded with a negative polarity (they did not use that term), but that it had just “felt right” for that particular combination of data and display dimensions.

Implications for Graphing Applications

The two key points from these conceptual data results are as follows: First, both positive and negative polarities may be preferred in some cases. Second, when it comes to the actual slope obtained for a mapping it really does matter what the data dimension is, as well as what the display dimension is. This has important implications for the graphical presentation of data, and a point perhaps narrowly overlooked by Williams

(1956). For instance, to create a visual graph such as a bar chart, the person (or computer) making the graph normally takes the available axis length (e.g., 20 cm) and simply divides that axis by the range in the data (e.g., 100 degrees of temperature) to determine the number of degrees per centimeter. However, better comprehension may result if the number of degrees per centimeter were determined using the slope of a magnitude estimation plot, such as those obtained in the present experiment. Unfortunately, a long history of visual graphing is not likely to be altered at this point in time. Besides, the present research is more interested in the paradigm of magnitude estimation itself, rather than specific results with visual stimuli. In that regard, it seems likely that auditory display dimensions, like frequency and tempo, could also have non-linear relationships to the data dimensions that they would be representing. Thus I pressed ahead into the domain of auditory display dimensions.

Experiment 3: Magnitude Estimation with Sound Stimuli

Purpose

Experiment 2 provided a validation not only of the utility of the magnitude estimation paradigm, but also of the importance of using it in data display applications. Apparently there are interesting differences between the psychophysical scaling functions for different conceptual data dimensions. The next stage in this line of research was to investigate the scaling functions for auditory display dimensions, as a first step in designing successful sonifications.

The present experiment employed the auditory dimensions of frequency and tempo, for the reasons outlined in Experiment 1. In addition to the data dimensions used in Experiment 2 (temperature, pressure, velocity, and size), I added the concept of “number of dollars.” This is a broad-use data dimension, as are the previous ones. In addition, it is likely of interest in many fields, such as economics, where both students and researchers are becoming more interested in using sonification to discover trends in their data.

Magnitude Estimation of Pitch

Just as size was a calibration of the magnitude estimation paradigm in Experiment 2, the perceptual dimensions of pitch and perceived tempo were included in this experiment as calibrations of their acoustic correlates. While the relationship between pitch and frequency has long been studied[10] the literature contains relatively few examples where the magnitude estimation paradigm has been employed. Some other techniques have been used to explore the perception of pitch, and several researchers have come to the conclusion that apparent pitch is not a simple function of the stimulus frequency (Stevens, 1975, p.164). Nevertheless, Stevens and his colleagues (Stevens, Volkmann, & Newman, 1937; Stevens & Volkmann, 1940) used the methods of fractionation and equisection and developed the Mel scale relating perceived pitch to frequency. In the way Stevens (1975, Figure 61) presented it, with a linear axis for pitch and a logarithmic axis for frequency over a frequency range from 10 to 10,000 Hz, the scale certainly does not follow a simple linear relationship (see Figure 19). However, the wide range of frequencies reported, in addition to Stevens's choice of axes, complicates the discussion as far as the present purposes are concerned. Sonification systems will utilize a smaller range of frequencies, and as seen in the previous experiments here, a log-log plot is preferred.

[10] The actual relationship between pitch and frequency was the subject of a classic controversy between Wilhelm Wundt and Carl Stumpf. More specifically, Wundt claimed that the frequency that bisects the apparent distance between two other frequencies should lie at the arithmetic mean. Stumpf argued that the bisecting frequency should lie at the geometric mean. It turns out that both are correct, given the right frequency interval (Boring, 1951; Stevens, 1975).

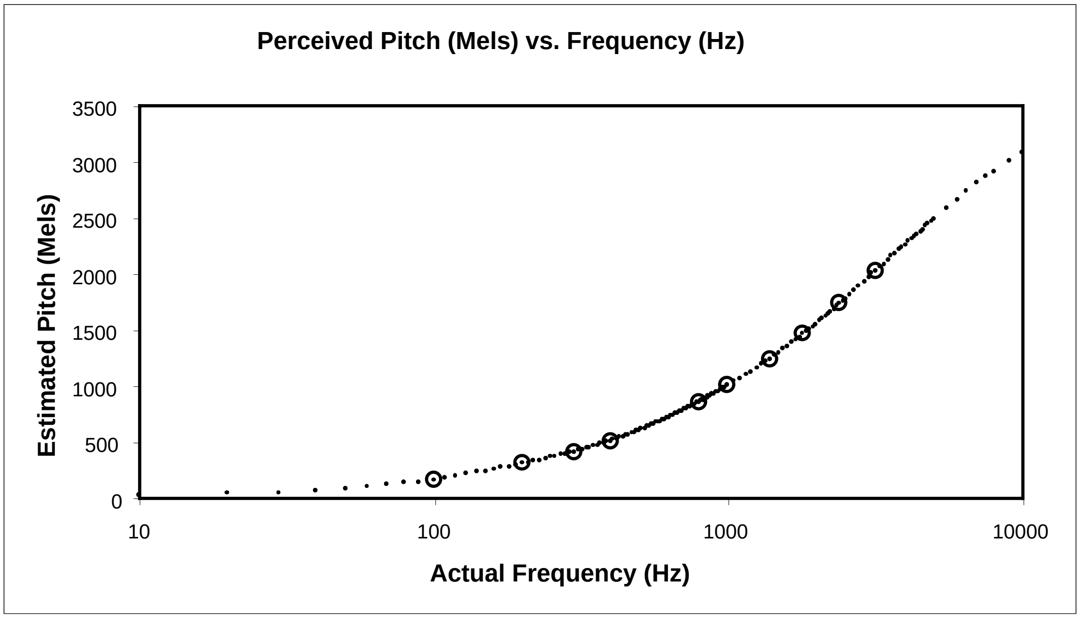


Figure 19. Mel Scale, relating perceived pitch to frequency. From data provided in Stevens, 1975, Appendix B. Note the log-linear axes and the frequency range from 10 – 10,000 Hz. Circles indicate frequencies of tones chosen for use in Experiment 3.

If Figure 19 is re-drawn with log-log axes, more in keeping with the magnitude estimation paradigm, and if the central region of the frequency spectrum is highlighted, a somewhat more familiar picture emerges. Figure 20 presents the Mel scale for the frequency region between 100 and 3200 Hz, which is central to the best human hearing range, appropriate for computer-generated sounds via headphones or moderate-quality speakers, and corresponds to the central frequency range of the piano keyboard. The dashed line through the points is a power-fit with the equation shown in the graph. Note that r^2 is nearly 1.0 and the slope is 0.73. The data lie in a nearly linear pattern, but there is a slight bowing of the data about the fit line.

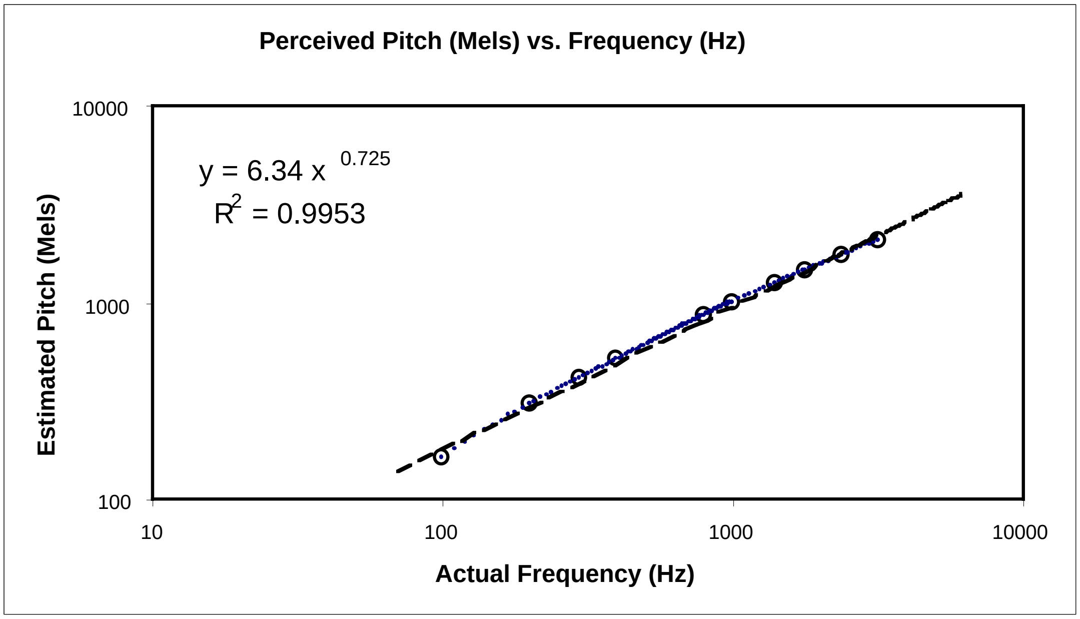


Figure 20. Mel Scale for frequency range from 100 to 3200 Hz. The circles indicate the frequencies used in Experiment 3. The dashed line is a power-fit, with its equation provided in the graph.

Painton, Cullinan, and Mencke (1977) summarize several studies involving pitch-frequency estimations. Overall the studies exhibit a slight downward concavity in the data, but in all cases the high r^2 values indicate that it is reasonable to use the linear regression equation as an excellent approximation. Hence, in the central frequency region one can expect to see linear or very nearly linear data, and one may use the slope of the regression as a measure of the change in pitch associated with the change in frequency. The experimentally-determined slopes should, therefore, be compared to the Mel scale plotted as in Figure 20.

The intention of the Mel scale is to relate perceptions of pitch to the underlying frequencies, and it has proven quite successful in describing observations about musical tones resulting from fractionation and equisection experiments (see, e.g., Moore, 1997). However, it is also important to note that when more modern magnitude estimation procedures have been used, and when less musical tones have served as stimuli, the results have tended to produce a slope just slightly steeper than the Mel scale (Beck & Shaw, 1961, 1962, 1963).[11] The slopes of the functions obtained by Beck and Shaw (1961, 1962, 1963) can only be estimated from their plots, as the full data were not provided in the text. However, a visual estimate of the data points confirms a slope slightly steeper than the Mel slope. Thus, if a free-modulus magnitude estimation procedure is used with frequencies between 100 and 3200 Hz, the expected relationship between perceived pitch and frequency should produce near-traditional psychophysical scaling plots, with slopes in the range of, say, 0.73 to 0.80.

Magnitude Estimations for Tempo

I am not aware of any magnitude estimation experiments involving tempo per se. However, Eisler (1976) has compiled a list of 111 studies, published on four continents over the span of more than a century, that have attempted to obtain scaling estimates for duration. Since the perception of non-syncopated tempo is highly related to the perception of the duration of the elements, Eisler’s compilation may be helpful here.

[11] In particular, slopes that are slightly steeper than the Mel scale have been obtained with a mid-frequency reference tone, and a mid-range modulus. Lower-frequency reference tones have produced different results, but that paradigm is no longer the recommended magnitude estimation approach (Stevens, 1975).

Eisler includes studies of duration across many modalities (e.g., lights, tones, vibrations), and with a variety of users (e.g., children, adults, schizophrenic youths). Across all of the studies, Eisler (1976, p.1157) concludes that, “time perception is not veridical; though the collected exponents straddle unity, most of them are smaller than 1…[A] value of .9 seems to come closest to the exponent of subjective duration.” Hence, we may assume that slopes for the estimations of tempo should also be slightly less than 1.0.

Method

Participants

From the same subject pool as Experiments 1 and 2, a total of 132 students participated (40 males and 92 females; mean age 19.5 years).

Apparatus

The apparatus was identical to the apparatus in Experiments 1 and 2. In this experiment the stimuli were presented through the headphones.

Stimuli

Auditory Stimuli

The auditory stimuli were presynthesized at 16-bit, 44.1 kHz using Csound to create AIFF files, as in Experiment 1. The Csound orchestra and score files used in the synthesis are included in Appendix B. The sounds were again composed of a one-beat long pure sine wave tone, followed by a half-beat of silence. There were two sets of stimuli. The 10 sounds in the Frequency Set were synthesized with frequencies of 100, 200, 300, 400, 800, 1000, 1400, 1800, 2400, and 3200 Hz, but were all played at a tempo of 60 bpm. The 10 sounds in the Tempo Set were all synthesized with a frequency of 1000 Hz but were repeated at tempos of 45, 60, 105, 150, 210, 270, 420, 500, 550, and 600 bpm.

The amplitude envelope of the tones included a 0.1-beat linear ramp onset (attack) and offset (decay). I normalized the stimuli in the Frequency Set for perceived loudness by starting with relative amplitude values from equal-loudness contours (Robinson & Dadson, 1956), then making minor adjustments based on pre-testing. The actual adjustments were made by changing the Csound synthesis parameters for the amplitude of each sound. To ensure the same relative levels, all of the sounds were synthesized into one long sound file and then separated into individual files using a sound-editing program. Table 5 contains the sound pressure level (SPL) values for the stimuli in the Frequency Set. All of the sounds in the Tempo Set were synthesized at 60 dB.

Table 5. Relative sound pressure level (SPL) for stimuli that varied in frequency in Experiment 3.

Frequency (Hz)	100	200	300	400	800	1000	1400	1800	2400	3200
SPL (dB)	69	66	65	62	61	60	60	60	58	56

Note: The sounds were adjusted relative to the reference tone of 1000 Hz at 60 dB. Loudness values were determined initially from equal-loudness contours (Robinson & Dadson, 1956), and then minor adjustments were made by ear.

Visual Stimuli: Horizontal Lines

For the block of subjects who saw visual line stimuli rather than hearing auditory stimuli, the horizontal lines were identical to those used in Experiment 2.

Design

The experimental design and implementation of this experiment was nearly identical to that of Experiment 2. The design again included the between-subjects factor of data dimension (temperature, pressure, velocity, size, number of dollars, pitch, tempo) and the within-subjects factor of display dimension (in this case, frequency and tempo). In this experiment temperature, pressure, velocity, size, and number of dollars were treated as conceptual data dimensions. The data dimensions of pitch and perceived tempo were treated as perceptual dimensions, and were included as a calibration of the procedure, much like the perceptual dimension of size (i.e., length of line) in Experiment 2.

Each data dimension (e.g., temperature) was paired with one of the display dimensions (e.g., frequency) for an entire block of trials (named, e.g., temperature+frequency). The perceptual dimension of pitch was only paired with frequency, and the perceptual dimension of perceived tempo was only paired with tempo since those data dimensions were calibrations of their respective auditory dimensions. Thus, there were 10 possible block types. Each participant again completed two blocks of trials separated by a brief break. The block types were assigned pseudo-randomly as before.

Table 6 presents the number of participants who completed each trial block type. Originally, 20 participants were planned for each trial block type, but to conserve subject hours this was scaled back to 16 per cell. However, some cells had already received more than 16 participants by that point. Some data from the velocity+tempo and tempo+tempo block types were discarded due to a computer error, resulting in fewer participants in those cells.

In addition to the new set of display and data dimensions, 16 participants in this experiment completed the magnitude estimation procedure between horizontal line length and perceived length simply as a replication of one of the visual trial block types from Experiment 2. The procedure, stimuli and word cues for this block of trials were identical to those in Experiment 2. These participants did not complete any blocks of trials using auditory stimuli. However, they completed a second block of trials with other visual stimuli (and no blocks with auditory stimuli). Those extra data were not included in this experiment.

Table 6. Number of participants who completed each block type in Experiment 3.

Display dimension	**Data dimension**						
	Temperature	Pressure	Velocity	Size	Dollars	Pitch	Tempo
Frequency	16	16	16	20	16	17	-
Tempo	20	20	12	16	16	-	13
Horizontal Lines	-	-	-	16	-	-	-

Trial Structure and Task

At the beginning of each block of trials, instructions like the following were displayed on the screen and read aloud by the experimenter:

> You will hear a series of sounds, one at a time, in random
> indicate what temperature they would represent, by assigning
> For the first sound, assign it any number of your choosin
> temperature. Then, for each of the remaining sounds, estimat
> relative to your subjective impression of the first one. For
> sound seems to represent a temperature that is 10 times as
> assign it a number that is 10 times the first number. If
> represent a temperature that is one-fifth as hot, assign it
> as large as the first number, and so on. You can use any
> fractions, or decimals that seem appropriate, as long as they

Figure 21 shows the screen layout for this experiment. It is nearly identical to the layout in Experiment 2. However, in this experiment there were no visual stimuli. In the center of the screen there was a button that the listener clicked to hear the sound. On each trial the participant heard one stimulus from the set being used for that block (e.g., sounds from the Frequency Set) and provided a number for the data dimension (e.g., temperature) in a text-entry box on the screen. Participants simply filled in their response and then clicked the “Next” button on the screen or pressed the “*return*” key on the keyboard to proceed to the next trial.

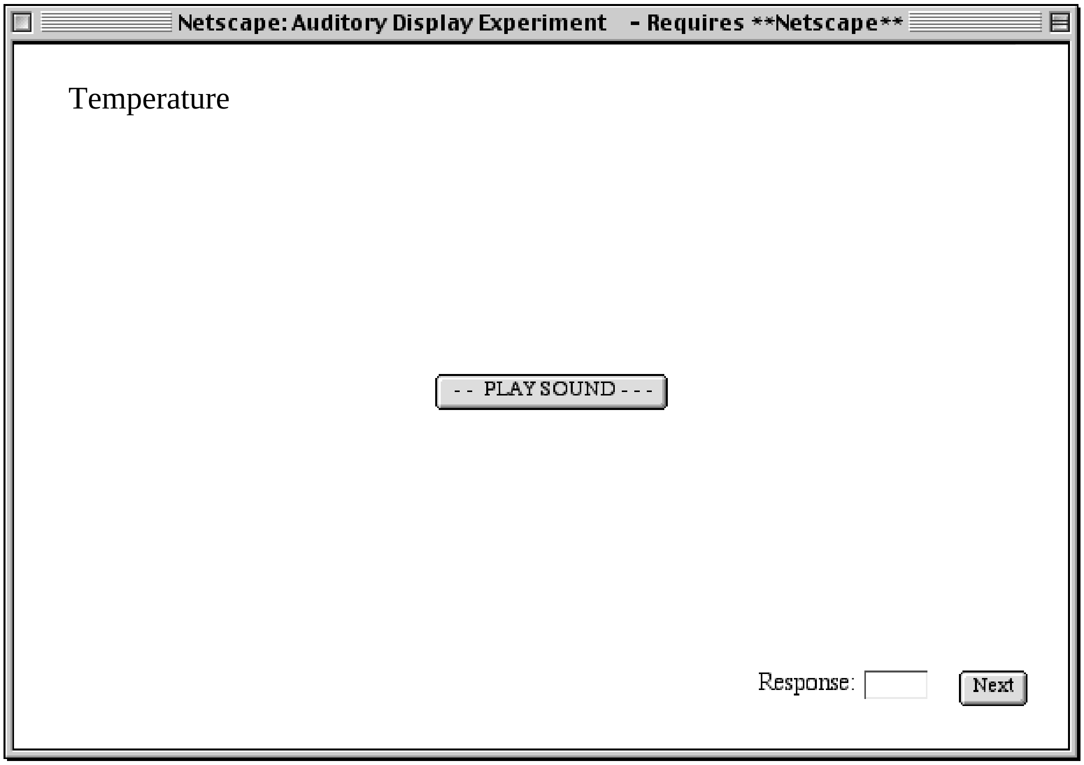

Netscape: Auditory Display Experiment - Requires **Netscape**
Temperature
- - PLAY SOUND - - -
Response:
Next

Figure 21. Image of the screen layout seen by participants in Experiment 3. The data dimension was displayed at the top left. The stimuli were played when the participant clicked the button in the center of the screen. Participants entered their responses in the text-entry box at the bottom right, and clicked the “Next” button to continue.

In a block of 20 trials, each stimulus was randomly presented twice, with the constraint that neither of the two extreme stimuli could occur first (as in Experiment 2). Following a brief rest, the participant began the second block with new instructions that introduced a different data dimension and the other display dimension.

Results

Horizontal Line Stimuli: A Replication

The block of trials comparing perceived size (length) with actual length of horizontal lines was a replication of part of Experiment 2, and the data were analyzed identically. Figure 22 shows the plot of the geometric mean of estimated length against the actual length of horizontal lines. The identical pattern resulted, with highly collinear points. In the case of length estimation versus horizontal line length in the present experiment, the regression slope m = 0.98 (SE_m = 0.02), and r^2 = 0.996.

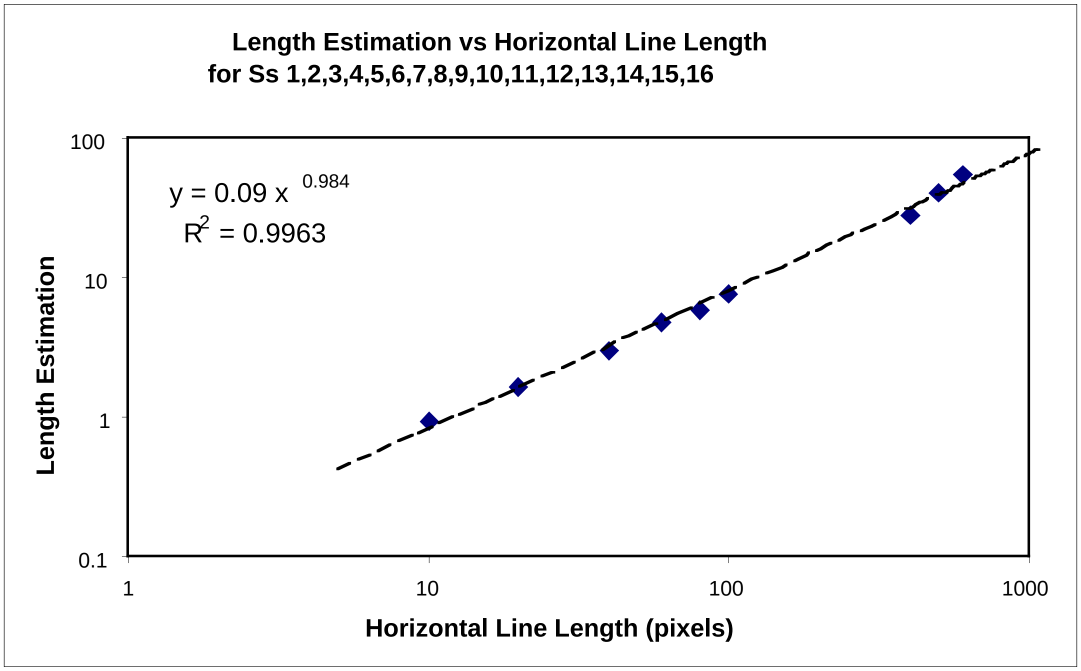


Figure 22. Geometric mean of estimated length against the actual length of horizontal lines in Experiment 3. This is a replication of the results obtained in Experiment 2. The slope of the fit line is m = 0.98 (SE_m = 0.02; r^2 = 0.996).

Auditory Stimuli: Individual Analyses

I planned to analyze the data as I had done in Experiment 2. However, since this was a new type of stimuli, I decided to look at the responses of the individual participants first. Most participants applied a consistent mapping polarity (be it positive or negative), and made fairly monotonic responses, so that, for example, low frequencies were given lower numbers and higher frequencies were given higher numbers. This monotonic performance had been followed by all of the participants in Experiment 2, where the stimuli were visual. However, in looking at the responses of some of the individual participants with auditory stimuli in the present experiment, it became clear that some were not as consistent as the others were. In any experiment one expects a few participants simply to “not get” the task. In this experiment, since none of the participants had ever likely contemplated “what temperature” or “how many dollars” a sound might represent, it is not surprising that several simply did not seem to get the idea of a monotonic mapping.

Since the goal of the study was to determine the slope of a (monotonic) mapping function of the data dimensions to the display dimensions, only data from the participants who responded in a somewhat consistent manner were used. To quantify the exclusion of some participants’ data I decided to take the simple Pearson correlation coefficient, *r,* between the perceived data value and the actual stimulus parameter (frequency or tempo).

Participants who responded more or less consistently (i.e., monotonically) would obtain a correlation coefficient approaching ±1.0. However, at this point I realized that a slightly more appropriate test would be the correlation coefficient between the log-transformed responses and the log-transformed actual values. After all, the magnitude estimation slopes would come from log-log plots. Thus, it was more important that responses be consistent in the log-log domain. Further, it turned out that if responses were consistent with the linear *r*, they were usually even more consistent with the log-based *r*. Conversely, low linear *r*-values usually resulted in even lower log-based *r*-values. This pattern made it even more diagnostic to use the log-based *r* values to decide whether to exclude a given subset of data. Thus, with two responses for each of the 10 stimuli (18 *df*), data sets with less than $r_{\text{critical}} = 0.444$ would not reach conventional levels of statistical significance for the correlation coefficient. Therefore, I excluded data (approximately 12%) where the absolute value of the correlation coefficient between the log of the responses and the log of the actual parameter values was less than 0.444.[12] Table 7 shows the number of participants whose data remained in each data-to-display block type, as a proportion of the participants who initially completed that block type.

[12] Note that this is a fairly generous limit. Teghtsoonian (1980, p.296) has used much more stringent requirements, such as excluding data which does not achieve an r^2 of 0.70 (i.e., 70% of the variance in log judgment accounted for by variation in log target value). However, Teghtsoonian and Teghtsoonian (1978, Note 1, p.313) have also pointed out that, “Since r^2 is statistically sensitive to range effects and is also monotonically related to the slope of the regression line, it is clear that to discard subjects with low r^2 is to invite a systematic bias in the resulting estimates of the exponent.”

Table 7. Proportion of participants whose data were collinear enough to be used in the analyses for Experiment 3.

Display dimension	**Data dimension**						
	Temperature	Pressure	Velocity	Size	Dollars	Pitch	Tempo
Frequency	13 / 16	12 / 16	16 / 16	19 / 20	11 / 16	16 / 17	-
Tempo	17 / 20	15 / 20	11 / 12	16 / 16	12 / 16	-	13 / 13
Horizontal Line	-	-	-	16 / 16	-	-	-

Note: Data were excluded if the correlation coefficient between the log of the responses and the log of the actual parameter values was less than $r = 0.444$.

I also investigated the individual data for any evidence that demographic variables might affect the mapping polarity or correlation coefficient. I tested for significant correlation between gender, handedness, and the number of years of music training, as provided during the demographics phase of the experiment. There were no significant correlations between any of the demographic variables and the consistency of responding, for any of the conditions. However, it should be noted that there were often only two or fewer left-handed participants, or three or fewer of either males or females in a given trial block type.

Auditory Stimuli: Aggregate Analyses

Perceptual Dimensions: Pitch and Perceived Tempo

As before, I calculated the geometric means of all judgments for all participants in a given data and display pair, and plotted them against the actual stimulus parameters.

Where necessary, I excluded inconsistent data as described above. First, the perceptual dimensions of pitch and tempo were plotted against their acoustic parameters of frequency and tempo, respectively. Figure 23 presents the plot of pitch (i.e., perceived frequency) vs. actual frequency. The slope of the regression line was m = 0.78 (SE_m = 0.05; r^2 = 0.96). Figure 24 presents the plot of perceived tempo vs. actual tempo, where the regression slope was m = 0.95 (SE_m = 0.05; r^2 = 0.98).

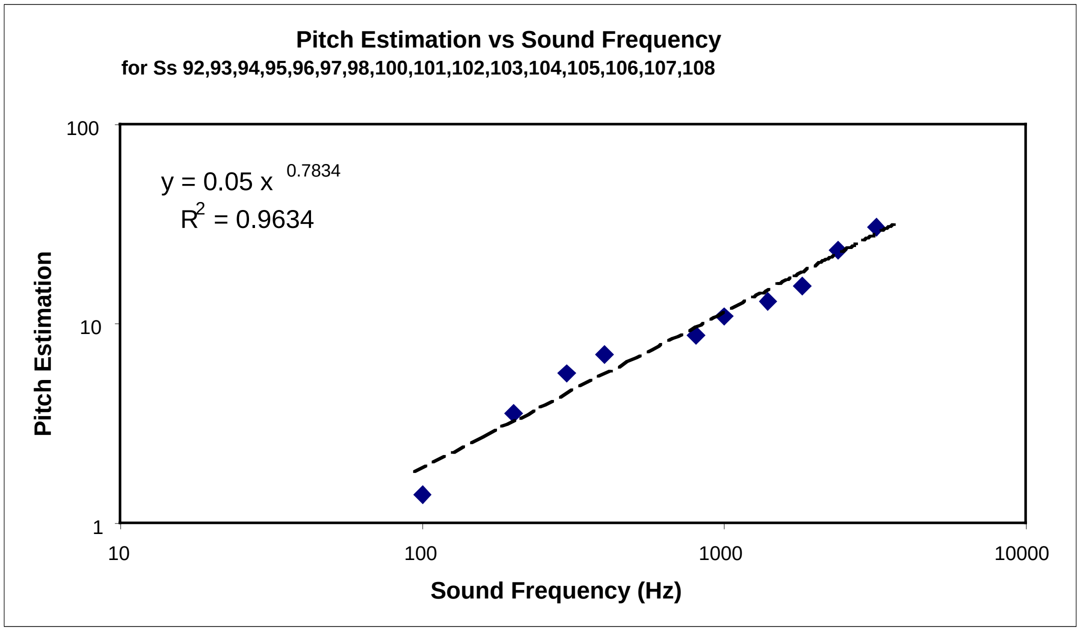


Figure 23. Geometric mean of estimated pitch against the actual frequency (in Hz) of sound stimuli in the Frequency Set in Experiment 3. The slope of the fit line is m = 0.78 (SE_m = 0.05; r^2 = 0.96).

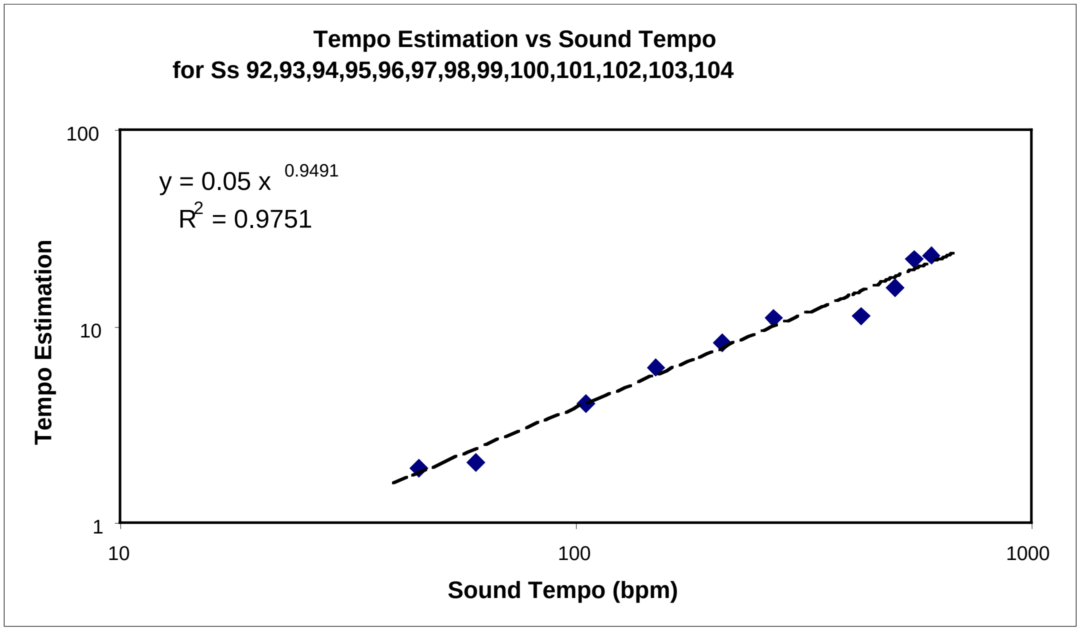


Figure 24. Geometric mean of estimated tempo against the actual tempo (in bpm) of sound stimuli in the Tempo Set in Experiment 3. The slope of the fit line is m = 0.95 (SE_m = 0.05; r^2 = 0.98).

Conceptual Dimensions

Following the perceptual dimensions, I began to examine the data for the conceptual dimensions of temperature, pressure, velocity, size, and number of dollars. As I was assessing the individual subject data for inclusion or exclusion, it was clear that some of the subjects had consistent positive mappings, whereas others in the same block type had consistent negative mappings. Both groups of subjects were to be included based on the absolute value of their correlation coefficients, however it was necessary to separate the positive and negative polarities for further analysis in those blocks.

As an example of a conceptual dimension, Figure 25 contains the plot of estimated pressure against actual frequency of the sound. The data follow the trend seen throughout Experiments 2 and 3, with highly collinear points along a slope less than 1. The majority of participants responded with a positive polarity, as well. In particular the regression slope for pressure versus frequency resulting from the majority of listeners was m = 0.78 (SE_m = 0.05; r^2 = 0.97). As indicated, a few participants responded with an inverse polarity. The corresponding plot for subjects responding with a negative polarity is included in Figure 26, later in this section. The pattern of results is very similar in that plot, just with a different (and negative) slope m = -0.49 (SE_m = 0.08; r^2 = 0.83). Note, also, that the data from that block type came from only four participants, resulting in a lower r^2 than for other block types.

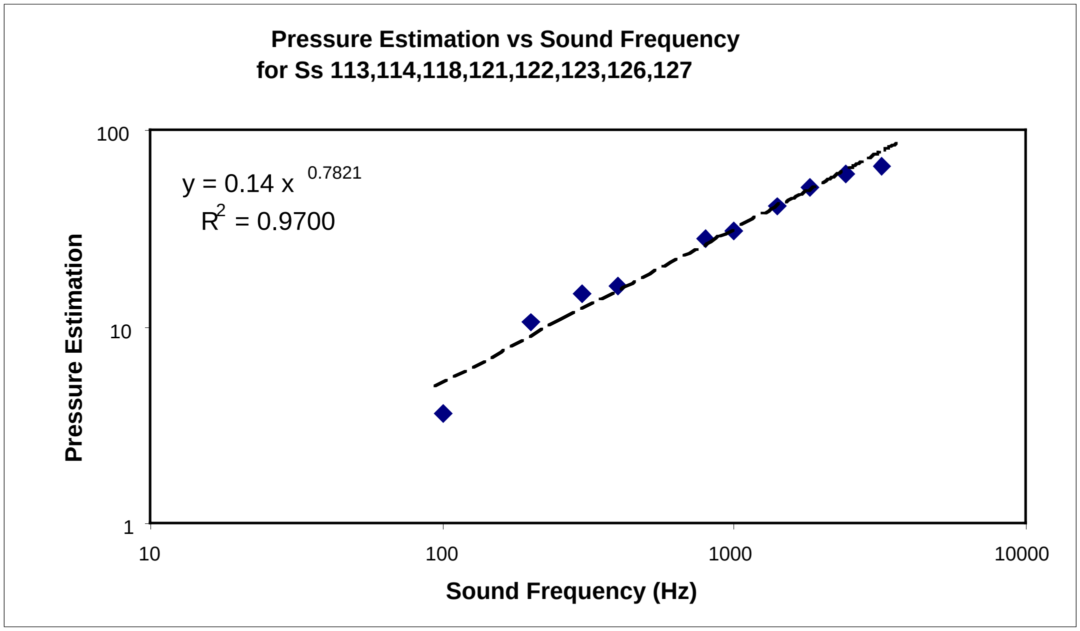


Figure 25. Geometric mean of estimated pressure against the actual frequency (in Hz) of sounds in the Frequency Set in Experiment 3. The slope of the fit line is m = 0.78 (SE_m = 0.05; r^2 = 0.97).

The example above reflects a majority of listeners following the positive polarity and a few responding with the negative polarity. That indicates a slightly ambiguous mapping preference. On the other hand, for some mappings the majority, or even unanimous, response pattern was in the negative polarity. For example, the majority of participants in the size+frequency block responded that increasing frequency corresponded to decreasing size. Likewise, all participants in the size+tempo block responded that increasing tempo corresponds to decreasing size.

The slopes of all of the data-to-display mappings are summarized in Table 8. The top half of a cell in the table indicates the positive polarity results, including the regression slopes enclosed in its 95% confidence interval and the number of subjects whose data contributed to that slope. The lower half of a cell indicates the same information for the negative polarity. The slope for the most common polarity is underlined.

The actual plots for each of the cells are presented in Figure 26 and Figure 27, with the positive polarities in the left side of the figure, and the negative polarities in the right side. Note that in some cases there were unanimous results, hence only one polarity is presented in the figure.

Table 8. Summary of results from Experiment 3.

Display dimension		**Slope of regression line (center, bold), inside 95% confidence interval (number of participants in that cell)**						
	Polarity	Temperature	Pressure	Velocity	Size	Dollars	Pitch	Tempo
Frequency	+^{ve}	**0.95** 0.68 ≤ m ≤ 1.22 (11)	**0.78** 0.66 ≤ m ≤ 0.89 (8)	**1.06** 0.99 ≤ m ≤ 1.13 (14)	**0.90** 0.78 ≤ m ≤ 1.02 (7)	**0.77** 0.539 ≤ m ≤ 1.00 (6)	**0.78** 0.67 ≤ m ≤ 0.89 (16)	
	-^{ve}	**-0.69** -0.18 ≤ m ≤ -1.19 (2)	**-0.49** -0.24 ≤ m ≤ -0.74 (4)	**-0.17** +0.21 ≤ m ≤ -0.551 (2)	**-0.76** -0.58 ≤ m ≤ -0.94 (12)	**-0.50** -0.39 ≤ m ≤ -0.61 (5)	(none)	
Tempo	+^{ve}	**0.43** 0.34 ≤ m ≤ 0.52 (11)	**0.68** 0.54 ≤ m ≤ 0.82 (10)	**1.04** 0.91 ≤ m ≤ 1.19 (11)	(none)	**0.66** 0.47 ≤ m ≤ 0.85 (8)		**0.95** 0.84 ≤ m ≤ 1.06 (13)
	-^{ve}	**-0.48** -0.35 ≤ m ≤ -0.61 (6)	**-0.72** -0.55 ≤ m ≤ -0.89 (5)	(none)	**-0.94** -0.79 ≤ m ≤ -1.09 (16)	**-0.46** -0.27 ≤ m ≤ -0.65 (4)		(none)
Horizontal Lines					**0.98** 0.94 ≤ m ≤ 1.02 (16)			

Note: The slope for the most popular polarity for each data-to-display pair is underlined.

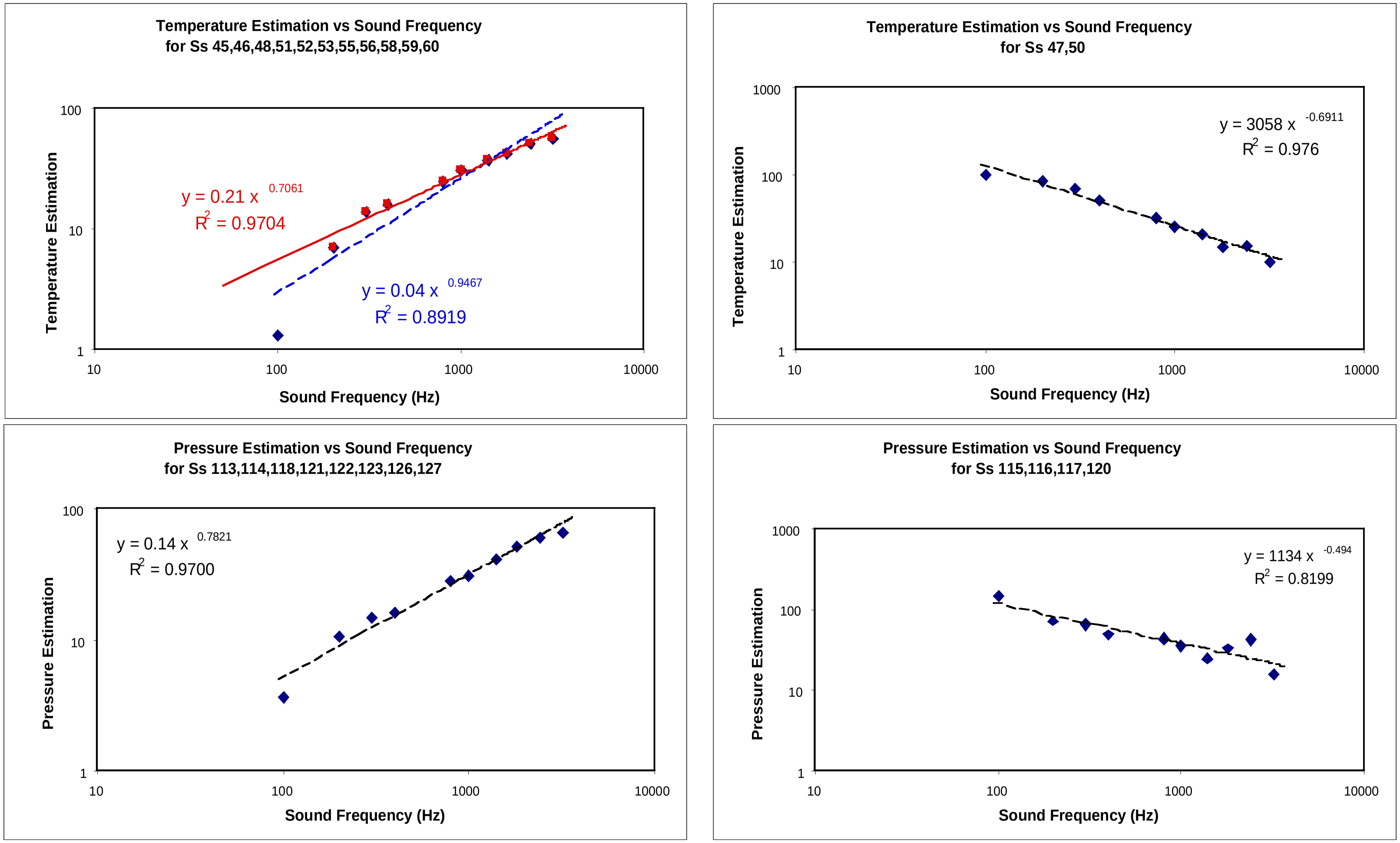


Figure 26. Plots of geometric mean of estimated values against the actual frequency of sound stimuli in Experiment 3. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The slope of the fit line is given by the exponent of x in the equation that is included in each plot. Continued next page…

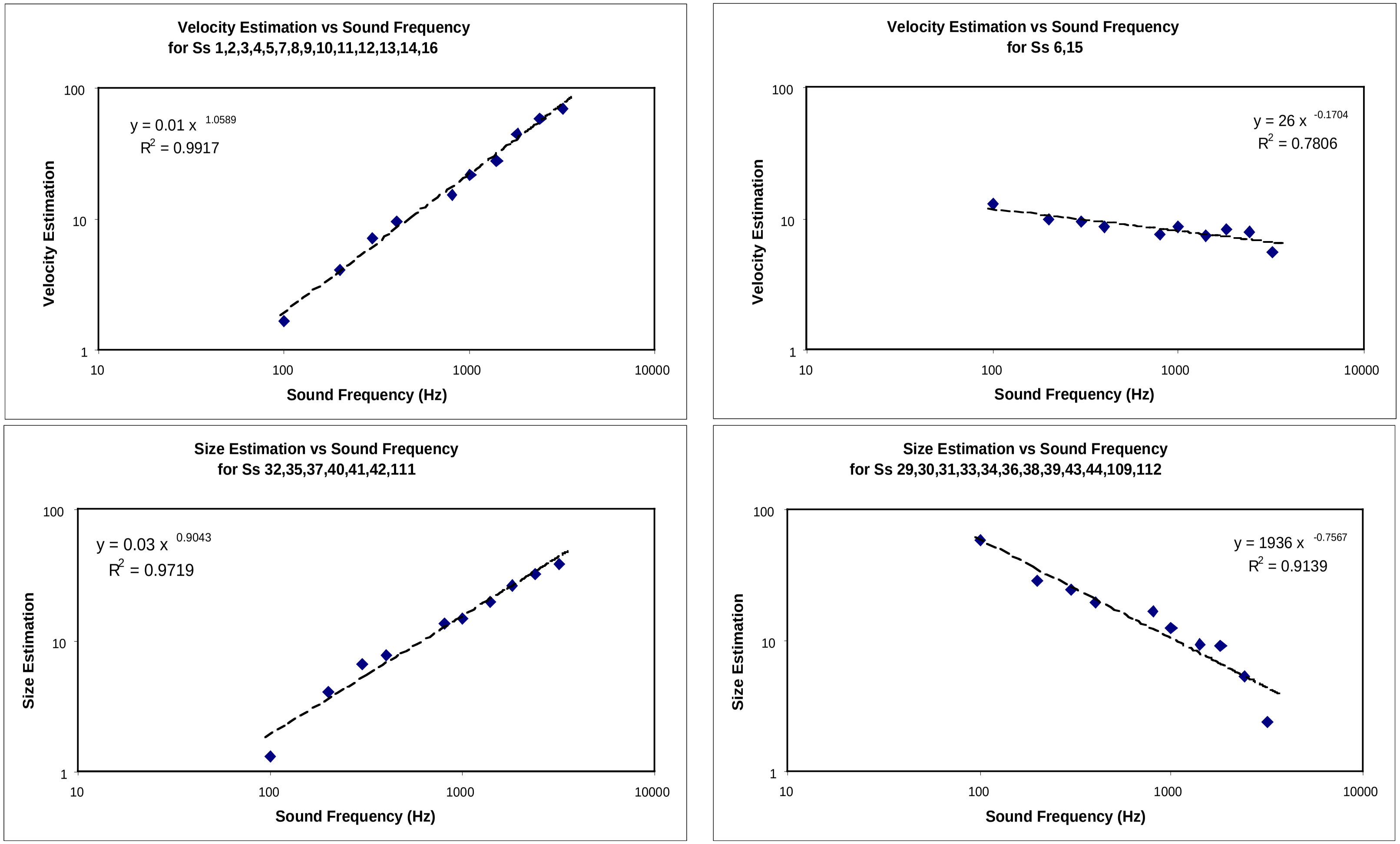


Figure 26, continued. Plots of geometric mean of estimated values against the actual frequency of sound stimuli in Experiment 3. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The slope of the fit line is given by the exponent of x in the equation that is included in each plot. Continued next page…

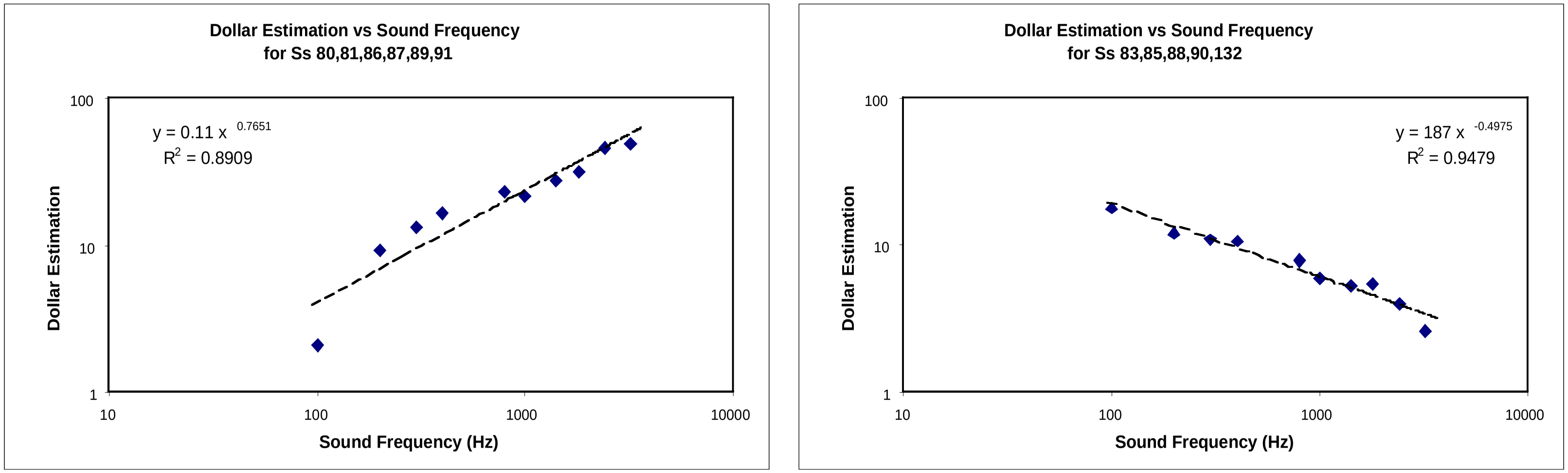


Figure 26, continued. Plots of geometric mean of estimated values against the actual frequency of sound stimuli in Experiment 3. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The slope of the fit line is given by the exponent of x in the equation that is included in each plot.

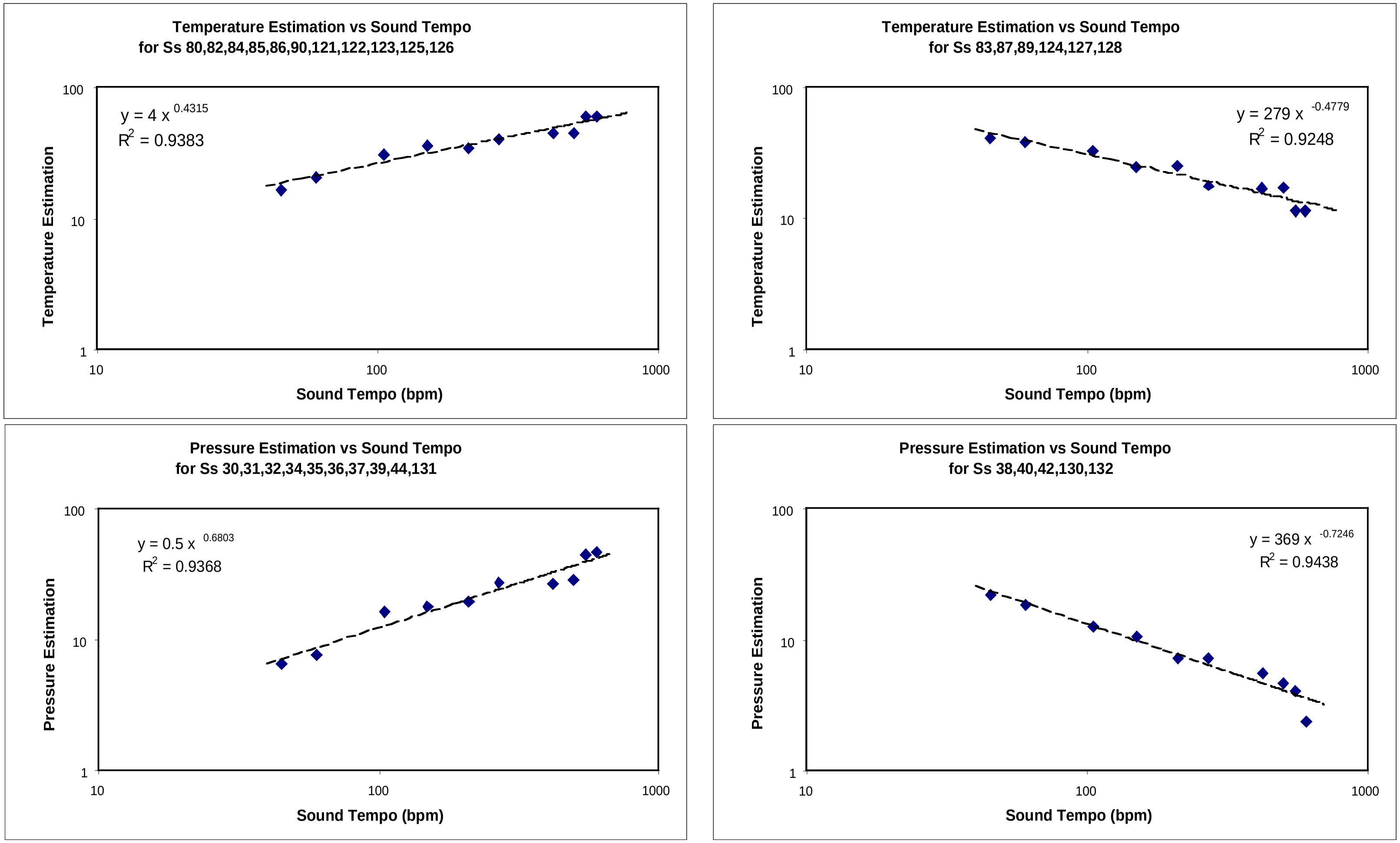


Figure 27. Plots of geometric mean of estimated values against the actual tempo of sound stimuli in Experiment 3. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The slope of the fit line is given by the exponent of x in the equation that is included in each plot. Continued next page…

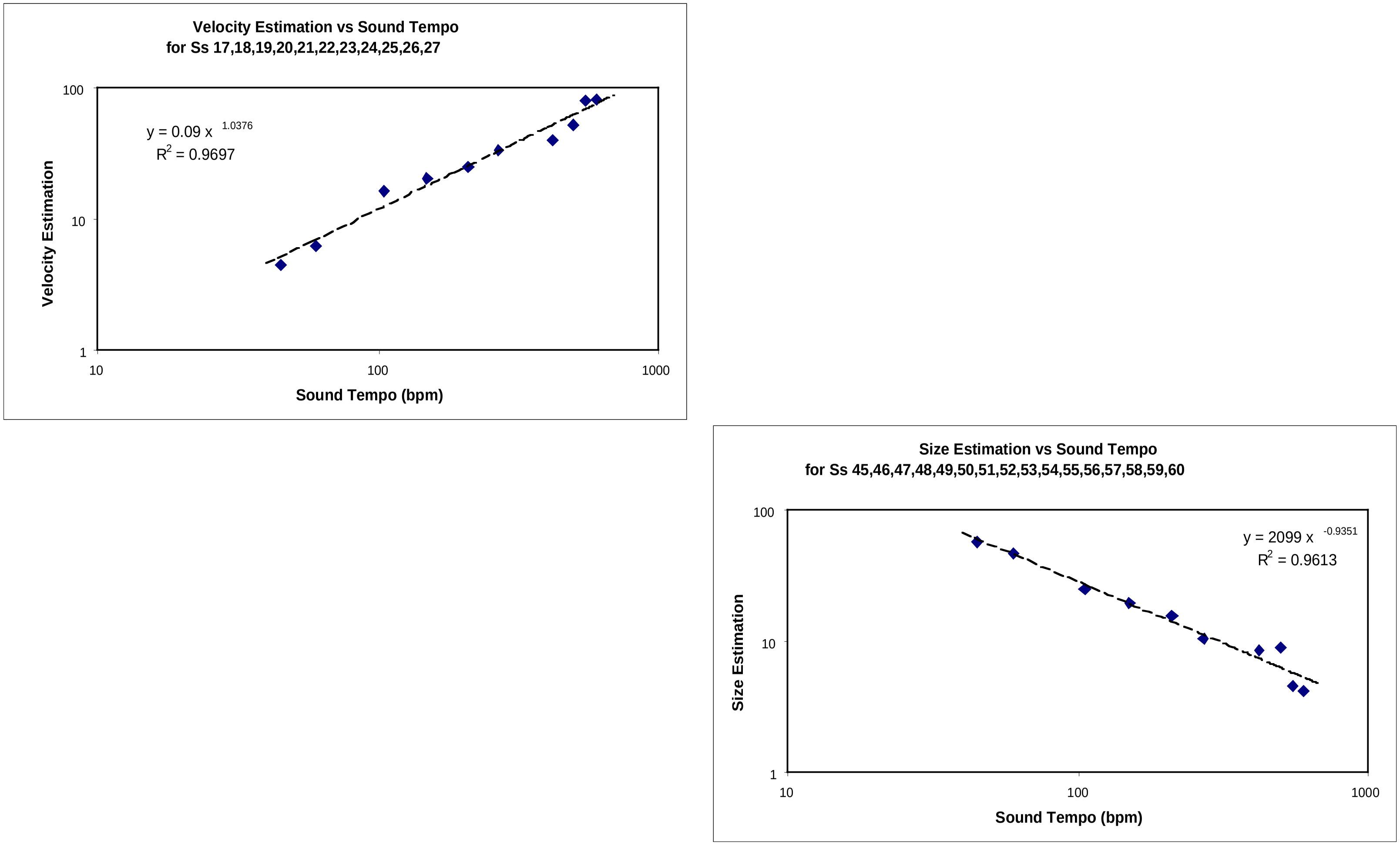


Figure 27, continued. Plots of geometric mean of estimated values against the actual tempo of sound stimuli in Experiment 3. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The slope of the fit line is given by the exponent of x in the equation that is included in each plot. Unanimous polarities result in only one plot. Continued next page…

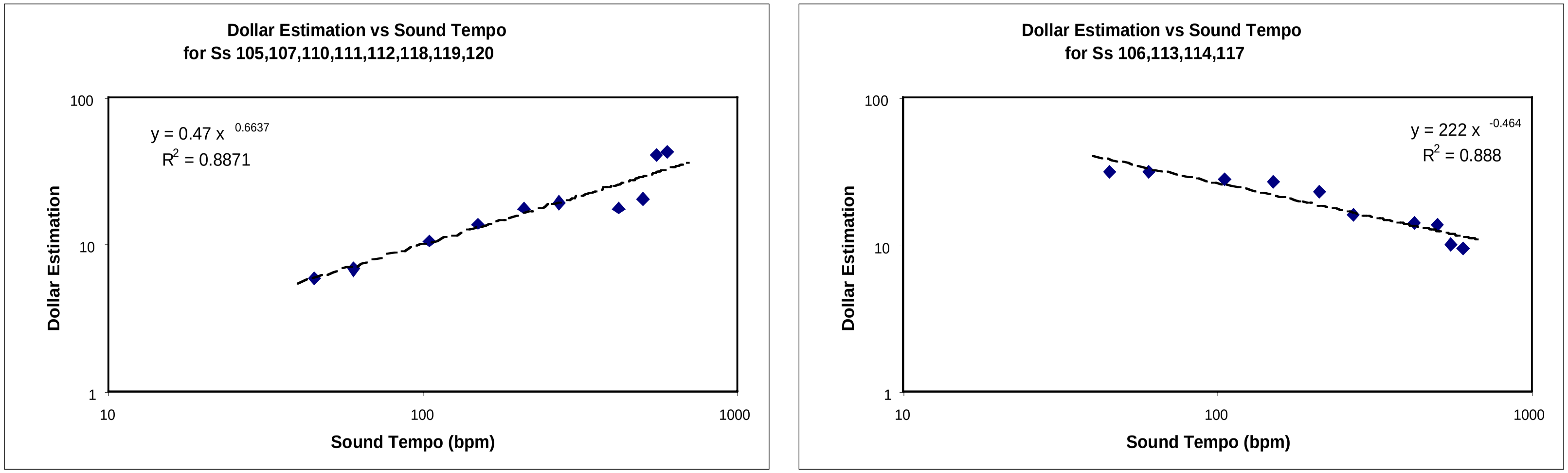


Figure 27, continued. Plots of geometric mean of estimated values against the actual tempo of sound stimuli in Experiment 3. Note that the axes are in log units. The participants whose responses contributed to the means are indicated at the top of the plot. The slope of the fit line is given by the exponent of x in the equation that is included in each plot.

Discussion of Experiment 3

Horizontal Lines: A Replication

The data for the participants who saw horizontal lines and estimated their lengths were again highly collinear points, this time yielding a slope of m = 0.98. This is a complete replication of both the previous experiment and other findings in the literature. This reconfirms the effectiveness of the Web-based approach to implementing the magnitude estimation paradigm.

Auditory Stimuli

Perceptual Dimensions

When participants listened to sounds that varied in frequency or in tempo, the estimated pitch or tempo followed the same data patterns as previous mappings: highly collinear points with a slope of 0.78 for frequency and 0.95 for tempo. These results are in keeping with expectations, and provide a baseline for the relationship between these two acoustic parameters and their perceived sensations. All of these verifications of the paradigm using perceptual dimensions (including the explorations with visual stimuli) have been necessary first steps, but have really been building the foundations for the investigation of conceptual dimensions.

Conceptual Dimensions

The magnitude estimations between auditory display dimensions and conceptual data dimensions is the main motivation for this entire research project. When listeners heard sounds that varied in frequency or in tempo, they were able to make judgments about how much change in the conceptual data dimension a given change in the sound dimension represented. This is a key task in interpreting a data sonification.

The responses were again highly collinear, with high r^2 values. There was more variability between participants with auditory stimuli than with visual stimuli, but once the listeners' data were sorted and separated appropriately, they provided reliable slope information. More participants in each block type might have helped to reduce some of the uncertainly in the magnitude estimations. Future experiments will now be able to take the variability within a block into account.

The variability between participants within a block, namely the number of listeners who responded in opposite polarities, should begin to provide a measure of how effective a sonification mapping will be. That is, more unanimous mappings, in terms of polarities, should be preferred over less unanimous mappings. For example, the dollars+frequency mapping yielded a 6 to 5 split in favor of a positive polarity. It may be that that pairing is less intuitive than, say, the dollars+tempo mapping which yielded an 8 to 4 split in favor of a positive mapping. A mapping with unanimous, or near unanimous support for a given polarity could be considered a good mapping, likely leading to fewer confusions. An example would be the temperature+frequency mapping, where 11 of 13 participants responded with a positive polarity. The size+tempo mapping showed that even a negative polarity can be unanimously preferred.

In this experiment some participants responded with a positive polarity in one block and a negative polarity in their other block. This demonstrates that the listeners are listening thoughtfully, and are not simply applying the same mapping polarity and ignoring the content of the dimensions. Comments from some participants about the mental models they used while listening indicated a cognitive translation was definitely occurring between the sounds and the meanings being attached to the sounds. The cognitive translation was not rigid, and listeners could switch from a positive polarity to a negative polarity, and vice versa. Such switching of conceptual mappings was one behavior that was anticipated in Experiment 1 but unfortunately never appeared. Of course, as discussed in relation to Experiment 1, more time, or more directions on the part of the experimenter and software might have helped in that “A versus B” technique. Also, in the present experiment this cognitive switching was only required between blocks, rather than between trials, as was the case in Experiment 1. Regardless, the magnitude estimation paradigm seems to be more useful than the “A versus B” approach, in that it provides not only polarity information, but also a scaling function relating the auditory parameter to the listener’s conception of a data dimension.

I had predicted, based on the results of Experiment 2, that the slopes of those scaling functions would vary depending on the actual data dimension in play. This was precisely the result obtained. Listeners who heard exactly the same sounds produced different magnitude estimation slopes, depending only on the conceptual name provided for the data dimension. The slopes are different from each other, and different from the slopes obtained for simple perceptions of the acoustic parameters. This means that the most effective way to use an auditory display parameter to represent changes in a data dimension is to take into account the slope of that scaling equation.[13]

The new idea developed here that conceptual data dimensions like temperature can be directly scaled to auditory parameters like frequency leads directly to this effective approach to determining both polarity and scaling information. The findings here for both polarity and slope should, in turn, inform the design of more effective sonifications.

[13] Eventual users of the slope information will want to determine the most precise values possible. Some of the slopes obtained have a slight curvature, which reduces the r^2, and therefore the certainty in the slope. This was anticipated, from the overview provided by Painton et al. (1977). An extreme example here was the frequency+temperature mapping, shown in Figure 26 (top left). A method for eliminating this curvature is to add a constant to each data point in the plot. In log coordinates, the effect is to make the points more collinear, yielding a smaller error in estimating the slope. The best constant is whatever leads to the highest r^2. Note that this can also change the value of the slope itself, often leading to a shallower slope. However, Stevens (1969, p.253) warns that it is only appropriate to consider any such adjustment when there is strong evidence that the underlying psychophysical function is really curved in log coordinates. In the current context this could only be considered after repeated experiments with different types of display and data dimensions that all suggest a persistent curvature in the magnitude estimation for a particular mapping. Besides, as Painton and his colleagues point out, the linear approximation is a good estimation. For that reason the slopes presented here, including the slope for the temperature+frequency mapping, are those obtained with all the data. A simple approximation to adding a constant may also be obtained by omitting the bottom and/or top data point, and using the remaining data to calculate the slope. The slope resulting from omitting the bottom data point is also shown in the temperature+frequency plot in Figure 26.

However, the ultimate goal of making sonifications “more effective” is predicated on two assumptions. The first assumption is that the results obtained here are, in some sense, “correct.” That is, the slopes obtained in this manner are the “real” underlying psychophysical scaling relations that listeners use. The fact that the experiment was “calibrated” (i.e., for perceptual dimensions the results replicated findings in the literature) adds further credibility to the conceptual-data slopes obtained. Of course, another way to address the issue is to simply replicate with new listeners and perhaps different variations of the display parameters as well. The repetition of the findings for length perception of horizontal lines shows that this paradigm does produce replications. Future investigators will need to examine these and other dimensions to gain a measure of the stability of the slope values.

The other assumption about these findings is that the use of unique polarities and slopes for each data and display pairing (as opposed to, say, a uniform linear data to display mapping with a slope of +1.0) will result in improved performance with a sonification system in a real task environment. If this turns out to be the case, the amount of any such improvement will need to be worth the effort involved in obtaining the slopes and implementing them in a sonification system. However, any benefit at all, however small, will likely be worth the effort since fast computers will handle all of the computation anyway. Regardless, it is still appropriate to begin to question the ultimate utility of these mappings for practical tasks. As sonification systems are developed they will need to be implemented with a variety of slopes and polarities, and those different mapping setups will need to be compared in methodical ways to examine the performance impacts. This is very similar to the approach Walker and Kramer (1996) took, implementing a variety of mappings in a simulated crystal factory sonification and examining the resulting speed and accuracy of the listeners.

Implementing a sonification system in order to assess the mapping slopes obtained here was beyond the scope of this project. However, I wanted at least to begin to validate the utility of these mappings. I decided that a good first step would be to play some “auditory graphs” (series of tones changing in frequency and tempo) and see if the listeners were interpreting those changes in accordance with the mapping slopes previously determined. This was the goal of the final experiment, presented next.

Experiment 4: Validation of Slopes

Purpose

The previous experiment has shown that the scaling function for the mapping of a data dimension to an auditory display parameter depends on the exact data dimension in play. The different magnitude estimation slopes obtained in Experiment 3 can be used to make sonifications conform more to both the perceptual *and* conceptual expectations of listeners. However, the exact values for the slopes for conceptual dimensions will need to be validated through replication as well as specific validation studies. The goal of the present experiment is to begin to take steps in that direction, and to determine ways to validate the slopes in a somewhat more task-like environment. This experiment presented mock auditory graphs and asked listeners to indicate which of two data patterns each sound pattern represented. One of the data patterns matched the slope determined in Experiment 3 whereas the other pattern did not match. If the sonification slopes determined in Experiment 3 are used in this sort of a data-analysis task, and if those slopes are preferred over both shallower and steeper slopes, that should provide evidence converging toward the validation of the slope values.

Method

Participants

From the same subject pool as Experiments 1, 2, and 3, a total of 60 students completed the experiment (28 males and 32 females; mean age 20.1 years). Twelve of these participants received $5 for participating, rather than course credit.

Apparatus

The apparatus was identical to the apparatus in Experiments 1 through 3. In this experiment the stimuli were again presented through the headphones.

Stimuli

Sound Stimuli: "Auditory Graphs"

The six multi-part auditory stimuli were pre-synthesized at 16-bit, 44.1 kHz using Csound, as in Experiments 1 and 3. The Csound orchestra and score files used in the synthesis are included in Appendix C. There were again two stimulus sets. The three stimuli in the Frequency Set were each made up of a series of five 1-s pure tones separated by 0.25 s of silence.[14] Each stimulus in this set sounded like a slow arpeggio played on an unusual scale. The frequency of the first tone for each stimulus was 200 Hz. The frequency increased over equal steps to final frequencies of 400, 600, or 1000 Hz.

Specifically, Stimulus F1 had frequency steps of 200, 250, 300, 350, and 400 Hz. Stimulus F2 had frequency steps of 200, 300, 400, 500, and 600 Hz. Stimulus F3 had frequency steps of 200, 400, 600, 800, and 1000 Hz. The amplitude envelope of the tones in this set of stimuli included a 0.1-s (0.1-beat) linear ramp onset (attack) and offset (decay), as in the previous experiment. Each of the steps was scaled for equal loudness, as described in the previous Experiments.

The three stimuli in the Tempo Set were also composed of five steps each. The steps in each of these stimuli increased in tempo rather than in frequency. The steps were composed of a repeating pattern of 0.200 beat of sound and 0.050 beat of silence (an on-off pattern, as in Experiments 1 and 3). That is, when repeated at 60 bpm, the sound was on for 200 ms, off for 50 ms. The pattern was repeated at a certain tempo (e.g., 60 bpm) for as many repetitions as were required to fill approximately 1 s per step.[15] Note that each step ended in a brief silence (the "off" portion of the last of the repeated patterns). The next steps were composed in the same manner, but the on-off pattern was repeated at progressively faster tempos, though still for a total of 1 s per step. These subsequent steps were appended directly to the end of the previous step, for a total of five steps. Specifically, Stimulus T1 had tempo steps of 60, 75, 90, 105, and 120 bpm. Stimulus T2

[14] To be explicit, the tone part of the sound was synthesized for 1 beat, followed by 0.25 beat of silence. At the Csound tempo of 60 bpm (see Appendix C) this resulted in a 1-s tone followed by 0.25 s of silence.

[15] The stimulus steps in the Tempo Set included whole multiples of the on-off patterns, so some of the steps were slightly longer or shorter than 1 s in duration.

had tempo steps of 60, 90, 120, 150, and 180 bpm. Stimulus T3 had tempo steps of 60, 120, 180, 240, and 300 bpm. Due to the fast repetition rates and brief on-off patterns, the amplitude envelope of the "on" part of the sound (the "tones") in this set of stimuli included a 0.01-beat linear ramp onset and offset, rather than the 0.1-beat ramp of the other stimuli. Note that the actual length of the onset and offset changed for each step, since it was relative to the tempo at which the pattern was repeated. The frequency of the sine-wave tone components was 800 Hz for all steps of each stimulus in the Tempo Set. Further, the stimuli in the Tempo Set were all synthesized at the same loudness as the 800-Hz component of the Frequency Set, giving all of the stimuli in both sets essentially equal perceived loudness.

Word Stimuli: "Data Patterns"

The idea of this experiment was to listen to an auditory graph (i.e., one of the sound stimuli described above), then determine which of two data patterns that auditory graph best represented. The data patterns were created as follows.

Consider the example where Stimulus T1 was played and the listener had to make a judgment about pressure. The starting frequency of the stimulus (f_1 = 200 Hz) was defined as being equal to an initial data (i.e., pressure) value of P_1 = 100 units. The final frequency of the stimulus was, in this case, f_5 = 400 Hz. To calculate the final pressure, P_5, one can use the equation

$$P_5 = P_1 \bullet (f_5 / f_1)^m \qquad (6)$$

where *m* is the slope from the regression equation determined in the previous experiment. The regression equation for estimated pressure *P* versus actual frequency *f*, obtained in Experiment 3 is

$$\log P = 0.14 \log f^{\,0.7812}. \qquad (7)$$

Substituting the data values from the present example, as well as the slope *m* from Equation 7, yields

$$P_5 = 100 \bullet (400 / 200)^{\,0.7812}. \qquad (8)$$

Solving this equation and rounding off to an integer yields

$$P_5 = 172. \qquad (9)$$

Thus, for the stimulus T1 (starting at 200 Hz and ending at 400 Hz), the calculated data values for pressure would start at 100 units of pressure and end at 172 units of pressure.

When the Tempo Set of stimuli was used, the equation used to calculate the final pressure values would be

$$P_5 = P_1 \bullet (t_5 / t_1)^m \qquad (10)$$

where t_1 and t_5 are the initial and final tempo of the stimulus being used, and m is the slope of the regression equation determined for pressure and tempo in Experiment 3.

For each stimulus the starting and ending frequencies or tempos were used in conjunction with the slope from the regression equations determined in Experiment 3 to calculate the “correct” starting and ending data values. Each of the intermediate values of pressure, temperature, and so on (corresponding to each of the steps in the stimulus) could have been calculated in the same manner. However, only the starting and ending values were necessary for use in the trials. To make the task somewhat easier for the participant, the starting value of each data dimension was 100 on every trial, for all participants. It is important to note that only positive slopes were used here, even if there were both positive and negative slopes obtained in Experiment 3. The reason for this was simply to reduce the number of experimental conditions. In one case, size+tempo, there was no positive slope obtained in Experiment 3. That is, every participant in the size+tempo group responded with a negative slope. In that case the slope used in the present experiment was +1.0.

For each pair of “correct” data points, two pairs of “incorrect” data points were created by multiplying the “correct” data endpoint by 0.80 and by 1.20. Thus, if the “correct” data set went from 100 to 172 units, the “incorrect” pairs would be from 100 to 138, and from 100 to 206 units (80% and 120% of the endpoint, respectively).

Design

The design again included the within-subjects factor of display dimension (frequency and tempo) and the between-subjects factor of data dimension (temperature, pressure, velocity, size, and number of dollars). Note that there were no perceptual data dimensions (pitch or perceived tempo) included in this experiment.

Each data dimension (e.g., temperature) was paired with one of the display dimensions (e.g., frequency) for an entire block of trials (e.g., temperature+frequency). Thus, there were 10 possible block types. Each participant again completed two blocks of trials separated by a brief rest. The block types were assigned pseudo-randomly as before. Twelve participants completed each block type, as summarized in Table 9.

Table 9. Number of participants who completed each block type in Experiment 4.

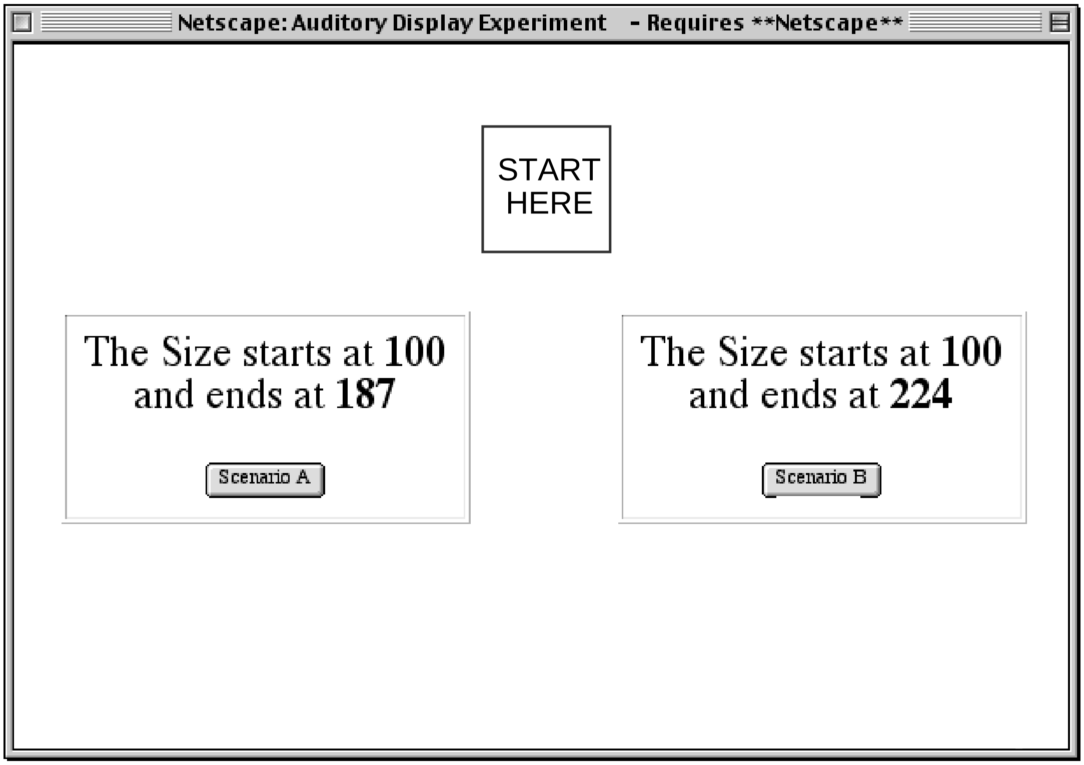
Display dimension	**Data dimension**				
	Temperature	Pressure	Velocity	Size	Dollars
Frequency	12	12	12	12	12
Tempo	12	12	12	12	12

Trial Structure and Task

At the beginning of each block of trials, instructions like the following were displayed on the screen and read aloud by the experimenter:

> You will hear some auditory graphs -- that is, patterns of sounds that are meant to represent **Pressure** data. Your task is to indicate which of two data descriptions the sound represents.
>
> Listen to the sound by moving the mouse over the sound squa
> button under the data description that best fits the sound.

Figure 28 shows the screen layout for this experiment. In the top center of the screen there was a 2.5-cm square graphic with the words, “START HERE.” The participant moved the cursor over the square to listen to the stimulus. Moving the cursor off and then back over the graphic would play the sound again.


Netscape: Auditory Display Experiment - Requires **Netscape**
START HERE
Scenario A
Scenario B

The Size starts at **100** and ends at **187**

The Size starts at **100** and ends at **224**

Figure 28. Image of the screen layout seen by participants in Experiment 4. Moving the mouse pointer over the “START HERE” square played the sound stimulus. Clicking on one of the JavaScript buttons indicated the participant’s response and proceeded to the next trial.

Below the sound square were two boxes, side-by-side on the screen. Each contained a set of data end points, calculated as previously described. On any given trial the “correct” set of data points could be paired with either of the corresponding sets of “incorrect” data points, or else both of the “incorrect” sets could be presented as a foil trial.

On each trial the participant heard one stimulus from the set being used for that block (e.g., sounds from the Frequency Set) and decided which of the two data descriptions that sound best represented. Participants simply clicked the JavaScript button underneath the data set that they felt “matched” the auditory graph. This was a forced-choice design, requiring responses even when there were two “incorrect” data sets, though, of course, the participants had no knowledge that any of the data patterns were considered “correct” or “incorrect.”

In a block of 18 trials each of the three stimuli in a set was played twice with every possible pairing of its corresponding “correct” and “incorrect” data sets. The location of the data sets was counterbalanced left-to-right, and all of the trials were presented in a randomized order. The experiment control program required the participant to listen to the sound stimulus at least once on every trial. Further, the listener could not respond until the auditory graph had stopped playing. Following a brief rest, the experimenter introduced the second block with new instructions that introduced a different data dimension and the other display dimension.

Results

Experimental Trials

I separated the data from the two display dimensions (frequency and tempo), then sorted by data dimension and participant. For each participant I computed a performance score as follows. For the 12 trials where the “correct” data pair had been present, I scored each response based on whether the listener had picked the “correct” data pair or not. I divided the number of correct responses by 12 to determine the percent correct for that participant. I then calculated the grand mean score across all participants where frequency was the display dimension, which showed that participants picked the “correct” data set significantly more often than would be expected by chance [*score* = 0.6069, *SD* = 0.1424; *t*(59) = 5.819, *p* < .0001]. When tempo was the display dimension participants also picked the “correct” data set significantly more often than would be expected by chance [*score* = 0.5958, *SD* = 0.1254; *t*(59) = 5.919, *p* < .0001].

I then calculated the grand mean across all participants within a given block type to determine if the participants in that block had responded better than chance (50%). Table 10 summarizes the grand mean for each of the trial block types, when frequency was the display dimension. As seen in the table, the grand mean for each block type was numerically larger than 50%, with this “better-than-chance” performance reaching statistical significance for temperature, velocity, and size. I performed the same analysis with the tempo data. Table 11 summarizes the grand mean across participants, for each of the block types, when tempo was the display dimension. The grand mean for each block type was, again, numerically larger than 50%, with all of the data dimensions but velocity reaching statistical significance.

Table 10. Grand mean (overall proportion correct) in each block type in Experiment 4 where frequency was the display dimension.

FREQUENCY SET	**All Data**	**Temperature**	**Pressure**	**Velocity**	**Size**	**Dollars**
Proportion Correct	0.607	0.639	0.549	0.639	0.653	0.556
Variance	0.020	0.012	0.022	0.017	0.031	0.016
***N* of Participants**	60	12	12	12	12	12
Value of *t*	5.819***	4.432***	1.134	3.708***	2.989**	1.542
Value of *p*	.0001	.0010	.2808	.0035	.0123	.1513

* Significant at p < .10 **Significant at p < .05 ***Significant at p < .01 (all two-tailed)

Table 11. Grand mean (overall proportion correct) in each block type in Experiment 4 where tempo was the display dimension.

TEMPO SET	**All Data**	**Temperature**	**Pressure**	**Velocity**	**Size**	**Dollars**
Proportion Correct	0.596	0.597	0.611	0.514	0.660	0.597
Variance	0.016	0.011	0.019	0.014	0.021	0.007
***N* of Participants**	60	12	12	12	12	12
Value of *t*	5.919***	3.189***	2.766**	0.411	3.838***	3.924***
Value of *p*	.0001	.0086	.0183	0.6887	.0028	.0024

* Significant at p < .10 **Significant at p < .05 ***Significant at p < .01 (all two-tailed)

Foil Trials

I also analyzed performance on the trials where neither one of the data sets was “correct.” Performance on these trials should be statistically equivalent to guessing. I again separated the data into the different block types. For each block type (i.e., each mapping type) I calculated the overall percentage of foil trials where participants had chosen the lower of the two “incorrect” data sets. This was a measure of whether one or the other of the “incorrect” data sets had been chosen more often than would be expected by chance. If it were statistically different from 0.50, then one of the foils had been preferred. Ten separate *t* tests, with Bonferroni correction of the significance levels, showed that in none of the block types was the mean percentage statistically different from 0.50. I also computed the overall grand mean of foil trials for each of the stimulus sets, across all block types. The overall mean for the Frequency Set was 0.49, which was not significantly different from 0.50, *t*(1, 14) = 0.2713, *p* = .7901. For the Tempo Set the overall grand mean was 0.43, which was again not significantly different from 0.50, *t*(1, 14) = 1.6976, *p* = .1117. Thus, the responses for the foil trials with both the Frequency and Tempo Sets of stimuli indicated no systematic preference for one or the other of the “incorrect” data sets.

Discussion of Experiment 4

This experiment was an initial attempt to validate some of the magnitude estimation slopes obtained in Experiment 3. Participants heard mock auditory graphs and for each one judged which of two data patterns the auditory graph represented. The “correct” data patterns were calculated using the slopes obtained in the previous experiment, while “incorrect” data patterns were also created using shallower and steeper slopes. When two “incorrect” data patterns were presented (i.e., neither of the data patterns was calculated from the slope obtained in Experiment 3) the participants showed no preference for one over the other. However, when one of the data patterns was calculated with the “correct” slope, that “correct” pattern was, overall, selected significantly more often than the “incorrect” pattern. That is, participants seemed to prefer the mappings that used the experimentally determined slopes. This provides evidence converging towards validation of the slopes obtained in Experiment 3.

In a few of the specific data-to-display pairings preference for the “correct” data pattern did not reach conventional levels of statistical significance. Specifically, the pressure+frequency, dollars+frequency, and velocity+tempo sets did not result in better-than-guessing performance. For both of the two frequency sets, though, there was not a strong degree of unanimity in the slopes when they were obtained in Experiment 3. Specifically, for the pressure+frequency set only 8 out of 12, or two thirds of the participants in that group, provided a positive slope, while one third provided a negative slope. For the dollars+frequency group, only 6 out of 11 participants (54%) responded with a positive slope in Experiment 3. The degree of unanimity for some slope polarities was presented in Experiment 3 as a possible measure of the “goodness” of a mapping. The findings here with the temperature data seems to support that. It is not clear why the velocity+tempo data, unanimous in Experiment 3, did not result in better-than-guessing performance here.

Thus the present experiment provides evidence supporting the results of Experiment 3, not only in terms of the effectiveness of the slopes obtained, but also in terms of the potential use of the unanimity of the slopes as a measure of the “goodness” of a mapping for use in sonification. This validation of the paradigm and results developed thus far in this project bode well for the continued development of a theory of sonification that is supported by experimental findings. That will be the best recipe for effective data sonification applications.

General Discussion

Fundamental pedagogical and research practices rely on determining patterns in data. The data sets of today are often large and complex, requiring the investigator to marshal every available perceptual and cognitive aid in the hunt for scientific meaning. However, the many graphing and analysis programs currently available are exclusively visual, thereby failing to exploit the excellent pattern recognition capabilities of the human auditory system. As a further result of this, students and researchers with visual disabilities remain marginalized.

Sonification is the use of non-speech audio to convey information about scientific data. That is, the data drives the creation of auditory graphs, effective for exploration in a range of fields, with researchers who are sighted or not. However, despite increasing popularity and the growing number of scientific success stories for sonification, there has been very little experimental evaluation of how to construct such auditory graphs. In particular, there has been no systematic evaluation of the way data values are mapped onto display values. The three main questions that such an evaluation needs to answer are: (1) What is the best sound parameter to use to represent a given data type? (2) Should an increase in the sound dimension (e.g., rising frequency) represent an increase or a decrease in the data dimension? (3) How much change in the sound dimension will represent a given change in the data dimension? In the research I have presented here I took these as guiding research questions.

Experiment 1: Which Sound is Hotter?

Over the past years I have casually asked may people which of two hummed, whistled, tapped, or sung sounds seemed to represent something that is hotter, faster, bigger, and so on. While there is not universal agreement, nearly everyone makes a quick response, often supported with ecologically-based or pseudo-physical mental models. For example, “As a teapot gets hotter, the pitch of its whistle rises.” With this in mind, I asked experiment participants to indicate which of two sounds represented something hotter, faster, and so on. The sound pairs differed only in frequency or only in tempo. I hoped to capture more formally the casual responses I had been observing from acquaintances.

While all of my participants were located on campus, I used the technologies associated with the World Wide Web to run the experiments. This was partly as a demonstration of the utility of such simple and widely-available resources, and partly to satisfy the goal of future replications and extensions of these experiments with participants from all parts of the world.

Unfortunately, it appears that listeners in this Experiment did not apply a separate cognitive analysis to each sound + data combination. They seemed to apply a single polarity to each of the sound dimensions, apparently disregarding the ecological or physical differences for the separate data dimensions. Probably this was due to the factorial interspersing of the sound and data types. Alas, hopes for a “quick fix” approach to sonification mappings were not realized. In the future, a more successful approach to that same experimental paradigm might include asking the listener to provide a verbal justification or explanation of the answer provided on each trial.

Psychophysical Scaling and Magnitude Estimation

Even if Experiment 1 had yielded more insights, it would still have been limited to preferences about mapping and polarity. The missing third element, scaling the mappings, is crucial for successful sonification. I turned to the method of magnitude estimation, as developed most extensively by S. S. Stevens and his colleagues (see Stevens, 1975 for a review), with the idea that it would provide answers to all three of the sonification research questions. In this paradigm participants observe a series of stimuli and make judgments about the magnitude of some aspect of the stimuli. Most such studies to date have asked for perceptions of a physical attribute of the stimulus, such as its perceived size, length, or pitch. I reasoned that this could be extended to include conceptual judgments about a stimulus, such as how much pressure or how many dollars it might represent. I predicted that the actual relationship between the physical parameter (e.g., size or frequency) and the conceptual dimension (e.g., pressure or number of dollars) would follow a power function. However, for conceptual dimensions the exact exponent of the function was not something I could predict in advance.

Experiment 2: Magnitude Estimation with Lines and Circles

Observers, again using Web technologies, made magnitude estimations about perceptual and conceptual attributes of visual stimuli. When participants saw horizontal lines, vertical lines, or filled circles and were asked simply to estimate the size of the stimuli, the resulting plots of perceived size versus actual size were straight lines in log-log coordinates, with high values of r^2. For line stimuli the regression slope was 0.96, in agreement with the results for similar stimuli reported in the literature. For circle stimuli the slope in the present experiment was 0.59, which is slightly lower than, but in the range of previous results. The collinearity of the data throughout the project, and the close agreement with previous results, validated the use of the Web-based experimental procedure to obtain magnitude estimation slopes.

When participants made conceptual magnitude estimations about the temperature, pressure, or velocity that the lines and circles would represent, the results were again very collinear, with high r^2 values for the regression lines. However, the slopes were nearly all different from each other, and, more importantly, different from the slopes obtained for the perceptual dimensions. Further, for the circle stimuli 25% of the participants responded with a different polarity. This resulted in both positive and negative slopes being computed for velocity and pressure estimations with circle stimuli. When asked about their responses, some participants provided explanations involving mental models of physical systems, showing that there is a cognitive translation involved in the mapping of a conceptual data dimension to a display dimension.

These results demonstrate the importance of obtaining the specific slopes relating a physical display dimension like the length of a line to each of the conceptual dimensions that it will be used to represent. Further, it provided evidence to suggest that similar results might even be found with the auditory display dimensions used in sonification.

Experiment 3: Magnitude Estimation with Sound Stimuli

The next stage in this line of research was to obtain magnitude estimations (both perceptual and conceptual) for auditory stimuli rather than visual stimuli. When participants listened to sounds that varied in frequency or in tempo, the estimated pitch or tempo followed the same data patterns as previous mappings: highly collinear points with a slope of 0.78 for frequency and 0.95 for tempo. These results are in keeping with expectations based on findings in the literature.

When listeners heard sounds that varied in frequency or in tempo, they were also able to make judgments about how much change in a conceptual data dimension a given change in the sound dimension represented. There was more variability between participants with auditory stimuli than with visual stimuli, but they still provided reliable slope information. Listeners who heard exactly the same sounds produced different magnitude estimation slopes, depending only on the conceptual name provided for the data dimension. Further, the slopes were different from the slopes obtained for simple perceptions of the acoustic parameters.

The additional fact that some participants responded with a positive polarity in one block and a negative polarity in their other block demonstrates that there was a cognitive translation occurring between the sounds and the meanings being attached to those sounds. The cognitive translation was not rigid, and listeners could switch from a positive polarity to a negative polarity, and vice versa.

For each data-to-display mapping, the number of listeners who responded in opposite polarities should begin to provide a measure of how effective a sonification mapping will be. For example, when representing the concept of dollars the dollars+frequency mapping, which yielded only a 6 to 5 split in favor of a positive polarity, may be less effective than the dollars+tempo mapping, which yielded an 8 to 4 split in favor of a positive mapping. A mapping with near unanimous support for a given polarity, such as the temperature+frequency mapping where 11 of 13 participants responded with a positive polarity, would likely lead to relatively little confusion. The size+tempo mapping made it clear that even a negative polarity can be unanimously preferred.

It is a novel finding that conceptual data dimensions like temperature can be directly scaled to auditory parameters like frequency. Using the mappings, polarities, and slopes obtained from magnitude estimation procedures will likely go a long way towards more effective sonifications. However, the ultimate goal of making sonifications “more effective” relies on the assumptions that: (1) the results obtained here reflect the “real” underlying psychophysical scaling relations that listeners use; and (2) the use of unique polarities and slopes for each data and display pairing will actually result in improved performance with a real-life sonification system task. The first of these assumptions is supported by the fact that the two magnitude estimation experiments both replicated findings in the literature for similar perceptual dimensions while finding different values for the conceptual dimensions. Future replication with new listeners and perhaps different variations of the display parameters should provide further support for the conceptual dimensions.

It still remains to be seen, however, if all the efforts required to obtain different slopes for different data-to-display mappings will translate into improved performance. However, the final experiment in this project served as an initial step along the validation path.

Experiment 4: Validation of Slopes

I presented mock auditory graphs to new listeners and asked which of two data patterns that sound pattern best represented. The change in the data values for one of the patterns matched the slope determined experimentally in Experiment 3 whereas the other pattern did not match. If the sonification slopes from Experiment 3 were preferred over both shallower and steeper slopes, that would provide evidence converging toward the validation of the slope values.

When two “incorrect” data patterns were presented (i.e., neither of the data patterns was calculated from the slope obtained in Experiment 3) the participants showed no preference for one over the other. However, when one of the data patterns was calculated with the “correct” slope, that “correct” pattern was selected more often than the “incorrect” pattern. That is, participants seemed to prefer the mappings that used the experimentally determined slopes. In addition, the findings with the temperature data seemed to support the use of a polarity’s unanimity as a measure of the “goodness” of a mapping for displaying a given data dimension. These results both lead to heightened confidence in the utility of the magnitude estimation procedure with auditory display and conceptual data dimensions. Further, these findings underscore the need to use the resulting polarities and slopes in sonification design, to avoid having listeners simply guess about the meaning of an auditory graph.

Research Questions Revisited

The research questions that drove this project have been successfully addressed through the experiments described here, but the issues are still too complex for simple or exhaustive answers. However, I will briefly discuss some of what the present research contributes to each issue.

Question 1: What sound parameter best represents a given data dimension?

It seems that the level of unanimity for the polarity is a good first approximation for the effectiveness of a data-to-display mapping. This was proposed based on the results of Experiment 3, and gained support with the results of Experiment 4. If the predicted effectiveness for the display dimensions of frequency and tempo[16] were based solely on the level of unanimity for a given polarity obtained with the magnitude estimation paradigm in Experiment 3, the results might look like those presented in Table 12. The percent unanimity for each mapping and polarity is indicated in the table, plus a symbol based on the cutoff values of 50 and 85%. These cutoffs come from examining a histogram including all of the percentages for the most-preferred polarities for each mapping. Below 50% unanimity any mapping is going against the majority and is therefore “bad” by definition. Several mediocre mappings fall between 50 and 70%. A third cluster appears above 85%, hence I made that the cutoff for a “good” mapping.

Table 12. Effectiveness of frequency and tempo for displaying various conceptual display dimensions, based on unanimity of polarities in Experiment 3.

[16] As defined and used in this project.

Display dimension		**Data dimension**				
	Polarity	Temperature	Pressure	Velocity	Size	Dollars
Frequency	$+^{ve}$	√ (85)	**~** (67)	√ (88)	***x*** (37)	**~** (55)
	$-^{ve}$	***x*** (15)	***x*** (33)	***x*** (12)	**~** (63)	***X*** (45)
Tempo	$+^{ve}$	**~** (65)	**~** (67)	√ (100)	***x*** (0)	**~** (67)
	$-^{ve}$	***x*** (35)	***x*** (33)	***x*** (0)	√ (100)	***X*** (33)

Note: The symbols represent only the level of unanimity of the mapping polarity. Borrowing the symbols from Walker and Kramer (1996), checks indicate a good mapping (> 85% unanimity in this case), x's indicate a poor mapping (< 50%), and tildes indicate a mapping that falls in the middle. The actual percentage values are provided in parentheses.

It is interesting to compare these results to the recommendations presented in Walker and Kramer (1996) which I have redrawn in Figure 2 in the Introduction of the present report. Recall that those recommendations were based on a combination of accuracy and reaction time results. Also note that they include fewer data dimensions, more display dimensions, and only positive polarities. The first two lines of Figure 2 are of relevance, here. Walker and Kramer (1996) seemed more "centralist" in their assessment, considering fewer mappings as either "good" or "bad." However, the results are in general agreement with the present findings, and there are no contradictions.

Question 2: What is the preferred polarity for a mapping?

The question of which polarity to use for a mapping is also addressed by the results presented in Table 12. The most popular polarity is, presumably, to be considered the better polarity to use in representing a data dimension with a given display dimension. However, several of the polarities were only slightly more popular than their inverse. In addition, this study has not addressed the issue of whether listeners can learn to use a different polarity (or a different mapping, for that matter) than the one that is found to be more unanimous. Consider, for example, a listener who responds to a magnitude estimation experiment using a positive polarity for size+frequency, where the negative polarity obtained the majority response. It may be that, given an explanation of the mental model in use by the majority, this listener could respond equally well on subsequent occasions using a negative polarity.

Research Questions 1 and 2 could be investigated further with a modified version of Experiment 1, where the listener is asked directly which sound represents something that is hotter, colder, bigger, faster, and so on. If an explanation or justification were required for every response (allowing for “I don’t know why”), the results might be compared to the results obtained with magnitude estimation. Further, such a direct exploration could even be used as a pretest for the magnitude estimation, encouraging the use of cognitive justifications and mental models during the actual magnitude estimation phase.

Question 3: What is the scaling function between the data and display dimension?

Magnitude estimation seems to provide an excellent way to obtain a function relating conceptual data dimensions to display dimensions. Theses functions can be used to scale the auditory display dimensions in a sonification appropriately for each data dimension. The slopes obtained also seem to hold for the mock auditory graphs used as an initial validation in Experiment 4. However, replication with different groups of users and different levels of the auditory parameters will be required before the reliability and stability of the slopes can be determined.

Limitations

There are three main areas of limitation to the present project: the sounds, the listeners, and the task. The first limitation is mostly about the small number of sounds used so far. Only a couple of different auditory stimulus sets were employed here. The possibilities are endless for creating different sounds. Many more acoustic parameters, resulting in many more sets of sounds, will need to be tested before more generalizable guidelines can be developed. However, the paradigm developed here, plus possible enhancements already discussed, should provide researchers with a quick and effective method for continuing the investigation. It is also possible that the community of sonification researchers could define a standard set (or sets) of test sounds as a testing palette, and examine those sounds extensively and systematically for subsequent use in sonification applications. Of course, all of the present research has used simple tones, and no mention has been made about the aesthetic quality of the sounds. If sonification is to continue its growth in popularity the auditory graphing tools will need to be carefully designed to combine effective and aesthetic qualities into the sounds. This is a huge area of research yet to be explored, though, thankfully, a long history of music composition theory will have much to contribute immediately.

The second limitation of the present study is that all of the participants were young-adult college students. While no differences were found within this group based on sex, handedness, or number of years of musical training, it is certainly possible that distinct listener populations would respond in different ways. This is particularly likely if the groups differ widely in terms of common musical experience. Listeners more used to the sophisticated rhythms of latin music, for example, might have different scaling factors for tempo than listeners unfamiliar with those beats. It is perhaps more relevant to scientific sonification to determine whether listeners with different conceptions of the data dimensions reflect those differences in scaling polarities or magnitudes. For example, the concept of “pressure” to a molecular chemist is presumably different from the concept of “pressure” to a social worker. This could require differences in their sonifications.

The third limitation has to do with the actual task that the sonification will be used to complete. The present project began to validate the actual slope values, but this does not constitute a rigorous test of the performance of a sonification task using different mappings, polarities, and scalings. The actual effect on performance resulting from a good rather than bad sonification needs thorough exploration. However, it has been shown repeatedly that poorly-designed visual displays can have serious effects on performance. Improvements in those displays can have measurable performance gains (see, e.g., Sanders & McCormick, 1993). Walker & Ehrenstein (2000; see also Proctor & Reeve, 1990) have shown performance degradation with auditory stimuli as a result of stimulus-response compatibility effects. There is every reason to believe that ignoring the preferred mappings for conceptual dimensions could lead to similar conflicts in the perceive-think-respond action chain involved in the use of auditory displays and sonifications (see, also, Sorkin, 1988). Further instantiation of sonification guidelines, and testing them with real tasks, will be a core requirement for further research in this area.

Future Directions

The short-term plans for this line of research include replication with participants from the same or similar population, and then with listeners with different cultural or scientific experience. In addition, a key replication population will include members of the blind and visually impaired community. There is no reason to expect any different performance with this group, but it will be important explicitly to consider blind users as sonification is developed further as both a scientific and pedagogical tool.

Another area of investigation is into the modulus effect. Different scientific applications will need to use the same concept, like temperature, with very different ranges of values. It may matter what range the conceptual data is in, such as temperatures from 0 to 100 ºC compared with 0 to 1,000,000 ºC.

A third and very important area of near-future research has to do with all of the many standard elements in visual graphs, other than the data points themselves. The mock auditory graphs used in Experiment 4 were the auditory equivalent of drawing a squiggle on a blank sheet of paper. There were no axes, no labels, no tickmarks, trendlines, or legends. A visual graph without these elements would certainly not be taken seriously. Nor should an auditory graph. However, few have even begun to consider these requirements in any practical way.

With a better handle on the mappings, scalings, and auditory graphing techniques I plan to implement a sonification and auditory graphing tool that will have real practical utility for researchers, teachers, and students, both sighted and blind, in all sorts of scientific disciplines. With a principled, careful design, based equally on aesthetic, theoretical, and experimental foundations, sonification tools will help us grapple with our ever-expanding data sets, and at the same time open the door of the scientific community to many who are only too eager to listen to their data sing its own story.

References

Andre, A. D., & Wickens, C. D. (1995). When users want what's not best for them. *Ergonomics in Design,* Oct. 10-13.

Atlas, R., Cornett, L., Lane, D. M., & Napier, H. A. (1997). The use of animation in software training: Pitfalls and benefits. In M. A. Quiñones & A. Ehrenstein (Eds.), *Training for a rapidly changing workplace* (pp. 281-302). Washington, DC: APA.

Banks, W. P., & Hill, D. K. (1974). The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology, 102(2),* 353-376.

Barrass, S. (1998). *Auditory information design*. Unpublished Ph.D. dissertation, CSIRO, Australia.

Beck, J., & Shaw, W. A. (1961). The scaling of pitch by the method of magnitude-estimation. *American Journal of Psychology, 74,* 242-251.

Beck, J., & Shaw, W. A. (1962). Magnitude estimations of pitch. *Journal of the Acoustical Society of America, 34(1),* 92-98.

Beck, J., & Shaw, W. A. (1963). Single estimates of pitch magnitude. *Journal of the Acoustical Society of America, 35(11)*, 1722-1724.

Begault, D. R., Wenzel, E., M., Shrum, R., & Miller, J. (1996). A virtual audio guidance and alert system for commercial aircraft operations. In S. Frysinger & G. Kramer (Eds.), *Proceedings of the Third International Conference on Auditory Display, ICAD '96*, (pp. 117-122). Palo Alto, CA: ICAD.

Bertin, J. (1981). *Graphics and graphic information processing*. Berlin: Walter de Gruyter.

Boring, E. G. (1951). *A history of experimental psychology*. New York: Appleton-Century-Crofts.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

Brooks, L. R. (1968). Spatial and verbal components in the act of recall. *Canadian Journal of Psychology, 22,* 349-368.

Cohen, J. (1994). Monitoring background activities. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 499-531). Reading, MA: Addison Wesley.

Dawson, W. E., & Brinker, R. P. (1971). Validation of ratio scales of opinion by multimodality matching. *Perception and Psychophysics, 9(5),* 413-417.

Deutsch, D. (Ed.) (1982). *The psychology of music*. New York: Academic Press.

Edworthy, J., Loxley, S., & Dennis, I. (1991). Improving auditory warning design: Relationship between warning sound parameters and perceived urgency. *Human Factors, 33*, 205-232.

Eisler, H. (1976). Experiments on subjective duration 1868-1975: A collection of power function exponents. *Psychological Bulletin, 83(6)*, 1154-1171.

Ekman, G. (1958). Two generalized ratio scaling methods. *Journal of Psychology, 45*, 287-295.

Falmagne, J.-C. (1985). *Elements of psychophysical theory*. New York: Oxford University Press.

Fitch, W. T., & Kramer, G. (1994). Sonifying the body electric: Superiority of an auditory over a visual display in a complex, multivariate system. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 307-326). Reading, MA: Addison Wesley.

Flowers, J. H., Buhman, D. C., & Turnage, K. D. (1996). Data sonification from the desktop: Should sound be a part of standard data analysis software? In S. Frysinger & G. Kramer (Eds.), *Proceedings of the Third International Conference on Auditory Display, ICAD '96,* (pp. 1-8). Palo Alto, CA: ICAD.

Flowers, J. H., Buhman, D. C., & Turnage, K. D. (1997). Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples. *Human Factors, 39(3)*, 341-351.

Flowers, J. H., & Hauer, T. A. (1992). The ear's versus the eye's potential to assess characteristics of numeric data. Are we too visuocentric? *Behavior Research Methods, Instruments, & Computers*, *24*, 258-264.

Flowers, J. H., & Hauer, T. A. (1993). "Sound" alternatives to visual graphics for exploratory data analysis. *Behavior Research Methods, Instruments, & Computers, 25*, 242-249.

Flowers, J. H., & Hauer, T. A. (1995). Musical versus visual graphs: Cross-modal equivalence in perception of time series data. Human Factors, 37, 553-569.

Forbes, T. W. (1946). Auditory signals for instrument flying. *Journal of the Aeronautical Society, May*, 255-258.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music* (pp. 149-181). New York: Academic Press

Gaver, W. W. & Smith, R. B. (1990). Auditory icons in large-scale collaborative environments. In D. Diaper et al. (Eds.), *Human-Computer Interaction - INTERACT'90* (pp. 735-740). Amsterdam: Elsevier Science Publishers.

Gaver, W. W., Smith, R. B., & O'Shea, T. (1991). Effective sounds in complex systems: The ARKola Simulation. *Proceedings of CHI '91*, held April 28-May2, 1991, in New Orleans. Reading, MA: ACM Press / Addison Wesley.

Hartmann, W. M. (1997). *Sounds, signals, and Sensation: Modern acoustics and signal processing*. New York, NY: Springer Verlag.

Hayward, C. (1994). Listening to the earth sing. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 369-404). Reading, MA: Addison Wesley.

Indow, T. (1961). [An example of motivation research applied to product design.] Published in Japanese in *Chosa To Gijutsu*, *102*, 45-60.

Kramer, G. (1993). Sonification of financial data: An overview of spreadsheet and database sonification. *Proceedings of Virtual Reality Systems '93, SIG Advanced Applications*, held 1993 in New York. New York: SIG.

Kramer, G. (1994a). Some organizing principles for representing data with sound. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 185-221). Reading, MA: Addison Wesley.

Kramer, G. (1994b). An introduction to auditory display. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 1-78). Reading, MA: Addison Wesley.

Kramer, G. (1996). Mapping a single data stream to multiple auditory variables: A subjective approach to creating a compelling design. In S. Frysinger & G. Kramer (Eds.), *Proceedings of the Third International Conference on Auditory Display, ICAD '96*. Palo Alto, CA: ICAD. Web publication by the International Community for Auditory Display. http://www.santafe.edu/~icad.

Kramer, G., & Ellison, S. (1991). Audification: The use of sound to display multivariate data. *Proceedings of the International Computer Music Conference*, 214-221. San Francisco, CA: ICMA.

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

Kubik, G. (1975). *The Kachamba Brothers' Band: A study of neo-traditional music in Malawi*. Zambian Papers, no. 9. Manchester: Manchester University Press for University of Zambia Institute for African Studies.

Marks, L. E. (1974). *Sensory processes*. New York: Academic Press.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 384-394.

Martins, A. C. G., & Rangayyan, R. M. (1997). Experimental evaluation of auditory display and sonification of textured images. In E. Mynatt & J. A. Ballas (Eds.), *Proceedings of the Fourth International Conference on Auditory Display, ICAD '97*, (pp. 129-134). Palo Alto, CA: ICAD.

McAdams, S., & Bigand, E. (Eds.) (1993). *Thinking in sound: The cognitive psychology of human audition*. Oxford: Clarendon Press.

McCabe, K., & Rangwalla, A. (1994). Auditory display of computational fluid dynamics data. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 327-340). Reading, MA: Addison Wesley.

Moore, B. C. J. (1997). *An introduction to the psychology of hearing*. (4th ed.). Orlando, FL: Academic Press.

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology, 66*, 347-352.

Painton, S. W., Cullinan, W. L., & Mencke, E. O. (1977). Individual pitch functions and pitch-duration cross-dimensional matching. *Perception & Psychophysics, 21(5)*, 469-476.

Papp, A. L., & Blattner, M. M. (1994). A centralized audio presentation system. In G. Kramer & S. Smith (Eds.), *Proceedings of the Second International Conference on Auditory Display, ICAD '94*, (pp. 129-134). Palo Alto, CA: ICAD.

Patterson, R. D. (1982). *Guidelines for auditory warning systems on civil aircraft*. Paper No. 82017, Civil Aviation Authority, London.

Pereverzev, S. V., Loshak, A., Backhaus, S., Davis, J. C., & Packard, R. E. (1997). Quantum oscillations between two weakly coupled reservoirs of superfluid 3He. *Nature, 388* (31 July 1997) 449-451.

Proctor, R. W., & Reeve, T. G. (Eds.) (1990). *Stimulus-response compatibility: An integrated perspective*. Amsterdam: North Holland.

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

Sanders, M. S., & McCormick, E. J. (1993). *Human factors in engineering and design* (7th ed.). New York: McGraw-Hill.

Smith, B., & Casati, R. (1995). Naive physics. *Philosophical Psychology, 7(2),* 227-247.

Smith, J. O. (1996). Physical modeling synthesis update. *Computer Music Journal, 20(2)*.

Sorkin, R. D. (1987). Design of auditory and tactile displays. In G. Salvendy (Ed.), *Handbook of human factors* (pp.549-576). New York: Wiley & Sons.

Sorkin, R. D. (1988). Why are people turning off our alarms? *Journal of the Acoustical Society of America, 84(3),* 1107-1108.

Speeth, S. D., (1961). Seismometer sound. *Journal of Acoustic Society of America, 33(7),* 909-916.

Stevens, J. C. (1957). A comparison of ratio scales for the loudness of white noise and the brightness of white light. Unpublished doctoral dissertation, Harvard University.

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

Stevens, S. S. (1955). The measurement of loudness. *Journal of the Acoustical Society of America, 27*, 815-820.

Stevens, S. S. (1966). A metric for the social consensus. *Science, 151*, 530-541.

Stevens, S. S. (1969). On predicting exponents for cross-modality matches. *Perception and Psychophysics, 1969, 6(4)*, 251-256.

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

Stevens, S. S., & Galanter, E., H. (1957). Ration scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54,* 377-411.

Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception and Psychophysics, 1966, 1,* 439-446.

Stevens, S. S., & Guirao, M. (1963). Subjective scaling of length and area and the matching of length to loudness and brightness. *Journal of Experimental Psychology, 66*, 177-186.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology, 53*, 329-353.

Stevens, S. S., & Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*, 185-190.

Teghtsoonian, M. A. (1965). The judgment of size. *American Journal of Psychology*, *78*, 392-402.

Teghtsoonian, M. A. (1980). Children's scales of length and loudness: A developmental application of cross-modal matching. *Journal of Experimental Child Psychology, 30*, 290-307.

Teghtsoonian, M., & Teghtsoonian, R. (1971). How repeatable are Stevens's power law exponents for individual subjects? *Perception and Psychophysics, 10(3)*, 147-149.

Teghtsoonian, M., & Teghtsoonian, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception and Psychophysics, 33(3)*, 203-214.

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception and Psychophysics, 24(4)*, 305-314.

Tzelgov, J., Srebro, R., Henik, A., & Kushelevsky, A. (1987). Radiation detection by ear and by eye. *Human Factors, 29(1)*, 87-98.

Valenzuela, M. L., Sansalone, M. J., Krumhansl, C. L., & Streett, W. B. (1997). Use of sound for the interpretation of impact-echo signals. In E. Mynatt & J. A. Ballas (Eds.), *Proceedings of the Fourth International Conference on Auditory Display, ICAD '99*, (pp. 47-56). Palo Alto, CA: ICAD.

Walker, B. N., & Ehrenstein, A. (2000). Pitch and pitch change interact in auditory displays. *Journal of Experimental Psychology: Applied*, *6*, 15-30.

Walker, B. N., & Kramer, G. (1996). 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*, (pp. 71-74). Palo Alto, CA: ICAD.

Walker, R. (1987). The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & Psychophysics, 42*, 491-502.

Wickens, C. D., & Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors, 30*, 599-616.

Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd. ed.). New York: Harper Collins.

Wickens, C. D., Gordon, S. E., & Liu, Y. (1998). *An introduction to human factors engineering*. New York: Addison Wesley.

Williams, R. L. (1956). *Statistical symbols for maps: Their design and relative values*. New Haven, CT: Map Laboratory, Yale University. 64-74.

Woolf, B. (1992). Hypermedia in education and training. In D. Kopec & R. B. Thompson (Eds.), *Artificial intelligence and intelligent tutoring systems: Knowledge-based systems for teaching and learning* (pp. 97-109). New York: Ellis Horwood.

Appendix A: Csound files for Experiment 1 sound stimuli.

The following two Csound files were used to synthesize the stimuli used in Experiment 1. The orchestra (.orc) file describes the instruments created to play the score (.sco) file. That is, the orchestra file describes the general sound-generation routine, and the score file includes the changing parameters for pitch, onset (attack), amplitude, and so on.

Note that some of the parameters were manually edited for each stimulus. For example, in the score file the tempo setting, t 0 60, would have been set to t 0 120, and so on. The file would be re-saved with the new value, and the synthesis re-run to obtain the new sound file.

Orchestra File

```
;******************************************************
;*  stim6.orc
;*  CSound orchestra file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;******************************************************

sr = 44100
kr = 2205
ksmps = 20
nchnls = 1

instr 1
     k1   linen   ampdb(p4), p6, p3, p7        ; p4=amp in db
     a1   foscili  k1, p5, 1, 1 ,0,1           ; p5=pitch or freq
          out     a1                           ; p6=attack time
   endin                                       ; p7=release time


;asig foscil amp, freq, carrier, modulator, index, function
```

Score File

```
;**************************************************************
;*  stim6.sco
;*  CSound score file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;***************************************************************

f1  0  8192  10  1               ; sine wave

t 0 600                          ; regular tempo at t=0 is 60 beat/s

;instr  start  duration  loud(p4)  pitch(p5)  attack(p6)  release(p7)
   i1     0     1        69       100         .1          .1
   f0   1.5
   i1   1.5     1        66       200         .1          .1
   f0   3.0
   i1   3.0     1        65       300         .1          .1
   f0   4.5
   i1   4.5     1        62       400         .1          .1
   f0   6.0
   i1   6.0     1        61       800         .1          .1
   f0   7.5
   i1   7.5     1        60       1000        .1          .1
   f0   9.0
   i1   9.0     1        60       1400        .1          .1
   f0   10.5
   i1   10.5    1        60       1800        .1          .1
   f0   12.0
   i1   12.0    1        58       2400        .1          .1
   f0   13.5
   i1   13.5    1        56       3200        .1          .1
   f0   15.0

 e
```

Appendix B: Csound files for Experiment 3 sound stimuli.

The following Csound files were used to synthesize the stimuli used in Experiment 3. The orchestra (.orc) file describes the instruments created to play the score (.sco) file. That is, the orchestra file describes the general sound-generation routine, and the score file includes the changing parameters for pitch, onset (attack), amplitude, and so on.

Note that some of the parameters were manually edited for each stimulus. For example, in the score file the tempo setting, t 0 60, would have been set to t 0 120, and so on. The file would be re-saved with the new value, and the synthesis re-run to obtain the new sound file.

Frequency Set

Orchestra File

```
;*****************************************************
;*  stim4.orc
;*  CSound orchestra file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;*****************************************************

sr = 44100
kr = 2205
ksmps = 20
nchnls = 1

instr 1
     k1   linen   ampdb(p4), p6, p3, p7        ; p4=amp in db
     a1   foscili  k1, p5, 1, 1 ,0,1          ; p5=pitch or freq
          out     a1                           ; p6=attack time
   endin                                       ; p7=release time

;asig foscil amp, freq, carrier, modulator, index, function
```

Score File

```
;***************************************************************
;*  stim4-adjusted.sco
;*  CSound score file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;****************************************************************

f1  0  8192  10  1              ; sine wave

t 0 60                          ; regular tempo at t=0 is 60 beat/s

;instr  start  duration  loud(p4)  pitch(p5)  attack(p6)  release(p7)
   i1     0     2        69       100        .1          .1
   i1     +     .        66       200        .           .
   i1     .     .        65       300        .           .
   i1     .     .        62       400        .           .
   i1     .     .        61       500        .           .
   i1     .     .        61       600        .           .
   i1     .     .        60       700        .           .
   i1     .     .        61       800        .           .
   i1     .     .        61       900        .           .
   i1     .     .        60       1000       .           .
   i1     .     .        .        1200       .           .
   i1     .     .        .        1400       .           .
   i1     .     .        .        1600       .           .
   i1     .     .        .        1800       .           .
   i1     .     .        59       2000       .           .
   i1     .     .        58       2400       .           .
   i1     .     .        56       2800       .           .
   i1     .     .        56       3200       .           .
   i1     .     .        57       3600       .           .
   i1     .     .        57       4000       .           .
   i1     .     .        58       4400       .           .
   i1     .     .        58       4800       .           .

 e
```

Tempo Set

Orchestra File

```
;******************************************************
;*  stim5.orc
;*  CSound orchestra file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;******************************************************

sr = 44100
kr = 2205
ksmps = 20
nchnls = 1

instr 1
     k1   linen   ampdb(p4), p6, p3, p7        ; p4=amp in db
     a1   foscili  k1, p5, 1, 1 ,0,1        ; p5=pitch or freq
          out     a1                           ; p6=attack time
   endin                                       ; p7=release time

;asig foscil amp, freq, carrier, modulator, index, function
```

Score File

```
;***************************************************************
;*  stim5.sco
;*  CSound score file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;****************************************************************

f1  0  8192  10  1               ; sine wave

t 0 700                          ; regular tempo at t=0 is 60 beat/s

;instr  start  duration  loud(p4)  pitch(p5)  attack(p6)  release(p7)
   i1     0     1        60       1000        .1          .1
   f0     1.5    ;makes the whole sound 1.5 beats long

 e
```

Appendix C: Csound files for Experiment 4 sound stimuli.

The following Csound files were used to synthesize the stimuli used in Experiment 4. The orchestra (.orc) file describes the instruments created to play the score (.sco) file. That is, the orchestra file describes the general sound-generation routine, and the score file includes the changing parameters for pitch, onset (attack), amplitude, and so on.

Note that some of the parameters were manually edited for each stimulus. For example, in the score file the tempo setting, t 0 60, would have been set to t 0 120, and so on. The file would be re-saved with the new value, and the synthesis re-run to obtain the new sound file.

Frequency Set

Orchestra File

```
;******************************************************
;*  stim7.orc
;*  CSound orchestra file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;******************************************************

sr = 44100
kr = 2205
ksmps = 20
nchnls = 1

instr 1
     k1   linen   ampdb(p4), p6, p3, p7        ; p4=amp in db
     a1   foscili  k1, p5, 1, 1 ,0,1           ; p5=pitch or freq
          out     a1                           ; p6=attack time
   endin                                       ; p7=release time

;asig foscil amp, freq, carrier, modulator, index, function
```

Score File

```
;***************************************************************
;*  stim7.sco
;*  CSound score file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;****************************************************************

f1  0  8192  10  1              ; sine wave

t 0 60                          ; regular tempo at t=0 is 60 beat/s

;instr  start  duration  loud(p4)  pitch(p5)  attack(p6)  release(p7)
   i1     0     1        66       200         .1          .1
   f0   1.25
   i1   1.25    1        62       400         .1          .1
   f0   2.50
   i1   2.50    1        61       600         .1          .1
   f0   3.75
   i1   3.75    1        61       800         .1          .1
   f0   5.0
   i1   5.0     1        60       1000        .1          .1


 e
```

Tempo Set

Orchestra File

```
;******************************************************
;*  stim8.orc
;*  v1.0  19 April 2000
;*  CSound orchestra file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;******************************************************

sr = 44100
kr = 2205
ksmps = 20
nchnls = 1

instr 1
     k1   linen   ampdb(p4), p6, p3, p7        ; p4=amp in db
     a1   foscili  k1, p5, 1, 1 ,0,1           ; p5=pitch or freq
          out     a1                           ; p6=attack time
   endin                                       ; p7=release time

;asig foscil amp, freq, carrier, modulator, index, function
```

Score File

```
;***************************************************************
;*  stim8.sco
;*  CSound score file for dissertation stimuli
;*  Bruce Walker
;*  Rice University
;****************************************************************

f1  0  8192  10  1              ; sine wave

t 0 60                          ; regular tempo at t=0 is 60 beat/s

;instr  start  duration  loud(p4)  pitch(p5)  attack(p6)  release(p7)
   i1     0     0.200       60       800         .01          .01
   f0   .250
   i1   .250    0.200       60       800         .01          .01
   f0   .500
   i1   .500    0.200       60       800         .01          .01
   f0   .750
   i1   .750    0.200       60       800         .01          .01
   f0  1.000

 e
```