Data Collection and Analysis Techniques for Evaluating the Perceptual Qualities of Auditory Stimuli

Terri L. Bonebright

Dept of Psychology, DePauw University
Greencastle, IN USA

Nadine E. Miner

Sandia National Laboratories
Albuquerque, NM USA

Timothy E. Goldsmith

Dept of Psychology, University of NM
Albuquerque, NM USA

Thomas P. Caudell

Dept of EECE, University of NM
Albuquerque, NM USA

Abstract

This paper describes a general methodological framework for evaluating the perceptual properties of auditory stimuli. The framework provides analysis techniques that can ensure the effective use of sound for a variety of applications including virtual reality and data sonification systems. Specifically, we discuss data collection techniques for the perceptual qualities of single auditory stimuli including identification tasks, context-based ratings, and attribute ratings. In addition, we present methods for comparing auditory stimuli, such as discrimination tasks, similarity ratings, and sorting tasks. Finally, we discuss statistical techniques that focus on the perceptual relations among stimuli, such as Multidimensional Scaling (MDS) and Pathfinder Analysis. These methods are presented as a starting point for an organized and systematic approach for non-experts in perceptual experimental methods, rather than as a complete manual for performing the statistical techniques and data collection methods. It is our hope that this paper will help foster further interdisciplinary collaboration among perceptual researchers, designers, engineers, and others in the development of effective auditory displays.

1 Introduction

The evaluation and validation of auditory stimuli for experimental or application use is an important component to the successful completion of a project that utilizes sound. More often than not, sound stimuli are chosen in an ad hoc manner and are integrated into a project without conducting appropriate perceptual studies. As a result, positive or negative effects of the sound stimuli on the users are unknown. The purpose of this paper is to present a battery of data collection methods and analysis techniques for evaluating and characterizing auditory stimuli. We stress the importance of testing human subjects with auditory stimuli during both the development phase and the active use of a system. We also emphasize that the choice of methods to test the perceptual properties of auditory stimuli depends on the goals of the specific system. For example, when evaluating the veracity of a sound produced by a sound synthesis algorithm, active use experiments may not be necessary. On the other hand, when selecting auditory stimuli for use in data sonification applications, active use experiments along with discrimination and identification tests are critical. Obviously, the effort involved in conducting such tests is extensive. However, if sound is a critical system component, we believe rigorous methods of evaluation are justified.

The methods described in this paper are presented as a starting point for an organized and systematic approach to evaluating the perceptual properties of acoustic signals. The methods aim to ensure effective use of sound in a variety of applications including virtual reality and data sonification systems. These methods are presented at a level for those with little or no formal training in perceptual experimental methods. They include guidelines for subject selection, sample size, number of stimuli, pilot testing, number and type of practice trials, duration of data collection sessions, and examples of computer software that can be used to automate data collection procedures. With feedback from readers, reviewers and users, we hope to continuously improve and refine this battery of methods.

1.1 General Experimental Procedures

This section presents general guidelines for the design of empirical studies to investigate the perceptual characteristics of auditory stimuli. In the following sections, we discuss more specific recommendations that are relevant for particular types of data collection tasks. Most of the methods presented are aimed at exploring the perceptual characteristics of auditory stimuli rather than testing specific experimental hypotheses. However, in either case, the first step in designing a study is to have a clear idea of the questions to be answered and what types of data and statistical analyses are required to answer them. It is also important to recognize during the design stage that the choice of statistical analyses often determines certain characteristics about the type of data (e.g., interval level) collected and the numbers of cases and subjects required.

1.1.1 Subject Selection and Sample Size

One of the most important considerations in designing a study is to select subjects that will be representative of the population to which the findings will eventually apply. Most often we are interested in the normal adult population, with normal hearing; however, there may be times when we are designing for specialized groups, such as the aged or disabled. Obviously, our subjects need to reflect the relevant characteristics of this population. There may be other times when general subject characteristics such as gender, age, and intelligence are of interest. In addition, each research project should be considered individually using number and type of stimuli, design type, and required statistical analyses to determine the appropriate number of subjects for the sample.

1.1.2 Number and Order of Stimuli

Another design consideration is the number of stimuli and the manner of presentation. In general, the number of stimulus variables (e.g., pitch, intensity, etc.) and the number of values per variable we wish to examine will dictate the number of unique stimuli needed for a study. The order in which stimuli are presented to subjects should be controlled to eliminate or minimize order effects. For example, people are more likely to perceive a sound as loud if immediately followed by a quiet sound. Providing all possible orderings of stimuli may be possible for a small number of stimulus conditions, but often it is most effective to randomize the stimuli order. Computer control of testing makes such randomization easy to implement. If stimuli must be presented in a fixed order (e.g., with a cassette deck), then three or four random orders should be used in order to eliminate the possibility of idiosyncratic order effects in any one order.

To measure subject reliability, the presentation of a small number of randomly selected stimuli can be repeated. Cross correlation coefficients are calculated between the data from both presentations of the repeated stimuli to compute subject reliability. Subjects should be unaware that the repeated trials are different from the original trials. The number of repeated stimuli should be large enough to obtain a reliable correlation coefficient (i.e., in most cases, at least 15).

1.1.3 Experimental Sessions and Pilot Testing

The testing conditions under which data are collected are important. For auditory judgment studies, we suggest eliminating as many extraneous variables as possible (e.g., noise, visual stimuli), and keeping the environmental conditions constant across task conditions for those variables that cannot be eliminated. Also, the instructions for the procedures should be standardized in content and presentation across all subjects. However, there may be exceptions to controlling these environmental conditions, such as with active use or applied testing. In these cases, the study should be conducted in similar conditions to the actual use environment.

Regardless of the experimental procedure employed, we highly recommend pilot testing to validate experimental procedures, to help ensure that instructions are clear and specific, and to test the equipment and software. A relatively small number (e.g., three to five) of subjects is usually sufficient for pilot testing; however, additional pilot sessions may be needed if problems are discovered in the procedures. Time spent piloting and debugging the procedures pays off later in higher quality data.

The time it takes subjects to complete the study is also important. Perceptual tasks tend to be attentionally demanding and subjects may become fatigued or lose motivation over time. Most studies should try to limit the actual task time to no more than 30 minutes; however, the complete session, including instructions, debriefing, practice trials, etc., might run an hour or more. Pauses or breaks to help reduce subject fatigue can be included in the experimental session if deemed necessary from feedback during pilot sessions. Time between stimuli should be standardized and should be sufficient for subjects to perceptually separate consecutive stimuli from each other. If several hours of testing are required for a particular experiment, the study should be broken up into multiple sessions. However, for some experimental approaches (e.g., sorting of sounds) it is critical that all of the stimuli be considered in one session.

1.1.4 Practice Trials

A sufficient number of practice trials should be given to ensure subjects are familiar with the test procedures. Practice stimuli should be similar, but not identical to the actual stimuli used in the study. A small number of practice trials is generally needed to ensure that subjects understand the task; researchers can use feedback from pilot testing to determine the optimum number. It may also be important for subjects to first listen to the full set of stimuli if they will be asked to perform any type of comparative task (i.e. paired comparisons and sorting tasks). Subjects then know the complete reference set of stimuli prior to judging the relation among specific stimuli pairs.

1.1.5 Final General Recommendations

As a final general recommendation, experimenters should realize that they are often seeking subjects' subjective perceptions of the stimuli. Subjects should be instructed to respond as they deem appropriate and that there are no absolutely right or wrong responses. Further, every attempt should be made to motivate and enlist subjects to actively participate in the task.

The techniques described in this paper can be conducted with a standard computer with audio support. Several software packages provide experiment graphical user interface design platforms such as Hypercard [1], Authorware [2], and Matlab [3]. Software automation of the experiment has the advantages of random stimuli presentation, consistent subject instruction delivery and permanent data storage that is convenient for later data analysis. In some cases, the use of an audio cassette player may be more appropriate than a computer. When using either a computer or a cassette player, it is best if the auditory stimuli are delivered through high-quality headphones.

Examples of experimental procedures for collecting data using auditory stimuli can be found at various websites. For Matlab graphical user interface examples to conduct identification tasks and similarity rating experiments (including instructions, practice trials, timing, data collection and debriefing), access http://members.aol.com/nadine505/slab/matlab_gui.

2 Data Collection Methods for Evaluating Perceptual Qualities of Single Auditory Stimuli

There are three basic methods that can be used for determining the perceptual qualities of single auditory events: identification tasks, context-based ratings and attribute ratings. The researcher should consider each technique in relation to the goals of the project to determine which ones are appropriate.

2.1 Identification Tasks

Identification tasks for auditory stimuli answer the question: What audio image comes to mind when listening to this sound? [4-6]. For the purposes of the present discussion, this method provides three contributions to the understanding of sound identification. First, this technique can determine whether subjects can correctly identify objects or events by their associated sounds as well as which sounds are systematically confused with one another. Second, identification tasks reveal whether synthesized sound stimuli resemble the "intended sound" strongly enough to elicit a free-form identification without any verbal or visual context. Third, perceptually related sound labels obtained from systematically confused sounds can be identified.

Typically, identification tasks include trials where subjects listen to an auditory stimulus and respond in a free-form format with a written description. In some circumstances, sounds can be played as many times as subjects desire with no time limit. In order to obtain intuitive responses, subjects are typically not permitted to change their descriptions. Identification descriptions can include a noun and any relevant descriptive adjectives. Depending on the project, it is often helpful to ask subjects to think of the identification phrases in terms of the object(s) creating the sound. In addition to the descriptions, response times can be recorded to provide information about how difficult the identification decision is for the subjects.

The first step in data analysis for this technique is typically response content analysis. Content analysis is used to determine if the response content or meaning is identical among subjects. The subjects' responses should be aggregated and counted. For example, when Miner [6] asked subjects to identify a synthesized sound, the responses "running river water," "water running in a river", and "river noise" were aggregated into a single term with a response count of three. The terms "rain falling against a window" and "rain" were not aggregated because the first term provided additional information that would be lost if it were combined with the simpler term "rain".

Simple examination of content labels and response frequencies can provide information directly relevant to the goals of the experiment. The response content can be used to determine whether subjects correctly identified the sound stimuli. The response frequency indicates the significance of these identifications. For example, to claim successful synthesis of a target sound there should be a matching identification response label with a significant response frequency count across subjects. This indicates that the synthesized sound resembled the target sound strongly enough to elicit a matching response. Examining the next several most frequent responses provides information about the perceptual relatedness of the stimuli and can reveal systematically confused sounds. This is important data for user interface design, data sonification applications and sound synthesis. In sound synthesis, systematically confused sounds can be used as a basis to simplify and speed up the production of the synthesized sounds for use in computer software and virtual reality environments

Examination of response times can provide a measure for the quality of the auditory image formed and relevant information about the validity of subject responses. A short response time may indicate that the stimulus represents a well-known and/or quickly identifiable sound, as well as that a subject made false starts. Conversely, a long reaction time may indicate that a sound is unfamiliar or, in the case of sound synthesis, that the auditory stimulus is not a convincing replica. Additionally, long reaction times can also indicate a lapse in the subject's attention. Thus, it is extremely important for the researcher to thoroughly examine the data to determine whether the patterns of reaction times indicate information relevant to the stimulus quality or to the validity of individual subject's data. Such extreme cutoff times can be identified with standard statistical methods for identifying outliers.

2.2 Context-based Ratings

Context-based rating experiments provide a metric for measuring the perceptual sound veracity within a context. The context can be motivated by text/phrases, previous sound sequences, visual stimuli or some combination of these factors. Subjects typically rate how well the phrase and the sound match on a fixed scale, such as a 5-, 7-, or 9-point scale. Such rating data are useful for researchers interested in using earcons, sounds in video games, warning messages, data sonification and sound synthesis. This type of study can also provide a metric for quantifying the success of various sounds in an application.

One approach for obtaining a context is to use phrases obtained from the previously described identification experiment. The phrases can be systematically paired with all sound stimuli, providing both matching and non-matching pairings for subject evaluation. To obtain intuitive subject responses, each sound-phrase pair is typically played only once. Some of the sound-phrase pairs can be repeated to obtain a subject reliability measure. It is typical to provide a fixed amount of time (e.g., two seconds) for reading the context-based phrases.

The mean rating across subjects for each sound-phrase will give an overall index of how well the sound matched the phrase. Depending on application requirements, various levels of success can be determined from the average rating data. The mean ratings can also quantify the quality of the perceptual labels obtained from the identification experiment. Finally, average response times can be examined. Low average response times typically indicate well-suited phrase and sound matches.

Additional context-based rating experiments can provide further metrics for quantifying the veracity of sounds. Context can be provided by still-pictures, video images, computer graphics or an immersive virtual environment, depending on what is appropriate for the application. Some researchers hypothesize that sound stimuli would be perceived as more compelling when coupled with a realistic visual stimulus [6]. Finally, sound stimuli can be compared across various contexts, including verbal and visual contexts, in order to test for perceptual congruence.

2.3 Attribute Ratings

Attribute ratings provide information about the perceptually salient qualities of auditory stimuli. This information is important not only for understanding basic perceptual aspects of sound, but also for use with other analysis techniques, such as factor analysis and multidimensional scaling.

This procedure begins by determining the appropriate attributes to use for the rating task. This will depend on the type of stimuli used and the purposes of the application. A standard set for attributes would include measures of perceived loudness and pitch, although many other attributes can be used as well, such as roughness, annoyance, or pleasantness. The choice of the rating scale varies among researchers but typically semantic differential scales of 5, 7, or 9 points are preferred. During the procedure, subjects experience practice trials and may also be exposed to the complete range of the stimulus set prior to beginning the actual test trials. Stimuli are presented in random order across subjects, and some researchers also present the rating scales in random orders as well to help keep the subjects' attention focused on the task. For each trial, subjects listen to a stimulus and then make the ratings on the desired attributes. A pilot study should be conducted for each stimulus set to determine the maximum number of attributes subjects can effectively consider on each trial.

Typical analysis procedures consist of standard descriptive statistics as well as correlational analysis or analysis of variance [7], factor analysis [8, 9], and as an additional measure for interpreting multidimensional scaling solution spaces [10]

3 Data Collection Methods for Evaluating Relations Among Auditory Stimuli

For many sonification projects, once the researcher has determined the relevant perceptual characteristics of individual stimuli by using the methods described in the previous section, the associations among auditory events must be examined. We are recommending three techniques (discrimination trials, similarity ratings and sorting tasks) to provide data for the statistical techniques discussed in Section 4.

3.1 Discrimination Trials

For almost any application using multiple audio signals, it is important to determine if the auditory stimuli are distinguishable from one another and to measure the extent to which subjects can discriminate among the stimuli. This can be accomplished with a simple discrimination task. In this procedure, subjects are presented with two sequential comparison stimuli (A and B), which are then followed by a third stimulus (X) [4, 11]. Subjects are asked to determine if X is the same as A, B or neither. The instructions notify subjects that there will be a number of "catch" trials on which the correct response would be neither. These are included in the procedure to ensure that subjects are attending to both stimuli A and B before making their judgments rather than adopting the simpler strategy of ignoring A, attending to B, and making a same-different judgment for the B-X pair [12].

Typical analysis procedures consist of comparisons of percentage correct responses and errors for the stimuli using descriptive statistics, such as means, standard deviations and ranges. Types of errors among subjects, as well as within individual subject's responses, can be examined for patterns which can indicate perceptual similarity among specific stimuli (see [4] for an example with auditory stimuli).

3.2 Similarity Ratings

Similarity ratings, also known as proximity ratings, paired comparisons, or similarity judgments, is a common data collection method in perceptual studies (see [13] for a complete description of this technique). This method provides a means for examining the perceptual structure of a set of stimuli without imposing experimenter bias. Similarity data, along with statistical tools for analyzing these data (e.g., multidimensional scaling) have proved very fruitful in psychological research. With these methods, it is possible to obtain a visual depiction of the human perceptual space that appears to underlie the representation of a set of stimuli. This type of information can be quite useful in understanding how, and perhaps even why, subjects confuse certain stimuli.

Sound pair stimuli combinations are obtained by pairing each sound with every other sound, without regard to order. The result of this type of data collection is a symmetric data matrix where the rows and columns represent individual sound stimuli and the intersection of a row and column is the rating for that sound pair. So, for example, if there are 20 stimuli, subjects would judge 190 stimulus pairs, (= $N(N-1)/2$, where N is the number of stimuli). Judgments from each subject would be placed in a data matrix with 190 entries filling half of the symmetric data matrix. A composite matrix is obtained by averaging sound pair stimuli ratings across all subjects.

During a similarity ratings task, sounds are played in pairs, and subjects are asked to rate how similar the sounds are to each other. Subjects typically give their judgments by indicating a value along a rating scale (e.g., "1" meaning dissimilar to "5" for similar). Scales of various points have been used, but typically an odd number of points in the range of 5 to 9 is used. It is also possible to have subjects make a mark along a continuous line with labeled endpoints to indicate degree of similarity. There does not appear to be any strong advantages for any one type of judgment method. Perhaps most important is to realize that people are limited in their ability to reliably discriminate among levels of similarity, with the number of discriminable levels probably being around seven plus or minus two [14].

Subjects are usually instructed to respond quickly to each pair, not spending more than a few seconds to evaluate their similarity. Even so, because of the large number of pairs this technique places a practical limit on the number of stimuli that can be evaluated during a single session. Subjects can typically rate between 20 and 25 stimuli in one session if the sound stimuli duration is short. We recommend a silent pause between consecutive stimuli pairs with duration of at least 2-3 times the length of the stimuli in order to provide sufficient perceptual separation of the sound pairs. The upper limit on the number of stimuli is determined by the maximum recommend testing duration of 30 minutes.

Ideally, the set of stimulus pairs is randomly ordered for each individual subject. Each sound pair is played once followed immediately by a subject's rating. Subjects are not allowed to change their ratings once entered. Sometimes the amount of time it takes to respond to the stimulus pairs turns out to be useful data. These response times often indicate the subjects' certainty about their ratings. Finally, as mentioned in the earlier section on general experimental procedures, it is often a good idea to collect a second set of ratings on a small subset of the pairs. These repeated ratings provide a measure of subject reliability, which in turn can be used to determine if the subject is an outlier. Other statistical indices such as mean rating, standard deviation of ratings, and mean response time can also be calculated to determine outliers.

3.3 Sorting Tasks

Sorting tasks are another method for collecting similarity data that provides information about the perceptual relations among stimuli and can be used for multidimensional scaling and Pathfinder analysis. Traditionally, such methods have been used for visual and tactile stimuli [10]; however recent studies indicate their utility in investigating auditory stimuli as well [15, 16]. This technique is particularly useful when the researcher is interested in investigating the perceptual structure of a large number of stimuli (e.g., > 25) and wishes to collect responses for all stimuli for each subject in a single session. It is also much more efficient in terms of time for the procedure and the number of subjects required compared with similarity ratings.

During the procedure, subjects are tested individually and begin by performing a practice sorting task on a set of visual stimuli, such as color cards. They are instructed to sort the stimuli into groups according to how they perceive the stimuli relate to one another. Additionally, there are typically constraints placed on the sorting procedure, such as there must be two stimuli per group and there must be a minimum number of groups for the set. Subjects are encouraged to make changes in the groups and to take as much time as necessary to make certain their final groupings reflect the relations they feel exist among the stimuli. When the subjects are finished, the researcher asks if they have any questions about the sorting task to make certain they understand the basic sorting procedure.

Next, subjects are trained to use a software package [17] for sorting sounds. Subjects see movable icons which are placed on one side of the computer screen. They practice opening and playing the sound files and then perform a practice sorting task with a small set of auditory stimuli (less than 10) that is different from the actual set tested. Finally, subjects perform the sorting task for the target set of auditory stimuli. They begin by listening to the entire stimulus set in order to become familiar with the range and type of stimuli, and then proceed using the sorting procedure they learned from the practice sorting task.

The instructions for the sorting task can specify a specific attribute (e.g., pitch) that the subjects should use for forming their groups or they can ask subjects to use whatever attributes of the stimuli they think are important. As mentioned above, subjects are provided with constraints on the number of stimuli per group and the total number of groups they can form. Subjects are also reminded that there are no right or wrong answers for the task since the purpose of the procedure is to determine their perceptual structure for the stimuli. Subjects are allowed to listen to the stimuli as many times as they desire to achieve their final sort. Subjects are also asked to perform a final check on their groups by listening to all the stimuli and making any changes required to produce groups that best represent the relations among the stimuli.

Data from the sorting task are compiled into individual dissimilarity matrices, using the same basic procedure for compiling the matrices outlined in the previous section on similarity ratings, with 0 indicating that the subject sorted the pair into the same group and 1 representing that the subject placed the members of the pair into separate groups [10]. Individual matrices are then aggregated into a composite matrix that can be submitted to multidimensional scaling or Pathfinder analysis.

4 Statistical Techniques for Examining the Perceptual Relations Among Auditory Stimuli

Researchers investigating perceptual phenomena have a variety of statistical tools available. Techniques commonly used include regression analysis, factor analysis, analysis of variance, and cluster analysis [7-9]. However, for the present discussion, we are focusing on two statistical techniques that provide a visual representation or perceptual map of the relations among a set of stimuli: multidimensional scaling (MDS) and Pathfinder analysis [18, 19]. For auditory stimuli, these methods are used to determine how people relate or group sounds on particular dimensions or attributes.

The input to these analysis techniques is similarity data obtained from the similarity ratings or sorting tasks described in the previous section. The similarity responses are converted to dissimilarities, proximities or distances. For example, with a 5-point rating scale, the conversion is accomplished by taking six minus the similarity rating response. Sorting data are also converted in dissimilarities using the method described in the previous section. Both the multidimensional scaling and Pathfinder analysis techniques convert the similarity data to distances and then use the average distance data in the form of a matrix as input.

4.1 Multidimensional Scaling (MDS) Analysis

Multidimensional scaling (MDS) is a mathematical tool that helps systematize data in areas where organizing concepts and underlying perceptual dimensions are not well understood [10, 20, 21]. An MDS analysis represents each stimulus as a point in a multidimensional space. Similar stimuli are close to one another in the space, and dissimilar stimuli are distant from one another. MDS spatially represents stimulus similarities to reveal underlying structure in the data as a map of the perceptual space. One main advantage of MDS analysis is that it does not require a-priori knowledge of the perceptual characteristics of the stimuli and thus helps eliminate experimenter imposed constraints, which results in an analysis that is low in experimenter contamination.

MDS constructs a perceptual space by using dissimilarity data from similarity ratings or sorting tasks to assign stimuli to locations in an n-dimensional space. An iterative procedure is used that maximizes the fit of the space to the dissimilarity data. Examining several dimensionalities (i.e. number of axes) helps to determine the best solution space for a stimulus set, although three dimensions or fewer are typical in perceptual research. Knowledge and familiarity with the stimuli are necessary not only for determining the number of expected dimensions but also the dimension labels. In addition, statistical techniques can be used for determining the appropriate number of dimensions. The experimenter can use the "elbow" in scree plots constructed by using the measures of fit from the MDS analysis (R2 and stress values) for dimensions 1 through 6. Also additional measures, such as attribute ratings and acoustic parameter measurements (e.g., amplitude and frequency), can be used as criterion variables with the dimensional coordinates as predictor variables in regression analyses. The results can be used to determine which acoustic measurements or attribute ratings account for sufficient variance within the solution space to justify adding them as vectors which help to identify the dimensions (see [23, 24] for two excellent articles with specific examples of MDS analysis using vectors from regression analyses for tactile stimuli).

The MDS Alternating Least-Squares scaling (ALSCAL) algorithm [24]contained in SPSS for Windows and Macintosh platforms, by SPSS Inc., is a popular program for analyzing distance data, although several other software programs also perform MDS analysis, such as MINISSA, POLYCON, KYST and MULTISCALE [10].

4.2 Pathfinder Analysis

Pathfinder [18] is a statistical scaling procedure that takes as input a set of distance data and provides as output a network. Each node in the network corresponds to a stimulus in the data set and the pattern of direct links in the network reflects the pattern of similarity in the distance data. Pathfinder differs from MDS both in the nature of the representation that is produced (network vs. space) and in the method by which the ratings are analyzed. However, both have the common goals of minimizing the effects of noise that is inherent in most sets of proximity data and in uncovering the underlying structure that is presumed to exist in such data sets.

The Pathfinder algorithm creates undirected networks based on proximity (or distance) information between pairs of stimuli [18]. Because of the nature of its algorithm, Pathfinder networks tend to emphasize the most related pairs in the data matrix. As a result, specific local relations in the data are maintained in the network. In contrast, MDS is equally influenced by all pairs, and so solutions are globally based. Whereas MDS provides a definition of the perceptual parameter dimensions, Pathfinder networks provide a means for analyzing the conceptual relatedness of the sound stimuli. Pathfinder analysis can answer many questions about the sound stimuli including: what, if any, clusters exist within the perceptual sound space and how are the sounds related to one another. For sound synthesis algorithms, this information is useful for refining and extending sound models to include synthesis of sounds that are closely related to the sounds generated by an existing model.

The Pathfinder algorithm functions as follows. First, all stimulus pairs are connected to one another to form a completely linked network. Each link is assigned a weight based on the raw proximity value for that pair of objects. Next, Pathfinder examines each link weight and repeatedly asks the question, "can this link be removed?" To remove a link, Pathfinder compares the direct link weight between two objects to the distances for all other paths involving at most q links. If the direct path weight is greater than or equal to any other path distance, the direct path is redundant and is removed. Two parameters affect the algorithm: the q-parameter and the r-parameter. The q-parameter , $2 <= q <= n-1$, constrains the search to only those pairs of nodes that are connected by q links or fewer. The maximum q value allows the search to extend over all pairs, hence yielding the sparsest graph. The r-parameter specifies how path distance is computed, in much the same way as it does in the Minkowski distance function. For ordinal data, the r-parameter is set to infinity.

It is often informative to examine several different Pathfinder networks of varying complexities. Simpler networks (fewer links) indicate the strongest relationships between sound stimuli. Denser networks (more links) show weaker interrelationships between sounds. Which sounds are directly connected and how multiple sounds cluster together offer useful information about how subjects conceptually relate sounds to one another. This type of information may help to extend the synthesis scope of sound synthesis models [6].

When Pathfinder networks are combined with MDS results, a more complete perceptual and conceptual picture of the interrelationships among the sound stimuli results. Both analysis methods provide guidance for sound model refinement and synthesis extensions to a broader class of sounds.

5 Active Use Experimental Procedures

Assessment of sound applications needs to continue into the actual use of the product or application in the “real-world” environment. This type of assessment is becoming more routine for computer hardware and software design since it has been found to help maximize the development of products with high user satisfaction [25]. Thus, researchers who evaluate sound applications should follow the lead of the computer industry and incorporate user satisfaction and acceptance measures using techniques such as surveys and verbal protocols. To maximize external validity, usability laboratories can be set up to provide a comparable environment to the one where the product will actually be used [26, 27]. In these laboratories, subjects can work with the product and provide feedback to the researchers. In this type of testing, it is imperative that the subjects realize that they are not being tested, but rather that it is the application or product that is under investigation.

5.1 Surveys

Surveys are designed to collect data after the subject has worked with the application. The questions in the survey are dictated by the particular application and concerns of the researchers. General questions about annoyance and distraction levels, overall satisfaction of interaction, and whether the subject would use such a product would be particularly pertinent for sound applications. These questions can be asked by using either closed format items, such as rating scales, true or false questions, and check boxes for relevant properties, or free response formats. Data from closed format items are easier to analyze, but free responses may provide a richer source of data. Researchers interested in using surveys should plan to pilot several versions of their questionnaires to ensure that subjects understand them and can provide appropriate answers (for a general reference on surveys, see [28]).

5.2 Verbal Protocols

Verbal protocols require subjects to "talk aloud" while they work with an application . The subject’s statements can be recorded on videotape [29] or an experimenter can take notes and cue the subject to elaborate on their comments during the session. The advantages of this type of procedure are that subjects do not need to rely on memory in order to report their responses at a later time and that subjects can many times provide spontaneous comments about improvements or problems while they are working with the application. Some researchers have pairs of subjects work together since this leads to more information for the researcher while the users explain aspects of the program to one another [25].

6 Conclusions

This paper provides a general framework for data collection and analysis techniques appropriate for evaluating the perceptual properties of auditory stimuli. Methods presented range from simple identification tasks, attribute ratings and discrimination tasks, to complex analysis methods such as MDS and Pathfinder analysis. Practical and hands-on references are provided as a guide for the non-expert to begin the consideration of which tests would be most appropriate to answer his or her questions, although we would recommend consulting someone with expertise in these methods before embarking on a specific project.

We emphasized that the experimental procedures are important to conduct during the product/project development phase in order to characterize the audio imagery and perceptual effectiveness of the sounds used. Furthermore, active use evaluation methods were described to enable evaluation of the sound application in the actual use environment and to provide information for continuous product improvement.

Finally, we hope that this set of techniques will lead to further collaborative work among researchers in disciplines interested in developing effective auditory displays.

7 References

1. HyperCard 2.4 [Computer software] Apple, Cupertino, CA, 1998.

2. AuthorWare 4.03 [Computer software] Allen Interactions, Minneapolis, MN, 1998.

3. Matlab 5.1 [Computer software] Building GUIs with Matlab. MathWorks, Inc., Natick, MA, 1996.

4. Ballas J. A. Common factors in the identification of an assortment of everyday sounds. Journal of Experimental Psychology: Human perception and Performance 1993; 19: 250-267.

5. Mynatt E. D. Designing with auditory icons. In: Proceedings of the Second International Conference on Auditory Display (ICAD), Santa Fe, NM ,1994, pp 109-119.

6. Miner N. E. Creating wavelet-based models for real-time synthesis of perceptually convincing environmental sounds. Ph.D. Dissertation, University of New Mexico, 1998.

7. Maxwell S. E., Delaney H. D. Designing experiments and analyzing data: A model comparison perspective. Wadsworth, Belmont, CA, 1990.

8. Dillon W. R., Goldstein M. Multivariate analysis: Methods and applications. John Wiley & Sons, New York, 1984.

9. Tabachnick B. G., Fidell L. S. Using multivariate statistics. Harper & Row, New York, 1989.

10. Schiffman S. S., Reynolds M. L., Young F.W. Introduction to multidimensional scaling: Theory, methods and applications. Academic Press, New York, 1981.

11. Turnage K. D., Bonebright T. L., Buhman D. C., Flowers J. H. The effects of task demands on the equivalence of visual and auditory representations of periodic numerical data. Behavior Research Methods, Instruments, & Computers 1996; 28: 270-274.

12. Garbin C. P. Visual-haptic perceptual nonequivalence for shape information and its impact upon cross-modal performance. Journal of Experimental Psychology: Human Perception and Performance 1988; 14: 547-553.

13. Nunnally J. C., Bernstein I. H.. Psychometric theory. McGraw-Hill, New York, 1994.

14. Miller G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 1956; 63: 81-97.

15. Bonebright T. L. Vocal affect expression: A comparison of multidimensional scaling solutions for paired comparisons and computer sorting tasks using perceptual and acoustic measures. Ph.D. Dissertation, University of Nebraska, 1996.

16. Bonebright T. L.. An investigation of data collection methods for auditory stimuli: Paired comparisons versus a computer sorting task. Behavior Research Methods, Instruments&Computers 1996,28:275-278.

17. Canary Bioacoustics Research Program 1.2 [Computer software] Cornell University, Laboratory of Ornithology, Ithaca, NY, 1995.

18. Schvaneveldt R. W. Pathfinder associative networks: Studies in knowledge organization.: Ablex Publishing, New Jersey, 1990.

19. Goldsmith T. E., Johnson P. J., Acton W. H. Assessing structural knowledge. Journal of Educational Psychology 1991; 83: 88-96.

20. Davison M. L. Multidimensional scaling. John Wiley & Sons, New York, 1983.

21. Kruskal J. B., Wish M. Multidimensional scaling. Sage, Beverly Hills, CA, 1978.

22. Garbin C. P., Berstein I. H. Visual and haptic perception of three-dimensional solid forms. Perception and Psychophysics 1984; 36: 101-110.

23. Hollins M., Faldowski R., Rao S., Young, F. Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis. Perception & Psychophysics 1993; 54: 697-705.

24. Young F. W., Lewyckyj R. ALSCAL-4 User's Guide (2nd ed.). Data Analysis and Theory Associates, Carrboro, NC, 1979.

25. Schneiderman B. Designing the user interface: Strategies for effective human-computer interaction. Addison-Wesley, Reading, MA, 1998.

26. Dumas J., Redish J. A practical guide to usability testing. Ablex, Norwood, NJ, 1993.

27. Nielsen J. Usability engineering. Academic Press, New York, 1993.

28. Oppenheim A. N. Questionnaire design, interviewing, and attitude measurement. Pinter Publishers, New York, 1992.

29. Harrison B. L. Video annotation and multimedia interfaces: From theory to practice. In: Human Factors Society Thirty-Fifth Annual Meeting, 1991, pp 319-322.

Portions of this work were funded by Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.