# USING DSP-BASED PARAMETRIC PHYSICAL SYNTHESIS MODELS TO STUDY HUMAN SOUND PERCEPTION

*Perry R. Cook*

Princeton University CS Department
(also jointly in Music)
35 Olden St., Princeton, NJ 08544 USA
prc@cs.princeton.edu

*Stephen Lakatos*

Washington State University Psychology Dept.
14204 NE Salmon Creek Ave.
Vancouver, WA 98686 USA
lakatos@vancouver.wsu.edu

**ABSTRACT**

A series of studies in human auditory perception, memory, attention, discrimination, and learning are described, based on the use of physical synthesis models. The use of realistic sounding DSP-based physical models allows individual parameters to be isolated and tested, while still keeping an "ecological" approach to the experiments. Partly due to the flexibility of the model parameters, and due to the large numbers of subjects tested, and also due to the large collections of sounds we have been able to use, the studies described have provided much new evidence about the nature of human cognitive auditory mechanisms. A new research agenda has been born out of this work, endeavoring to answer difficult questions about the subjective nature of sonic "reality," "naturalness," "presence," "immersion," etc.

## 1. INTRODUCTION

The advent of computer sound synthesis and DSP-based sound processing in the early 1960's ushered in a new era for the study of human audio perception. Accurate control over the amplitude and phase of sine components, precisely timed clicks, noise bands of arbitrary width and shape, accurate stereo and multi-channel playback, and other such stimuli became much easier (this paradigm of research was inherited from traditions dating back to the invention of the vacuum tube), and were exploited in many important psychoacoustic studies. Much has been learned from the ability to generate accurate temporal and spectral auditory stimuli, but unfortunately the unnatural bleeps and clicks became a dominant norm for psychophysical hearing research for 40 years, and still dominate many studies today.

This paper briefly describes a series of studies using physical DSP models and processing to understand complex auditory perception. The work represents a new trend in "ecological psychoacoustics," where the stimuli are natural sounds, yet the rigorous scientific control characteristics of traditional psychoacoustic research are retained.

Our first studies used parametric physically-inspired models (computationally compact, but with all parameters directly related to the basic sound production physics) to probe the nature of human auditory learning, sensitivity, attention, and memory. A selective attention study was conducted to determine whether listeners could select model parameters on which to focus their auditory attention. In a larger subsequent study, we presented different interactive parameter interfaces to listeners in order to measure the effect that exploratory sound control environments with varying structure have on listeners' ability to discriminate and remember sounds they have explored in such environments.

Another set of studies asked listeners to provide similarity ratings for a large collection (150) of real-world sounds by means of a new interactive graphical computer program that gave listeners considerable flexibility to move, label, and sort the sounds. Analysis of the collected similarity data revealed various perceptually salient features of sounds, that can be used to inform machine classification and clustering algorithms [1].

The real-world sound collection study motivated our most recent work, which is just beginning as a large research project. This third research area is a series of pilot studies aimed at developing a psychoacoustic tool to begin to ask the essentially intractable (scientifically) question, "does this sound seem real to you?" Literature reviews from fields as diverse as philosophy, virtual reality, user interface design, art, music, psychoacoustics, and many others were compiled to collect a potential set of terms related to auditory realism. These terms, along with corresponding sets of sounds, were given to listeners so that they could rate how well each term characterized each sound. Results from this and other research projects were combined to isolate a core set of descriptive terms that best represent different facets of auditory realism. We are currently using this set of terms with a variety of sound recordings and synthesis models to further validate our realism measures.

## 2. LEARNING BY EXPLORING PHYSICAL MODELS

Our first studies probed the relation between the parameters of a sound source and listeners' ability to attend to, remember, and discriminate these parameters. Both passive (sound presentation only) and active (direct manipulation of the model) conditions were tested. The next section briefly describes the Physically Inspired Stochastic Event Model (PhISEM) family used for these experiments.

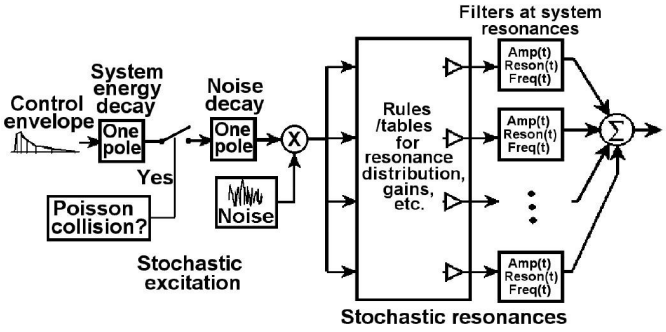
### 2.1. PhISEM Models

Physically Inspired Stochastic Event Modeling (PhISEM) [2] is a DSP algorithm that was devised to synthesize the variety of sounds that are created naturally by random particle systems. The basis is a stochastic calculation of collisions of many independent objects, and the application of system resonance(s) to simple collision sounds. Models of particles in containers

(beans within the gourds of virtual maracas) were solved numerically using the fundamental Newtonian equations of motion. Simulations with varying numbers of particles and damping (energy loss when particles collide with each other and the container) were run, and statistics were collected about the likelihood of sound-producing collisions, the overall decay in sound energy, and other properties of the simulated system.

The PhISEM synthesis algorithm reduces the behavior of the particle systems to a statistical process in which parameters relate directly to the parameters collected in the simulations. System energy (total kinetic energy in the moving particles) decays exponentially, and the decay is faster for systems with higher damping. System energy is increased by shaking (scraping, bowing) or other external activities that introduce energy into the system. There is a Poisson probability of sound-producing collisions, with a high waiting time (low probability) between collisions for systems with few objects, and a low waiting time for many objects. Sound-producing events are modeled as short exponentially-decaying bursts of white noise, and the system resonances are modeled using simple resonant filters.

Even though the original models studied were particles within spheres, the PhISEM model extends well to any system with multiple independent semi-random sound-producing objects. This includes maracas, tambourines, sleighbells, coins in a pocket, wind chimes, and even gravel/grass/snow beneath walking feet [3]. Figure 1 shows the PhISEM synthesis block diagram, along with a pseudo-C code listing of the algorithm for simple maraca synthesis. As shown in the diagram, the filters can be fixed (as is the case for the maraca gourd), or variable (as for wind chimes).




```
#define SOUND_DECAY 0.95
#define SYSTEM_DECAY 0.999
                                        //  EACH SAMPLE:
shakeEnergy *= SYSTEM_DECAY;    // Exponential system decay
if (random(1024) < num_beans)    {    // If collision
     sndLevel += gain * shakeEnergy;   //    Add energy to sound
     //  Could also reallocate resonances here for some models
}
input = sndLevel * noise_tick();        // Sound is random
sndLevel *= SOUND_DECAY;        // Exponential sound decay
input -= output[0]*coeffs[0];           // Do one or more simple
input -= output[1]*coeffs[1];           //      system resonance
output[1] = output[0];                  //          filter
output[0] = input;                      //              calculation(s)
```

Figure 1. *Physically Inspired Stochastic Event Modeling (PhISEM) block diagram and C Code for simple maraca.*

### 2.2. Selective Attention to Physical Model Parameters

Using a probe signal paradigm, our first study endeavored to determine which model parameters were most salient for detecting an object sound in the presence of noise (a measure of selective attention to a parameter in a controlled condition). The bamboo wind chime model was used to synthesize four sounds with high (4.2Khz) and low (1.6Khz.) average resonant frequency, and few (4-6) and many (30) objects. For the experiment, a baseline task first measured each subject's absolute sensitivity to each sound in noise. Sixteen listeners were tested. A baseline task first determined each listener's detection threshold for each sound against a white-noise background using a convergent staircase method (i.e., ascending and descending tracks, with threshold taken as the average of 12 reversals after track convergence). In the main task, sounds served as both cue (12 dB above a listener's threshold) and target (at the listener's threshold) in random order. On each trial, listeners first heard a 625 ms. cue consisting of noise plus a wind chime sound at 15dB above the subject's threshold for that sound, then one second of silence, then two 625 ms. noise bursts presented separated by 0.5 seconds of silence. The listeners' task was to determine which of the two observation intervals contained the target two-alternative forced-choice paradigm). Trials were blocked by the sound features being varied (i.e., resonant frequency and object number).

On so-called "attended" trials, both cue and target possessed either the same resonant frequency or the same number of objects, while on "unattended" trials, cue and target differed in one of these features. Impoved detection performance on attended trials would imply the presence of attention "bands" for the sound features being attended. In fact, detection was better for attended (65%) than for unattended (47%) targets, indicating that listeners could attend selectively to object number. A second experiment using different instruments, rather than the same instrument for cue and target on a given trial, also yielded improved detection for object number, but not resonant frequency, suggesting that listeners' abilities to attend to object number is relatively independent of, or abstracted from, a specific sound source.

### 2.3. Learning Physical Models by Exploration

Building on the success of the attention study, we designed an experiment in which listeners could actually interact with the parameters of the virtual sound-producing objects. Five different pools (for five experimental conditions) of 15 listeners were allowed to use a real-time graphical user interface computer program (GUI) to explore a total of eight physical shaker/scraper models. The models used were maraca, cabasa (afuche), guiro, bamboo wind chimes, tambourine, sleighbells, ratchet, and Coke-can (scraped).

The first exploration condition was called "highly structured," and is shown in Figure 2. In this condition, a picture of each instrument was shown, grouped with four sliders beneath it. Each block is labeled with the name of the instrument, and each slider is labeled with the parameter it manipulates (Excitation energy, Decay, Number of Objects, and Resonance Frequency). Listeners for this condition were also provided with a written description of each instrument. Feedback was provided to subjects during testing.

The second experimental condition (Figure 3) was called "moderately structured," and presented labeled and ordered sliders in blocks for each instrument, but without the written descriptions, labels, or pictures.

The third experimental condition (Figure 4) was called "weakly structured," presenting sliders for each instrument in a block, but with no labels, pictures, or descriptions.

The fourth experimental condition (Figure 5) was called "unstructured," and presented all 32 sliders (8 instruments times 4 sliders) in completely random order, with no grouping or labels. Thus in this condition, no two adjacent sliders affected the same instrument and there were no clues provided at all as to the effects or links of any of the sliders.

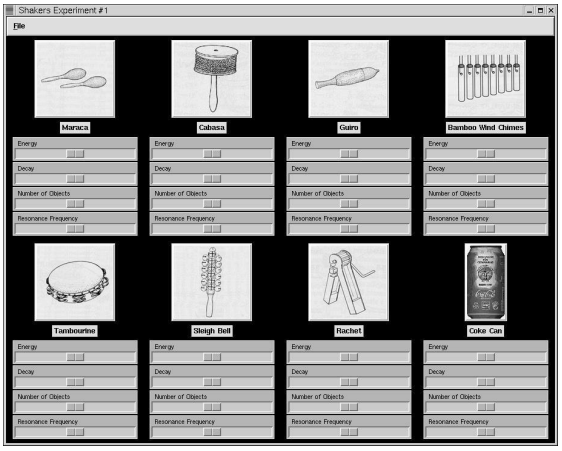



Figure 2. *Interface for "highly structured" experiment.*

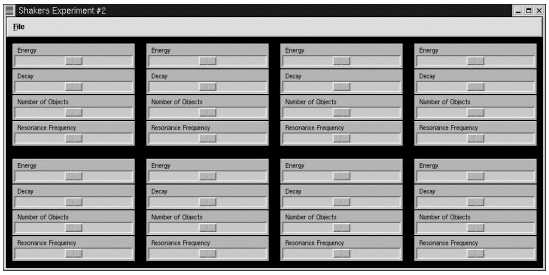



Figure 3. *Interface for "moderately structured" experiment.*

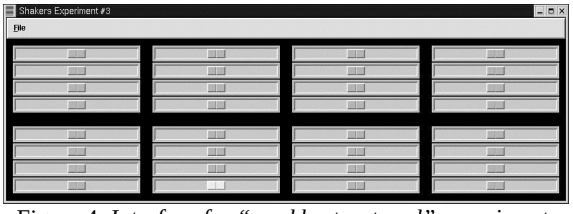



Figure 4. *Interface for "weakly structured" experiment.*

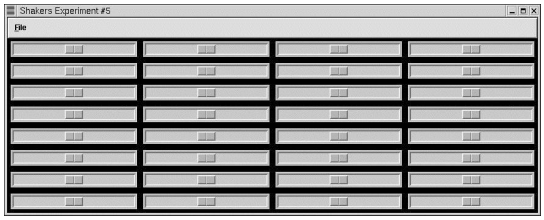



Figure 5. *Interface for "unstructured" experiment.*

The fifth experimental condition (control) group was not allowed to explore the instruments at all.

After all subjects completed a 15 minute learning session in which they explored (or didn't, in the case of the control group) the instruments using the appropriate GUI, they then completed a 15 minute testing session consisting of a discrimination task, a memory task, and a similarity-rating task (randomized within testing sessions).

The testing phase comprised three listener tasks. In the discrimination task, listeners judged whether two sounds presented in sequence were the same or different. The sound pair always came from the same instrument, but sounds were varied along one of two parameters – damping or number of objects. In the memory task, listeners heard a sequence of sounds (between three and seven total), then a target sound afterward. Listeners then judged whether the target sound had occurred in the prior sequence. Finally, a similarity-rating task had listeners estimate the similarity of all pairwise comparisons of 12 sounds (the same sounds used in the discrimination task). Each listener completed three alternating sessions of learning and testing.

Learning improvements across testing sessions were found only for the object number parameter. The most significant results from the study were that by the third testing session, both the memory and discrimination performances had improved proportional to the richness (highly structured to unstructured) of the learning interface. Figure 6 shows the performance in the discrimination task for trials 1 and 3, and Figure 7 shows the performance in the memory task for trials 1 and 3. There were no significant differences across learning interfaces for the similarity-rating test. More complete descriptions of these experiments are described in [4].

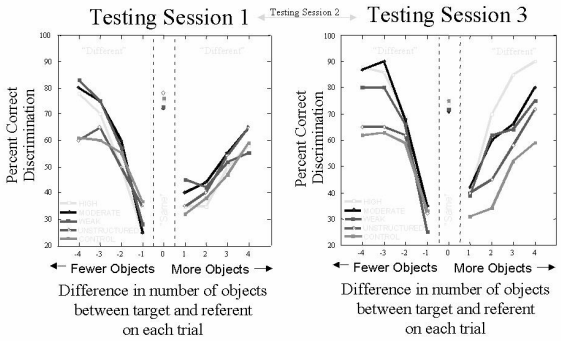



Figure 6. *Discrimination results show improvements proportional to structure of learning condition.*

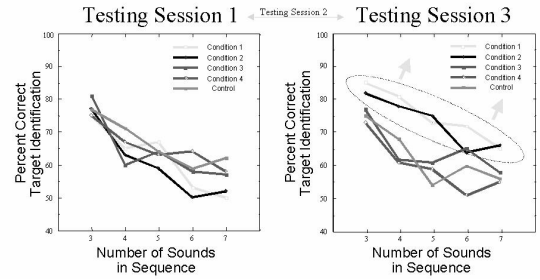



Figure 7. *Memory results show improvements proportional to structure of learning condition.*

## 3. SIMILARITY SORTING OF REAL-WORLD SOUNDS

Our interest in sound effects synthesis led us to a study of the characteristics of the perception of recorded sound effects. We selected a set of 150 representative sound effects from library CDs, with the main criterion being that the sounds are those that we generate and control with our gestures in real life. No background sounds, music, or speech were used. The difficulty of conducting pair-wise comparison tasks on 150 sounds led us to create a new interface, called the Sonic Mapper (shown in Figure 8) [5]. This program allowed subjects to select, play, move, group, and label sounds by similarity. The interface required users to listen to each sound a minimum number of times, and also enforced pair-wise comparisons of a subset of the sounds. The subset pair-wise comparison data was used to verify distance results from Multi-Dimensional Scaling analysis of the sorting data. Two similarity testing conditions and subject groups were used. One group was instructed to think about the object(s) and gesture(s) that would have made the sounds. The other group was specifically instructed to not think about these factors, and instead concentrate only on the sounds themselves. One main result from these studies is that the MDS distances correlated well with the subset of pair-wise distances, indicating that the Sonic Mapper tool can be used for similarity ranking tasks on large stimulus sets. Full results of these studies are reported in [6].

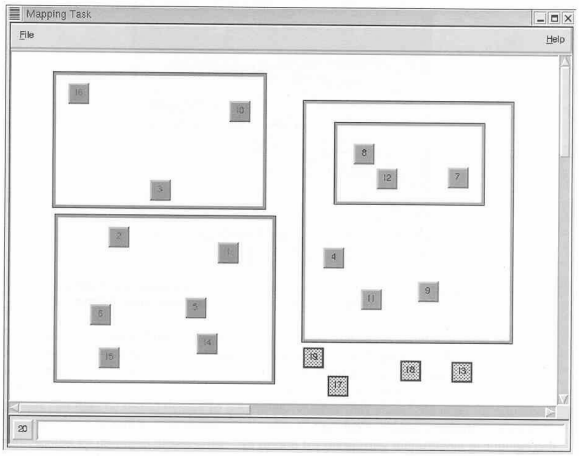



Figure 8. *Interface to the "Sonic Mapper," which allows the user to move, play, group, and sort sounds by similarity*.

## 4. QUANTIFYING AUDITORY "REALISM"

Working with the real-world sounds, and attempting to develop and calibrate parametric synthesis models for them, led us to our newest area of research, which attempts to determine the realism of recorded or synthesized sounds. This is a historically pesky problem, pondered by philosophers, engineers, sound designers, and many others. Our initial studies led us to propose a set of eight orthogonal factors: detail, physical plausibility (biological and non-biological), temporal consistency, vividness, presence, and whether a sound evokes sensory images and related memories. Sets of 8-9 declarative statements were selected by committee for each factor. In an exploratory study, 82 participants rated how representative each exemplar was of each statement. Statistically significant differences were found between exemplars for all factors.

A subsequent study using new sounds, coupled with compressed, degraded, and parametrically synthesized versions of those sounds is currently underway. Results will be presented at the conference.

## 5. CONCLUSIONS AND FUTURE DIRECTIONS

We have learned much from our DSP-based perceptual studies, but much work remains to be done. We are currently focusing most of our efforts on the continued development of the assessment tool for auditory realism. However, we are also adding two conditions to the learning study. In one condition (passive listening) listeners will hear recordings made while other listeners explored the instruments in the highly structured case, allowing us determine if the structure can be perceived simply by listening to the sounds of the structured exploration (a highly structured user is more likely to deal with the parameters of one model at a time). Another group of listeners will be allowed to explore the models by "playing" them using a hand-held shaker (an accelerometer mounted in a small wooden container, sending energy into the computer synthesis model). We anticipate that this more direct haptic control over the shaking of each virtual instrument will allow listeners to learn even more about the physical characteristics implied by the models.

## 6. ACKNOWLEDGEMENTS

We thank our collaborators; Colin Harbke, Gary Scavone, Candice Lindsay, Georg Essl, and Georgos Tzanetakis. This research was sponsored under Air Force Grant F49620-99-1-0293, NSF CAREER Grant 9984087, and New Jersey Council on Science and Technology Grant 01-2042-007-22.

## 7. REFERENCES

[1] Tzanetakis, G. and Cook P. "MARSYAS 3D: A Prototype Audio Browser-Editor Using a Large Scale Immersive Visual and Auditory Display," *Proc. Int. Conf. Aud. Display*, Helsinki, 2001

[2] Cook P., "Physically Inspired Sonic Modeling (PhISM): Synthesis of Percussive Sounds," *Computer Music Journal*, 21: 38-49.

[3] Cook P., "Modeling Bill's Gait: Analysis and Synthesis of Walking Sounds," *Proc. AES 22nd Intl. Conf. on Virtual, Synthetic, & Entertainment Audio*, 73-78. 2002.

[4] Lakatos, S., Scavone G., and Cook P. "Selective Attention to the Parameters of a Physically Informed Sonic Model," *J. Acoust. Soc. Amer.* 107(pt.1), L31-L36, 2000.

[5] Lakatos, S., Scavone G. and Cook P. "An Interactive Similarity Rating Program for Large Timbre Sets," *141 Acoust. Soc. Am.* (Poster), 2001.

[6] Scavone, G., Lakatos S. and Harbke, C. "Perceptual Spaces for Sound Effects Obtained With an Interactive Similarity Rating Program," *Intl. Symposium on Musical Acoustics*, Perugia, Italy, 2001.