MODELING BILL'S GAIT: ANALYSIS AND PARAMETRIC SYNTHESIS OF WALKING SOUNDS

PERRY R. COOK

Princeton University Dept. of Computer Science (also Music), 35 Olden St., Princeton, NJ, USA, 08544
prc@cs.princeton.edu

This paper presents algorithms and systems for automatic analysis and parametric synthesis of walking and other (gesture-rate) periodically modulated noisy sounds. A recording of walking is analyzed, extracting the gait (tempo and left/right asymmetries), heel-toe events, etc. Linear prediction is used to extract the basic resonances. Wavelet decomposition is performed, and a high frequency-subband is used to calculate statistics for a particle resynthesis model. Control envelopes are extracted from the original sound. A real-time synthesis program allows flexible resynthesis of walking sounds, controlled by a score extracted from a sound file, a graphical user interface, or parameters from game/animation/VR data. Results for the analysis algorithm are presented for synthesisized data, and for hand-crafted real experimental sounds of gravel.

INTRODUCTION

A staple of production for movies, stage performance, television, and radio dramas is the addition (often last minute) of artificial and natural sound effects. For movies, "Foley" artists put shoes on their hands and "walk" in boxes of gravel, leaves, cornstarch (for the sound of snow), etc. in real time, effectively "acting" the sounds as they watch the scenes. Radio and stage sound effects performer/engineers use tape cartridges or CDs of pre-recorded sounds, and physical noisemakers to add sounds in real time. For offline production and real-time dramas, these might indeed be the best means to add sounds. However, for virtual reality, training simulations, and games, the lack of parametric control, along with the need for large libraries of pre-recorded sounds and complex software for selecting the right sounds, forces inherent compromises in responsiveness and sonic quality.

One clear benefit of parametrized analysis/synthesis of sound effects is the possibility of significantly reducing the memory storage required for a large library of digital walking sounds. Greater benefits, however, include the ability to economically synthesize sounds in real time, coupled to parameters generated by sensors (virtual reality and training simulations), or data from games, animations, etc [1].

The synthesis method presented here is based on prior work by the author, but the primary novelty presented in this paper is the presentation of an algorithm for estimating the sonic granularity parameter (previously it had to be adjusted manually by ear), and the integration of the entire system into an interactive Graphical User Interface workbench. Related work [2][3] has investigated various background and noise-like sounds such as walking, but the work presented here is unique in the fundamental algorithm and parametric architecture, testing the multi-step analysis algorithm on real and synthesized data, and the integration of the analysis/resynthesis algorithms into a complete interactive system (for which all source code is made publicly available for free).

1 SYSTEM ARCHITECTURE

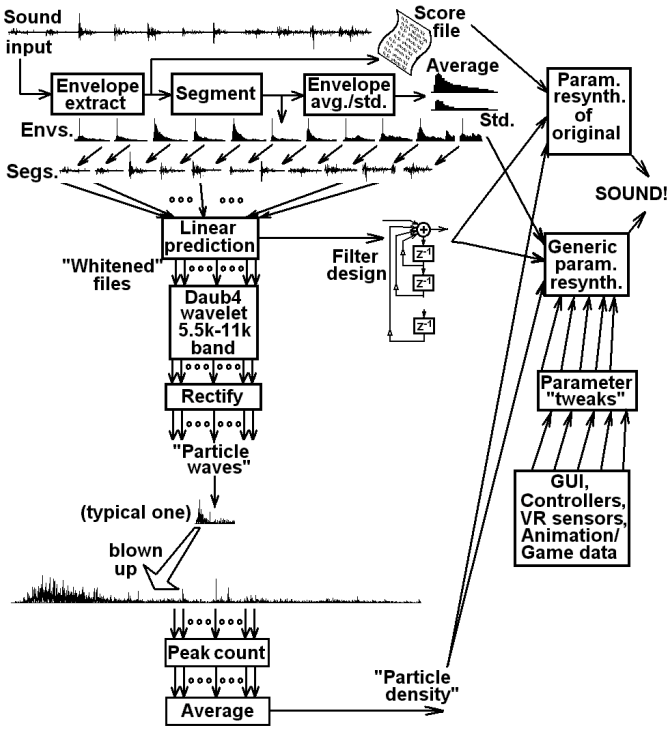

Figure 1: Walking sound system architecture.

Figure 1 shows the overall system architecture. A sound file is analysed to determine the individual footfalls, and the average gait (tempo plus left/right asymmetries). A control envelope scorefile is written. The sound file is marked and segmented into individual footstep events. The individual envelopes are averaged to yield an average control envelope and a standard envelope deviation. Linear prediction is performed on each footstep sound to determine the overall resonances for a resynthesis filter, and to yield "whitened" (spectrally flat) step sounds. Wavelet extraction of a high frequency band is performed, and the (rectified) peaks are used to estimate a "particle density." New files with arbitrary lengths, gaits, materials etc. can be generated automatically, using the extracted average and deviation envelopes. A real-time walking synthesis program can be controlled by parameters from a score file, a Graphical User Interface (GUI), data from foot/ground collision detections in a game or simulation, sensors in a virtual reality system, etc.

2 ENVELOPE ANALYSIS

The first stage of signal processing involves extracting and analysing the overall envelope of a walking sound. Figure 2 shows a simple non-linear low-pass filter for performing envelope extraction. The two signal-dependent pole positions are similar to rise/fall times on audio compressors and many historical analog envelope followers [4], allowing the output to rise faster when the signal is rising, and fall slower when the signal is falling. This helps to ensure that peaks are tracked accurately, while still eliminating spurious high-frequency components. Due to the low-pass filtering effects of the envelope follower, and also the generally slow nature of walking gestures and steps themselves, the extracted envelope can be stored at a much lower sample rate than the original sound.

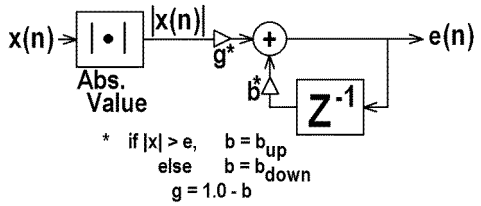

Figure 2: Digital Envelope follower.

An autocorrelation of the envelope signal is performed, to get a rough periodicity estimate. Significant peaks (local maxima over a threshold) are marked in the envelope, then a set of "best peaks" is selected, to meet various criteria of periodicity and expected walking tempos. At this point the event-marked envelope can be inspected, and peaks can be edited by hand if desired, though the algorithm is quite reliable over a fairly large class of walking sounds. Figure 3 shows some envelopes, automatically marked by the system.

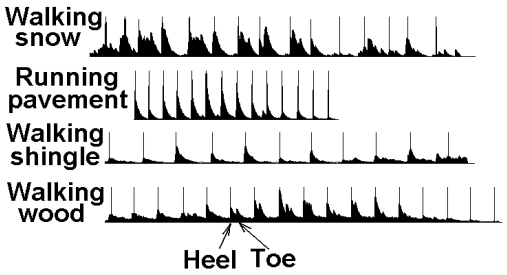

Figure 3: Some extracted/marked envelopes.

The envelope can be written out as a constant update rate "gestural" scorefile (30 Hz. updates of MIDI aftertouch, for example). To make the representation parametric, the envelope is also segmented into individual step events, bounded by a set of lowest significant values coming just before marked peaks. Again the same constraints on periodicity used to mark the original peaks are applied to mark the cut points. The original footfall sound segments are stored as individual soundfiles to allow identity resynthesis later. Next, the individual envelopes are compared to yield an average and standard deviation envelope, as shown in Figure 4. Significant sub-events can sometimes be found within the envelopes, such as the Heel/Toe events marked in Figure 3. Events like this have been shown to be important in the perception of identity, age, sex, etc. purely from walking sounds [5]. The parametric prototype envelopes are stored as multiple (6-10) break-point linear envelopes based on ordered triplets of (time, amplitude, deviation), chosen to best fit the peaks, valleys, and slope changes. 8 breakpoints correspond well to the important walking events of heel and toe landings and lifts.

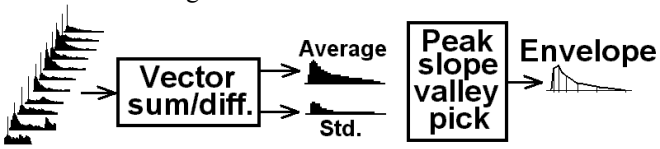

Figure 4: Envelope prototyping.

3 PARTICLE MODELING

The heart of the parameterization and resynthesis of walking sounds described here is based on PhISEM (Physically Inspired Stochastic Event Modeling) [6]. This type of modeling is quite natural since much of the sound produced by walking on various textures involves particle interactions, excited by friction/ pressure from the feet. Stochastic parameterization means that we don't have to model all of the particles explicitly, but rather only the probability that particles will make noise. For many particle systems, this is modeled well by a simple Poisson distribution, where sound probability is constant at each time step, giving rise to an exponential probability waiting time between events as shown in Figure 5.

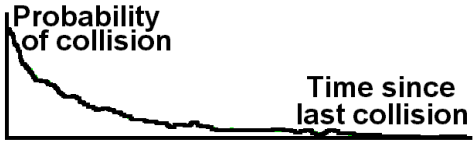

Figure 5: Poisson sound event probability.

The sound of each particle is assumed to be a noisy excitation of whatever system resonances are present. This could be the resonances of the actual particles themselves (rocks for instance), or of one or more larger resonating structures (such as planks in the floor excited by scraping of walking feet).

3.1 Resonance Modeling

As a first sound analysis step in PhISEM, the resonant structure is removed using Linear Prediction [7], keeping the "whitened" residual as the "raw" particle sound. To determine the proper filter order, the LPC prediction order can be incrementally increased until the residual power does not decrease significantly. Figure 6 shows the waveforms and spectra of a particle system before and after 2nd order LPC processing. LPC is performed on a per-footstep basis, writing the filter coefficients into the score file along with the excitation envelope. For a given walking material, and for generic resynthesis later, an average LPC filter is also computed and stored.

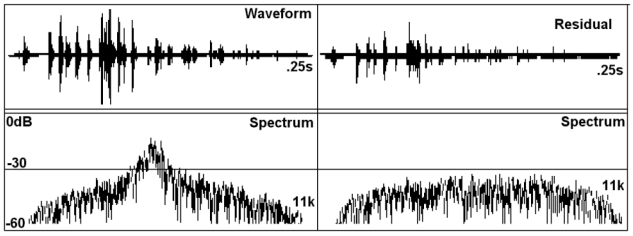

Figure 6: Removing resonance by Linear Prediction.

3.2 Particle Modeling

Each particle noise event can be modeled as an impulse, but to better "fill in" the sonic events , a short exponentially decaying (α_c) noise burst is used. To model the total sound, we do not explicitly overlap and add these events, because the sum of exponentially decaying indepenent noises (with the same exponential time constant) is an exponentially decaying noise (with the same time constant). On each collision, we simply add energy to the exponential sound state, corresponding to the current system energy (net kinetic energy of all particles in the system). We keep and calculate the exponentially decaying (α_s) state variable representing the system energy, modified (added to) by the control envelope. So the collisions are modeled by calculating a random number, and if that number is < the particle constant N, the switch in Figure 7 closes for one sample, exciting the exponential noise source with energy equal to the current system kinetic energy. At this point the synthesis algorithm is complete (Fig 7).

Both exponential functions are modeled using simple one-pole recursive filters (poles at α_c and α_s). Two random number calculations are required per time step, one for the Poisson collision decision, and one for the excitation noise source. The algorithm is efficient, depending mostly on the order of the resonant filter.

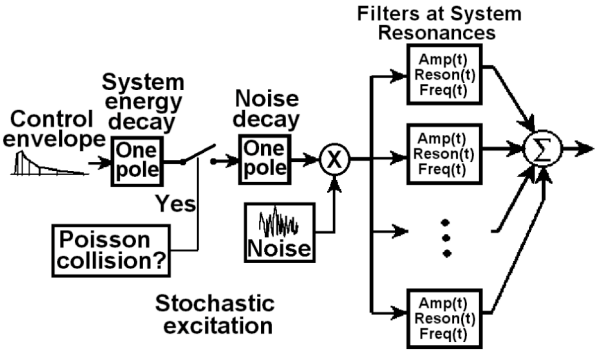

Figure 7: PhISEM synthesis.

4 PHISEM ANALYSIS

So far we have developed a simple synthesis model, but we still lack a complete technique for analysing sounds, specifically to determine N, the Poisson probability constant. The next two sections will outline and test a technique for estimating N.

4.1 Estimating N

N is estimated by inspecting a high-frequency band (5.5 – 11kHz) of the whitened footstep sounds. A Daubechies 4 wavelet filterbank [8] is used to split the signal into subbands, and these subbands are rectified. As can be seen by inspecting the top three rectified subband outputs in Figure 8, the 5.5-11kHz band seems to capture the rapid collision events best. The method can be used to analyze soundfiles of any sample rate, but 22.05kHz or 44.1kHz files are best because of the presence of the 5-11k band (best band for any sample rate above 22.05k.). For sample rates below 22k (of interest because of a growing number of lower bandwidth sound effects files available on the Web), the highest octave subband is used for forming the best estimate of N.

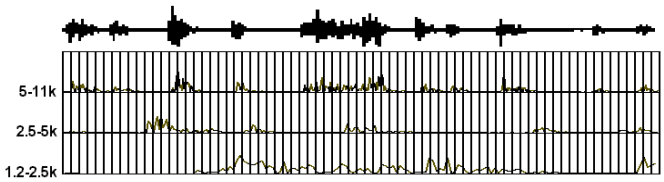

Figure 8: Rectified wavelet subbands.

Next a threshold value is computed from the maximum peak and average values of the rectified subband signal, and the number of peaks per second over the threshold are counted. This estimate of N, the density of collisions, tends to miss collisions as the probability N increases. To correct for this, the actual N estimate is calculated to be:

$$N_{est} = 2^{avgPeaks/blocks/10} \quad (1)$$

Where "blocks" means that the average number of peaks are only counted in blocks where significant peaks occur. Equation 1 was arrived at by trial and error, in a set of experiments involving extracting known N values from simulations, as described in the next subsection.

4.2 Verifying the System ID Techniques

Using the basic synthesis model shown in Figure 7, 1050 soundfiles were synthesized using a simple raised cosine excitation envelope. System parameters were:

α_c = constant at 0.95 (60 dB decay of 130 samples)

N = 2,4,8,16,64,256,1024

α_s = 0.95,0.99,0.995,0.999,0.9995,0.9999

r = 0.7,0.8,0.9,0.95,0.99

f = 1000,2000,3000,5000,8000

where r and f are the pole radius and center frequency of a single 2nd order resonant filter. The steps of envelope extraction, 2nd order Linear Prediction, rectified high-frequency 5.5-11kHz subband extraction, and N estimation were performed on all 1050 synthesized files. As Figure 9 shows, the average frequencies and resonances yielded by Linear prediction are more accurate for higher resonances, and for middle frequencies. Most of the frequency errors were experienced on files with very low resonance (pole = 0.7 or 0.8). In those cases, frequency mis-adjustment is not so important anyway.

Synth. N	Avg. Estimate	StdDev.
2	2.04	0.82
4	3.06	1.31
8	5.33	2.91
16	13.0	8.26
64	99.1	79.8
256	396	291
1024	699	532

Table 1: Particle density estimates for synthesized data.

Table 1 and Figure 10 show the results of calculating estimates of N (using Equation 1) for all 1050 files.

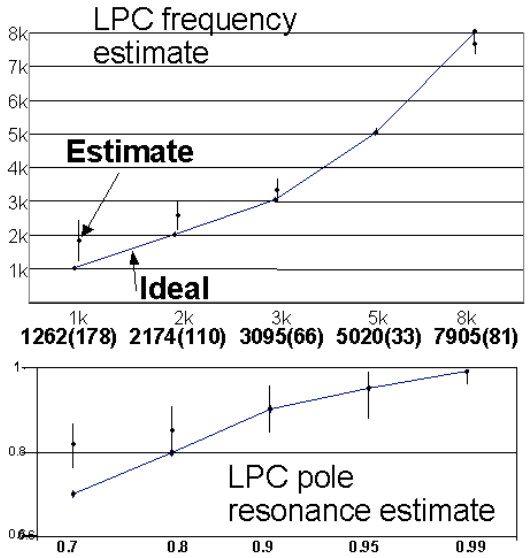

Figure 9: Resonance estimates for synthesized data.

Given that the synthesis is stochastic, the estimated averages for N are quite good. However, the relatively high standard deviations suggest that we should collect and average as many sound files as possible for any given walking condition. This means that the more individual footsteps we have on a recording, the better the overall estimate of particle density N for the material will be.

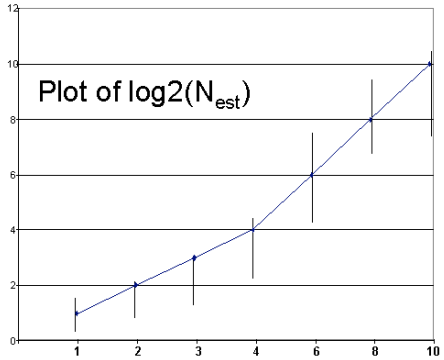

Figure 10: Particle density estimates, synthesized data.

4.3 "Real World" Analysis

To test the analysis techniques on non-synthesized data, two types of gravel were analysed. Large gravel rocks, averaging 1.5 cm diameter and 14 gm. weight, and small gravel rocks, averaging 0.3 cm diameter and 1 gm. weight were analysed. Ten shaking sounds of different gravel samples were analysed for each gravel type. For the large rocks, the center frequency and resonance was estimated by LPC to be 6460 +/- 701 Hz, and r = 0.932. N was estimated to be 23.22, with a standard deviation of 14.53. For the small rocks, the

center frequency and resonance were estimated to be 12670 +/- 3264 Hz. and 0.843, and N was estimated to be 1068 with a standard deviation of 755. Figure 11 shows the superimposed power spectra of the multiple shakes of the two gravels.

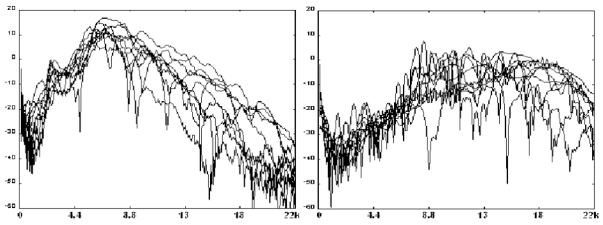

Figure 11: Large (left) and small gravel shake spectra.

5 CONTROLLING THE RESYNTHESIS

Once analysis of a walking sound is complete, the original soundfile can be exactly reassembled by concatenating the stored wave segments. Depending on the amount and nature of the background noise in the original recording, often a highly realistic semi-parametric resynthesis of arbitrary length and tempo can be made by randomly editing together the individual original footstep waves.

For truly flexible parametric resynthesis, the extracted envelope parameters can be used to generate new step envelopes, using the prototype average multi-breakpoint envelope, perturbing it by the standard deviation parameters. These new envelopes drive the PhISEM model (Figure 7) set to the analysed resonance and particle density, yielding infinite possible new syntheses. Any analyzed parameter can be interactively edited and changed as desired to modify the resynthesis. Figure 12 shows "Bill's GaitLab," a TCL/Tk Graphical User Interface for controlling synthesis/resynthesis in real time.

Much of the power of parametric synthesis comes from driving the parameters from controllers, or from other algorithmic processes. Examples might include automatically synthesizing footsteps for various materials and characters in a video game, based on the same physical processes that are calculated to animate the walking of those characters.

A specific example involves using a pressure sensitive floor surface in an immersive virtual environment to automatically provide walking sounds, programmatically tied to the environment being simulated and displayed. Figure 13 shows the Princeton "PhOLIEMat" (Physically Oriented Library of Interactive Effects). The base mat senses the location and pressure of each of a variety of moveable tiles. Each tile feeds envelope parameters in real time to the appropriate walking sound, responsive to walking pressure.

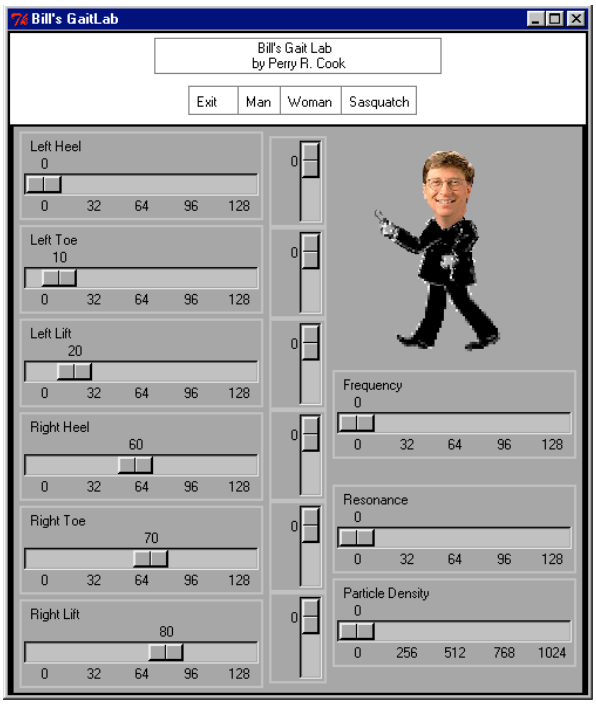

Figure 12 Bill's GaitLab

Figure 13 Princeton PhOLIEMat

6 CONCLUSIONS, FUTURE WORK AND A NOTE ON PERCEPTUAL REALISM

Work remains to be done in refining the envelope, resonance, and particle parameter extraction subsystems. The system fails on some sound files, but still provides a good starting point for hand-crafting the synthesis parameters. Clearly the largest amount of work remains to be done in evaluating the perceptual quality of the synthesized sounds.

The purpose of this paper has been to describe a system for analysis and synthesis of noisy, enveloped, periodic sounds such as those created by walking. Also this paper endeavoured to analyze (via signal processing and statistical measures) the techniques used for estimating the parameters. Validation of the "realness"

of the synthesized interactive environmental sounds lies in the realm of a huge psychoacoustic research agenda on the perception of "realistic sound" itself. In fact, there is not yet a proposed or accepted experimental paradigm for asking questions like "is this sound real?" or "is this sound more real than that one?" Further, perception of sound effects, like much of perception, is inherently a multi-modal phenomenon. For example, presenting the sound of breaking glass while displaying video of a dropped glass bowl breaking could improve the perceived quality of sound, or the sound could change the perception of the image. Of course, Foley artists know this instinctively and exploit the interaction between sound and image for emotion effects.

Interaction with the sound synthesis adds even more perceptual modes, most importantly the gestures and the responsive "feel" of the controllers. Thus a true understanding of the "sonic realness" of synthesized sounds needs a large amount of perceptual research to be designed and carried out.

There have been efforts in the past [9][10] and various new efforts are underway [11][12] [13], to evaluate the salience, dispensability vs. indispensability, and other aspects of perceptual and synthesis control features of environmental and interactive sounds. The author looks forward to performing more perceptual experiments in this area, and to integrating any new findings from sound perception research into future signal-processing and systems work.

REFERENCES

[1] P. R. Cook, "Toward Physically-Informed Parametric Synthesis of Sound Effects," Invited Keynote Address, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October, 1999.

[2] Casey, M., *Auditory Group Theory: with Applications to Statistical Basis Methods for Structured Audio*, Ph.D. Thesis, MIT Media Lab, February 1998.

[3] N. Miner, 1998, *Creating Wavelet-based Models for Real-time Synthesis of Perceptually Convincing Environmental Sounds*, Ph.D. Diss., University of New Mexico.

[4] C. Anderton, *Electronic Projects for Musicians*, Guitar Player Books, Saratoga, CA, 1978.

[5] X. F. Li, R. J. Logan, and R. E. Pastore, (1991). "Perception of acoustic source characteristics: Walking sounds." *Journal of the Acoustical Society of America,* (America Institute of Physics, Vol. 90, No. 6 1991) pp. 3036-3049.

[6] P. R. Cook, "Physically Informed Sonic Modeling (PhISM): Synthesis of Percussive Sounds," Computer Music Journal, 21:3, 1997.

[7] J. Markel and A. Gray, 1976, *Linear Prediction of Speech*, New York, Springer.

[8] I. Daubechies "Orthonormal Bases of Compactly Supported Wavelets" Communications on Pure and Applied Math. Vol.41 1988, pp. 909-996.

[9] Bregman, A., 1990, *Auditory Scene Analysis: The Perceptual Organization of sound*. Cambridge, Massachusetts: The MIT Press.

[10] Cipra, B. 1992. "You can't always hear the shape of a drum," *Science* 255(March 27):pp 1642-1643.

[11] "The Sounding Object Project," Multi-lab EU project as part of "The Disappearing Computer," http://www.soundobject.org.

[12] Lakatos, S., Cook, P. R., and Scavone, G. P. 2000, "Selective attention to the parameters of a physically informed sonic model," Acoustic Research Letters Online, Acoustical Society of America.

[13] Scavone, G., Lakatos, S., Cook, P. and C. Harbke, 2001, "Perceptual spaces for sound effects obtained with an interactive similarity rating program" International Symposium on Musical Acoustics, Perugia, Italy.