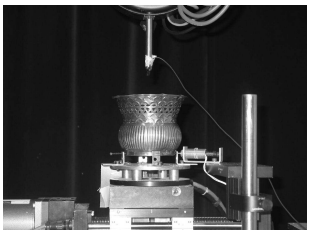
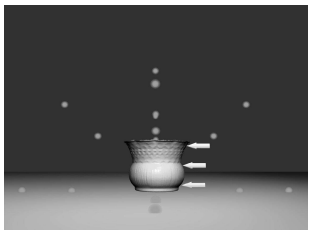
TimbreFields — 3D Interactive Sound Models for Real-Time Audio

Richard Corbett, Kees van den Doel, John E. Lloyd, Wolfgang Heidrich
The University of British Columbia

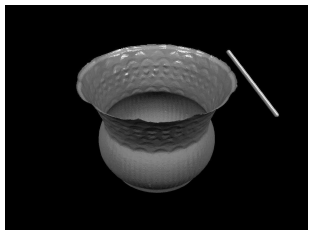
(a) Experience the real thing



(b) Measure it



(c) Model it



(d) Experience the virtual thing

Figure 1: (a) The timbre of the sound a physical object makes depends on where we touch it and where the object is. (b) To model this we measure the sounds around the object. (c) From the data we extract a unified interactive sound model. (d) The model is rendered in real-time in a simulator incorporating haptic, audio, and graphical display, recapturing the interactive experience of the original.

Abstract

We describe a methodology for Virtual Reality designers to capture and resynthesize the variations in sound made by objects when we interact with them through contact such as touch. The timbre of contact sounds can vary greatly, depending on both the listener's location relative to the object, and the interaction point on the object itself. We believe that an accurate rendering of this variation greatly enhances the feeling of immersion in a simulation. To do this we model the variation with an efficient algorithm based on modal synthesis. This model contains a vector field that is defined on the product space of contact locations and listening positions around the object. The modal data are sampled on this high dimensional space using an automated measuring platform. A parameter-fitting algorithm is presented that recovers the parameters from a large set of sound recordings around objects and creates a continuous timbre field by interpolation. The model is subsequently rendered in a real-time simulation with integrated haptic, graphic, and audio display. We describe our experience with an implementation of this system and an informal evaluation of the results.

1 Introduction

The human inspection of a physical object involves a simultaneous consideration of information through all five sensory channels. While investigating an unfamiliar object, a person finds it natural to hold it at different orientations to gather both haptic and visual data to classify the object. Simultaneously, audio information created by this haptic interaction is also considered, offering more information about the object's material and density. Often, a single picture, haptic texture, or tapping sound contains enough information for someone to roughly classify an object. However, an intimate virtual interaction with an object, which creates a feeling of presence, immersion, and engagement, requires much more.

An intimate interaction, which involves interaction through multiple senses, requires that the user be able to inspect the object from all angles and with varying haptic queries. To get a *real feeling* for something, however, a person will simultaneously consider the sounds generated from physical contact with the object. The perceived sounds will vary depending on how the object is excited by contact, and on the difference in position and orientation between the excitation point and the listener. We believe that a stronger sympathetic human response is possible with a model of the intimate sound differences an object will make as a response to basic human interaction.

A particular deficiency in current multi-modal systems is the inability to model the spatial dependency of the timbral audio information an object produces. There are approaches (reviewed below) that model the important localization cues of loudness variance and head related transfer function (HRTF) scattering. These approaches model the most apparent differences in the sound that can be heard when an object is moved with respect to the listener. This allows an observer to get a strong sense of where the sound is coming from. There are also techniques to compute sound sources as well as techniques to spatialize distant sources so that they appear to come from a particular direction, in a particular acoustic environment. All these methods fail to model the *near-field* acoustic properties of objects. It is not only the localization and spatialization cues but also the timbre of a sounding object that varies with changes in observation point. The sounds we hear when we tap or scrape an object, such as the one depicted in Figure 2, depend on our ear's location in the three-dimensional sound field around the object. The importance of these variations in sound for musical instruments is well-known. For example, Weinreich (Weinreich, 1996) has argued that "Directional Tone Color" plays an essential role in the perception of the sound of the solo violin. The premise of the work presented here is that modeling these subtle variations in timbre will also substantially enhance the presence and realism of virtual interactions with everyday objects, as depicted in Figure 2.

Our work aims to create a pipeline for measuring and simulating these effects. In particular, the novel components of our work are:

- Modeling the 3D changes in timbre of interaction sounds around objects.
- An automated system to acquire model data.
- A novel parameter-fitting technique to extract modal data from recorded sounds.
- The implementation of a multi-sensory simulation of a virtual object with integrated haptics, graphics, and spatially varying contact sounds.

Figure 2: When we interact with an object, the sounds we hear depend, sometimes greatly, on where we are listening from.

Overview

This paper presents *TimbreFields*, a novel interactive sound synthesis system that supports intimate user interaction. TimbreFields encompasses contact sounds that depend on the surface contact location, the mode of interaction, as well as the position of the listener in three-dimensional space. If the contact surface is two-dimensional, this gives us a total of five dimensions in which timbral variation occurs. Additional variations are caused by different interaction modes such as tapping, scraping, or sliding, but they are simply different excitations of a modal soundmodel which is five dimensional[1]

With this system, real-time audio can be synthesized for different types of user interaction, such as scraping and tapping, for varying locations on an object as well as varying observer positions. The audio model contains a "timbre field" defined on the product space of the object's two-dimensional surface and the three-dimensional listening space around the object. This model allows for the reconstruction of the timbre of the contact sounds, depending on the observer's location as well as on the contact location.[2] We extend the real-time modal synthesis methods described by Doel et al. (Doel, Kry, & Pai, 2001) to allow for the modeling and synthesis of timbre variations in this extended five-dimensional space. We describe how to extract the parameters of the model from the measurements of sounds around the object and how to interpolate the model parameters to create a continuous timbre space.

[1]A consistent definition of "timbre" does not exist. We are referring here to qualities of the sound that are not pitch and loudness. Below we define the timbre of a contact sound precisely within the context of modal synthesis as the relative magnitudes of the modal frequencies.

[2]Cook (Cook, 2002) argues that the angle at which an object is struck could also affect the timbre. Such an extension is easily accommodated into our method.

The remainder of this paper is organized as follows. Section 2 describes the audio model used to model the timbral variations in real-time. Section 3 describes how we acquire the data for our models from measurements and parameter fitting. Section 4 describes the implementation of a system designed to demonstrate the effectiveness of our ideas and, finally, we present our conclusions in Section 5.

Related Work

The importance of sound in the interaction with virtual objects is well known. Gaver (Gaver, 1988, 1993) pioneered the use of synthetic contact sounds to accompany direct human interactions with a computer. Takala and Hahn (Takala & Hahn, 1992) first constructed a general framework for producing sound effects synchronized with animation. In their framework, sounds are attached to graphical objects, and events can trigger visual, as well as, sonic events. Hahn et al. (Hahn, Fouad, Gritz, & Lee, 1995) introduced a number of synthesis algorithms for contact sounds. The musically motivated real-time synthesis techniques pioneered by Cook (Cook, 1995, 1996) can also be used to create sound effects for contact interactions with virtual objects.

Recently, major progress has been made in the integration of animation, sound, and graphics. O'Brien (O'Brien, Cook, & Essl, 2001) describes an off-line system to compute both sound and motion from a single physical model of deformable bodies. The FoleyAutomatic system (Doel et al., 2001) uses modal resonance models to create real-time sound effects to accompany interactions with virtual (simulated) objects. The model data is acquired systematically through measurements as described by Pai et al. (Pai et al., 2001). O'Brien et al. (O'Brien, Chen, & Gatchalian, 2002) present methods to compute, offline, and from first principles, the modal data for real-time modal models. This approach could be considered complementary to the one we present, allowing us to compute the TimbreField parameters from geometric and material models instead of measuring the TimbreField responses. Physically based models for the excitations during collision and sliding contacts have also been investigated (Avanzini & Rocchesso, 2001; Avanzini, Serafin, & Rocchesso, 2002), as have physical models for aerodynamic sounds (Dobashi, Yamamoto, & Nishita, 2003, 2004).

The placement of sound sources in acoustic environments to create the illusion of directional sound and reverberation has been investigated widely. Several examples are available, (Begault, 1994; Tsingos, Funkhouser, Ngan, & Carlbom, 2001; Funkhouser, Min, & Carlbom, 1999; Savioja, Huopaniemi, Lokki, & Vnnen, 1997; Tsingos, Gallo, & Drettakis, 2004). Note, however, that these methods deal exclusively with the far-field sound. That is, all sound sources are assumed to be distant enough to be treated as point sources, though extended sound sources can be modeled with a large number of point sources using the techniques described by Tsingos et al. (Tsingos et al., 2004). Cook's NBody system (Cook & Trueman, 1998) is the only work on capturing the directional characteristics of real-time synthesized sounds that we are aware of. Impulse responses were measured around stringed musical instruments and a multi-speaker system was then used to resynthesize the directional sound field.

Our method is also related to other works in computer graphics that strive to capture realistic models from real-world objects. In many cases, the measurement of *impulse-response* has proven particularly useful. For example, when capturing the visual appearance of a model, one can ac-

quire the light field for a given illumination (e.g., (Levoy & Hanrahan, 1996; Gortler, Grzeszczuk, Szelinski, & Cohen, 1996; Wood et al., 2000)). However, this limits rendering to the exact same lighting situation as during capturing. Alternatively, one can measure the bi-directional reflectance distribution function (BRDF, e.g., (Marschner, Westin, Lafortune, & D. Greenberg, 1999; Yu, Debevec, Malik, & Hawkins, 1999)), which models the impulse response of reflectance characteristics for a material, i.e., the distribution of reflected light under a thin pencil of incident light. This, then, allows the use of the captured data in arbitrary illumination environments by convolving the impulse response over the actual illumination. Similarly, deformation properties of objects have been modeled as impulse-response, which allows for simulation of actual deformations by convolving with the actual contacts taking place between objects (James & Pai, 1999). Our work is very much in the same spirit, in that we capture the impulse response of contact sounds obtained by very brief excitations with a small tip. Rather than just playing back the recorded sounds, we can then simulate arbitrary contacts such as scraping or knocking by convolving the measured model with the actual area and duration of contact. (Doel et al., 2001)

2 Interactive Spatial Audio Models

We extend the synthesis model described by Doel et al. (Doel et al., 2001) to incorporate timbral variation in contact sound depending on the location of the observer as well as that of the interaction point. A vibrating object is modeled by a bank of damped harmonic oscillators. These oscillators are excited by an appropriate external stimulus derived from the contact interaction. By varying the stimulus in real-time, we can create a great variety of sound effects.

For interactions at a fixed point and a fixed observer, a modal model $\mathcal{M} = \{\boldsymbol{f}, \boldsymbol{d}, \boldsymbol{a}\}$ consists of three vectors of length N, where N is the number of modes modeled. The modal frequencies are $\boldsymbol{f}$, their decay rates are $\boldsymbol{d}$, and their gains (amplitudes) are $\boldsymbol{a}$. The impulse response $y(t)$ is given by

$$y(t) = \sum_{n=1}^{N} a_n e^{-d_n t} \sin(2\pi f_n t). \tag{1}$$

The frequencies and dampings of the oscillators are intrinsic properties of the object. They remain constant, and are independent of where the object is excited or where the listening point is. The gains $\boldsymbol{a}(\boldsymbol{w})$, however, depend on the contact point as well as on the listener's location. These gains determine the timbre of the resulting sound. They can be represented as a vector field on the five-dimensional product space whose points are denoted by

$$\boldsymbol{w} = (u, v, x, y, z), \tag{2}$$

with (u, v) the excitation coordinates on the object surface and (x, y, z) the (not necessarily Cartesian) coordinates of the listener position in space. The sounds of a modal model can be interactively rendered in real-time by convolving the impulse response with an excitation (Doel et al., 2001). By varying $\boldsymbol{w}$ we can create the correct timbres for the various interaction points and listener locations.

In reality, when we are listening to sounds of nearby objects, we are in a great majority of cases, listening with both ears. The scattering of sound around the head, upper torso, and, most importantly, the pinna, causes differences in the sounds entering the two ear canals, giving us critical information about the spatial location of objects.

Synthesis techniques such as presented in this paper are concerned only with computing the pressure fluctuations at particular observer points. When rendered in complete simulation environments the binaural sounds can then be computed by a subsequent stage in the audio processing pipeline. The simulation of these effects for distant point sources has received considerable attention recently (see the review in Section 1).

To apply similar "spatialization" techniques to the audio synthesis models described here, it is however necessary to also model the variation in timbre caused by the difference in location of the two ears in the sound field. This can be done by synthesizing the sound field at a number of locations simultaneously, and processing these point sources with an appropriate near-field scattering model to compute a binaural stereo signal. As a crude approximation, one could simply compute the convolution at the locations of the two ears and use an appropriate near-field HRTF (see for example (Begault, 1994)), to further process the sounds.

This can be done with very little extra computational cost within the modal synthesis model. The convolution is implemented most efficiently by modeling a bank of damped oscillators with frequencies and dampings determined by the modal model. The contact force is applied as an input force to this mechanical system, which is computed using a forward Euler ODE discretization with time step $1/S$ where $S = 44100Hz$ is the sampling rate. The audio signal is then computed at a particular location in the 5D timbre space by adding the oscillators with weight factors a_k. To compute the sound in parallel at different locations in the timbre space, we just sum the same oscillators with the coefficients from the $\boldsymbol{a}$ field corresponding to each point, at the cost of N extra multiplications per location. Experimentally we found that computing two sources takes only 2% more computing resources than computing only one.

We note that the field $\boldsymbol{a}(\boldsymbol{w})$ can be computed from first principles using techniques as described by O'Brien et al. (O'Brien et al., 2002). Instead, we advocate acquiring the field from the measurements of real sounds around a real object, as described in Section 3. Capturing these real-world sounds allows for the immediate capture of accurate data that is not dependent on approximated geometry, or material information. In addition, it is extremely difficult to get precise results using first-principles modeling, particularly for higher order effects such as we consider here. Nevertheless, this would be an interesting, complementary, approach.

After we obtain the field $\boldsymbol{a}$ on a grid in our five-dimensional timbre space, we interpolate linearly at runtime to get continuously varying timbral variations. For a given point $\boldsymbol{w}$ in five-dimensions we obtain the interpolated values $\boldsymbol{a}(\boldsymbol{w})$ by linearly interpolating the nearest grid points in each dimension separately. The resulting quinti-linear interpolation is a tensor product of the tri-linear interpolation for the listener position in space and the bi-linear interpolation of the excitation point in the object parameter space. It should be noted that the coordinates and grid arrangement for the listener position do not have to be Cartesian; in particular, spherical coordinates were used for the results described in Section 4.

3 TimbreField Capturing and Creation

To construct a modal model we recorded a large set of sounds emitted by an object when excited by an impulsive force

provided by a solenoid-activated piston.[3] We analyzed these sounds as representing the impulse responses of the object as specified by Eq. 1.

Sounds were captured by controlling both the excitation point on the object and the microphone location. The excitation point was changed by moving the solenoid manually to a predetermined set of discrete locations on the object's surface. For each impact location, sound samples were captured at discrete locations in the space around the object. The microphone location was controlled at the end of a robotic arm, allowing for precise positioning of the microphone. For objects with a complex and rapidly varying timbre field, a relatively dense set of impulse responses would be required. Our measurement setup is depicted in Figure 3.

Impulse responses were recorded digitally as 16 bit audio at a sampling rate of 44.1Khz. The microphone was mounted on a programmable robotic arm, which automatically moved the microphone to a variety of listening locations for a given contact location. The solenoid was synchronized with the robotic arm to expedite the recording process. All the microphone and excitation control programming was done in advance. The strategy was to gather all the impulse responses within a robotic environment so as to minimize errors introduced through human control.

Before each impulse response recording, 0.5 seconds of ambient sound was recorded to help identify spurious modes in the background noise. This precaution was necessary because the measuring environment contained a lot of acoustic noise from machine fans and power motor control equipment.

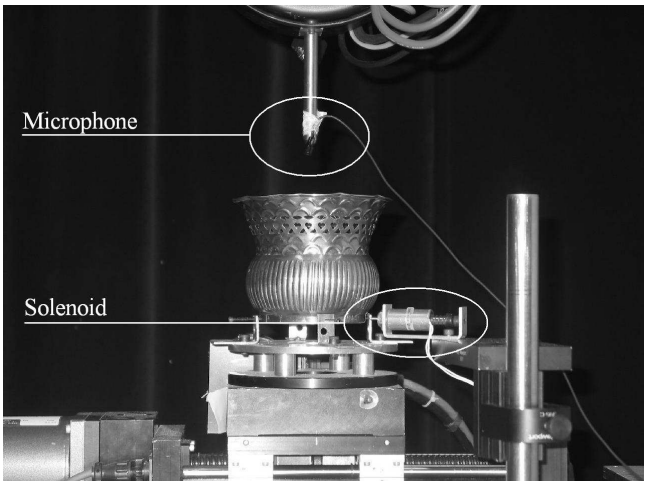


Figure 3: An automated measuring system was used to acquire the data. A robot arm moves the microphone to a preprogrammed set of locations. An impulse force was then automatically applied to the object with the solenoid and sounds were recorded for subsequent analysis.

This process produced a set of acoustic response samples for different contact and listener locations distributed across the 5D timbre space. The modal model for the entire object was then constructed in several phases.

[3]A solenoid is a coil of wire around an iron core that becomes a magnet when current passes through its coil. In this case the solenoid is used to control a piston that taps the object.

Estimation of the Modes

The estimation of the modal model parameters from the measurements was achieved in the following three phases.

First, we estimated the modal model parameters for each sound sample separately. This is done by computing the windowed discrete Fourier transform (Gabor transform) of the signal - extracting the frequencies, dampings, and amplitudes by fitting each frequency bin with a sum of a small number (4 for example) of damped complex exponentials. The parameter fitting method is capable of very accurate frequency reconstructions and is able to resolve very close modes. Close modes are very common in artificial objects that have approximate symmetries resulting in mode degeneracy. Manufacturing impurities break this symmetry, splitting the frequencies by a small amount. These nearby frequencies are distinctly audible as beating, or "shimmering" sounds and significantly enhance the realism of the synthesized sound.

Second, we determined which modes are actually audible, using a rough model of auditory masking (Doel, Pai, Adam, Kortchmar, & Pichora-Fuller, 2002; Doel, Knott, & Pai, 2004), and retained only the audible modes. This step is necessary because the parameter fitting algorithm produces many spurious modes.

Third, we merged all the modal models, using a simple model of human frequency discrimination, which results in a single frequency and damping model for the entire object, and a discrete sampling of the timbre field $\boldsymbol{a}$ on the 5D interaction space. In theory all the models should share the same set of frequencies and dampings, but due to noise they will not be precisely the same, motivating this third step.

We now describe the details of the parameter fitting.

To estimate the modal content of a single recorded impulse response $s(t)$, we first compute the windowed Fourier transform. We use a Blackman-Harris window of length T_w, and $N_{overlap} = 4$ windows are taken to be overlapping, giving a "hopsize" T_H of $T_H = T_w/N_{overlap}$. (The hopsize is the amount by which consecutive windows are apart.) The window size T_w is chosen to be $46ms$, which is appropriate for audio analysis. Let us denote the windowing function by $w(t)$, with support $[0, T_w]$. The windowed Fourier transform is given by

$$\tilde{s}(t, F_k) = \int_{-\infty}^{\infty} e^{-2\pi i F_k \tau} w(\tau - t) s(\tau) d\tau, \qquad (3)$$

where the discrete frequencies take on the values $F_k = k/T_w$, with $k = 0, 1, \ldots$. In practice, Eq. 3 is computed at sampling rate S on a finite set of overlapping windows, defined by the support of the windowing functions $w(\tau) = w(\tau - t_j)$, with j labeling the windows. We get $t_j = jT_w/N_{overlap}$, with $j = 0, \ldots, N_w$, and N_w the largest value generating a window within the domain of the signal. The discrete frequencies F_k are now limited to the Nyquist rate $S/2$, so $k = 0, \ldots, N_F - 1$, with $N_F = ST_w$ the number of samples in a window, which is a power of 2.

From the recorded audio signal $s(t)$ we compute the $N_w \times N_F$ complex matrix of Windowed Discrete Fourier Transform (WDFT) $\tilde{s}(j, k)$ as

$$\tilde{s}(j, k) = \sum_{n=0}^{N_F - 1} e^{-2\pi i F_k \tau_n} w(\tau_n - t_j) s(\tau_n)/S, \qquad (4)$$

with $\tau_n = n/S$, and $t_j = jT_H$.

We cannot simply estimate the modes from the power spectrum (the norm of the WDFT averaged over all windows), because frequencies closer together than $2\Delta F$, with

$$\Delta F = 1/T_w, \tag{5}$$

cannot be resolved in this manner. (A moderate improvement can be obtained by higher order interpolation (Sullivan, 1997).)

To obtain higher accuracy and to resolve nearby frequencies, we fit the complex frequency trajectories, obtained by viewing the WDTF for a fixed bin as a complex function of time, with the sum of a small number of damped exponentials using the Steiglitz-McBride algorithm (Steiglitz & McBride, 1965; Brown, 1996). We do this for every bin. This will provide us with the estimated gains $\boldsymbol{a}$, the dampings $\boldsymbol{d}$, and the frequencies $\boldsymbol{f}$. This "phase unwrapping" procedure has been used before to obtain a very accurate single frequency estimate (Brown & Puckette, 1993). Our method differs in that we also estimate the dampings and gains, and are able to reconstruct very closely spaced frequencies.

The WDFT of $y_M(t)$ as defined in Eq. 1 can be written as

$$\begin{aligned}\tilde{y}_M(t, F_k) &= \int_{-\infty}^{\infty} e^{-2\pi i F_k \tau} w(\tau - t) y_M(\tau) d\tau = \\ &\sum_{m=1}^{M} [R(f_m) + R(-f_m)],\end{aligned} \tag{6}$$

where

$$R(f_m) = a_m e^{[2\pi i (f_m - F_k) - d_m]t} B_{mk},$$

with

$$B_{mk} = \int_0^{T_w} w(\tau) e^{[2\pi i (f_m - F_k) - d_m]\tau} d\tau.$$

The key observation is that B_{mk} peaks strongly when $|f_m - F_k|$ is small, i.e., when a bin frequency F_k is close to a mode f_m. If we consider Eq. 6 for a fixed value of k where we expect a mode, we can reorder the sum in Eq. 6 as

$$\tilde{y}_M(t, F_k) = \sum_{\text{m s.t. } \mathrm{f_m} \approx \mathrm{f_k}} R(F_m) + \text{distant modes}.$$

Assuming the "distant modes" can be ignored, we conclude that the measured WDFT given in Eq.4, viewed as a function of discrete time t_m, has the approximate form

$$\tilde{s}(t_j, k) = \sum_{m=1}^{N_d} C_m e^{[2\pi i \cdot g_m - d_m]t_j}, \tag{7}$$

with C_m, g_m, and d_m to be fitted to the data. The number of terms in the expansion N_d is chosen depending on the expected density of the modes. We usually set this to a small number (2 or 4) in order to capture the effect of beating modes. We can regard $\tilde{s}(t_j, k)$ as a function of discrete time $t_j = jT_H$, i.e., at effective sampling rate of $1/T_H$, and fit it to the measured data using the Steiglitz-McBride algorithm (Steiglitz & McBride, 1965), fitting with $N_d > 1$ terms. We then recover the modal gains from

$$a_m = C_m / B_{mk}$$

and the frequencies from

$$f_m = F_k + g_m.$$

Because of the presence of noise, and because we tried to find modes in every frequency bin, most of the modes found will have very small gains and are inaudible. We remove the inaudible modes by using the perception-inspired pruning method described by Doel et al. (Doel et al., 2002, 2004).

Mode Merging

We now have a separate modal model (with one gain vector) per measurement. We could modify the model and allow the frequencies and dampings to change depending on location, but this approach would make the synthesis extremely inefficient. In addition, the different frequencies are really artifacts of the parameter fitting and therefore should be removed from the model.

We accomplish the mode merging by identifying "close" frequencies in the modal models and, when they originate from separately measured impulse responses, combine them into a single frequency by averaging them.

Two frequencies, reconstructed from separate impulse responses, which are Δf apart are merged if the frequency difference Δf is judged to be inaudible. Frequency discrimination is a well studied subject in audiology; see for example (Moore, 1986). Experimental data (Shower & Biddulph, 1931) has been gathered on the threshold of audibility of frequency modulation by Δf at a rate of $2 - 4$Hz as a function of the frequency. We adapted this and made a crude fit to experimental data from this study as follows:

$$\Delta f = 4$$

for frequencies f below 2000Hz,

$$\Delta f = (f - 2000)/250 + 4$$

for frequencies f above 2000Hz.

In addition to the frequencies, we also have to merge the dampings of the close modes. Theoretically they should be close if the frequencies are close but due to noise this is often not so. If we simply took the average of the dampings, the resulting model would be too damped. We found the over-damping to be caused by modes that are weak at certain locations, which sometimes results in an overestimated value for the damping d. This is because these modes rapidly disappear in the background noise, leading to an overestimation of the decay rate. We solve this problem by selecting the mode with the maximum gain a and by taking the merged d to be its value. This is reasonable, as the mode with maximum a will presumably have the most accurate fit, since it has the highest signal-to-noise ratio.

4 Results

We verified the effectiveness of the TimbreFields system by modeling the contact sounds of a metal vase (depicted in Figure 3). We then used this model to provide acoustic rendering in a real-time simulation integrated with haptics and graphics.

4.1 Measurement and Modeling

To create a TimbreField for the vase, we decided to exploit its axial symmetry and consider only the height h of the surface excitation point, hence reducing the TimbreField dimension by one. For the observer location, we used a spherical coordinate system (r, θ, ϕ) centered on the bottom of the

vase, and so the coordinate space (Eq. 2) for our TimbreField reduced to

$$\boldsymbol{w} = (h, r, \theta, \phi).$$

A discrete grid was defined on this space (Figure 4) by three excitation heights h (top, center, and bottom), two radii values r, four azimuthal values θ (0, $\pi/2$, π, $-\pi/2$), and three elevations ϕ (0, $\pi/2$, π). Also, one observer point was located inside the vase (giving an additional value of r for $\phi = 0$). Since points at $\phi = 0$ are invariant with respect to θ, this defines a total of 57 grid points in all.

The robotic capture environment was used to measure sound samples at each of these grid points. Because of symmetry, points at $\theta = \pm\pi/2$ were considered to be identical, and so measurements were only performed for θ values of 0, $\pi/2$, and π. To ensure control of the geometry during the sound capture, the vase was mounted on a rotary stage. Care was taken to mount the vase in a way that would minimize the damping caused by the fixtures. (It is noted, however, that any fixture arrangement will cause deviation in the acoustic response of the object being held in place.)

The process of capturing the sounds was to first set the solenoid position on the outside of the vase. For such a solenoid position the mechanical armature carrying the microphone would be moved to all desired listening positions where sound samples were recorded. Once responses at all listening positions had been captured for a solenoid position, the solenoid was moved to its next location, and the recording process repeated.

Once all impulse responses were recorded, the model parameters of the TimbreField were calculated using the algorithm presented in Section 3. It followed that the original 2692 modes aggregated from all 57 impulse responses were merged into a TimbreField that described only 386 unique frequencies. In this case, the final number of frequencies is very manageable in a real time system.

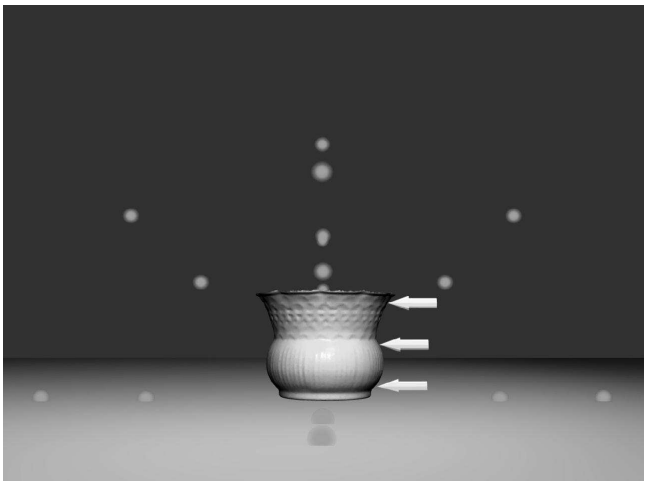


Figure 4: A discrete TimbreField grid. The white arrows show excitation locations. For any point of excitation the impulse response is recorded at all observer points (shown by the gray markers).

The resulting timbral variations can be visualized to a certain degree by mapping tonal brightness onto visual contrast. We use a tonal brightness measure that is calculated from a weighted average of the frequencies at any point. This measure is used just for simple visualization purposes. Naturally, the corresponding amplitude vector is used for weighting. In Figure 5 we show two slices of the TimbreField of the vase. The points shown are calculated via interpolation from the original data configuration shown in Figure 4. For both images, the point of excitation on the vase was near the right hand side of the base.

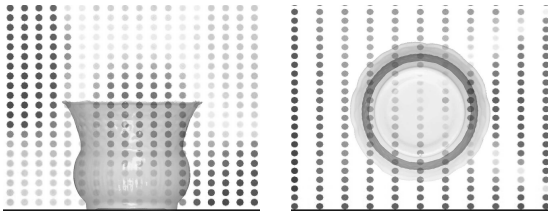


Figure 5: The variations in timbre around the object are mapped to grayscale (black is brightest, white is dullest) to visualize the spatial variations of sounds. Both images show the TimbreField produced by hitting the vase at the bottom right. The picture to the left depicts the brightness of the sound on a plane through the center of the vase. The picture to the right depicts the variations in a cross-sectional plane at the bottom.

4.2 Simulation

We have used TimbreFields to effect acoustic rendering in a multi-sensory virtual environment with a Sensable Technologies Desktop Phantom haptic device and a graphical display (Figure 6). The Phantom controls a simulated probe with which the user may touch the virtual object and create contact sounds representative of both tapping and scraping.

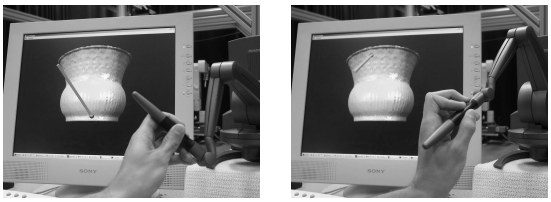


Figure 6: The vase rendered in our multi-sensory display. The user is interacting with it through a Phantom force feedback device that controls a virtual probe. Contact and collision detection feeds data to the audio and haptic simulator. The sound depends on the locations of both the contact point and viewpoint.

Contact simulation between the object and the probe is implemented using a dual-loop approach. The collision detection and contact forces are calculated at 100 Hz, and the contact forces are then being used to drive the Phantom, which is connected to the probe by a virtual 6 DOF spring/damper implemented at 1 KHz. The contact object is represented using a triangular mesh, collision detection is done using an OBBTree (Gottschalk, Lin, & Manocha, 1996), and the contact forces are computed using an optimized version of Lemke's algorithm (Lloyd, 2005). Except for the 1 KHz control loop, all software is implemented in Java, and the system runs on a 1.1 GHz AMD Athlon processor.

The 100 Hz loop computes and reports contact point between the probe and the object, along with the associated forces and tangential velocities. This provides contact location coordinates for the TimbreField. Observer locations

are determined by transforming the eye viewpoint coordinates into object coordinates. As with any immersive display technology, effective realism requires careful matching of the viewing frustum parameters with the actual location of the display screen and the user's eyes. Otherwise, perceptual anomalies may occur, such as the vase sounding "close to the ear" when in fact it appears far away on the screen.

For a particular contact point and observer location, an amplitude vector for the object's modal model is determined by linear interpolation with the nearest grid points in the TimbreField, as described in Section 2. This set of modal amplitudes, along with the contact force and velocity, is then used to create the contact sound via the JASS API for real-time acoustic rendering (Doel & Pai, 2001).

This simulation provided strong results showing how the timbre can change, depending on both the viewpoint and the contact point. Using JASS it was relatively easy to incorporate realistic interactive sounds like scraping and tapping in real-time.

There are, of course, combinations of contact point and viewpoint that don't generate strongly differing timbres. With a method to predict where the most interesting changes will occur, one could save time in sound capturing and produce an equally impressive result for the average ear. The most drastic and impressive variations in timbre which we found in our vase model were the differences observed when going in or out of the vase as well as when the contact point was alternated between the front and the back.

5 Conclusions

We have described a method to model the near-field timbral sound variation that occurs when interacting with physical objects through contact such as touch. Our model extends modal synthesis to allow for the spatial variations of timbre in the 3D space around the object, as well as variation due to the changes in contact point and contact type.

The model allows for efficient real-time audio synthesis in multimodal interactive simulations. We describe how we acquire the model parameters from measurements using a novel parameter fitting algorithm, in conjunction with an automated robotic remote controlled measuring system which allows the accurate recovery of the modal frequencies, dampings, and gains. A detailed model of a metal vase was constructed and rendered in a multi-sensory simulation with integrated haptics, graphics, and audio.

Our model can be easily extended to feed audio data to a "spatialization" stage in a complete audio pipeline, because the sounds at different locations can be computed simultaneously with very little extra cost. The integration of *TimbreFields* into a full audio pipeline that would consider spatialization effects such as HRTF scattering is the next step in our system's verification. We expect that such an integration would provide a very good approximation to the real world audio experienced when interacting with an object. Furthermore, as noted in Section 2, both the interpolation and the synthesis could be extended to produce binaural output with little added computational cost.

An investigation into how the timbral quality of an object's sound can vary with angle of excitation as reported in Cook (Cook, 2002) could be done. Again, the extension of our system to incorporate this extra dimension would be minimal.

References

Avanzini, F., & Rocchesso, D. (2001). Modeling Collision Sounds: Non-linear Contact Force. In *Proc. COST-G6 Conf. Digital Audio Effects (DAFx-01)* (pp. 61–66). Limerick, Ireland.

Avanzini, F., Serafin, S., & Rocchesso, D. (2002). Modeling Interactions Between Rubbed Dry Surfaces Using an Elasto-Plastic Friction Model. In *Proc. COST-G6 Conf. Digital Audio Effects (DAFx-02)* (pp. 111–116). Hamburg, Germany.

Begault, D. R. (1994). *3-d sound for virtual reality and multimedia*. London: Academic Press.

Brown, J. C. (1996). Frequency ratios of spectral components of musical sounds. *J. Acoust. Soc. Am.*, *99*(2), 1210–1218.

Brown, J. C., & Puckette, M. S. (1993). A high resolution fundamental frequency determination based on phase changes of the fourier transform. *J. Acoust. Soc. Am.*, *94*(2), 662–667.

Cook, P. R. (1995). Integration of physical modeling for synthesis and animation. In *Proceedings of the international computer music conference* (pp. 525–528). Banff.

Cook, P. R. (1996). Physically informed sonic modeling (PhISM): Percussive synthesis. In *Proceedings of the international computer music conference* (pp. 228–231). Hong Kong.

Cook, P. R. (2002). *Real Sound Synthesis for Interactive Applications*. Natick, MA: A. K. Peters, ltd.

Cook, P. R., & Trueman, D. (1998). NBody: Interactive Multidirectional Musical Instrument Body Radiation Simulations, and a Database of Measured Impulse Responses. In *Proceedings of the International Computer Music Conference*. San Francisco.

Dobashi, Y., Yamamoto, T., & Nishita. (2003). Real-time Rendering of Aerodynamic Sound Using Sound Textures based on Computational Fluid Dynamics. In *Computer Graphics (ACM SIGGRAPH 03 Conference Proceedings)* (pp. 732–740). Los Angeles.

Dobashi, Y., Yamamoto, T., & Nishita. (2004). Synthesizing Sound from Turbulent Field using Sound Textures for Interactive Fluid Simulation. *Eurographics 2004*, *23*(3), 539–546.

Doel, K. v. d., Knott, D., & Pai, D. K. (2004). Interactive Simulation of Complex Audio-Visual Scenes. *Presence*, *13*(1).

Doel, K. v. d., Kry, P. G., & Pai, D. K. (2001). FoleyAutomatic: Physically-based Sound Effects for Interactive Simulation and Animation. In *Computer graphics (acm siggraph 01 conference proceedings)* (pp. 537–544). Los Angeles.

Doel, K. v. d., & Pai, D. K. (2001). JASS: A Java Audio Synthesis System for Programmers. In *Proceedings of the International Conference on Auditory Display 2001*. Helsinki, Finland.

Doel, K. v. d., Pai, D. K., Adam, T., Kortchmar, L., & Pichora-Fuller, K. (2002). Measurements of Perceptual Quality of Contact Sound Models. In *Proceedings of the International Conference on Auditory Display 2002*. Kyoto, Japan.

Funkhouser, T. A., Min, P., & Carlbom, I. (1999). Real-time acoustic modeling for distributed virtual environments. *Proc. SIGGRAPH 99, ACM Computer Graphics*.

Gaver, W. W. (1988). *Everyday listening and auditory icons*.

Unpublished doctoral dissertation, University of California in San Diego.

Gaver, W. W. (1993). Synthesizing Auditory Icons. In *Proceedings of the ACM INTERCHI 1993* (pp. 228–235).

Gortler, S. J., Grzeszczuk, R., Szelinski, R., & Cohen, M. F. (1996, August). The Lumigraph. In *Computer graphics (siggraph '96 proceedings)* (pp. 43–54).

Gottschalk, S., Lin, M. C., & Manocha, D. (1996, August). Obbtree: A hierarchical structure for rapid interference detection. In *Siggraph '96 conference proceedings* (pp. 171–180). New Orleans.

Hahn, J. K., Fouad, H., Gritz, L., & Lee, J. W. (1995). Integrating sounds and motions in virtual environments. In *Sound for Animation and Virtual Reality, SIGGRAPH 95 Course 10 Notes.* Los Angeles.

James, D. L., & Pai, D. K. (1999). ArtDefo: accurate real time deformable objects. In *Computer graphics (siggraph '99 proceedings)* (pp. 65–72).

Levoy, M., & Hanrahan, P. (1996, August). Light field rendering. In *Computer graphics (siggraph '96 proceedings)* (pp. 31–42).

Lloyd, J. E. (2005, April). Fast implementation of Lemke's algorithm for rigid body contact simulation. In *Proceedings of the 2005 ieee international conference on robotics and automation* (pp. 4549–4554). Barcelona.

Marschner, S., Westin, S., Lafortune, E., & D. Greenberg, K. T. an d. (1999). Image-based BRDF Measurement Including Human Skin. In *10th eurographics workshop on rendering* (p. 131-144).

Moore, B. C. J. (1986). *An introduction to the psychology of hearing.* London: Academic Press.

O'Brien, J. F., Chen, C., & Gatchalian, C. M. (2002). Synthesizing Sounds from Rigid-Body Simulations. In *Siggraph 02.*

O'Brien, J. F., Cook, P. R., & Essl, G. (2001). Synthesizing Sounds from Physically Based Motion. In *Siggraph 01* (p. 529-536). Los Angeles.

Pai, D. K., Doel, K. v. d., James, D. L., Lang, J., Lloyd, J. E., Richmond, J. L., et al. (2001). Scanning physical interaction behavior of 3D objects. In *Computer Graphics (ACM SIGGRAPH 01 Conference Proceedings)* (pp. 87–96). Los Angeles.

Savioja, L., Huopaniemi, J., Lokki, T., & Vnnen, R. (1997). Virtual environment simulation - Advances in the DIVA project. In *Proc. int. conf. auditory display.* Palo Alto, USA.

Shower, E. G., & Biddulph, R. (1931). Differential pitch sensitivity of the human ear. *J. Acoust. Soc. Am.*, *2*, 275–287.

Steiglitz, K., & McBride, L. (1965). A technique for the identification of linear system. *IEEE Trans. Automatic Control*, *AC-10*, 461–464.

Sullivan, D. L. (1997). Accurate frequency tracking of timpani spectral lines. *J. Acoust. Soc. Am.*, *101*(1), 530–538.

Takala, T., & Hahn, J. (1992). Sound rendering. *Proc. SIGGRAPH 92, ACM Computer Graphics*, *26*(2), 211–220.

Tsingos, N., Funkhouser, T., Ngan, A., & Carlbom, I. (2001). Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Siggraph 01.*

Tsingos, N., Gallo, E., & Drettakis, G. (2004). Perceptual Audio Rendering of Complex Virtual Environments. In *Computer Graphics (ACM SIGGRAPH 04 Conference Proceedings).* Los Angeles.

Weinreich, G. (1996). Directional tone color. *J. Acoust. Soc. Am.*, *101*(4), 2338–2347.

Wood, D., Azuma, D., Aldinger, K., Curless, B., Duchamp, T., Salesin, D., et al. (2000). Surface Light Fields for 3D Photography. In *Computer graphics (siggraph '2000 proceedings)* (pp. 287–296).

Yu, Y., Debevec, P., Malik, J., & Hawkins, T. (1999). Inverse Global Illumination: Recovering Reflectance Models of Real Scenes From Photographs. In *Computer graphics (siggraph '99 proceedings)* (pp. 215–224).