Analysis / Synthesis of Sounds Generated by Sustained Contact between Rigid Objects

Mathieu Lagrange, Gary Scavone, Philippe Depalle

***Abstract*— This paper introduces an analysis/synthesis scheme that aims at reproducing sounds generated by sustained contact between rigid bodies. This scheme is rooted in a source/filter decomposition of the sound where the filter is described as a set of poles and the source is described as a set of impulses representing the energy transfer between the interacting objects.**

Compared to single impacts, sustained contact interactions like rolling and sliding make the estimation of the parameters of the source/filter model challenging because of two issues. First, the objects are almost continuously interacting, making the estimation of the poles of the filter more difficult. The second is that the source is generally unknown and has therefore to be modeled in a generic way. The proposed analysis/synthesis scheme combines advanced analysis techniques for the estimation of the poles and a flexible model of the source. It allows the modeling of a wide range of sounds. Examples are presented for objects of various shapes and sizes, rolling or sliding over plates of different materials. In order to demonstrate the versatility of the approach, the system is also considered for the modeling of sounds produced by percussive musical instruments.

***Index Terms*— AUD_ANSY,AUD_AUMM (EDICS),Environmental Sounds Modeling, Source/Filter Model, High-Resolution Analysis**

I. Introduction

THE synthesis of sounds generated by sustained contact between rigid objects is of interest for many application areas, such as game development and sound design for music and audio. With the widespread use of virtual reality applications, such as animation movies, video games and on-line training systems, together with the realism of the physical engines that are describing the virtual scenes, one can expect realistic sonification of the interactions between different objects to be available where the sound would be driven by the physical interaction of the objects within the scene.

Synthesis techniques based on the Source/Filter decomposition [1], [2], [3] are available to create such sounds but most of them have to be driven empirically, namely most of their parameters has to be set by an expert, making comparative studies hard to conduct and reducing the applicability of the synthesis approaches in realistic scenarios. Consequently, one would like the parameters of the source/filter model to be estimated from actual recordings. The ability to describe such sounds would also be of potential use in analysis-oriented areas such as environmental or musical sound recognition and classification [4], [5]. In order to achieve such a task, one has to tackle issues that have not, to the best of our knowledge, been considered in the literature.

The first issue is that the objects are almost continuously interacting, making the estimation of the parameters of the filter difficult. In the case of rolling or sliding interactions, multiple excitation and damping phases can alternate almost randomly. In this paper, High-Resolution (HR) techniques [6] are shown to allow the estimation of the filter parameters under such constraints by considering observation interval of short duration after a significant hit, thus maximizing the probability of the observed system to be in "free-regime". The second issue is that the numerous types of interactions between the two objects have to be adequately described by the model of the source. For that purpose, a statistical model that is able to realistically render various types of interactions is proposed. It assumes that many types of sustained interactions can be decomposed into a series of micro-impacts between the asperities of the surfaces of the interacting objects. The source is then modeled as a series of events and it is shown that the timing and amplitude of those events can be compactly modeled using statistical distributions whose parameters can be conveniently controlled in order to generate specific interactions.

The remainder of the paper is organized as follows: The global scope of the paper is described in Section II. The proposed sound model is introduced, motivated and compared to existing systems in Section III. The estimation of the modal parameters and the source of the proposed model from actual recordings are respectively described in Sections IV and V. This model is next applied in Section VI to sounds issued from various types of sustained contact between rigid bodies. The rolling and sliding of objects over plates is considered as well as percussive musical instruments. In light of those experiments, we finally discuss the benefits and drawbacks of such an analysis scheme as well as potential improvements.

II. Previous Work

Following Li&al approach [7], in order to convey some properties, an acoustic event must produce sounds with a non arbitrary acoustic structure. This acoustic structure must be recovered by the listener and successfully mapped to some auditory source categories defined in terms of the properties of the acoustic event. Consequently, when designing an analysis/synthesis algorithm for audio rendering purposes, one would like to root the system on a model of the acoustic structure such that variations of the parameters of the representation lead to the synthesis of sounds that will be successfully mapped by the listener to the corresponding categories.

This work was supported by the National Science and Research Council of Canada (NSERC). M. Lagrange, G. Scavone, and P. Depalle are with the Music Technology Area, Schulich School of Music of the McGill University, Canada (email: {mathieu.lagrange,gary.scavone, philippe.depalle}@mcgill.ca).

One approach for the synthesis of sustained contact sounds could include a rigorous physics-based analysis as taken in [8] where the simulation is defined directly in terms of the physical properties of the acoustic event. However, for real-time interactive scenarios, such an approach is computationally complex.

Alternatively, we propose in this paper a representation of the acoustic structure that is applicable to a wide range of contact interactions. This representation, based on the source/filter model is useful for synthesizing signals where the perception of the type of interaction and the type of material is nicely conveyed. In order to motivate this approach, we now study previous work on source/filter modeling.

A. Impact Sounds

Assuming certain conditions [9], the solution to the equation for motion for the struck-clamped bar is a sum of exponentially dampened sinusoids whose individual frequencies, amplitudes, and decay moduli are joint functions of the elasticity and mass density of the bar, its specific geometry, and the manner in which the bar is struck [10]. This sound model have been used in psycho-acoustical studies [9], [11], [12] to generate synthetic acoustic structures.

Considering this model is consequently natural for rendering purposes [13], [14], [15], [16], [1], [17]. The filter parameters encapsulate the perceptually relevant characteristics of the physical structure. In this simplest form, a modal synthesizer can be viewed as a source/filter model whose source is a Dirac impulse. Depending on the targeted application, several strategies can be considered to better model the source. One previously reported approach considers all interaction forces to be directed normal to the surface [18]. The resulting system can be modeled as a mass-spring system, enabling the generation of hits, bounces and breaking-like events [19]. The parameters can be set using physical descriptions of the interacting objects. An extension of this work has recently been proposed which leads to convincing rolling sound synthesis [20].

Alternately, one might be interested in estimating the parameters of the source/filter model from actual sound recordings to eventually achieve a perceptively coherent synthesis without extensive parameter tuning. Some methods targeting this goal have been proposed for the modeling of impact sounds [21], [22]. In [21], the modal parameters are derived based on Fourier analysis and the Energy Decay Relief method. The estimation of the source is then carried-out by inverse filtering of the sound. Since the source is assumed to be of short duration and therefore cheap to store in memory, the authors do not address the need for a parametric representation of it.

B. Sustained Contact Sounds

Rolling or scraping behaviors involve numerous contact events between objects. The source is therefore not of short duration. In synthesis schemes such as the FoleyAutomatic [1], the authors consider that the source excitation consists of the interaction of the moving objects with the protrusions of the surface. It is assumed that the interacting surfaces are fractal

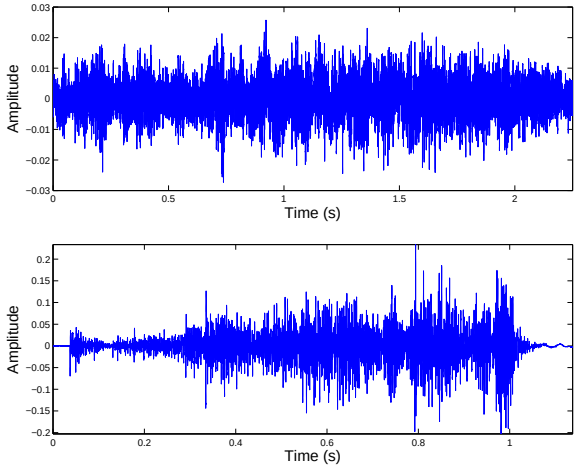

Fig. 1. *Signals of a bottle respectively sliding (top) and rolling (bottom) over a plate.*

and that the source has the same property. Such a distribution is efficiently approximated using $1/f$-shaped white noise. This shaping is carried out using a bank of filters tuned to obtain a power spectrum proportional to ω^{β}, where ω is the frequency and β is the fractal dimension [23]. Another filtering step is carried out using a resonating filter whose frequency is tuned according to the speed of the object within a low frequency range. While satisfying for scraping sounds, the authors felt the need for further processing in the case of rolling sounds.

When comparing sliding and rolling sounds, two main differences arise. First, the amplitude envelope of the rolling sound exhibits more pronounced long-term modulation, see Fig. 1. Second, the rolling sound is of lower frequency content, see Fig. 2. Although the physics behind the rolling phenomena are not fully understood [8], it is speculated in [1] that the first difference is due to the fact that the rolling object bounces more and therefore hits the surface less often, leading to a source signal of lower frequency. The second difference is explained in [24] by the asymmetry of the rolling object with respect to its center of mass. This cue is very important for perception, since nearly perfect spherical shapes with nearly perfect mass distribution are not easily perceived as rolling objects.

It is proposed in [24] that the oscillating height and center of mass can be approximated by a sinusoidal evolution as the excitation force. Alternatively, the FoleyAutomatic tackles those two issues by considering another low-pass filtering step. The overall spectrum is of much lower frequency and, as a side effect, the source exhibits long-term amplitude modulations due to the enhancement of the beating of some modes within the low frequency region.

III. Proposed Approach

The synthesis model proposed in [3] provides convincing cartoon-like rolling sounds. However, the physical model used to generate the source signal is not flexible enough to allow arbitrary shapes and surface types without specific adaptations.

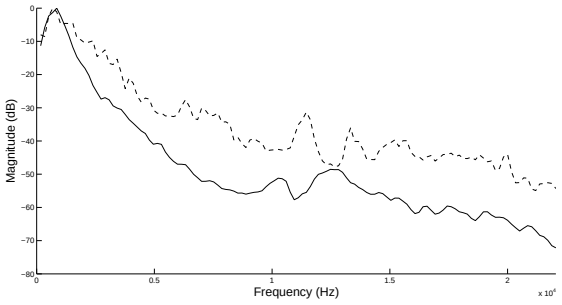

Fig. 2. *Fourier spectrum of the sound of a bottle respectively rolling (solid line) and sliding (dashed line) over a plate.*

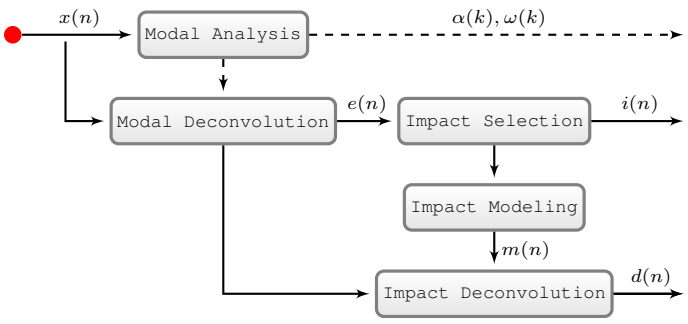

Fig. 3. *Block-diagram of the proposed analysis scheme.*

The FoleyAutomatic [1] also captures most of the aspects of rolling behaviors but does not provide any systematic approach for the estimation of the parameters of the different filters considered.

A. Source/Filter Modeling

We propose in this paper a synthesis scheme that is relevant for the synthesis of sustained contact sounds generated by interactions between rigid bodies of different shapes and surface properties. This is achieved by means of an analysis method that allows the user to estimate the parameters of the model from recordings of the sustained contact between the two bodies. As for most source/filter models, it is assumed that is possible to define a source and a filter such that the filter models perceptively important spectral properties of the sound whereas the source models perceptively important properties of the temporal envelope of the sound.

Source/filter models have been studied extensively, mainly in the domain of speech processing [25], but also for musical signals [26], [27] and more generally for generating sounds caused by the interactions between rigid bodies for rendering purposes [14], [1]. From a physical point of view, this model can be interpreted as follows: the vibrating structures are represented by the resonant filter, and the interaction between the vibrating structure and the physical exciter by an excitation signal.

In order to cast our problem into this convenient model, we consider that the resonances of the exciting object - if any - can be considered part of the filter as its resonances will be captured at the analysis stage. With this assumption, we can model sounds generated by sustained contact between rigid objects as the output of an infinite impulse response filter (IIR) excited by the source signal. This source signal represents the multiple contact events between the two objects by encoding the amplitude and the phase of the modes and the resonant filter encodes their frequencies and damping factors.

B. Multiple-Impulse Modeling of the Source

Significant previous work in the speech domain is relevant to our proposed excitation model. The vocoder [28], [29], [30] is rooted on the assumption that the voice is either “voiced” or “unvoiced”. Depending on this status, the source is modeled as a pulse train with period tuned according to the pitch or white noise. For real speech signals, however, such a dichotomy is less clear. Atal&al proposed in [31] a new model of speech excitation for producing more natural-sounding speech at low bit rates. They allow a few pulses per pitch period to model voiced and unvoiced speech with the same approach.

Similarly, by having the same scheme as the vocoder for interacting bodies, one could have a situation where the object is bouncing and the excitation would then be more or less regularly spaced pulses. On the contrary, if the object is rolling or sliding, the excitation could be modeled by shaped noise as in the FoleyAutomatic [1]. One can understand the limitation of such a scheme when considering sounds created by objects with irregular shapes and surfaces. We therefore opt for an unconstrained excitation model, where we assume that many kinds of complex interactions can be decomposed into elementary contact events. A similar approach has been followed for the modeling of musical percussive-like maracas in [16], where the enveloped noise pulses are randomly generated using empirically defined statistical distributions.

We propose to model the source of sustained excitation sounds as a series of triggers of an impact signal. The shape of this impact signal typically encodes physical properties of the interacting objects. For example, a glass marble hitting a metallic plate has a sharp impact, whereas a rubber hammer will induce a smoother impact. We will describe in Section V how this impact can be derived from sound recordings. Alternatively, this impact signal could also be generated using the physical properties of the interacting objects as in [18] for morphing purposes.

The issue is then to estimate the time location and intensity of those events given a recorded sound, which has been proved to be difficult in the case of speech signals. In [31], the location and amplitudes of the pulses are found by means of an iterative analysis-by-synthesis method, which requires an empirically defined cost function. Fortunately, we do not consider vibrating structures that are changing shape over time like the vocal tract, so the process of estimating those pulses can be done by inverse filtering as described in Section V.

C. Overall Description of the Analysis Algorithm

We now give an overview of the analysis scheme, whose block-diagram is shown on Fig. 3. Considering $x(n)$, the recorded sound of the interaction between two objects, an

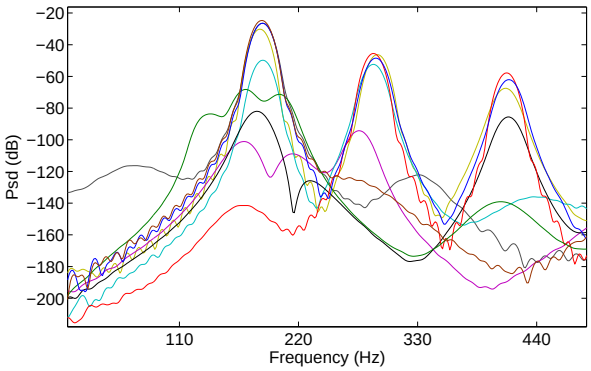

Fig. 4. *Amplitude spectra of modal filters with parameters chosen according to High Resolution analysis of a plate hit with a hammer at nine different locations recorded using an accelerometer.*

analysis interval is defined, where the modal analysis is performed in order to estimate the modal parameters using the HR method as described in Section IV. The impulse response of this filter $s(n)$ is defined in 2.

As described in Section V, this filter is considered to estimate the source signal by means of a modal deconvolution carried out by inverse filtering. A representative impact $i(n)$ is then extracted from this source signal. Its amplitude envelope is modeled using a parametric shape $m(n)$ that has previously been considered for the modeling of transients in the audio coding area [32]. This parametric shape allows us to perform an impact deconvolution in order to obtain a spike-like excitation signal which is modeled as a series of amplitude modulated Dirac Functions $d(n)$. The resulting synthesis scheme is a simple two-step convolution:

$$\hat{x} = d * i * s \tag{1}$$

where $*$ denotes the convolution. For the modeling of rolling and sliding sounds we considered in [33] several approaches to process and parametrize further the representative impact signal $i(n)$. Those schemes were evaluated in [34] and [35] where the unprocessed impact is shown to be the most reliable solution as far as perceptual similarity with the analyzed sound is concerned. Therefore, $i(n)$ is considered as-is in this paper.

IV. Estimation of the Filter Parameters

Once excited, an object vibrates according to its own set of modes [10]. The frequencies and damping factors of those modes are determined by the physical properties of the resonating body, mainly its dimensions and stiffness and do not depends on the locations of the excitations. Fig. 4, which illustrates amplitude spectra of modal filters with parameters chosen according to HR analysis of signals resulting from strokes at different locations recorded using an accelerometer placed on one end of the plate.

If we assume that an exciting impact is of short duration, the resulting sound will be primarily composed of a linear combination of "free regime" vibrations of the object (after a short transient regime). At this stage, the sound can be conveniently described as a set of exponentially damped sinusoids:

$$\begin{aligned} x(n) &= \sum_{k=1}^{K} A_k e^{z_k n} \\ z_k &= \alpha_k + j\omega_k \\ A_k &= g_k e^{j\phi_k} \end{aligned} \tag{2}$$

where g_k, ϕ_k, α_k, and ω_k are respectively the gain, the phase, the damping factor and the frequency of the pole k. Those parameters are the parameters of the filter. The parameters α_k and ω_k are considered as fixed while the g_k may evolve according to the displacement of the location of the contact.

Psycoacoustical studies [11] show that even in the relatively simple case of struck metallic bars, it is not straightforward to relate the estimated poles to the actual modes of the vibrating objects. We therefore only focus in this paper at the estimation of the filter parameters that lead to a resynthesis of satisfying quality and do not make any attempt at relating the estimated poles to the actual modes of the vibrating object.

Though, the estimation of the parameters of the filters should be performed when the vibrating structure we want to model is - as much as possible - vibrating freely without interacting with the other object. In [33], this difficulty is avoided by assuming the availability of the recording of an impact between the two objects. However, this approach reduces considerably the scope of applicability of the method.

Alternatively, we propose in this paper to estimate the filer parameters directly from the sustained contact sound. Consequently, the observation of the modal parameters should be performed when the vibrating structure we want to model is as much as possible vibrating freely without interacting with the other object. The algorithm considered for finding the boundaries of such an interval is detailed in Appendix A. In order to satisfy as much as possible those constraints,we propose to set the analysis interval as a short time interval following a significant impact.

A. Spectral Analysis

Given an appropriate analysis interval, one needs to estimate the parameters of (2). Many methods are available to achieve this task. A common nonparametric method is the Fourier transform. Although quite robust and considered in many previous studies [21], [1], it requires a long observation window in order to achieve a good frequency resolution as well as a precise estimation of the damping factor (see Appendix B for details). Though, our analysis has to be performed over a very short-time observation interval with a potentially low signal-to-noise ratio, due to the residual influence of the numerous impacts that occurs. From a theoretical point of view, parametric methods with careful pre-processing seem best adapted [36].

This choice is confirmed by the comparative study detailed in Appendix B, from which we chose the ESPRIT algorithm [37] as the best option in our application context. This algorithm belongs to the family of subspace-based HR spectral estimation techniques. This means that an eigenanalysis of the autocorrelation matrix is performed. With the assumption

that the observed signal is of the type as defined in (2), it can be shown that the K highest eigenvalues will correspond to the powers of the components of the model plus the power of the residual. The eigenvectors associated with the K highest eigenvalues then form a base of the so-called "signal subspace". By considering (2), the frequencies and damping factors of the modes are computed as follows:

$$f_k = \frac{\Im(\log(z_k))}{2\pi} \text{ and } \delta_k = \Re(\log(z_k)). \tag{3}$$

The amplitude and phase of each component is then obtained via a projection of the observed signal on the estimated signal model [36]. It is important to note that an extensive pre-processing is performed on the observed signal before it is fed into the ESPRIT algorithm: the signal is first split into subbands, down-sampled in each of these bands, and whitened. All these steps are used in order to ensure a better conformance of the data to the model [6]. The interested reader is referred to [6] for a precise outline of the steps in the ESPRIT technique as well as extensions to that technique.

Nevertheless, a quantitative comparison of the HR method with Fourier-based and LP-based methods is provided in Appendix B, where synthetic scenarios considering impact-like and sustained excitation of realistic modal structures are considered. With impact-like excitations, it is shown that the HR method performs best. With stochastic excitations, it compares favorably for the estimation of the frequency but not for the estimation of the damping parameter, as the underlying model is no longer verified. Closer investigations showed that the Autoregressive (AR) method consistently under-estimates the damping, whereas the HR method tends to overestimate the dampings when the excitation energy is still significant in the observation interval. Since a precise frequency estimate is crucial for our purposes and the Fourier based method cannot be considered, the HR method is selected.

The estimation of the optimal number of modes given a signal is yet another difficult problem. In our approach, we assume a large number of modes, 20 for each of the eight sub-bands and the desired number of modes is selected as the modes with positive dampings and the highest amplitude. For the sake of generality, the number of modes is limited to 80. However, this number could be highly reduced depending on the type of objects and interactions we are considering. For example, in the case of a glass ball sliding over a wooden plate, a good quality synthesis can be achieved with ten modes.

B. Modeling the change of the location of the excitation

As extensively studied by Stoelinga&al [38], gradually varying ripples can be observed on the spectrograms of rolling sounds. It can be shown that this pattern arises from the interference between the sound directly generated at the point of contact between ball and plate, and the sound reflected at the edge of the plate. This phenomenon is important for the perception of movement. The "phasing" effect illustrated with this theoretical simulation can also be observed in real rolling sounds. Although less pronounced in real cases, this effect is clearly audible and perceptively important.

The estimation of the parameters of this comb filter from actual recordings is left for future research. However, assuming that the location of the rolling object is known, that the trajectory is linear, and that the two edges of the plate are clamped or supported, we can analytically compute the fundamental frequency of the comb as a function of the relative location. While being an approximation, it was found that this processing step provides an enhanced perception of movement.

V. Estimation of the Source Parameters

As described in Section III, sustained excitation signals are modeled as a convolution of a signal $d(n)$ that encodes the location and amplitude of the impacts and an impact signal $i(n)$ that encodes the initial deformation of the vibrating body induced by a single contact between the two bodies.

The signals $d(n)$ and $i(n)$ are computed from the excitation signal, which itself is estimated by deconvolution of the modal parameters from the recorded signal. This deconvolution process is achieved by inverse filtering. A selected section of the resulting estimated source will be considered as the impact signal $i(n)$. This signal will be parameterized in order to identify $d(n)$ using yet another deconvolution step.

A. Estimation of the Source Signal

In most situations, only the output of the source/filter system is available. Several methods are available to estimate the excitation signal (the source) from the output. Some do not use any explicit Source/Filter model [39]. Those methods usually try to "whiten" the output by removing resonances in the spectral domain by assuming that the resonant part is only due to the resonances. Using the same assumption, others consider a sinusoids+noise model such as the one proposed in [40] where the result of the subtraction of the synthesis of the sinusoidal part from the original is considered as the excitation.

In our approach, we assume knowledge of the filter structure which allows us to this deconvolution process by performing an inverse filtering step. More precisely, we consider a set of second-order cosine filters to model the filter part of our model:

$$S(z) = \frac{1}{2}\sum_{k=1}^{K/2} \frac{A_k}{1 - z_k z^{-1}} + \frac{A_k^*}{1 - z_k^* z^{-1}} \tag{4}$$

The excitation $e(n)$ is then computed using inverse filtering in the frequency domain:

$$E(\omega) = X(\omega)/S(\omega) \tag{5}$$

where $S(\omega)$ is the Fourier spectrum of the output of the filter $s(n)$. Frequency-domain deconvolution is more reliable because of numerical instabilities in the computation of the filter coefficients [21].

B. Modeling the Impact Excitation

The impact excitation serves two purposes. It is used as is for synthesis purposes, and a smooth version is used for estimating $d(n)$. We propose to choose a representative impact by

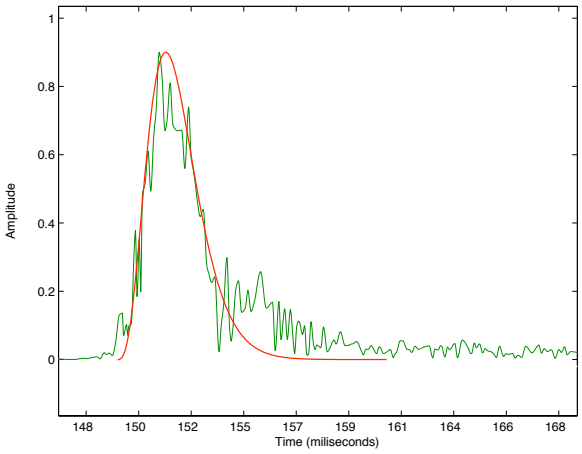

Fig. 5. *Meixner window (smooth curve) fitted over the envelope of the excitation signal of a hit of a glass bottle over a MDF plate.*

identifying a time interval where a significant impulse occurs in the excitation signal $e(n)$. This procedure is explained in details in Appendix A.

We next need to define at which time and at which amplitude this impact should be triggered. As it will be described in the next section, the estimation of the time and amplitude properties of the series of triggers is done by means of a deconvolution of the amplitude envelope of the excitation signal. However, it was not found practical to define a stable filter for every kind of envelope of the impact shape. We therefore need to abstract the impact shape using a parametric curve that will always lead to a stable filter.

For this purpose, we propose to approximate the amplitude envelope of the impact using a "Meixner" temporal envelope that can be adapted to many different impact types by means of time scaling. This shape has previously been used for the modeling of transients in a low-bit rate hybrid audio coder [32]. We considered this window because it has an overall shape compatible with various impact-like events and is continuous at the boundaries. The "Meixner" envelope is computed as:

$$w(n) = (1-\gamma^2)^{\beta/2}\sqrt{\frac{h(n)}{n!}}\,\gamma^n \quad (6)$$

$$h(n) = \beta\cdot(\beta+1)\cdot\cdots\cdot(\beta+n-1) \quad (7)$$

$$h(0) = 1 \quad (8)$$

with $\beta > 0$, $0 < \gamma < 1$ and $n = 0, 1, 2, ...$ The attack is controlled by β and the exponential decay is controlled by γ. In this paper, we made use of one particular shape, computed with β and γ empirically set to 10 and 0.89 respectively. This shape is next truncated to be 200 samples long, delayed, scaled and stretched to fit the envelope of the selected impact, see Figure 5.

C. Location of the Impacts

It is assumed that one object "strikes" the surface of the other many times, producing a sustained excitation composed of many individual contact events at various times and with various gains. Given an excitation signal $e(t)$ and an impact envelope model w of length l_w, we want to estimate $\mathbb{T} = \{t_m, a_m\}$, the set of triggers, where each trigger is defined by its time index t_m and amplitude a_m.

The envelope $\mathcal{E}$ of the excitation signal $e(t)$ is first computed as a spline interpolation between the maximal values of $|e|$. It is proposed in [33] to iteratively identify the locations of the maximum cross-correlation between $\mathcal{E}$ and w to identify the location of the events t_m. This method performs well in the case of bouncing objects but it was not as satisfying in the case of more complex behaviors like rolling.

To improve the identification of the location of the impacts in those cases, a signal $\mathcal{T}$ so that $\mathcal{E} = \mathcal{T} * w$ is considered. The estimation of $\mathcal{T}$ is achieved by means of a two step inverse filtering approach. In order to provide us with stable filters, w is separated in two sections: $w_1 = w(0, m_w - 1)$ and $w_2 = w(m_w, l_w - 1)$, where m_w is the index of the maximal value of w. $\mathcal{E}$ is first filtered using w_2 from the beginning to the end and next filtered using w_1 from the end to the beginning.

The resulting signal is composed of peaks whose dominant maxima are considered as candidates for impact locations. To do so, an indicator vector $m(n)$ is built:

$$m(n) = \begin{cases} \beta\, m(n-1) + (1-\beta)\,\mathcal{T}(n) & \text{if } \mathcal{T}(n) > m(n-1) \\ \alpha\, m(n-1) + (1-\alpha)\,\mathcal{T}(n) & \text{otherwise} \end{cases} \quad (9)$$

where $m(0) = 0$. The parameters β and α have been empirically set to 0.999 and 0.3, respectively. Every local maximum whose amplitude is above the indicator vector is inserted in the list of triggers $\mathbb{T}$. All the peaks from $\mathcal{T}$ that have a highest value greater than $c = \max(\mathcal{T}/2)$ are next discarded. This process is carried out until c is below a given threshold empirically set to $\max(\mathcal{T})/5$.

D. Excitation Encoding

With the assumed linearity of the synthesis scheme, another advantage of such an approach is the simplicity of the parameter set. The number of modal parameters is bounded, can be relatively low in many synthesis scenarios and can be encoded using well known quantization techniques [41]. The impact excitation is also relatively cheap to encode due to its short duration. Furthermore, in a scenario that requires transmission of data such as on-line virtual reality, both pieces of information can be stored statically on the decoder side. This is not the case of the set of triggers $\mathbb{T}$ that should to be transmitted efficiently. The proposal and evaluation of a complete encoding scheme for $\mathbb{T}$ is out of the scope of this paper. Instead, we demonstrate how this set can be statistically described at a microscopic level in order to be compactly represented without a significant loss of fidelity.

Let us consider an alternative representation of $\mathbb{T}$ where time indices are represented using a differential scheme $\tilde{\mathbb{T}} = \{\delta_m, n_m\}$ where $\delta_0 = t_0$ and $\delta_m = t_m - t_{m-1}$. We considered the statistical distribution of those parameters over the databases of sounds of sustained contact between solids described in the next section, see Figure 6. It was found that

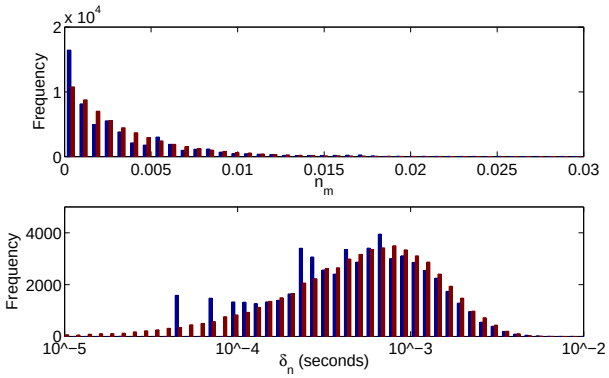

Fig. 6. *Measured (left bar) and modeled (right bar) frequency distributions of the parameters n_m and δ_m which are respectively modeled by exponential and gamma distributions.*

the parameter δ_m follows a gamma distribution whereas an exponential distribution is relevant for the parameter n_m.

In order to encode $\mathbb{T}$, the set is first split in several subsets $\mathbb{T}_n$, such that $t_m \in \mathbb{T}_k$ are in the $[k\Delta,(k+1)\Delta]$ interval, where Δ is a number of samples corresponding to $\approx 0.25s$. Within each subset, the 20 elements with the highest amplitude are modeled as is, in order to retain precise modeling of the macroscopic events. The remainder is modeled using some statistical distributions. For each remaining element of the subset $\mathbb{T}_k$, the parameters of the two distributions are estimated to generate $\hat{\mathbb{T}}_k$. Those sets can finally be considered to synthesize an approximate vector of triggers as follows:

$$\hat{d}(n) = \begin{cases} = \hat{n}_k^i & \text{if } \exists i | \mod \left(\sum_{p=1}^{i} \hat{\delta}_k^p, \Delta\right) = n \\ = 0 & \text{otherwise} \end{cases} \quad (10)$$

As it will be discussed in the next section, no significant perceptual difference has been found between the estimated and approximate set of triggers in the case of rolling and sliding sounds, although the bit-rate is reduced by about 100 times compared to a straightforward 16-bit encoding of $\mathcal{T}$.

VI. Application to the Modeling of Recorded Audio Signals

The proposed algorithm is considered for approximating of a variety of sounds. Impact sounds are first considered followed by sounds with sustained excitation, like bouncing, sliding and rolling. Sounds of musical instruments that are excited in a sustained fashion, *i.e.* hit at multiple time more or less rapidly, shaken, and bowed are finally considered. All the sounds discussed in this section are available online[1]. The same set of parameters given in the previous sections have been considered for all the examples reported here.

A. Impact Sounds

In order to validate the estimation of the filter parameters in short duration excitation scenarios, a database of impact sounds recorded by Bruno Giordano [42] is first considered. The accuracy of the estimation can be perceptually asserted by considering a Dirac impulse as the source. It was found that both the filter parameters estimation and the quality of the whole synthesis chain is good for most of the sounds provided that the number of estimated modes is not too low. If this condition is not met, some modes are left in the estimated excitation, inducing artifacts during the estimation of the set of events. Iterative approaches such as the one proposed in [43] may lead to a better estimation of the modes and consequently a better estimation of the parameters, leading to a significant overall improvement.

B. Rolling and Sliding Sounds

We consider here three set of sounds. The first database consists of objects sliding and rolling across inclined boards made of wood and melamine. The objects are respectively an empty glass bottle rolling on its side, a glass marble and a wooden croquet ball. Details about the recording settings can be found in [33]. It was found that the difference of shape of objects as well as the granularity of the surface is nicely conveyed by the synthesis algorithm. However, the type of material of the board can sometimes be hard to perceive, probably due to a relative inaccuracy in the estimation of the damping factors.

The second set of sounds are issued from the demonstration sounds of the FoleyAutomatic [1]. Since this algorithm is not data-driven, we used their synthesis sounds as the input to our algorithm in order to be able to compare our approach.

Probably due to the facts discussed in Section III-B, it was found that the transition between bouncing and rolling behaviors are more subtle using the proposed scheme, leading to a more realistic perception of rolling. This ability to model different types of interactions in a flexible way can also be observed with the third database recorded by Christopher Stoelinga [8], where some wooden balls of different sizes are rolling over wooden plates of different thicknesses at constant speed.

The parametrization described in the paper provides the best synthesis quality in our opinion and follows perceptual validation that have been conducted while designing the processing algorithm. Indeed, perceptual experiments have been carried out [34] that evaluate the ability of the tested approach to generate sounds that are perceived as rolling sounds. A more in depth evaluation compares the synthesized sounds with respect to the original recordings [35] where several alternative algorithmic designs are evaluated. This last work gives useful insights about which design choices are relevant and demonstrates the usefulness of a complete analysis/synthesis scheme to conduct meaningful perceptual evaluations of such synthesis algorithms.

C. Sounds of Percussive Musical Instruments

We also considered some sounds of musical instruments of the McGill University Master Samples [44] to evaluate the ability of the proposed algorithm to model diverse timbres and type of interactions. The proposed scheme was found to provide good fidelity for wood block, castanet and timbals.

[1] http://perso.telecom-paristech.fr/~lagrange/demos/ciqs.

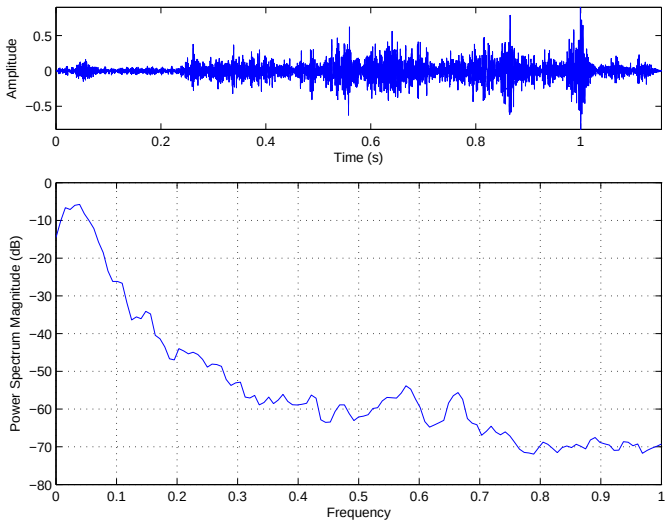

Fig. 7. *Time and frequency domain representation of the synthesized sound of a bottle rolling over a plate with encoding of the set of triggers.*

Instruments like maracas and bamboo chimes are also correctly modeled, making the proposed algorithm an interesting alternative to the PhISEM approach proposed by Perry Cook [2].

To a certain extent, sounds produced by snare drums, high hats and Chinese cymbal are correctly modeled as far as timbre is concerned. However, the time properties, like the location of the hits are not correctly conveyed. Another interesting type of sustained excitation is bowing. Even though the interaction of the bow and the string is highly complex [45], it is worth noticing that the synthesis provided by the proposed approach is satisfying in terms of timbre and long-term amplitude modulation.

VII. Conclusion

We introduced in this paper an analysis / synthesis algorithm of sustained excitation sounds. This scheme is, to the best of our knowledge, the first method that allows the complete estimation of the synthesis parameters from actual recordings. We believe that such a method could be of use for many application areas from human/computer interaction, sonic and haptic feedback, virtual reality, as well as new sound creation for musical purposes. We are currently investigating the use of this method in interaction scenarios and have found it to provide satisfying synthesis quality and good flexibility.

Some limitations highlighted in the previous section suggest several directions for future research. First, an improvement of the estimation of the modes is highly desirable since it is the root of the whole modeling chain. We are currently investigating iterative approaches for the refinement of the modal parameters [43]. Secondly, examples such as the stone bouncing and rolling inside a wok illustrates clearly that for such an irregular shape, the use of only one impact excitation profile can be perceptively costly. We plan on studying the use of many impact excitation signals as well as a parametric modeling of those impact excitations to increase the flexibility of the model. Lastly, a more in-depth study of the statistical properties of the set of triggers would be desirable for the estimation of physical parameters such as the speed or the shape of the moving objects as well as to propose efficient encoding schemes that would benefit application areas such as distributed virtual reality.

VIII. Acknowledgments

The authors would like to thank Roland Badeau for useful discussion and for kindly providing us with the implementation of the HR analysis method as well as Bruno Giordano and Christopher Stoelinga for respectively providing us with impact and rolling sound databases. The authors are also grateful to the anonymous reviewers for their valuable comments. This work was supported by a Special Research Opportunity grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

A. Analysis Interval

The method described in this appendix aims at identifying the time interval where the vibrating object is as much as possible in a free regime, *i.e.* when the object is not excited.

Once reaching a peak intensity during the excitation, the amplitude envelope of a signal s following the model expressed in (2) will decrease until the next contact. Considering this cue, the representative impact and the analysis interval boundaries n_b and n_e are iteratively searched as follows. We first identify the time indices n_s, n_m and n_e which respectively indicates the beginning , the location of the maximal value and the end of the representative impact. To do so, an indicator vector $a(n)$ is first defined:

$$a(n) = \frac{1}{\delta} \sum_{m=\lfloor n/\delta \rfloor \delta}^{(\lfloor n/\delta \rfloor + 1)\delta} 20 \log_{10}(|s(m)| + \epsilon) \qquad (11)$$

where $\lfloor x \rfloor$ denotes the rounded value of x towards 0, δ is a number of samples corresponding to five ms and ϵ is an arbitrary small and positive value. We then look for an interval matching the amplitude decay after a significant impact. The first boundary is set so that $|s(n_m)| = \max |s(n)|$ and n_e is initialized to n_m and incremented until the following condition is no longer met:

$$\frac{\sum_{n=b}^{e} |a(n) - l(n)|}{(e-b) \max_{n \in [b,e]} a(n)} < \tau \qquad (12)$$

where $b = n_m$, $e = n_e$, $l(n)$ is a linear approximation of $a(n)$ computed within the $[b, e]$ interval and τ is a threshold parameter set to 1.8. The index n_s is searched in a similar fashion.

In order to focus the analysis on the interval where the influence of the excitation is no longer dominant and at the same time where the energy of the signal is still significant, one would like to have the analysis performed at the center of the selected interval. To do so, n_l and n_r are first initialized to n_m and n_e. They are next respectively incremented and decremented in order to center the analysis interval if the interval is larger than the requirements of the analysis module.

Finally, in order to ensure continuity at the boundaries, n_l and n_r are respectively incremented and decremented so that $x(n_r)$ and $x(n_l)$ are close to 0, *i.e.*:

$$\text{sign } x(n_r) \neq \text{sign } x(n_r + 1) \quad (13)$$

$$\text{sign } x(n_l) \neq \text{sign } x(n_l + 1) \quad (14)$$

. If the interval is not sufficiently large for the minimal requirements of the analysis module, this interval is discarded and another interval is searched. If no satisfying interval can be found, the largest interval found is padded with trailing zeros prior to analysis.

B. Evaluation of the Modal Analysis Methods

We provide here some numerical motivations for the use of the High-Resolution analysis method for the estimation of the damping and frequency parameters of the modes of the vibrating structure whether the source is of short duration like a single impact or a sustained one like rolling or sliding.

We consider 1000 random sets of five modes excited using a Dirac pulse or white noise. Their frequencies are in the $[0, 0.2]$ normalized frequency range. The distribution of the frequency and damping parameters are randomly chosen within realistic ranges. Those ranges are set by computing statistics over 100 sets of modes extracted from an impact sounds database comprising various plates and hammer materials [42].

The precision of a given method is evaluated by its ability to estimate the frequency and damping parameter of the modes. For the frequency, all the permutations of the set of estimated frequencies $\hat{f}_k$ are compared to f_k and the estimation error e_f is computed as:

$$e_f = \min_p e_f(p) \quad (15)$$

where

$$e_f(p) = \sum_{k=1}^{5} |\hat{f}_{p_k} - f_k| \quad (16)$$

and p_k are the permuted indices of the p permutation. The damping estimation error e_d is computed similarly. The mean and standard deviation (in parenthesis) of the estimation errors over 1000 trials are reported in Table I.

Three analysis methods are considered. For reference, we first considered a Fourier-based estimator. Numerous methods, from quadratic interpolation [21] to phase-based methods [46] are available. We considered here the "vocoder" phase-based method for the estimation of the phase, amplitude and frequency.

The damping parameter has to be estimated by other means. The use of the DFT over a finite duration assumes a signal that is periodic over the interval of observation. The damping parameter is estimated using the Energy-Decay Rate (EDR) method [21] requiring a much larger time interval for such a method. Several modal amplitude estimates have to be performed over several frames in order to obtain reasonable estimates of the damping factors. In the experiments described here, we considered an analysis window of ≈ 92ms and four frames with an interval between successive frames of ≈ 1.5ms.

Closer to our approach, we also considered an LP-based method [47]. The covariance method is used to identify the frequencies and damping factors of the model over the same time interval as the one considered for the HR analysis ($\approx$ 23ms). The N poles, z_n of the sound are computed by finding the roots of the autoregressive part of the signal model. The frequency and damping factor of the mode associated with one of the pole pairs is obtained using (3). For the three evaluated methods, the selection of the five most prominent modes over the $N = 20$ extracted modes is done by decreasing magnitude. As can seen on Table I that shows the errors as relative values, the HR method compares favorably for the estimation of the frequency, a major prerequisite for the approach proposed in this paper.

TABLE I

RELATIVE ESTIMATION ERROR OF THE AR, FOURIER AND HR ANALYSIS METHODS FOR THE ESTIMATION OF THE FREQUENCY (A) AND DAMPING (B) PARAMETERS.

parameter	excitation	AR	Fourier	HR
frequency	impact	0.18 (0.012)	0.22 (0.015)	0.17 (0.013)
	sustained	0.26 (0.033)	0.18 (0.02)	0.17 (0.016)
damping	impact	0.11 (0.032)	0.1 (0.013)	0.09 (0.012)
	sustained	0.13 (0.07)	0.16 (0.016)	0.14 (0.016)

REFERENCES

[1] K. van den Doel, P. G. Kry, and D. K. Pai, "Foleyautomatic: physically-based sound effects for interactive simulation and animation," *International Conference on Computer graphics and interactive techniques (SIGGRAPH)*, pp. 537–544, 2001.

[2] P. R. Cook, ""physically informed sonic modeling (PhISM): Synthesis of percussive sounds"," *Computer Music Journal*, vol. 21, no. 3, pp. 38–49, 1997.

[3] M. Rath and D. Rochesso, "Continuous sonic feedback from a rolling ball," *IEEE Multimedia*, vol. 12, pp. 60–69, 2005.

[4] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio Speech and Language Processing*, 2006.

[5] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, Vancouver, Canada, 2006.

[6] R. Badeau, "High resolution methods for estimating and tracking modulated sinusoids. Application to music signals." Ph.D. dissertation, Telecom Paris, France, 2005, in French.

[7] X. Li, R. J. Logan, and R. E. Pastor, "Perception of acoustic source characteristics: Walking sound," *Journal of Acoustical Society of America*, vol. 90, no. 6, pp. 3036–3049, 1991.

[8] C. N. Stoelinga, "A psychomechanical study of rolling sounds," Ph.D. dissertation, University of Eindhoven, The Netherlands, 2007.

[9] R. A. Lufti and E. L. Oh, "Auditory discrimination of material changes in a struck-clamped bar," *Journal of the Acoustical Society of America*, vol. 102, pp. 3647–3656, 1997.

[10] P. M. Morse and K. U. Ingard, *Theroretical Acoustics*. Princeton University Press, 1968.

[11] S. Lakatos, S. McAdams, and R. Causse, "The representation of auditory source characteristics : Simple geometric form," *Perception & Psychophysics*, vol. 59, no. 101, pp. 1180–1190, 1997.

[12] S. McAdams, A. Chaigne, and V. Roussarie, "The psychomechanics of simulated sound sources: Material properties of impacted bars," *Journal of Acoustical Society of America*, vol. 115, no. 3, pp. 1306–1320, 2004.

[13] W. W. Gaver, "Everyday listening and auditory icons," Ph.D. dissertation, University of California, San Diego, 1988.

[14] ——, "The sonic finder: An interface that use auditory icons," *Journal of Human Computer Interaction*, vol. 4, pp. 67–94, 1989.

[15] J. M. Adrien, *The Missing Link: Modal Synthesis*, G. dePoli, A. Piccialli, and C. R. Eds, Eds. MIT Press, 1991.

[16] P. Cook, *Real Sound Synthesis for Interactive Applications*. A. K. Peters Ltd., 2002.

[17] D. K. Pai, K. van den Doel, D. L. James, J. Lang, J. E. Lloyd, J. L. Richmond, and S. H. Yau, "Scanning physical interaction behavior of 3D objects," *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2005.

[18] F. Avanzini and M. Rath, *"Modal Resonators", The Sounding Object*, D. Rochesso and F. Fontana, Eds. Mondo Estremo, 2003.

[19] M. Rath, *"Modal Synthesis", The Sounding Object*, D. Rochesso and F. Fontana, Eds. Mondo Estremo, 2003.

[20] ——, "Energy-stable modelling of contacting modal objects with piecewise linear interaction force." International Conference on Digital Audio Effects (DAFx'08), 2008.

[21] J. Laroche and J. L. Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 329–344, 1994.

[22] M. Aramaki and R. Kronland-Martinet, "Analysis-synthesis of impact sounds by real-time dynamic filtering," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, pp. 695–705, 2006.

[23] K. Steiglitz, *A Digital Signal Processing Primer with Applications to Digital Audio and Computer Music*. Addison-Wesley, New York, 1996.

[24] M. Rath, "An expressive real-time sound model of rolling," in *International Conference on Digital Audio Effects*, 2002, pp. 165–168.

[25] J. D. Markel and A. M. Gray, *Linear Prediction of Speech*. Berlin: Sringer-Verlag, 1976.

[26] X. Rodet, P. Depalle, and G. Poirot, "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitations functions," in *International Computer Music Conference*, 1988.

[27] P. Depalle, "Analyse, modélisation, et synthèse des sons basés sur le modèle source/filtre," Ph.D. dissertation, Université du Maine, 1991, in French.

[28] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE*, vol. 54, pp. 720–734, 1966.

[29] J. Flanagan, *Speech Analysis Synthesis and Perception*. New-York: Springer-Verlag, 1972.

[30] J. Markel and A. Gray, "A linear prediction vocoder simulation based upon the autocorrelation method," *IEEE Transactions on Audio Speech and Signal Processing*, vol. 22, pp. 124–134, 1976.

[31] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *IEE International conference on Audio, Speech and Signal Processing*, 1982, pp. 614–617.

[32] B. D. Brinker, E. Schuijers, and W. Oomen, "Parametric Coding for High-Quality Audio," in *112th Convention of the Audio Engineering Society*, May 2002.

[33] M. Lagrange, G. Scavone, and P. Depalle, "Time-domain analysis / synthesis of the excitation signal in a source / filter model of contact sounds," *Proceedings of the International Conference on Auditory Display (ICAD)*, 2008.

[34] E. Murphy, M. Lagrange, G. Scavone, P. Depalle, and C. Guastavino, "Perceptual evaluation of a real-time synthesis technique for rolling sounds," in *5th International Conference on Enactive Interfaces*, Pisa, Italy, 2008.

[35] ——, "Perceptual validation for the development of a sound synthesis algorithm for rolling sounds," *Submitted to IEEE Transactions on Speech, Audio and Language Processing*, 2008.

[36] J. Laroche, "The use of the matrix pencil method for the spectrum analysis of musical signals," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 1958–1965, 1993. [Online]. Available: http://link.aip.org/link/?JAS/94/1958/1

[37] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics. Speech, and Signal Processing*, vol. 37, pp. 984–995, 1989.

[38] C. N. J. Stoelinga, D. J. Hermes, A. Hirschberg, and A. Houtsma, "Temporal aspects of rolling sounds: A smooth ball approaching the edge of a plate," in *Acta Acustica united with Acustica*, vol. 89, 2003, pp. 809–817.

[39] N. Lee, Z. Duan, and J. O. S. III, "Excitation signal extraction for guitar tones," *International Computer Music Conference*, 2007.

[40] X. Serra, *Musical Signal Processing*, ser. Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997, ch. Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122.

[41] R. Salami, C. Laflamme, J. Adoul, and A. Kataoka, "Design and description of cs-acelp: a toll quality 8 kb/s speech coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 116–130, 1998.

[42] B. L. Giordano, "Sound source perception in impact sounds," Ph.D. dissertation, University of Padua, Italy, June 2005.

[43] M. Lagrange and B. Scherrer, "Two-step modal identification for increased resolution analysis for percussive sounds," *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2008.

[44] F. Opolko and J. Wapnick, "Mcgill university master samples." [Online]. Available: http://www.music.mcgill.ca/resources/mums/html

[45] N. Fletcher and T. Rossing, *The Physics of Musical Instruments*. New York: Springer-Verlag, 1991.

[46] M. Lagrange and S. Marchand, "Estimating the instantaneous frequency of sinusoidal components using phase-based methods," *in J. of the Audio Eng. Soc.*, 2007.

[47] S. M. Kay, *Modern Spectral Estimation*. Prentice Hall, 1988, ch. Autoregressive Spectral Estimation : Methods, pp. 228–231.

Dr. Mathieu Lagrange (M'07) is a Cnrs researcher at Ircam. He obtained his Ph.D. degree in 2004 at the LaBRI (Computer Science Laboratory). After two years in Canada (University of Victoria and McGill) where he studied model based approaches for the processing of musical and environmental sounds, he joined Telecom PariTech as a Research Assistant within the Quaero Project. His research focuses on structured modeling of audio signals applied to the indexing, browsing, and retrieval of multimedia.

Dr. Gary Scavone is an Assistant Professor of Music Technology at McGill University, where he directs the Computational Acoustic Modeling Laboratory. From 1997-2003, he was Technical Director and Research Associate at the Center for Computer Research in Music and Acoustics at Stanford University, where he received a Ph.D in "Computer-Based Music Theory & Acoustics" and a Master of Science degree in Electrical Engineering. His research includes acoustic modeling, analysis, and synthesis of musical systems and the development of sound synthesis software. Dr. Scavone is also a professional saxophonist specializing in the performance of contemporary concert music.

Dr. Philippe Depalle is an Assistant Professor of Music Technology at McGill University, where he directs the Sound Processing and Control Laboratory.