Perceptual Evaluation of a Real-time Synthesis Technique for Rolling Sounds

Emma Murphy[*], Mathieu Lagrange[*], Gary Scavone[*] , Philippe Depalle[*] ,Catherine Guastavino[*]
(*) Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT),
McGill University, Canada.
E-mail: emma.murphy@mcgill.ca, mathieu.lagrange@mcgill.ca, gary@music.mcgill.ca, depalle@music.mcgill.ca, catherine.guastavino@ mcgill.ca

Abstract

In this study 6 different versions of a new real-time synthesizer for contact sounds have been evaluated in order to identify the most effective algorithm to create a realistic sound for rolling objects. 18 participants took part in a perceptual evaluation experiment. Results are presented in terms of both statistical analysis of the most effective synthesis algorithm and qualitative user comments. Finally recommendations for future implementations of synthesis techniques and subsequent perceptual evaluations are presented and discussed.

1. Introduction

The role of perceptual validation as part of any sound real-time sound synthesis process is crucial for the development of useful algorithms to create sounds with “natural” or “realistic” attributes. The application of contact sounds to virtual reality systems should assume a level of realism in order to create convincing environments for the user. Furthermore in the creation of real-time sound synthesis techniques for new interfaces for musical expression or for the design of enactive interfaces, the perception of naturalness is crucial to create intuitive interactions. In this study a number of synthesis technique versions have been evaluated to identify the algorithm that generated the most natural or realistic sound for a rolling object.

As part of the field of ecological acoustics William Gaver has explored issues in the analysis and synthesis of physical sounds to create effective algorithms for the synthesis of basic-level events such as impact, scrapping and dripping as well as more complex events such as bouncing, spilling and machinery [1, 2]. Furthermore the real-time synthesis of contact sounds has received much attention in the auditory display community and some convincing results have been achieved [3, 4].

There is a significant body of research concerning the identification [5] and classification [6] of everyday and environmental sounds. More specifically, Warren and Verbrugge [7] have investigated the perceptual attributes of breaking and bouncing events from a temporal perspective. Van den Doel et al. [8] have investigated measurements of the perceptual quality of sound synthesis for contact sounds. Furthermore Stoelinga [9] conducted auditory perception experiments investigating the perceptual understanding and evaluation of the direction, size and speed of rolling objects.

This present study details the evaluation of different versions of a sound synthesis technique, which has recently been proposed in [10]. Firstly the synthesis techniques are described in relation to the sound set stimuli under evaluation in the present study. The experimental design and method are presented followed by a discussion in terms of qualitative and quantitative results and recommendations for future work.

2. Real-time Synthesis Technique

Algorithms that have been proposed for the purpose of real-time rolling sounds synthesis [11, 12, 4] are based on empirical settings of the parameters of the algorithm. The algorithm evaluated here is an attempt to overcome this limitation by estimating the synthesis parameters from actual recordings. The analysis method follows a standard source/filter approach where the filter parameters are estimated using a High-Resolution modal analysis technique [13]. This method is processed over a recording of the plate hit by the rolling object. The estimated filter parameters are then considered to parametrize a deconvolution filter implemented as a structure of cosine cells [14]. The rolling sound is then deconvolved using this filter in order to estimate the source signal. A more detailed description of the analysis scheme can found in [10]. This signal is then modeled as a series of triggers of an amplitude modulated impact signal. Determining whether the filter should encode only the damping and frequency of the modes alone or also include the gains is a design issue that is difficult to gage during the synthesis implementation process. Therefore the two cases (referred to as gain/ no gain) are considered in the experiment presented in this study in order to evaluate whether one alternative is better than the other from a perceptual perspective.

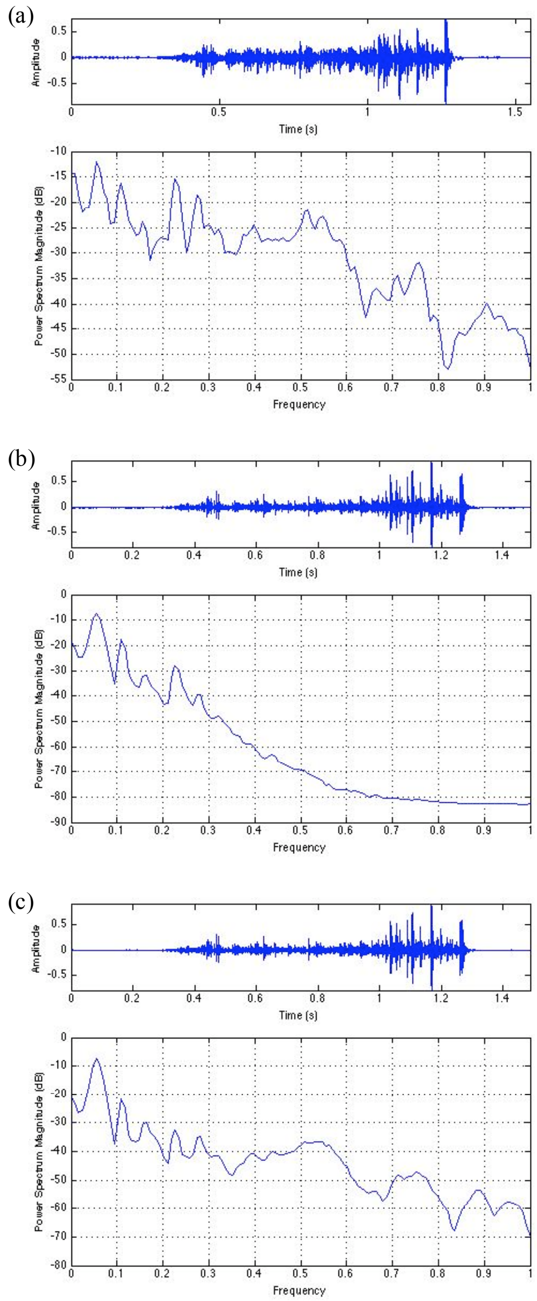

Fig 2: Synthesized sounds of a marble rolling over a highly inclined MDF plate using the three evaluated schemes: (a) impact excitation signal (Raw) (b) Meixner window (c) a combination (Mix)

This impact signal can be of different shapes. One, termed "Raw", is the selected section of the deconvolved signal. Another, termed "Meixner" is the fit of a parametric shape commonly used for the modeling of attacks in the audio coding area, namely the Meixner window [15]. The final signal shape, termed "Mix" is an additive combination of the two previous ones. The intent behind the Mix approach is to balance the properties of the two previous shapes (Raw and Meixner). Indeed, by inspection, it was found that the Raw induces too much high frequency content, whereas the Meixner too much low frequencies, see Figure 1. As shown on the spectral plot of the Mix synthesis, the spectral content is more balanced, approximating more closely the spectral content of the original recording displayed in Figure 2. It was therefore expected that the Mix would be rated rated highest by participants in the perceptual evaluation.

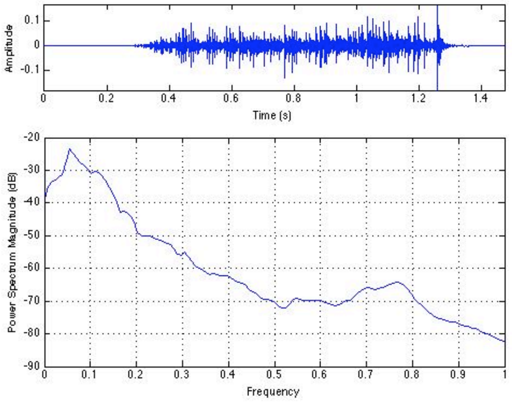

Fig 2: Original recording of a marble rolling over a highly inclined MDF plate

3. Perceptual Validation Experiment

3.1 Stimuli

Impact and rolling sounds were recorded and analyzed to obtain specific model parameters. We considered three different rolling objects: a half liter bottle made of glass (rolling on its side), a small glass marble, and a croquet ball made of wood. The rolling surface was either a medium density fiberboard (referred to as MDF) plate of 95 by 25 by 2 centimeters melaminated or a medium density fiberboard (referred to as Medium) of 80 by 30 by 2 centimeters non-melaminated. Both contact and rolling sounds were recorded, with three different plate inclinations (to vary speed). The sounds were recorded in an IAC double-walled sound isolated booth using both external Behringer omni-directional microphones (ECM 8000).

Table 1: Rolling Stimuli

Synthesis Shape	Filter	Rolling Speed	Material
Raw	Gain	Fast	Bottle, Glass, Wood
	No Gain		
	Gain	Medium	Bottle, Glass, Wood
	No Gain		
	Gain	Slow	Bottle, Glass, Wood
	No Gain		
Mix	Gain	Fast	Bottle, Glass, Wood
	No Gain		
	Gain	Medium	Bottle, Glass, Wood
	No Gain		
	Gain	Slow	Bottle, Glass, Wood
	No Gain		
Meixner	Gain	Fast	Bottle, Glass, Wood
	No Gain		
	Gain	Medium	Bottle, Glass, Wood
	No Gain		
	Gain	Slow	Bottle, Glass, Wood
	No Gain		

Six synthetic versions (3 synthesis shapes with and without gain filters) were evaluated for 2 different plates (MDF and Medium), on 3 different Materials (Bottle, Glass and Wood) at 3 speeds (Fast, Medium and Slow). This gave a total sounds set of 108 sounds; 6 synthesis versions across 18 trials. However one sound set (the bottle rolling on the medium plate at a slow speed) had a significantly lower gain in comparison to the remaining sound sets across the other trials. It was considered that normalizing all sounds could have had an impact on the perceptual judgments of the synthesis techniques under evaluation. Therefore this sound set was excluded from the experiment and users were presented with a total of 102 sounds, six sounds across 17 trials. [1]

The experiment took place in an acoustically treated room and sounds were played back through closed headphones (AKG K271) at a comfortable level on a Mac Pro through a MOTU 828MKII audio interface.

3.2 Experimental Design

18 participants between the ages of 21 and 47 (AV: 27, S.D. 7), students or staff at McGill University were recruited for the perceptual evaluation experiment. On each trial, participants were presented with six sounds. They were asked to first of all listen to all of the sounds and then to rate the extent to which each version sounded like a rolling object. Participants could repeat individual sounds as many times as desired and play counts were recorded. Presentation of sounds within and across trials was randomized.

To indicate a rating for each sound, users were asked to move a slider button for the corresponding sound, over a continuous scale with both numeric (ranging from 0 to 100) and verbal descriptors. Verbal descriptors were adapted from those used in a previous study [16] conducted by van den Doel et al. to evaluate perceptual attributes of liquid sounds. In this present study verbal descriptors for “rollingness” ranged from “Not at all like rolling” (0), “A little bit like rolling” (25), “Somewhat like rolling” (50), “Close to Rolling” (75) to “Exactly like rolling” (100). Participants were encouraged to use the full range of the rating scale. Users were also encouraged to describe the sound or justify their rating in “comments” text boxes provided for each sound (see figure 3).

As many of the synthesized versions for evaluation were not close to the original recordings it was decided that the originals would not be presented for comparison or included as hidden reference signals. Furthermore the aim of the study was to evaluate the sounds in terms of realism rather than replication or quality in comparison to the original. Therefore only the various synthesized versions were presented to participants for evaluation. However participants were also asked to rate the original versions of the sounds as a separate task. This step was taken in case users did not use the full scale to judge sounds then their ratings could then be normalized against their rating of the originals. The testing interface was developed based on a modified version of the MUSHRAM interface [17]). However it should be clarified that the MUSHRA method was not used for perceptual quality evaluation for this experiment. Rather participants were presented with a straightforward comparison task.

In addition to providing verbal comments, participants were asked to complete a post-task questionnaire in which they were asked to describe the differences in the sounds presented in each trial and explain their strategy for rating the sounds.

4. Results

All participants made use of the full range of the rating scales from 0 to 100 and therefore it was not necessary to normalize the ratings using the perceptual data from the original recordings.

The mean results of participant ratings according to synthesis version with standard deviations are

[1] Sounds are available at http://mt.music.mcgill.ca/~mlagrange/enactive/deliverables/2/rolling

illustrated in figure 4. The Meixner shape was rated highest amongst participants and the Raw shape was perceived as lowest on the scale of "rollingness". Interestingly the mean perceptual ratings for the Mix synthesis technique fall between the Meixner and Raw mean ratings. This in fact confirms the actual synthesis process as the Mix shape is a combination of Meixner and Raw shapes.

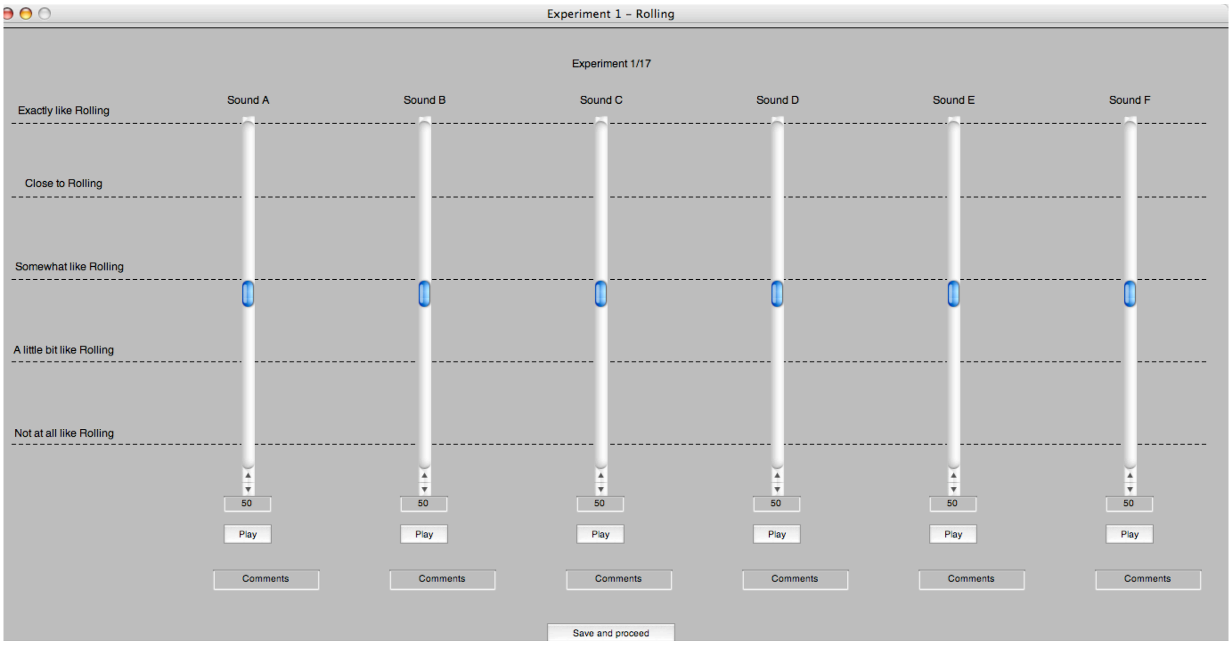

Fig 3: Perceptual Evaluation Testing Interface

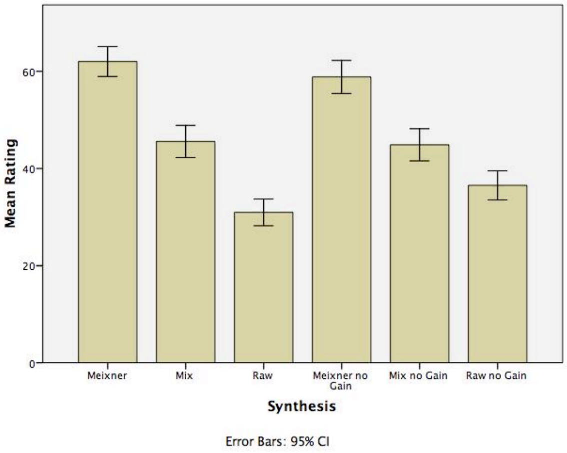

Fig 4: Mean ratings according to synthesis version collapsed over participants, speed and material. A main effect of synthesis shape was observed (p = <0.001) but no significant effect of filter type (gain/no gain).

A 3 (synthesis shape) x 2 (filter – gain, no gain) x 3 (speed) x 3 (material) x 2 (plate) factorial ANOVA revealed main effects of synthesis shape (F(2,1833)=159.4, p = <0.001), material (F(2,1833)=63.4 , p = <0.001), and speed (F(2,1833)=10.7 , p = <0.001) but the effect of both plate and filter (gain/no gain) were non-significant. The following interaction effects were observed among variables: synthesis shape * material (F(4, 3668)=10.5, p = <0.001), synthesis shape * speed (F(4,3668)=4.0, p = <0.01), speed * material (F(4,3668)=6.5, p = <0.001), filter * speed (F(2,3669)=17.5, p = <0.001) and filter * material (F(2,3669)=7.0, p = <0.01). No other interaction effects were observed, specifically no interaction of plate with any other variable was observed.

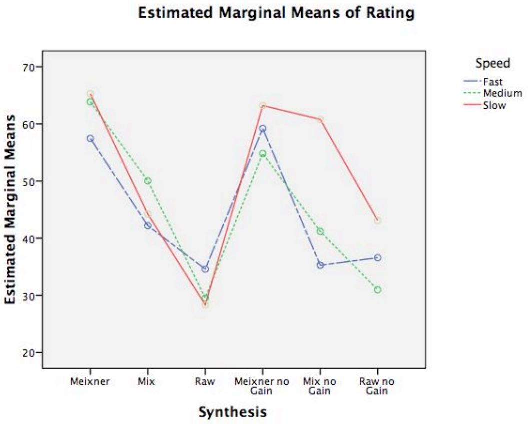

Fig 5: Main effect of speed (p<0.001) on participants ratings grouped by synthesis versions.

Figure 5 illustrates the effect of speed on synthesis version. The slow speed was ranked significantly higher for the Mix shape with no gain. The material type also affected participants' ratings of synthesis version (see figure 6). The wooden material was rated higher for "rollingness" across synthesis versions.

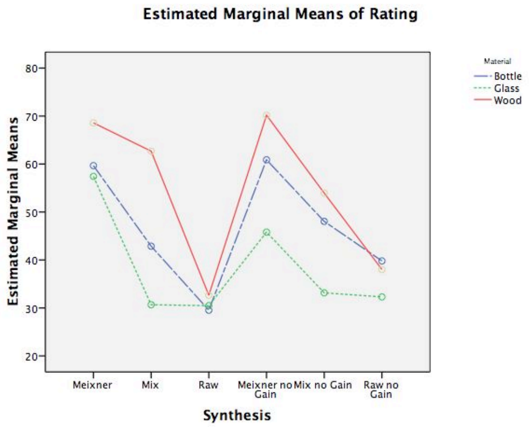

Fig 6: Main effect of material (p<0.001) on participant ratings grouped by synthesis versions.

While there was no overall effect of the gain/no gain filters on participant ratings, the interaction effects between the gain filters and the speed of the objects reveal significantly higher ratings from participants for the no gain synthesis versions at slow speeds (see figure 7).

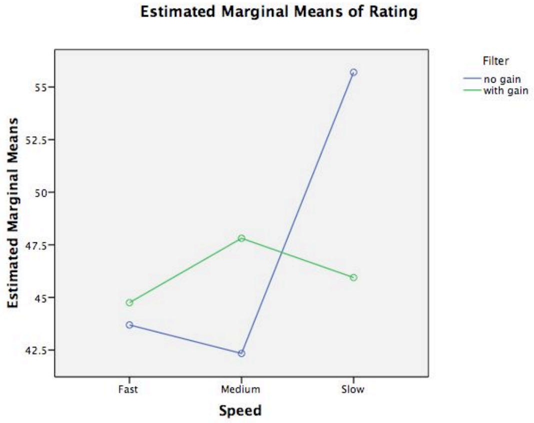

Fig 7: Interaction effect of speed * filter (p<0.001) on participants ratings grouped by rolling speed.

4.1 Qualitative Feedback

While the testing interface enforced participants to listen to each sound and provide a corresponding rating, the provision of individual comments for sounds was optional. Users tended to provide verbal feedback for individual sounds to explain a low ranking usually by commenting on distortions or unusual features perceived. For most sounds presented, participant comments were sparse, however there were a number of particular sounds that elicited verbal comments from a number of participants. For example the glass object rolling at both medium and slow speeds were described by 3 participants as "small objects dropped on a surface", "raindrops", and "sparse". This can be explained by the nature of the excitation generated by the glass object. Being very stiff and small, the glass object bounces, creating approximately periodic high amplitude impulses dominating the smaller ones due to the rolling behavior. Our synthesis algorithm only identifies the prominent impulses and discards the smaller one, which may account for the perceptual attribute as too "sparse".

The wooden object rolling at a slow speed was described as "static", "Random activity (electric activity)" "…too noise-like". At very low speeds, our algorithm fails to distinguish between dominant impacts and smaller ones, modeling them in a similar way. This generates an excitation close to being randomized, hardening the perception of a rolling object. "Static" was also a descriptor associated by 3 participants with the Raw shape.

5. Discussion of Results

It had been hypothesized that the Mix shape would be rated higher by participants in terms of realism as this synthesis technique created more accuracy in higher frequencies. Yet the shape rated significantly higher in the perceptual evaluation was in fact the Meixner shape. Furthermore post-task questionnaires revealed that users attributed higher ratings to sounds that they identified as being "lower pitched". As users were not presented with an original source for comparison their perception of the size and shape of the object was subjective and dependant on the type of object the user identified with individual sounds. Perhaps had the task been a comparison to the original recordings, the outcome may have differed. However the purpose of this perceptual evaluation was to find the best synthesis match for realism and naturalness rather than a comparative analysis to original recordings. The preference for lower frequency rolling sounds could be attributed to participants visualizing larger, heavier rolling objects, such as vehicles.

Sound stimuli with any form of distortion were attributed low rankings by users and automatically excluded from having properties of rolling. The raw shape was ranked lowest by participants and was associated with largest number of verbal descriptors for distortions. This can be attributed to the process of extracting the excitation for the Raw synthesis shape which could be improved for the next implementation.

There was no significant difference between the gain and no gain conditions for all 3 shapes (Raw, Mix, Meixner). This illustrates that, at this stage, there is no clear indication whether the gains of the modes should be modeled in the source or in the filter, which is often a design issue while dealing with source/filter models.

6. Future Work

Since there was a significant interaction effect of both speed and material it would be interesting to conduct a more detailed evaluation in these areas. Participants' preference for the no gain filter technique at slow speeds could also be an issue for further exploration. With a view to applying the synthesis rolling models to a physical task in a virtual environment, it would be interesting to evaluate participants' ability to discriminate between synthesis versions in terms of speed and material type.

As an extension of this study we intend to compare the Meixner synthesis shape, which was rated highest in this perceptual evaluation, with a revised synthesis algorithm. The interface and perceptual method will remain the same for this next auditory perception study in order to refine the synthesis technique. Adding haptic feedback will then be considered to create a more immersive virtual environment and enable exploration into both perception and action interactions for rolling across audio-haptic modalities.

References

[1] W. W. Gaver, What in the world do we hear? An ecological approach to auditory source perception. Ecological Psychology, 5(1), pp. 1-29, 1993.

[2] W. W. Gaver, How do we hear in the world? Explorations in ecological acoustics. *Ecological Psychology*, 5(4), pp. 285-313, 1993.

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

[4] K. van den Doel, P. G. Kry, and D. K. Pai, Foleyautomatic: physically-based sound effects for interactive simulation and animation, *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 537–544, 2001.

[5] J. A. Ballas, Common factors in the identification of an assortment of brief everyday sounds, *Journal of Experimental Psychology: Human Perception and Performance*, Aug Vol 19(4), pp. 250-267 1993.

[6] C. Guastavino, Categorization of environmental sounds, *Canadian Journal of Experimental Psychology*, 60(1), pp. 54-63, 2007.

[7] H. Warren, and R. R. Verbrugge, Auditory perception of breaking and bouncing events: A case study in ecological acoustics, *Journal of Experimental Psychology; Human Perception and Performance*, 10, pp. 704-712, 1984.

[8] K. van den Doel, D. K. Pai, T. Adam, L. Kortchmar and K. Pichora-Fuller, Measurements of perceptual quality of contact sound models. *Proceedings of the International Conference on Auditory Display*, Kyoto, Japan, pp. 345—349, 2002.

[9] C. Stoelinga, A psychomechanical study of rolling sound, *PhD Thesis*, Technische Universiteit Eindhoven, 2007. Available: http://pastel.paristech.org/2368

[10] 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,* Paris, France, 2008.

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

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

[13] R. Badeau, G. Richard, and B. David, Performance of esprit for estimating mixtures of complex exponentials modulated by polynomials, IEEE Transactions on Signal Processing, vol. 56, pp. 492–504, 2008.

[14] 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. Effects Conference, 2008.

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

[16] K. van den Doel, Physically based models for liquid sounds, *ACM Transactions on Applied Perception.* 2 (4) 2005.

[17] E. Vincent, M. G. Jafari, and M. D. Plumbley, Preliminary guidelines for subjective evaluation of audio source separation algorithms, *Proceedings of the ICA Research Network International Workshop*, Liverpool, UK, 2006.