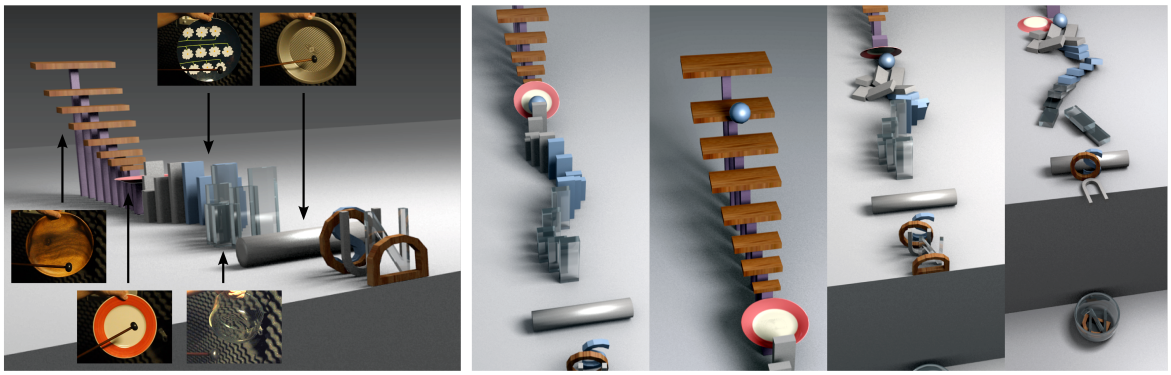
AudioCloning: Extracting Material Fingerprints from Example Audio Recording



(a) (b)

Figure 1: *From real-world recordings the material parameters are estimated. The estimated parameters are applied to virtual objects of various sizes and shapes, generating sounds corresponding to all kinds of interactions such as colliding, rolling, and sliding.*

1 Introduction

Incorporating sound in a virtual environment application is critical, given its diverse applications such as video games, computer animation, and feature films. Sound effects are usually produced from either audio recordings or physically-based simulations. The latter offers greater flexibility, but one key challenge is to determine the material parameters, such as stiffness, density, or damping coefficients. Finding a satisfactory set of simulation parameters that recreate realistic audio quality of sounding materials is a time-consuming and non-intuitive process. In the case of highly complex scene consisting of many different sounding materials, the parameter selection procedure quickly becomes prohibitively expensive and unmanageable.

In this work, we propose the first complete system that is able to automatically estimate the material parameters from a single recorded audio clip. The estimated material parameters can be directly used in an existing sound synthesis framework to generate sounds that preserve the intrinsinc audio quality of the original recording of the sounding material, while naturally varying with different geometries and physical interactions. We also present a method to compute the residual, i.e. the differences between the real-world recording and modal-synthesized sounds, and transfer it to various virtual objects at run time, thereby allowing automatic generation of more realistic sounds. Both the estimated material parameters and residuals can be stored in a 'material database' for future reuse and provide an excellent starting point for foley (sound) artists to further fine-tune desirable sound quality of materials for creative expression.

2 Our Approach

The complete pipeline of our approach consists of the following stages.

Feature Extraction: Given a recorded impact audio clip, we first extract some high-level *features*, namely, the frequencies, dampings, and initial amplitudes of a set of damped sinusoids that collectively represent the impact sound.

Parameter Estimation: A virtual object of the same size and shape as the real-world physical object used in the audio recording is constructed, and an impact is applied at the same location. By assuming a set of material parameters, the sound generated by the impacted virtual object, as well as the feature information of the resonance modes can be determined by modal synthesis techniques (see the supplementary document for detail).

We then use a difference metric designed based on *psychoacoustic* principles to compare the synthesized sound and its features with the recorded sound and extracted features. The optimal set of material parameters is thereby determined by minimizing the error metric function. Using the same set of material parameters, a different set of modes and excitations can be found for objects of different geometries and run-time dynamics.

Residual Compensation: The final stage accounts for the residual to increase realism. First, the residual is computed from the example recording and the synthesized audio using the material parameters found in the previous stage. Then at run time, the residual is transfered to various virtual objects based on the transferring of frequency modes. The final synthesized sounds (modal components plus residual difference) are slightly different from the example recordings, due to geometric discretization required for sound simulation on computers, but they preserve the intrinsic audio quality of the original sounding material.

3 Results

Figure 1 demonstrates our framework applied to a complex dynamic scene consisting of several virtual objects. From audio recordings of striking five real-world objects of different materials (clockwise from top: a plastic plate, a metal plate, a glass bowl, a porcelain plate, and a wood plate), the material parameters are estimated and extracted. Once estimated, these material parameters can be used in an existing sound synthesis framework and automatically create realistic sound effects in real time. The generated sounds preserve the intrinsic audio qualities of these different materials, while capturing the variety of sizes and shapes of the virtual objects, as well as their rich interactions such as colliding, rolling, and sliding (Figure 1b). Please refer to the supplementary video and document for more detail.