CHAPTER 1: BACKGROUND

In this chapter, I review work related to the main aspects of this research. I also provide a mathematical background of modal sound synthesis.

1.1 Sound Synthesis

Sound synthesis techniques recreate natural sounds for virtual environments. Sounds are dynamic and can be created by a variety of sound sources. Different types of sound sources produce different types of sounds, so different models are needed. Examples of sound sources that have been modeled are liquids (Langlois et al., 2016; Moss et al., 2010), paper (Schreck et al., 2016), and fire (Chadwick and James, 2011).

In this dissertation, the focus is on sounds created by rigid objects. Strings and drums can be simulated through physical models such as the Karplus-Strong algorithm (Karplus and Strong, 1983) and digital waveguide synthesis (Smith, 1992). Simple objects with known analytical vibration patterns can be simulated through additive synthesis, where individual sine waves are added together to create more complex sounds (van den Doel and Pai, 1996). Arbitrary rigid objects use the same additive synthesis method, but to determine their frequencies of vibration, or modes of vibration, discretized models of the objects need to be analyzed first (Morrison and Adrien, 1993; O’Brien et al., 2002). This is referred to as modal sound synthesis, consisting of a precomputation step called “modal analysis” and a runtime synthesis step called “modal synthesis”. We now review the details of this method and explain the need for accurate damping parameters.

1.1.1 Modal Analysis

When a rigid object is struck, it vibrates in response, though these vibrations may be imperceptible to the eye. As the surface of the object vibrates and deforms, the surrounding air is rapidly compressed and expanded, creating pressure waves which propagate through the environment. Our ears perceive the variation in air pressure as sound. The standard range of human hearing covers sound waves between 20 Hz and
20 kHz. In modal analysis, the shape and material properties of the object are analyzed to decompose the vibrations into a set of *modes of vibration*. Each mode of vibration describes one independent component of the overall vibration as the object oscillates sinusoidally over time. Each object has a different set of modes depending on the object’s shape and material. Vibrations from an impact can roughly be represented as a linear combination of normal modes with different amplitudes, frequencies, and phases.

Modal analysis is often performed numerically, where the object is represented using a discretized model such as a FEM mesh or spring-mass thin-shell system. Regardless of the choice of discretization, we can consider the dynamics of the system as it vibrates using a system of equations:

\[
M \ddot{r} + C \dot{r} + K r = f
\]

(1.1)

Here, \( r \) is a vector of vertex displacements, where a vector of all zeros represents the object at rest. Since we usually work with three-dimensional objects, an object with \( n \) discrete elements would have a \( r \in \mathbb{R}^{3n} \). \( f \) is the vector of forces applied to each element, inducing vibrations. \( M \) is the mass matrix, which describes the distribution of mass throughout the object. \( C \) is the viscous damping matrix, which describes how the velocity of the elements \( \dot{r} \) decays over time. \( K \) is the stiffness matrix, in which the connectivity of the elements is defined. Given these matrices, we can properly simulate the vibration of the object in response to an impulse. \( M \) and \( K \) can be constructed through knowledge of the shape of the object and its material properties, notably its density, Poisson’s ratio, and Young’s modulus. The damping matrix \( C \), is not as simple to construct.

Modal analysis examines the eigenvalues and eigenvectors of the system in free vibration, that is, with \( f = 0 \) after some initial impulse has been applied. Temporarily ignoring damping, we can set up a generalized eigenvalue problem of the form:

\[
K \nu = \lambda M \nu
\]

(1.2)

Finding this eigendecomposition and combining the eigenvectors into a matrix \( \Phi \) allows the matrices \( M \) and \( K \) to be diagonalized. Specifically, the eigenvectors are mass-normalized such that:

\[
\Phi^T M \Phi = I \quad \text{and} \quad \Phi^T K \Phi = \Omega^2
\]

(1.3)
The matrix $\Phi$ can be intuitively described as a matrix which transforms between object space and mode space: each column of $\Phi$ contains the shape of a normal mode, while $\Phi^T f$ converts forces on elements to normal mode amplitudes. The natural undamped frequencies of the system are contained in the diagonal matrix $\Omega$, while their squares in $\Omega^2$ are the eigenvalues of the system. We can continue the decoupling by considering the system in mode space $z = \Phi^T r$:

$$\Phi^T M \ddot{z} + \Phi^T C \dot{\Phi} \dot{z} + \Phi^T K \Phi z = \Phi^T f$$

$$(1.4)$$

$$\ddot{z} + \Phi^T C \dot{\Phi} \dot{z} + \Omega^2 z = \Phi^T f$$

$$(1.5)$$

Equation (1.5) now runs into problems with the damping matrix $C$. While $\Phi$ diagonalizes $M$ and $K$, if it does not diagonalize $C$ then the system does not properly decouple and the resulting modes are not linearly independent. The linearly dependent modes are called complex modes, and accurately modeling them is much more difficult compared to the linearly independent normal modes (Imregun and Ewins, 1995). We must now consider methods for modeling damping behavior and constructing appropriate $C$ matrices.

### 1.1.2 Damping Modeling

Damping has long been a concern in analysis of vibrations of buildings and other structures (Nashif et al., 1985; Adhikari and Woodhouse, 2001). There are a number of ways to model material-based damping to varying degrees of accuracy (Woodhouse, 1998; Slater et al., 1993), and standard tests have been designed to consistently measure damping in materials (E756, 2017). Complex models are often required to produce accurate fits to observed damping behavior (Adhikari, 2001).

To construct appropriate $C$ matrices, for sound synthesis purposes we restrict ourselves to classical damping with only normal modes, which means all of our damping matrices must be diagonalizable by $\Phi$. Various damping models have been developed which guarantee only normal modes (Caughey, 1960). These damping models typically have real-valued parameters that vary between materials. In this dissertation, $\alpha_j$ is used to represent these damping parameters. However, be aware that each damping model has a different definition of $\alpha_j$.

The most popular model is Rayleigh damping (Rayleigh, 1896), in which the damping is a linear combination of mass and stiffness:

$$C = \alpha_1 M + \alpha_2 K$$

$$(1.6)$$
$\alpha_1$ and $\alpha_2$ are the real-valued parameters in this Rayleigh damping model. Rayleigh damping has been, to the best of our knowledge, the only damping model used for sound synthesis in computer graphics.

Caughey and O’Kelly proposed a more general model, now known as Caughey damping or a Caughey series (Caughey and O’Kelly, 1965), which they proved to be a necessary and sufficient condition for normal modes:

$$C = M \sum_{j=0}^{n-1} \alpha_j (M^{-1} K)^j$$

All $\alpha_j$ are real-valued parameters for Caughey damping models. In practice, the series could truncated after a few terms.

For a given damping model, the real-valued parameters $\alpha_j$ are the damping parameters which define the damping of each mode. By varying these values, the same object can be made to sound like a wide range of materials. Damping parameters have been shown to be perceptually geometry-invariant for a wide range of geometries under the Rayleigh damping model (Ren et al., 2013a); it is reasonable to assume this holds for other damping models as well. Thus, if damping parameters can be estimated for a metal bowl, synthesizing sound for a solid cube with those parameters will produce a metallic sound. However, the geometry-invariance assumption has only been thoroughly tested on thick, very rigid objects (Ren et al., 2013a), and the assumption may fail for thin-shelled objects (Chadwick et al., 2009), less rigid objects, objects with loosely-coupled points of self-collision, or objects demonstrating nonlinear vibrational behavior.

1.1.3 Modal Sound Synthesis

With these damping models, we have a damping matrix guaranteed to be diagonalizable by $\Phi$. With the system diagonalized, the free-vibration form is now decoupled into independent second order differential equations:

$$\ddot{z}_i + c_i \dot{z}_i + \omega_{in}^2 z_i = 0$$

(1.8)

c_i is an entry in the diagonalized damping matrix corresponding to the $i$’th mode of vibration, and is discussed in more detail in Section 1.1.4. These equations each have known analytical solutions as damped sinusoids:

$$z_i(t) = a_i e^{-d_i t} \cos(\omega_i d t)$$

(1.9)
\( a_i \) is the amplitude of the sinusoid, while the damping coefficient \( d_i = c_i / 2 \) defines the rate at which the amplitude decreases. \( \omega_{in} \) in Equation (1.8) is the natural undamped frequency of oscillation, but in the presence of damping we use the damped frequency \( \omega_{id} \):

\[
\omega_{id} = \sqrt{\omega_{in}^2 - d_i^2}
\]  

(1.10)

1.1.4 Obtaining Damping Coefficients

In practice, we do not actually want to perform the matrix operations in the damping models. Through heavy use of Equation (1.3), we can find analytical solutions for how \( C \) is diagonalized and compute \( c_i \) in terms of the corresponding eigenvalue \( \omega_{in}^2 \). The solution for Rayleigh damping is common in modal sound synthesis work:

\[
\Phi^T C \Phi = \Phi^T \alpha_1 M \Phi + \Phi^T \alpha_2 K \Phi
\]

\[
= \alpha_1 I + \alpha_2 \Omega^2
\]

\[
c_i = \alpha_1 + \alpha_2 \omega_{in}^2 \]  

(1.11)

Caughey damping is slightly more involved, but leads to a fairly intuitive solution:

\[
\Phi^T C \Phi = \Phi^T M \sum_{j=0}^{n-1} \alpha_j (M^{-1}K)^j \Phi
\]

\[
= \Phi^{-1} \sum_{j=0}^{n-1} \alpha_j (\Phi \Omega^2 \Phi^{-1})^j \Phi
\]

\[
= \sum_{j=0}^{n-1} \alpha_j \Omega^{2j}
\]

\[
c_i = \sum_{j=0}^{n-1} \alpha_j \omega_{in}^2 j
\]  

(1.12)

Using these solutions, the damping rates for each mode of vibration can be determined.

1.1.5 Modal Synthesis

For real-time synthesis, a preprocessing step is first performed for a given object and material. In this step, the eigendecomposition is performed and the resulting \( \Phi^T \) and each mode’s \( d_i \) and \( \omega_{id} \) are saved.
At runtime, an applied force $f$ is transformed to mode space by $\Phi^T$, and the resulting vector contains the amplitudes with which to excite each mode. The resulting damped sinusoids can be combined and sampled at 44.1 kHz to produce the sound itself. Tools for performing additive synthesis and modal sound synthesis are plentiful; examples include the Synthesis ToolKit (Cook and Scavone, 1999) and the Faust programming language (Michon et al., 2017; Michon and Smith, 2011).

For interactive applications, as a user performs actions to create sounds, sound synthesis algorithms must run fast enough to generate sound in real time. The computation requirements at runtime are proportional to the complexity of the analyzed input shape, making some objects’ sounds too slow for real-time applications without optimizations. Vibration modes can be culled based on psychoacoustic principles, for example, humans cannot tell the difference between two frequencies very close to one another, so those modes can be combined into one (Raghuvanshi and Lin, 2006). If an object has any geometric symmetries, these can be exploited to reduce memory usage and caching requirements (Langlois et al., 2014). Synthesis can be done in frequency space to further improve performance (Bonneel et al., 2008). When performing real-time synthesis, vectorization (van Walstijn and Mehes, 2017) and parallelism on CPUs (Bilbao et al., 2013) and GPGPUs (Webb, 2014) are effective, as each mode of vibration can be synthesized independently.

1.1.5.1 Additional Factors

Modal sound synthesis roughly simulates the sounds produced by rigid, vibrating objects, but in the real world more factors influence the final sound we hear; four such examples are acoustic radiance, sound propagation effects, contacts with other objects, and acceleration sound.

**Acoustic radiance** is the efficiency of propagation for each mode: depending on the shape of an object some modes radiate in different directions with different strengths (James et al., 2006; Li et al., 2015). Once the vibrations transfer to the surrounding air, sound waves bounce around the environment before reaching a listener’s ears.

**Sound propagation** can be simulated most realistically with wave-based simulation (Mehra et al., 2015), but can be simulated more quickly with geometric methods (Chandak et al., 2008; Schissler and Manocha, 2016) and can be coupled with modal sound synthesis (Rungta et al., 2016).

**Contacts with other objects** are common as objects rarely float in midair. These contacts with other objects modify the produced sound and can be accounted for with contact models (Zheng and James, 2011). Interactions between a sounding object and a striking tool can be modeled to better simulate the attack of
the sound (Avanzini and Rocchesso, 2001; Bilbao et al., 2015). Contact modeling can be exploited to create real objects that vibrate only at desired frequencies. An object can be placed on foam blocks, specifically positioned to damp out the undesired frequencies while leaving the desired frequencies alone (Bharaj et al., 2015).

**Acceleration sound** is produced when an object is rapidly accelerated through air, and is perceptually noticeable for very small objects such as dice and keys (Chadwick et al., 2012).

### 1.2 Object Understanding Through Sound

The inverse of the modal sound synthesis problem is to use impact sounds as input to understand the original objects. While material properties can be estimated experimentally with specialized measurement equipment (E756, 2017), impact sounds do not require specialized equipment or trained personnel. A common application is to learn properties of a real-world object in order to resynthesize similar sounds in a virtual environment. Ideally, sounds from struck real-world objects could be used to recreate the shape and material properties of the objects. While the ideal case of using one sound to reconstruct an entire object is known to be underconstrained (Kac, 1966), prior research has explored what information can be estimated under different constraints.

Some methods use a single recorded sound, then apply modifications to create realistic variety in resynthesized sounds. Deterministic features of a sound can be extracted, then stochastic noise can be added to those features to model slight variations (Serra and Smith, 1990). Alternatively, the modal content of a sound can be extracted, then resynthesized, slightly modifying mode amplitudes to create variations (Lloyd et al., 2011).

Other methods use multiple input sounds for a single object, generated by striking the object in known locations. The sounds’ spectral content can be interpolated spatially to approximate hit points at new locations (Pai et al., 2001). The Young’s modulus for small parts of the object can be individually optimized to best match the input sounds, estimating more fundamental material parameters (Yamamoto and Igarashi, 2016).

By estimating material parameters, such as the Young’s modulus, those parameters can be applied to synthesis of sounds for any object with that material. If the exact shape of the struck object is known, material parameters can be estimated from a single recorded impact sound (Ren et al., 2013b).
These methods are often limited in their robustness by using constrained models that do not account for environmental factors or multimodal input. Methods that are capable of estimating material damping parameters only support one material damping model. All methods assume that properties of the recording environment are known or are assumed to be minimal. If both video and audio of the object are available, there are not currently methods available to improve the visual reconstruction of the object using the audio. These limitations are addressed as part of this dissertation.

1.3 Multimodal Interaction with Virtual Objects

Multimodal interaction, in the context of this dissertation, refers to interaction using multiple senses simultaneously. The senses of sight, hearing, and touch are each different interaction modalities. Methods for rendering content for each sense has been independently researched. Since this work focuses on audio, rendering for the sense of hearing has been discussed in the earlier Section 1.1.

Realistic visual rendering has been the focus of the computer graphics field for many decades, and photorealistic visual appearances are possible given talented artists and sufficient computational resources. Many books provide an introduction to the field (Akenine-Moller et al., 2002; Foley et al., 1990). Creating realistic visual appearances in interactive environments in real time is more challenging, but can be accomplished using optimizations. For example, texture mapping uses low-resolution 3D triangle meshes with higher-resolution 2D textures to model detailed objects. Using normal mapping (Blinn, 1978), depth mapping (Oliveira et al., 2000; Policarpo et al., 2005), and other methods (Szirmay-Kalos and Umenhoffer, 2008), performance can be improved.

Haptics refers to interaction using the sense of touch, and focuses on the textures of surfaces. Haptic rendering lets a user feel an object’s shape and textures by applying forces based on point-contacts with the object (Ho et al., 1997, 1999), given appropriate hardware. Normal maps, though originally designed for visual rendering, can be used for haptic rendering (Theoktisto et al., 2010).

For realistic multimodal interaction, it is important that content is not only rendered well for individual senses, but that each sense is consistent with one another. Depth maps can modify contacts between objects, coupling the visual appearances of the objects with their physical movements (NykI et al., 2013). Modal sound synthesis can be coupled with physics simulations to couple the movements of objects and their resulting sounds (O’Brien et al., 2002; Zheng and James, 2011). When multiple objects are in contact, the
long-lasting contacts produce continuous sounds which depend heavily on the objects’ textures, further
coupling motion and sound (Ren et al., 2010). These methods only involve a few interaction modalities each,
and do not use a single representation of surface detail to inform all modalities.

1.4 Visual Object Reconstruction

It is useful to obtain the 3D shape and visual surface texture of a real-world object. This is an important
step in virtualizing an object. This information can be obtained by reconstructing the object from a series of
images looking at the object from multiple angles.

Structure from Motion (SFM) (Westoby et al., 2012; Snavely et al., 2006), Multi-View Stereo (MVS) (Goesele et al., 2007; Seitz et al., 2006b,a), and Shape from Shading (Zhang et al., 1999) are classes of techniques
for obtaining 3D shape information from a set of 2D images. Although these methods alone do not achieve a
segmented representation of the objects within the scene, they serve as a foundation for many algorithms.
Bundle adjustment is used to jointly optimize poses when many images are used as input (Triggs et al., 2000).
RGB-D depth-based, active reconstruction methods can also be used to generate 3D geometrical models of
static (Newcombe et al., 2011; Golodetz* et al., 2015) and dynamic (Newcombe et al., 2015; Dai et al., 2017)
scenes using commodity sensors such as the Microsoft Kinect and GPU hardware in real-time.

Some objects and materials are difficult to accurately reconstruct. Reflective objects have glare which
change in location with the movement of the viewer, while transparent objects make it difficult to determine
depth. Additional input modalities may improve results for these difficult objects. A time-of-flight camera
can correct estimated depth of transparent objects (Tanaka et al., 2017). The dip transform for 3D shape
reconstruction (Aberman et al., 2017) uses fluid displacement of an object to obtain shape information.

Impact sound provides an additional input modality, containing cues about the internal structure of an
object. Environmental scene classification is a related task approached through spectral analysis (Büchner et al., 2005) or convolutional neural networks (Piczak, 2015; Salamon and Bello, 2017), but produces broad
classifications of an entire environment. However, no current methods use impact sounds in particular to aid
in complete shape reconstruction. One goal of this research is to use impact sounds to help determine the
object shape and material in cases where visual methods struggle.
BIBLIOGRAPHY


