INSTITUTE OF PHYSICS PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **69** (2006) 1–78 doi:10.1088/0034-4885/69/1/R01

Discrete-time modelling of musical instruments

Vesa Välimäki, Jyri Pakarinen, Cumhur Erkut and Matti Karjalainen

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, PO Box 3000, FI-02015 TKK, Espoo, Finland

E-mail: Vesa.Valimaki@tkk.fi

Received 31 March 2005, in final form 1 September 2005
Published 17 October 2005
Online at stacks.iop.org/RoPP/69/1

Abstract

This article describes physical modelling techniques that can be used for simulating musical instruments. The methods are closely related to digital signal processing. They discretize the system with respect to time, because the aim is to run the simulation using a computer. The physics-based modelling methods can be classified as mass–spring, modal, wave digital, finite difference, digital waveguide and source–filter models. We present the basic theory and a discussion on possible extensions for each modelling technique. For some methods, a simple model example is chosen from the existing literature demonstrating a typical use of the method. For instance, in the case of the digital waveguide modelling technique a vibrating string model is discussed, and in the case of the wave digital filter technique we present a classical piano hammer model. We tackle some nonlinear and time-varying models and include new results on the digital waveguide modelling of a nonlinear string. Current trends and future directions in physical modelling of musical instruments are discussed.

0034-4885/06/010001+78$90.00

Contents

1. Introduction

Musical instruments have historically been among the most complicated mechanical systems made by humans. They have been a topic of interest for physicists and acousticians for over a century. The modelling of musical instruments using computers is the newest approach to understanding how these instruments work.

This paper presents an overview of physics-based modelling of musical instruments. Specifically, this paper focuses on sound synthesis methods derived using the physical modelling approach. Several previously published tutorial and review papers discussed physical modelling synthesis techniques for musical instrument sounds [73, 129, 251, 255, 256, 274, 284, 294]. The purpose of this paper is to give a unified introduction to six main classes of discrete-time physical modelling methods, namely mass–spring, modal, wave digital, finite difference, digital waveguide and source–filter models. This review also tackles the mixed and hybrid models in which usually two different modelling techniques are combined.

Physical models of musical instruments have been developed for two main purposes: research of acoustical properties and sound synthesis. The methods discussed in this paper can be applied to both purposes, but here the main focus is sound synthesis. The basic idea of physics-based sound synthesis is to build a simulation model of the sound production mechanism of a musical instrument and to generate sound with a computer program or signal processing hardware that implements that model. The motto of physical modelling synthesis is that when a model has been designed properly, so that it behaves much like the actual acoustic instrument, the synthetic sound will automatically be natural in response to performance. In practice, various simplifications of the model cause the sound output to be similar to, but still clearly different from, the original sound. The simplifications may be caused by intentional approximations that reduce the computational cost or by inadequate knowledge of what is actually happening in the acoustic instrument. A typical and desirable simplification is the linearization of slightly nonlinear phenomena, which may avert unnecessary complexities, and hence may improve computational efficiency.

In speech technology, the idea of accounting for the physics of the sound source, the human voice production organs, is an old tradition, which has led to useful results in speech coding and synthesis. While the first experiments on physics-based musical sound synthesis were documented several decades ago, the first commercial products based on physical modelling synthesis were introduced in the 1990s. Thus, the topic is still relatively young. The research in the field has been very active in recent years.

One of the motivations for developing a physically based sound synthesis is that musicians, composers and other users of electronic musical instruments have a constant hunger for better digital instruments and for new tools for organizing sonic events. A major problem in digital musical instruments has always been how to control them. For some time, researchers of physical models have hoped that these models would offer more intuitive, and in some ways better, controllability than previous sound synthesis methods. In addition to its practical applications, the physical modelling of musical instruments is an interesting research topic for other reasons. It helps to resolve old open questions, such as which specific features in a musical instrument's sound make it recognizable to human listeners or why some musical instruments sound sophisticated while others sound cheap. Yet another fascinating aspect of this field is that when physical principles are converted into computational methods, it is possible to discover new algorithms. This way it is possible to learn new signal processing methods from nature.

2. Brief history

The modelling of musical instruments is fundamentally based on the understanding of their sound production principles. The first person attempting to understand how musical instruments work might have been Pythagoras, who lived in ancient Greece around 500 BC. At that time, understanding of musical acoustics was very limited and investigations focused on the tuning of string instruments. Only after the late 18th century, when rigorous mathematical methods such as partial differential equations were developed, was it possible to build formal models of vibrating strings and plates.

The earliest work on physics-based discrete-time sound synthesis was probably conducted by Kelly and Lochbaum in the context of vocal-tract modelling [145]. A famous early musical example is 'Bicycle Built for Two' (1961), where the singing voice was produced using a discrete-time model of the human vocal tract. This was the result of collaboration between Mathews, Kelly and Lochbaum [43]. The first vibrating string simulations were conducted in the early 1970s by Hiller and Ruiz [113, 114], who discretized the wave equation to calculate the waveform of a single point of a vibrating string. Computing 1 s of sampled waveform took minutes. A few years later, Cadoz and his colleagues developed discrete-time mass–spring models and built dedicated computing hardware to run real-time simulations [38].

In late 1970s and early 1980s, McIntyre, Woodhouse and Schumacher made important contributions by introducing simplified discrete-time models of bowed strings, the clarinet and the flute [173, 174, 235], and Karplus and Strong [144] invented a simple algorithm that produces string-instrument-like sounds with few arithmetic operations. Based on these ideas and their generalizations, Smith and Jaffe introduced a signal-processing oriented simulation technique for vibrating strings [120, 244]. Soon thereafter, Smith proposed the term 'digital waveguide' and developed the general theory [247, 249, 253].

The first commercial product based on physical modelling synthesis, an electronic keyboard instrument by Yamaha, was introduced in 1994 [168]; it used digital waveguide techniques. More recently, digital waveguide techniques have been also employed in MIDI synthesizers on personal computer soundcards. Currently, much of the practical sound synthesis is based on software, and there are many commercial and freely available pieces of synthesis software that apply one or more physical modelling methods.

3. General concepts of physics-based modelling

In this section, we discuss a number of physical and signal processing concepts and terminology that are important in understanding the modelling paradigms discussed in the subsequent sections. Each paradigm is also characterized briefly in the end of this section. A reader familiar with the basic concepts in the context of physical modelling and sound synthesis may go directly to section 4.

3.1. Physical domains, variables and parameters

Physical phenomena can be categorized as belonging to different 'physical domains'. The most important ones for sound sources such as musical instruments are the acoustical and the mechanical domains. In addition, the electrical domain is needed for electroacoustic instruments and as a domain to which phenomena from other domains are often mapped. The domains may interact with one another, or they can be used as analogies (equivalent models) of each other. Electrical circuits and networks are often applied as analogies to describe phenomena of other physical domains.

Quantitative description of a physical system is obtained through measurable quantities that typically come in pairs of variables, such as force and velocity in the mechanical domain, pressure and volume velocity in the acoustical domain or voltage and current in the electrical domain. The members of such dual variable pairs are categorized generically as 'across variable' or 'potential variable', such as voltage, force or pressure, and 'through variable' or 'kinetic variable', such as current, velocity or volume velocity. If there is a linear relationship between the dual variables, this relation can be expressed as a parameter, such as impedance $Z = U/I$ being the ratio of voltage U and current I, or by its inverse, admittance $Y = I/U$. An example from the mechanical domain is mobility (mechanical admittance) defined as the ratio of velocity and force. When using such parameters, only one of the dual variables is needed explicitly, because the other one is achieved through the constraint rule.

The modelling methods discussed in this paper use two types of variables for computation, 'K-variables' and 'wave variables' (also denoted as 'W-variables'). 'K' comes from Kirchhoff and refers to the Kirchhoff continuity rules of quantities in electric circuits and networks [185]. 'W' is the shortform for wave, referring to wave components of physical variables. Instead of pairs of across and through as with K-variables, the wave variables come in pairs of incident and reflected wave components. The details of wave modelling are discussed in sections 7 and 8, while K-modelling is discussed particularly in sections 4 and 10. It will become obvious that these are different formulations of the same phenomenon, and the possibility to combine both approaches in hybrid modelling will be discussed in section 10.

The decomposition into wave components is prominent in such wave propagation phenomena where opposite-travelling waves add up to the actual observable K-quantities. A wave quantity is directly observable only when there is no other counterpart. It is, however, a highly useful abstraction to apply wave components to any physical case, since this helps in solving computability (causality) problems in discrete-time modelling.

3.2. Modelling of physical structure and interaction

Physical phenomena are observed as structures and processes in space and time. In sound source modelling, we are interested in dynamic behaviour that is modelled by variables, while slowly varying or constant properties are parameters. Physical interaction between entities in space always propagates with a finite velocity, which may differ by orders of magnitude in different physical domains, the speed of light being the upper limit.

'Causality' is a fundamental physical property that follows from the finite velocity of interaction from a cause to the corresponding effect. In many mathematical relations used in physical models the causality is not directly observable. For example, the relation of voltage across and current through an impedance is only a constraint, and the variables can be solved only within the context of the whole circuit. The requirement of causality (more precisely the temporal order of the cause preceding the effect) introduces special computability problems in discrete-time simulation, because two-way interaction with a delay shorter than a unit delay (sampling period) leads to the 'delay-free loop problem'. The use of wave variables is advantageous, since the incident and reflected waves have a causal relationship. In particular, the wave digital filter (WDF) theory, discussed in section 8, carefully treats this problem through the use of wave variables and specific scheduling of computation operations.

Taking the finite propagation speed into account requires using a spatially distributed model. Depending on the case at hand, this can be a full three-dimensional (3D) model such as used for room acoustics, a 2D model such as for a drum membrane (discarding air loading) or a 1D model such as for a vibrating string. If the object to be modelled behaves homogeneously enough as a whole, for example due to its small size compared with the wavelength of wave propagation, it can be considered a lumped entity that does not need a description of spatial dimensions.

3.3. Signals, signal processing and discrete-time modelling

In signal processing, signal relationships are typically represented as one-directional cause–effect chains. Contrary to this, bi-directional interaction is common in (passive) physical systems, for example in systems where the reciprocity principle is valid. In true physics-based modelling, the two-way interaction must be taken into account. This means that, from the signal processing viewpoint, such models are full of feedback loops, which further implicates that the concepts of computability (causality) and stability become crucial.

In this paper, we apply the digital signal processing (DSP) approach to physics-based modelling whenever possible. The motivation for this is that DSP is an advanced theory and tool that emphasizes computational issues, particularly maximal efficiency. This efficiency is crucial for real-time simulation and sound synthesis. Signal flow diagrams are also a good graphical means to illustrate the algorithms underlying the simulations. We assume that the reader is familiar with the fundamentals of DSP, such as the sampling theorem [242] to avoid aliasing (also spatial aliasing) due to sampling in time and space as well as quantization effects due to finite numerical precision.

An important class of systems is those that are linear and time invariant (LTI). They can be modelled and simulated efficiently by digital filters. They can be analysed and processed in the frequency domain through linear transforms, particularly by the Z-transform and the discrete Fourier transform (DFT) in the discrete-time case. While DFT processing through fast Fourier transform (FFT) is a powerful tool, it introduces a block delay and does not easily fit to sample-by-sample simulation, particularly when bi-directional physical interaction is modelled.

Nonlinear and time-varying systems bring several complications to modelling. Nonlinearities create new signal frequencies that easily spread beyond the Nyquist limit, thus causing aliasing, which is perceived as very disturbing distortion. In addition to aliasing, the delay-free loop problem and stability problems can become worse than they are in linear systems. If the nonlinearities in a system to be modelled are spatially distributed, the modelling task is even more difficult than with a localized nonlinearity. Nonlinearities will be discussed in several sections of this paper, most completely in section 11.

3.4. Energetic behaviour and stability

The product of dual variables such as voltage and current gives power, which, when integrated in time, yields energy. Conservation of energy in a closed system is a fundamental law of physics that should also be obeyed in true physics-based modelling. In musical instruments, the resonators are typically passive, i.e. they do not produce energy, while excitation (plucking, bowing, blowing, etc) is an active process that injects energy to the passive resonators.

The stability of a physical system is closely related to its energetic behaviour. Stability can be defined so that the energy of the system remains finite for finite energy excitations. From a signal processing viewpoint, stability may also be defined so that the variables, such as voltages, remain within a linear operating range for possible inputs in order to avoid signal clipping and distortion.

In signal processing systems with one-directional input–ouput connections between stable subblocks, an instability can appear only if there are feedback loops. In general, it is impossible to analyse such a system's stability without knowing its whole feedback structure. Contrary to this, in models with physical two-way interaction, if each element is passive, then any arbitrary network of such elements remains stable.

3.5. Modularity and locality of computation

For a computational realization, it is desirable to decompose a model systematically into blocks and their interconnections. Such an object-based approach helps manage complex models through the use of the modularity principle. Abstractions to macro blocks on the basis of more elementary ones helps hiding details when building excessively complex models.

For one-directional interactions used in signal processing, it is enough to provide input and output terminals for connecting the blocks. For physical interaction, the connections need to be done through ports, with each port having a pair of K- or wave variables depending on the modelling method used. This follows the mathematical principles used for electrical networks [185]. Details on the block-wise construction of models will be discussed in the following sections for each modelling paradigm.

Locality of interaction is a desirable modelling feature, which is also related to the concept of causality. For a physical system with a finite propagation speed of waves, it is enough that a block interacts only with its nearest neighbours; it does not need global connections to compute its task and the effect automatically propagates throughout the system.

In a discrete-time simulation with bi-directional interactions, delays shorter than a unit delay (including zero delay) introduce the delay-free loop problem that we face several times in this paper. While it is possible to realize fractional delays [154], delays shorter than the unit delay contain a delay-free component. There are ways to make such 'implicit' systems computable, but the cost in time (or accuracy) may become prohibitive for real-time processing.

3.6. Physics-based discrete-time modelling paradigms

This paper presents an overview of physics-based methods and techniques for modelling and synthesizing musical instruments. We have excluded some methods often used in acoustics, because they do not easily solve the task of efficient discrete-time modelling and synthesis. For example, the finite element and boundary element methods (FEM and BEM) are generic and powerful for solving system behaviour numerically, particularly for linear systems, but we focus on inherently time-domain methods for sample-by-sample computation.

The main paradigms in discrete-time modelling of musical instruments can be briefly characterized as follows.

3.6.1. Finite difference models. In section 4 finite difference models are the numerical replacement for solving partial differential equations. Differentials are approximated by finite differences so that time and position will be discretized. Through proper selection of discretization to regular meshes, the computational algorithms become simple and relatively efficient. Finite difference time domain (FDTD) schemes are K-modelling methods, since wave components are not explicitly utilized in computation. FDTD schemes have been applied successfully to 1D, 2D and 3D systems, although in linear 1D cases the digital waveguides are typically superior in computational efficiency and robustness. In multidimensional mesh structures, the FDTD approach is more efficient. It also shows potential to deal systematically with nonlinearities (see section 11). FDTD algorithms can be problematic due to lack of numerical robustness and stability, unless carefully designed.