Introduction

Structured Audio

Digital audio, as it is widely implemented at present, is not at all structured; the representation of audio data as a discrete bit stream is no more accessible or malleable than a recording that is frozen onto a compact disc or digital audio tape. There is little or no access to the actual structure of the sound, therefore there is little that can be done to search, browse or re-purpose the data for applications other than the original intended. What is needed is a method of representation and extraction of the internal content of audio events; thus, we seek to actually represent the salient structure in sound.

The goal of this thesis is to develop a mathematical framework for representing sound events from a structured perspective and to present techniques for the analysis and characterization of everyday sounds, such as those commonly used for sound effects in films, TV shows and computer-based entertainment. We therefore concentrate upon general audio event representation, analysis and synthesis in a manner that facilitates structured re-purposing and control.

Among the advantages of a structured audio representation are controllability, scalability and data compactness. In the following sections we outline the key issues surrounding the use of structured audio synthesis techniques.

Controllability

A structured audio representation is capable of generating audio signals for the many possible states of an object, this is because it is affords an object-based description of sound. For example, sounds in a game may be controlled, at the time of game play, to respond to changing object properties, such as materials, in the artificial environment; objects made from wood, glass and metal would respond differently if either struck by a large metal sword or kicked over by a heavy boot. These different sound actions are possible because structured audio represents sounds by a combinatoric composition of object properties such as large wooden object and small glass object, and action properties such as kick, strike, bounce, scrape and smash.

The representation of sound building blocks is the main difference between audio design using structured audio techniques and stream-based techniques. In a structured audio representation, sounds are produced by programs which are executed by an application. These programs represent the potential high-level structures for a set of elemental materials; for example, the behaviors of bouncing and hitting a ball are represented as different types of high-level structure, iterated versus impact, but their low-level structures are the same. Furthermore, these high and low level structures can be combined in novel ways in order to produce new sounds. Samples, or streams, generally offer little modification and control that can be used for the purposes of representing alternate physical states of an object therefore structured audio is in no way like a stream-based representation. There is a stronger relationship between the underlying physical properties of the modeled sound objects, hence there is control over the sound structure. This relationship is represented by elemental features in sound signals, that we call structural invariants, and modifications of these elemental structures, which is achieved by well-defined signal transformations.

Scalability

Since structured audio representations render a bit-stream from a description of object properties, i.e. the data is represented as audio building blocks rather than sound samples, it is possible to specify different rendering configurations for the final sounding result. For example, a high-end playback machine may be capable of rendering full CD-quality stereo audio with effects processing, and a low-end playback machine may be capable of rendering only mono, sub CD-quality audio. Even though these renderings differ in their bandwidth, they are both produced from exactly the same structured audio representation. Thus scalable rendering configurations are used to adjust the resolution of a structured audio sound track to best fit a particular end-user hardware configuration; therefore distinct multi-resolution audio formats are not required.

Compactness

A structured audio packet is far more compact than stream-based audio packet; in fact, it is generally several orders of magnitude more compact over the equivalent generated stream representation. The compactness of the representation stems from the fact that the data represents the fundamentally most important parts of sound structures. Very often this material is a small collection of filters with very few coefficients and a series of time-varying generator functions which create excitation signals and transformations of the filter structures. The compactness of the representation makes structured-audio a well-suited scheme for distributing audio data over low-bandwidth networks. This basic property can be exploited in order to represent high-quality sound with a very small amount of data. The low-bandwidth data representation is useful for transporting sound over modems or low-density media such as floppy disks and for representing a large amount of data with limited resources; it is standard industry practice for a CD-ROM-based game to restrict audio soundtracks to, say, 15%-20% of the available data space. With such limitations, alternate methods to stream-based audio representation are being sought.

Ideas to be Investigated

Structured Audio Event Representation

The major task for audio event structure representation is to find a method of representing the various parts of a sound event such that we can control the audio content of a given event class. Our representation seeks to identify structural invariants, such as material properties of sound objects, as well as signal transformations for creating new audio events from these elements, such as bouncing, scraping and smashing. Therefore we seek to identify signal models that represent the inherent structures in sound events. These signal models fall into several distinct classes of synthesis algorithms called auditory group models.

Feature Extraction

Natural sounds generally comprise superpositions of many noisy signal components. Often these components are separate parts of a sound generated by independent sub-processes within a sound structure; such elements are statistically independent components. We explore matrix decomposition methods for extracting statistically independent features from time-frequency representations of sound events, the resulting independent components are considered to be the invariant components sought by the sound structure representation methods discussed above.

Structured Re-purposing and Control

With a set of features and a well-defined control-structure representation for sound we then investigate the problem of audio re-purposing and control. We seek to create combinatoric compositions of spectral and temporal features from different sound-events in order to create novel sound events with predictable new features. Our structured audio event representation method, auditory group theory, provides the framework within which to develop the necessary algorithms.

Applications for Structured Audio

Automatic Foley

Among the applications for structured audio representations are sound synthesis engines that are capable of generating sound from scene descriptions. Figure 1 shows a scenario for an interactive game in which a model of a runner provides the parameters for synthesizing an appropriate audio stream to coincide with the action. A small collection of appropriate game parameters, such as ground materials and shoe materials, are passed to a synthesizer which then generates a corresponding sound track. Most audio synthesis techniques that are widely used at present are generally oriented toward speech or music applications. In order to engineer an automatic Foley application, the sound synthesis algorithms must be capable of a representing a much more general class of sounds than existing techniques allow.

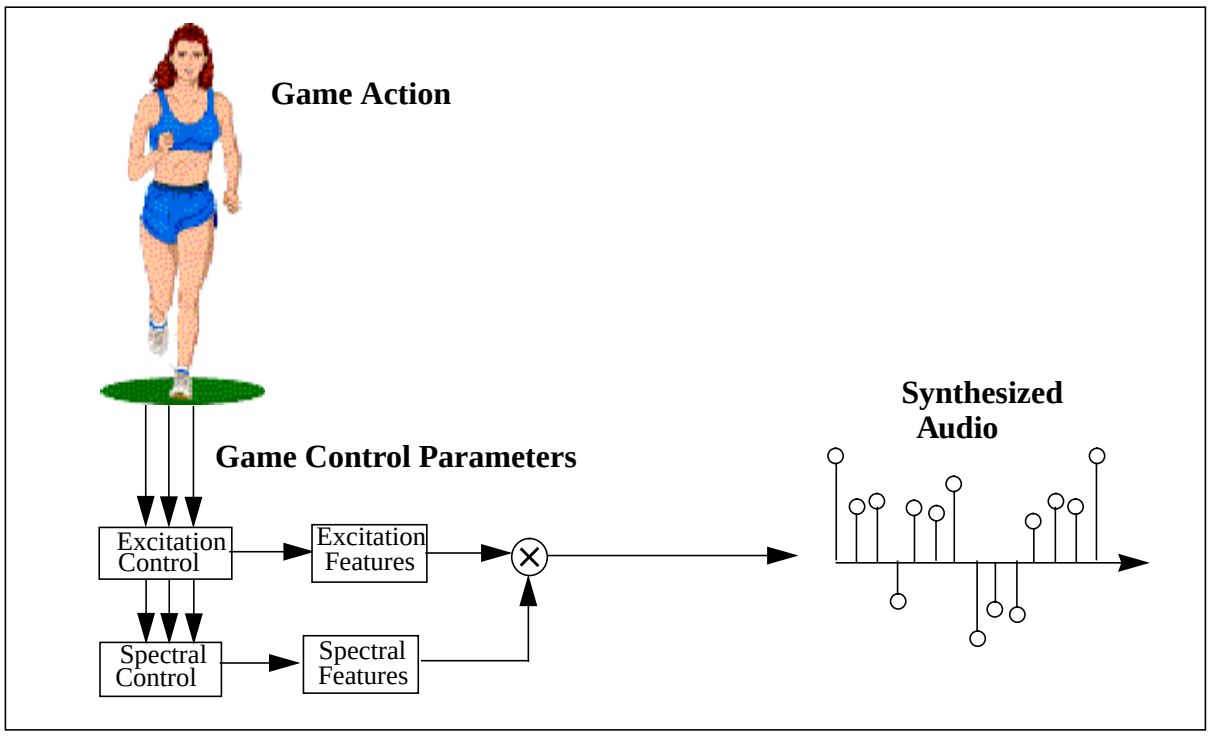

FIGURE 1. Automatic Foley Generation. The audio for an interactive game can be generated from a structured audio description of the materials and action parameters of a scene. This allows automatic synthesis of appropriate sounds such as the footsteps of a runner, which depend on the motion dynamics and mass of the runner, the shoe materials and the material properties of the terrain.

Producer and Sound Designer's Assistant

An extension of the automatic Foley application is the Producer's Assistant. The scenario is here a production environment, such as video or film editing, where a sound-designer creates and places appropriate sounds into an image-synchronized sound track. Instead of a computer program controlling the features of the sounds, a producer could use a set of control knobs to create the sound that best fits the action. The most desirable control pathways for such an application are those that offer physical object properties as handles on the sounds such as materials, size and shape properties.

Low-Bandwidth Audio Representations

Another application for structured audio representations is that of low-bandwidth encoding. A structured audio representation comprises a description of a set of algorithms and procedures that generate sounds as well as the necessary control data for the sounds. The collection of structured

audio packets can be rendered on the client side of a computer network in order to save large amounts of bandwidth during the transmission process. An example of this type of audio encoding is described in Casey and Smaragdis (1996). The use of structured audio packets to transmit audio data for ultra low-bandwidth transmission contrasts with the use of streaming audio packets which contain a time-locked series of compressed audio samples, see Figure 2 and Figure 3.

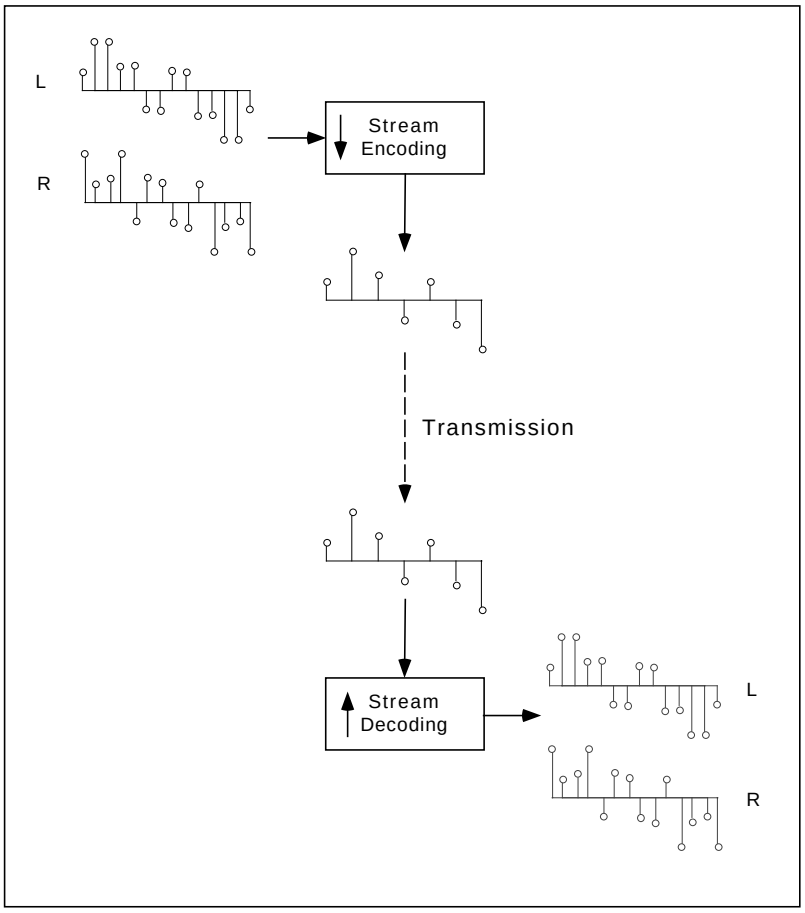

FIGURE 2. Streaming audio flow diagram. An audio source is compressed into a smaller representation using a stream encoder. Encoded streams must be decoded at the receiving end before being rendered.

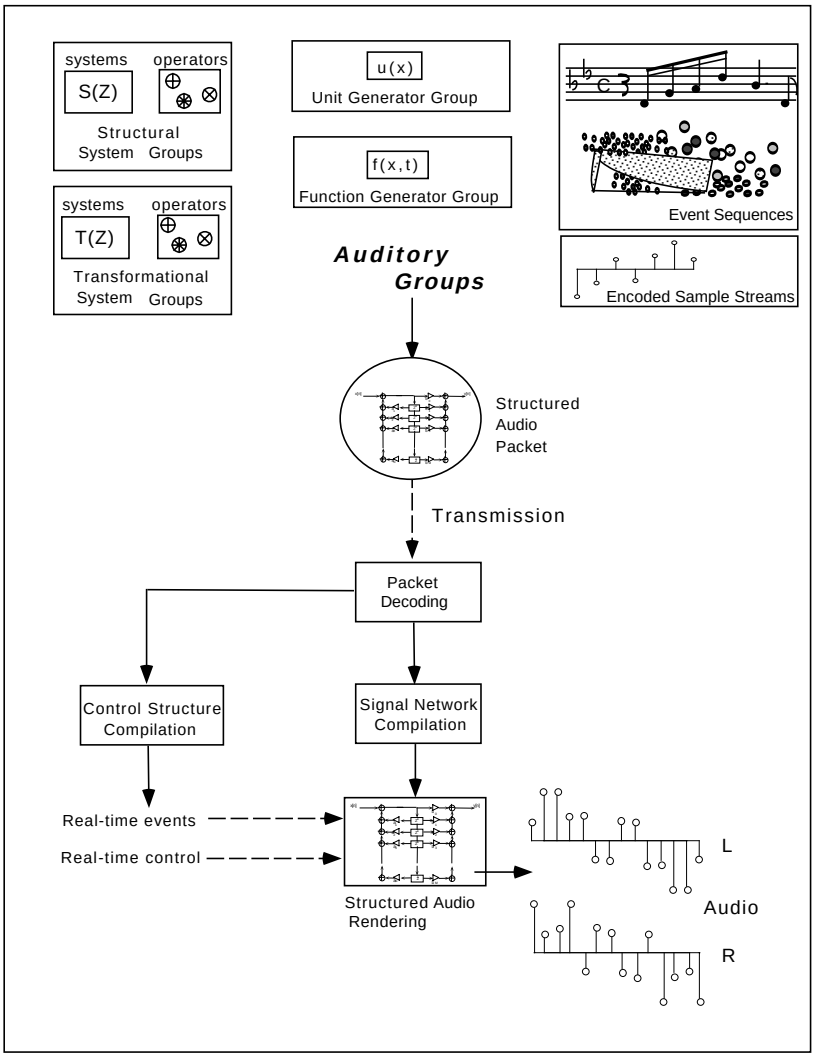

FIGURE 3. Structured audio flow diagram. Structured audio sound events are represented using a combination of elementary building blocks called auditory invariants which can be represented by mathematical groups for the purposes of formal description. It is this representation that is transmitted rather than actual audio data.

Auditory Invariants

To date, there has been no formal analysis of the invariant structure of sound objects and this has hindered progress in the development of new structured analysis/synthesis techniques for audio. A

formal definition of structured audio is necessitated by a growing trend of interactivity and control in media-based systems. Structured representation for the domain of natural sounds, such as a glass smashing or footsteps, for the purposes of sound synthesis have been inadequately addressed by the available literature. As a consequence, current sound production techniques are based on traditional technologies, used for synchronized sound-tracks in film and TV production, and these technologies do not in any way represent the underlying structure of sound objects.

It is our thesis that sound objects can be represented by the combination of archetypal signal building blocks, called auditory invariants and that these invariants and their well-defined transformations constitute a structured audio representation. Furthermore, it is considered that this representation is necessary for the production of perceptually plausible, synthetic sound objects. We define two groups of invariants for sound objects, structural and transformational, as well as operations that can be performed upon them which leave the invariant properties intact. Structural invariants are divided into two groups; *spectral* invariants represent physical system properties of source objects, such as materials, topology and size, *excitation* invariants represent energy-function couplings, such as striking, scraping, and bowing. Transformational invariants are functions that represent higher-order combinations of these structures such as collisions, scatterings, textures, and music. To date there is no analysis/synthesis scheme that is capable of simultaneously characterizing these different levels of auditory structure.

It is shown that, under very general conditions, auditory invariants have *the group property*; hence their combinatoric structures can be usefully subjected to group theoretic analysis. Auditory group theory (AGT) is, then, the analysis of sound-object transformations using auditory invariants. We demonstrate that AGT representation is generally applicable to the formal description of structured audio schemes and, as such, it can be used for the analysis and design of structured audio algorithms, programs and languages.

In addition to presenting the basics of auditory group theory, we also describe methods for the analysis and synthesis of real-world sound objects based on AGT models. The analysis approach is a higher-order statistical generalization of the singular value decomposition (SVD) and it is used to perform a decompositions of time-frequency distributions (TFDs) into statistically-independent components. These statistically-independent components correspond with the structural and transformational invariants of auditory group theory. Syntheses of novel sound objects is achieved directly from the AGT representation by transformations of the analyzed invariants, the resulting elements are signal sequences represented in the complex plane. Efficient filter techniques for implementing AGT synthesis models are investigated. These methods allow the efficient synthesis of novel sound objects by re-purposing of their invariant structures via AGT transformations.

The motivation for the AGT representation comes from observations in the field of ecological acoustics. Ecological acoustics is concerned with the identification of structural and transformational invariants in everyday sounds, such as walking, bouncing and smashing. Evidence for the invariance structure of sound objects is given by previous research on the perception of everyday sounds, the results of which suggest that higher-order structures play an important part in the perception of natural sound events.

Thesis Overview and Scope

As outlined above, the major goals of this work are to define some of the key components of a structured approach to synthetic audio representation and to develop a well-defined mathematical framework within which to implement natural sound-event models. The issues surrounding the representation of natural sound events are complex and subtle and there is, as yet, no prevailing framework within which to represent such audio structures in a general manner. Therefore, in the quest for a structured audio representation method, research from several disciplines is employed.

Chapter 1: Ecological Acoustics

We begin with an overview of work on auditory-event perception from the point of view of ecological acoustics. The framework of ecological perception is concerned with the identification of invariants in the physical world and forming hypotheses on the salience of such invariants from the perspective of human auditory perception. Several studies on the perception of everyday sounds are explored and their results are used to induce a useful background to a theory of natural sound event structures.

Our general approach is motivated, in large part, by previous work in ecological audio perception, the goal of which is to identify structural and transformational invariants among a broad range of sound classes. Previous attempts at characterizing the structure of natural sounds have not had the benefit of a unified mathematical framework within which signals and their transformation structures can be represented. Therefore we seek a precise definition of the signal structure and transformational structure of classes of natural sound events.

Chapter 2: Auditory Group Theory

The main goal of the second chapter is the development of a mathematical framework within which to identify the salient components of sound events. The components of the framework are introduced as those parts of a sound event that are invariant under classes of physical transformations. These signal classes and their corresponding transformations constitute mathematical groups that preserve specifiable structural features of signals under various transformations. We relate these group properties to symmetries in acoustic systems, examples of which are discussed early in the chapter. The relationship between physical symmetries and signal transformation structures provides a well-defined framework for transforming sound features for the purposes of synthesizing novel sounds.

Chapter 3: Statistical Basis Decomposition of Time-Frequency Distributions

The third chapter introduces analysis techniques that are capable of extracting structural invariants from sound recordings under the signal assumptions outlined in Chapter 2. Starting with the singular value decomposition (SVD) we develop an independent component analysis (ICA) algorithm that can be used to extract statistically-independent components from time-frequency distributions

of audio events. This algorithm is capable of revealing independent features in both the spectral domain and the temporal domain and we demonstrate its application to the analysis of several different classes of natural sound. The extracted features are shown to correspond to the components of the auditory group theory models developed in the previous chapter.

Chapter 4: Structured Sound Effects using Auditory Group Transforms

The fourth chapter introduces discrete-time signal processing techniques that enable efficient implementation of structured audio-event models. These models are obtained by estimation of signal parameters from the independent components extracted by statistical basis decompositions. We give several examples of implementations for modeling natural sound events and demonstrate that the structural and transformational properties of our modeling techniques are capable of synthesizing a combinatoric proliferation of plausible auditory events. Furthermore it is argued that these synthesized events are well-formed signals from the perspective of ecological event perception.

Scope of Current Work and Results / Findings

The scope and results of the current work are 1) a new modeling framework for describing auditory events, the application of which encompasses environmental audio, sound textures and the more widely-researched areas of music and spoken utterance, 2) the development of analysis techniques for extracting the salient content of natural sound events from recordings within the framework described above and 3) the implementation of efficient signal-modeling strategies for real-time synthesis models of natural sound events from parametric descriptions of objects and actions.