Chapter II: Auditory Group Theory

2.1 Exploitable Symmetries in Physical Acoustics

No matter how we probe the phenomena of sound, it is, ultimately, produced by physical systems of various assortments. Sound is often, in fact, a by-product of the interactions of many types of complex physical components of events in the natural world, so a complete understanding of the nature of a particular sound stimulus can only be gained from the analysis of the totality of physical interactions of a particular event.This type of analysis has been the subject of many studies in the mechanics of vibratory systems. These studies are applications of Newton's laws from the perspective of multi-dimensional arrays of coupled mass-spring systems whose time-varying deformations, often around an equilibrium state, are transmitted, via some from of boundary constraint, to a transmitting medium such as water, earth or air; for a detailed introduction to Newtonian mechanics and its applications to vibratory systems see, for example, (French 1971; French 1975).

The field of acoustics is primarily concerned with the generation and propagation of air-pressure waves using these mechanical methodologies. We draw from several sources in acoustics in the following section in order to present some examples of how acoustical analyses can shed light on the issue of identifying invariants in sound. For a detailed introduction to the field see for example (Rayleigh 1894; Helmholtz 1885; Fletcher and Rossing 1991).

2.1.1 Physical Modeling of Acoustic Systems

An example of a complex acoustic system is that of the human speech production system. The sound of speech is produced by a glottal source excitation signal, initiated by the forcing of air from the diaphragm-lung system through the windpipe and the vocal folds, and is coupled with the vocal tract, under the control of the speech articulators. The nature of this system can be characterized by arbitrarily complex physical interpretations. For example, the motion of articulators can be modeled, as well as the acoustic properties of the biological structures and tissues that comprise the diaphragm, lungs, wind pipe, glottis, vocal tract, tongue and lips. The air flowing through the system can be modeled as a fluid-dynamical system and the transmission characteristics from the lips to the receiver can also be characterized acoustically. The approach generally used, however, is

that of modeling the broad-band spectral envelope of vowel transitions and the narrow-band spectral envelope of the excitation signal due to glottal excitation.

With such complicated physical sources, as acoustic systems tend to be, it is necessary to perform some level of reduction in the analysis and construction of a suitable model. For the purpose of modeling sound for synthesis, it is often sufficient to identify the most salient degrees of freedom in the underlying system and collect the effects of the remaining components into a single correction component. This is the view taken in speech synthesis, in which the motion of the articulators is considered to bound a series of lossless-acoustic tube areas which are considered the salient components of the vocal tract.

Other examples of the reduction of physical degrees of freedom in an acoustic system are the physical modeling synthesis algorithms for musical instruments; see, for example, (Karplus and Strong 1983; Smith 1990; McIntyre *et al.* 1983). These models often begin with a consideration of the physical systems they are modeling, incorporating analyses of the physics of bowed and plucked strings with that of linear acoustic tubes and resonating bodies. Whilst physicists are concerned with the description of precise mechanisms and elements in acoustic modeling, research into the practical applications of these models is generally concerned with reduction to relatively simple linear systems driven by simple non-linear excitation functions, McIntyre *et al.* (1981). However, such reductions are still physical models and in order to implement a sound synthesis algorithm, all the important elements of a physical sounding system must be represented at some level. This type of modeling leads to an explosion in complexity when moving from, say, a plucked string to a bowed string; or, even more combinatorically implausible, moving from a simple violin model to modeling a Stradivarius.

The issues of complexity related to the modeling of a physical acoustic system are sometimes outweighed by issues of control. Once a physical model has been implemented, with the various degrees of freedom represented as variables, it is the task of the sound designer to essentially "perform" the physical model in order to generate sound. One cannot expect that a realistic violin sound could come from a physical model of a violin whose input parameters are not violin like. Thus a physically modeled system still leaves the task of mapping simple interface variables to more complex physical performance variables; this usually involves additional knowledge of musical performance practice or modeling motor control actions, see Casey (1993, 1994, 1996).

2.1.2 Non-Explicit Physical Characterization of Sound Objects

Whereas the systems cited above lend themselves to physical interpretation, they mediate a concern for the precise description of acoustic systems and the utilitarian needs of producers and musicians. The direct application of physical equations via efficient simulations of solutions to the wave equation leads to reasonable sounding systems for well-defined linear acoustic systems with non-linear excitation functions. However, the strict adherence to various types of wave-propagation schemes fails to recognize the importance of affordance structure in these acoustic systems. We propose that object affordance can be represented and queried in order to generate a plausible sound for a wide variety of physical object interactions. It is precisely this affordance structure that

is missed by the restriction of physical modeling to that of musical instrument systems. Consider the affordance structure of a violin for example, there are many more ways of playing it than most physical models allow for, a cursory glance at a Classical orchestral score shows directions to the string performers such as *sul tasto,* and *sul ponticelli*; indications that the performer is to use different parts of the bow, with different force actions such as bouncing and tremolo, to produce the different sounds. Ultimately, the physical equations can be seen as the most detailed form of investigation that we can analytically apply to an acoustic system, but they are not often applied to modeling the higher-level affordance structures of acoustic systems.

Our concern in this thesis is with the modeling of natural sound phenomena, i.e. non-speech and non-music sounds. As discussed in the previous chapter, there are many ways that an object can be made to create sound; hitting, bouncing, breaking, etc. For a given object, each of these actions results in physical equations that, for the most part, essentially remain the same. So how, then, can we account for the obvious structural differences? Here lies the central issue of this chapter: what is the relationship between the detailed, specific, description of the micro-structure of acoustic systems and the general higher-order behaviors that apply across many different systems? We choose to address this problem using the concepts of structural and transformational invariance that were developed in the last chapter.

2.1.3 Physical Evidence for Auditory Invariants

Wigner, the Nobel laureate in physics, expressed a view on the value of symmetry for the purposes of understanding nature: “There is a structure in the laws of nature which we call the laws of invariance. This structure is so far-reaching in some cases that laws of nature were guessed on the basis of the postulate that they fit into the invariance [symmetry] of structuring.” Shaw et al. (1974). With an appropriately domain-limited interpretation of this faith in the symmetry in natural laws we now offer some observations on sound-generating systems that will lead to a physical argument on the concept of auditory group invariance.

In order to demonstrate mathematical principles of invariance in sound-generating systems we first describe a number of contrasting physical acoustic systems and then proceed to systematically demonstrate principles of invariance across these systems. It is perhaps fitting that we start this section with the description of an acoustic system whose construction and mathematical analysis is attributed to Helmholtz. For, as Ernst Cassirer (1944) notes in his seminal article on the relationship between the group concept and perception, it was Helmholtz who provided “the first attempt to apply certain mathematical speculations concerning the *concept of group* to psychological problems of perception”, in his essay *Ueber die Tatsachen, die der Geometrie zu Grunde liegen* in 1868.

2.1.4 The Helmholtz Resonator

As an example of a well-known and non-trivial acoustic system we consider the Helmholtz resonator. This system is based on the principle of a “spring of air”, which is attributed to Bernoulli, (French 1971; Fletcher and Rossing 1991). The Helmholtz resonator is a system in which a piston of mass m, is free to move in a cylinder of area S and length L. The system vibrates in much the

same manner as the canonical form of a mass attached to a spring which is denoted by writing Newton's second law as:

$$-kx = ma \quad [1]$$

assuming that Hooke's law applies, the system has a restoring force $F = ma$, due to a displacement x, that is proportional to the total spring displacement by a constant factor k called the spring constant. For the Helmholtz resonator the spring constant is that of the confined air:

$$K = \gamma p_a \frac{S}{L}, \quad [2]$$

where γ denotes a constant that is 1.4 for air and p_a is atmospheric pressure. this system is thus a simple harmonic oscillator and its natural frequency is:

$$f_0 = \frac{1}{2\pi}\sqrt{\gamma p_a \frac{S}{mL}} \quad [3]$$

The mass of air m in the neck is a piston and the large volume of air V acts as a spring. The modifications to the cylindrical piston arrangement are given by the terms:

$$m = \rho SL \quad [4]$$

and

$$K = \frac{\rho S^2 c^2}{V} \quad [5]$$

where ρ is the density of air and c is the speed of sound in air. The natural frequency of vibration of the Helmholtz resonator is then given by:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{K}{m}} = \frac{c}{2\pi}\sqrt{\frac{S}{VL}} \quad [6]$$

2.1.5 Modes of an Edge-Supported Rectangular Plate

As an example of a contrasting system with many modes of oscillation we consider a rectangular plate with equal boundary conditions on all four sides. The equation of longitudinal motion of the plate is solved by writing the solution to the wave equation as a product of three functions of single variables, i.e. the planar displacement $Z(x, y, t)$ is written as $Z(x, y, t) = \mathrm{X(x)Y(y)T(t)}$. Following Fletcher and Rossing (1991), the displacement amplitude is given by:

$$Z = A\sin\frac{(m+1)\pi x}{L_x}\sin\frac{(n+1)\pi y}{L_y}, \quad [7]$$

where L_x and L_y are the plate dimensions, and m and n are integers starting at zero. The corresponding modal vibration frequencies are given by:

$$f_{mn} = 0.453 c_L h\left[\left(\frac{m+1}{L_x}\right)^2 + \left(\frac{n+1}{L_y}\right)^2\right], \quad [8]$$

where h is the initial displacement due to an initial force acting on the plate. The wave displacement is similar to that of a rectangular membrane, but the frequencies of vibration are not for the element of stiffness in the plate gives rise to the different modes of oscillation. This accounts for the c_L term in Equation 8 which is given, for longitudinal waves, by the expression:

$$c_L = \sqrt{\frac{E(1-\nu)}{\rho(1+\nu)(1-2\nu)}}, \quad [9]$$

where E denotes the Young's modulus of the plate material and ν is Poisson's ratio ($\nu = 0.3$ for most materials, see Fletcher and Rossing (1991).

2.1.6 The *General Law of Similarity* for Acoustic Systems

The Helmholtz resonator and rectangular plate systems described above are clearly very different types of resonating structures with seemingly little in common in the way of basic mechanical activity. However, there are common invariants across these systems. One such invariant is that produced by the re-scaling of linear dimensions of the acoustic system. This produces a scaling of the natural modes of vibration such that the pattern of relative frequency relations of partials are preserved under the translation, but the absolute values of the partials are shifted by inverse proportion to the scale factor. This makes the pattern of natural mode vibrations of the acoustic system a *structural invariant* and the shift in absolute frequency of modes a *transformational invariant* of the equations of motion for these vibrating systems. Using the systems described above as examples we now consider the action of re-scaling of linear dimensions.

The physical effect of applying a scaling of the linear dimensions of the Helmholtz resonator by a uniform factor K, is given by:

$$f''_0 = \frac{c}{2\pi}\sqrt{\frac{K^2 S}{(K^3 V)(KL)}} = \frac{c}{2\pi K}\sqrt{\frac{S}{VL}} = \frac{1}{K} f_0, \quad [10]$$

where f''_0 is the shifted vibrational mode of the resonator and f_0 is the original frequency of the mode.

This relation expresses a general principle of fundamental importance to the study of invariants in acoustical systems. Namely, that the relationship between the modes of vibration is *invariant*

under the *transformation* of uniform scaling of linear dimensions. Let us investigate this notion further by considering the effect of uniform scaling of the linear dimensions of the supported plate, shown in Equation 8:

$$f'_{mn} = 0.453 c_L h \left[\left(\frac{m+1}{KL_x} \right)^2 + \left(\frac{n+1}{KL_y} \right)^2 \right] = \frac{1}{K} f_{mn}. \quad \textbf{[11]}$$

Again, we see that the proportional relationships between the modes of vibration is preserved under the scaling operation. This transformation is invariant across all acoustic systems under the conditions that the materials remain the same, since the Young's modulus and Poisson' ratio of a material affects the speed of wave propagation in the medium, see (Cremer 1984; Fletcher and Rossing 1991).

So what, then, is the audible effect of this operation? We hear such a scaling as a shift in pitch, but the sound quality or *timbre* of the sound remains the same. We now survey a number of applications of this principle in an effort to demonstrate the broad applicability of timbral invariance to significantly different acoustic applications.

2.1.7 The New Family of Violins

The principle of scale-change invariance has been used to design a new family of violins, each with a different depth of timbre, but each preserving the essential auditory features of a reference violin. The composer Henry Brant suggested to Frederick Saunders and Carleen Hutchins, in 1958, that they design and construct a new family of violins based on scaling of the dimensions of existing violins. The new family would extend the range of the violin family in both frequency directions, high and low, and would cover the entire orchestral range thus creating an extended string-family orchestra- each having its own distinct *timbral depth* but preserving the same basic timbral qualities as the other instruments. The violins were designed and built and in 1965 a full set of eight was used in its first concert performance, (Cremer 1984; Fletcher and Rossing 1991).

2.1.8 Synthesis of Timbral Families by Warped Linear Prediction

A related effect was employed by the composer Paul Lansky for his 1985 piece *Pine Ridge.* As Lansky states, "the starting material for *Pine Ridge* was a tune of 10 notes lasting about 11 sec and played on a violin by Cyrus Stevens.", Lansky and Steiglitz (1981). Lansky built the remaining material for the piece by transforming the timbre of the starting melody via a frequency warping expression in the discrete signal processing domain. A set of filters was estimated at $17.9 msec$ intervals using the covariance method of linear prediction, Markhoul (1975). A unit-sample delay linear predictor over a discrete-time sequence can be expressed as the convolution of past samples with a set of linear constant coefficients:

$$\hat{y}[n] = \sum_{k=1}^{N} a[k] y[n-k]. \quad \textbf{[12]}$$

The coefficients a[k] are obtained by the solution to a set of linear equations in terms of the input and output covariance of the unit-sample delay linear prediction filter. The linear system of equations can be expressed in matrix form as:

$$\mathbf{K}\mathbf{a} = \mathbf{k} \qquad [13]$$

where

$$\mathbf{K} = \mathrm{E_n}\begin{bmatrix} y[n]y[n] & y[n]y[n+1] & \dots & y[n]y[n+N-1] \\ y[n+1]y[n] & y[n+1]y[n+1] & \dots & y[n+1]y[n+N-1] \\ \dots & \dots & \dots & \dots \\ y[n+N-1]y[n] & y[n+N-1]y[n+1] & \dots & y[n+N-1](y[n+N-1]) \end{bmatrix} \qquad [14]$$

is the covariance matrix at a time n over N samples generated by $\mathrm{E_n}[.]$ which denotes the element-wise expectation operator over the same time frame. The sample-delayed covariance vector, **k**, is given by:

$$\mathbf{k} = \mathrm{E_n}\begin{bmatrix} \mathrm{y[n]y[n+1]} \\ \mathrm{y[n]y[n+2]} \\ \dots \\ \mathrm{y[n]y[n+N]} \end{bmatrix}. \qquad [15]$$

Under the condition of the invertibility of the system of linear equations **K** the predictor coefficients a are given by:

$$\mathbf{a} = \mathbf{K}^{-1}\mathbf{k}. \qquad [16]$$

Now, the Z-transform of these coefficients produces the prediction-filter system function which is given by the expression for an N-th order all-pole model of the form:

$$\mathrm{H(Z)} = \frac{1}{\mathrm{A(Z)}} = \frac{1}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{a[k]Z^{-k}}}, \qquad [17]$$

which is a complex function of the complex variable Z. The roots of the denominator polynomial give the poles of the system.

Lansky used a series of filters of this type estimated over windowed portions of the original violin-melody signal at regular time frames of 17.9 msec. In order to reconstruct the violin sound for a given fundamental frequency f_0 an excitation signal with period determined by $\frac{f_s}{f_0}$ is generated using one of a number of generator functions. The simplest functions are those of a band-limited

impulse train with impulses spaced at the desired period, the Z-transform of the synthetic excitation signal is $X(Z)$. The system function for the synthesized violin sound at a particular frame in time for a particular input is now given by:

$$S(Z) = X(Z)H(Z), \quad [18]$$

where $S(Z)$ is the output of the linear-predictive synthesis filter $H(Z)$ in response to the excitation signal $X(Z)$. The fundamental frequency of the synthetic output is controlled by altering the time-structure of the excitation signal to reflect new periodicities. But, more importantly to our discussion, Lansky also used his system to synthesize signals for members of the string-instrument family other than the violin by applying a frequency-warping function to the linear prediction synthesis filter. This frequency-warping filter took the form:

$$W(Z) = \frac{d + Z^{-1}}{1 + dZ^{-1}}, \quad [19]$$

which is an allpass system function. The warped system function of the string-family linear prediction filter is now given by:

$$S(Z) = X(Z)H(W(Z)) = \frac{X(Z)}{\sum_{k=1}^{N} a[k]\left(\frac{d + Z^{-k}}{1 + dZ^{-k}}\right)} \quad [20]$$

The effect of this transformation is to warp the frequency axis by:

$$\phi(\omega) = \omega - 2\tan^{-1}\left(\frac{d\sin(\omega)}{1 + d\cos(\omega)}\right). \quad [21]$$

Since the solution to the prediction filter coefficients, Equation 16, is a least-squares solution of a polynomial function approximator in Z, the roots of the characteristic denominator polynomial occur at frequencies where strong vibrational modes occur in the underlying physical system; which is in this case a violin. Thus the effect of warping the frequency axis shifts the vibrational modes in such a way as to preserve the *relative* modal structure, in terms of products of a fundamental mode, but alter the absolute frequencies of the modes. So for a modal resonance at a frequency ω_0, the frequency warp operation produces a resonance at a new frequency ω'_0 such that:

$$\phi(\omega'_0) = \omega_0. \quad [22]$$

To second-order approximation, $\phi(\omega)$ is linear in the region of the origin of the frequency axis. Thus, in the limit of small ω, the frequency-warping function shifts a modal resonance by the relation:

$$\omega'_0 \sim \left[\frac{1 + d}{1 - d}\right]\omega_0. \quad [23]$$

Lansky chooses the relationship between ω'_0 and ω_0 for each modal resonance in the linear predictor system function to reflect the timbral shift of different members of the violin family. Specifically, filters for the different members of the violin family are obtained by the following table of relations (based on the tannings of the instruments with respect to the violin):

TABLE 1. Violin Modal-Frequency Warping for Lansky's String-Family LPC Filters.

String-family Instrument	Relative pitch to violin (semitones)	Warped pole frequencies $\phi(\omega_0)$	Warp coefficient d	Equivalent linear dimension re-scaling
Viola	-7	$2^{-\frac{7}{12}}\omega_0$	0.19946	1.498
'Cello	-19	$2^{-\frac{19}{12}}\omega_0$	–0.49958	3.000
Bass	-27	$2^{-\frac{27}{12}}\omega_0$	–0.65259	4.757

The last column of the table is our estimate of the linear-dimension re-scaling for the underlying physical system implied by the frequency-warping transform. This factor assumes that the underlying physical change is a simple uniform re-scaling of the linear dimensions of the violin. We can see that this re-scaling occurs in roughly equal additive factors of 1.5 with respect to the size of the reference violin. Therefore Lansky's operation of warping the frequency axis in order to produce new string-family timbres is an approximation of *exactly* the same transformation that was applied by Hutchins in order to create a new family of violins. The common principle they share is the general law of similarity of acoustic systems, Fletcher and Rossing (1991).

So far we have seen this principle applied to simple acoustic systems, such as the Helmholtz resonator, as well as more complex systems such as vibrating plates and string-family instruments. The result has been consistent for each of these systems; namely, the transformation produces a shift in frequency of the vibrational modes of the underlying system but leaves the shape of the *spectral envelope* of the system unaltered with respect to a $\log$ frequency axis.

2.1.9 Gender Transforms in Speech Synthesis

A similar principle has been applied in several studies on modifying speech analyses for the purposes of producing gender transforms in speech synthesis. The basic mechanism is similar to the transformations used for both approaches to violin re-purposing described above. The perceptual studies of Slawson (1968) suggest that the perception of vowels is slightly dependent upon the fundamental frequency of the excitation signal. Slawson reported that for a shift in fundamental frequency by a ratio of 2:1, vowel-quality is perceived as similar when a corresponding shift in formant center frequency of about 10% was introduced. Plomp (1970) interprets this result as a natural prominence for hearing slightly higher resonant modes for higher fundamental pitches due to

gender differences in speech. We can account for the slight transposition of formant center frequencies by a statistical shift in volume of resonant cavities between male and female speakers. Thus the law of general similarity of acoustic systems holds in the realm of speech perception as a cue to gender identification.

Several signal processing algorithms have been proposed that utilize the symmetry of transformations in order to manipulate speech analyses to give the impression of cross-gender or inter-age transforms, (see, for example, Rabiner and Schafer 1978). Whilst these signal processing strategies may appear *ad hoc* from the point of view of physical interpretation, we can see that their general form relates to shifting the modal frquencies of the resonant cavity component of speech signals that corresponds to the formant regions.

2.1.10 Practical Limits of Linear Dimension Scaling of Acoustic Systems

In the application of the re-scaling transformation to violin building, it was found that the dimensions could not practically be scaled to factors corresponding to the required excitation-frequency shift. Consider, for example, the consequences of a scaling in all dimensions by a factor of 3 - the resulting instrument would be 27 times the weight of the reference violin. Thus a compromise is made by altering the pitch structure by a ratio corresponding to the factor 3 but the actual re-scaling is only a factor of 1.5. Therefore the relation between the modes of the driving excitation signal, due to the motion of the string and the bridge, and the resonating body of the instrument is not preserved under the transformation. However, the resulting sound does exhibit the qualities of a timbre scaling. Now we can see that re-scaling of the resonator component of the system dominates the perception of a timbre transform, but re-scaling of the excitation component, in general, does not.

In the implementation of Lansky, an LPC filter was considered to represent the resonator component, or body, of a violin and a synthetic band-limited impulse-train signal represented the excitation signal due to bowing. The underlying assumption, then, in solving for the linear predictor coefficients is that the resonances of the violin body dominate the frequency-response function that the coefficients estimate. However, as Lanksy himself notes, at high frequencies the excitation signal due to bowing the string has strong modes of vibration which are passed-through by the body-resonance system of the violin. Thus the spectrum of the violin has strong modes of vibration due to its body resonance structure as well as high-frequency bowed excitations. The narrowband nature of periodic excitation functions leads to a thinning of broad spectral resonance at high frequencies. For Lansky, this resulted in an upper formant structure that tracked the resonance of the excitation signal as the pitch of excitation was altered.

2.1.11 Acoustical Invariants

The general law of similarity of acoustical systems is an example of an acoustical invariant. It is a re-structuring of a physical equation that affects the resulting sound waveform in a physically meaningful manner, but leaves a complimentary component of the signal unchanged. As we shall see later, the manner of these transformations permits them to be mathematically considered as

groups; we refer to the collection of such acoustical transformations as an *auditory group*. We now begin to generalize the concept of acoustical invariants and develop the concept of an auditory group.

2.1.12 Force Interactions in Acoustical Systems

Forces between objects in a natural event interact in such a manner as to produce a vibratory result. The manner of forces acting on a body produce a deformation in an object corresponding to both the nature and materials of a source. For the purposes of sound, several sources are common enough in their natural occurrence to comprise a substantial part of the force-interaction group of transformations. We find example listings of members of this group in Gibson (1966), Gaver (1993) and Winham and Steiglitz (1970). For each manner of force interaction we provide a short description in order to define terms.

1. Hitting - impulsive

Force interactions which involve a rapid, sharp discontinuity in the formation of a medium are called impulses. These are produced by actions such as hitting and striking materials. The nature of the deformation is such that the modes of vibration are excited maximally, thus producing a broad bandwidth of excitation in the resulting oscillations.

There are two quantitatively different methods of coupling for impulsive force interactions which are characterized by the basic forms of collisions: elastic and inelastic. An elastic impulsive driving force acts for a short time and produces rapid deformation without further influence on the subsequent vibrations. Conversely, inelastic collisions affect the damping and oscillatory behavior of the vibrating bodies.

Impulsive excitations are distinguished from driving-force oscillations in several important ways. Perhaps the one best known to those who are familiar with piano tuning is that of the hammering of a piano string. The hammer is adjusted to strike at roughly one seventh of the total length of the string. The purpose of this effect is not to excite the seventh harmonic, as one would expect by such an action, but in fact to suppress it by not delivering energy to that mode. This is effected because the position of one-seventh is a node for the seventh harmonic, thus no impulsive displacement produces motion in that mode. The seventh harmonic, especially in lower strings, if very prevalent would be considered in-consonant with the diatonic key structures of western music in equal temperament, French (1971).

2. Rolling - angular friction-based continuant motion

The deformations in a material due to rolling, say a steel rod, are not as impactive as the hitting action described above and they are continuant in time. The vibratory action caused by rolling is the result of complex interactions between the surfaces of contact. In the case of a spherical surface on a flat plate the rolling excitation function is caused by angular forces acting against the friction of surface contact.The driving force of rolling is a continuous input of small surface deformations due to continuous angular momentum and forces from vibrations at the point of contact in the meeting surfaces.

3. Scraping - medium-friction-continuant

Unlike rolling, scraping is produced by a constant-coupled linear deformation of the surface of contact. The presence of friction is enough to cause the build-up and release of potential energies in the motion of the scraping surface. These energies are manifest as small irregular pulses applied to the scraped surfaces. The texture of the surfaces has an effect on the nature of scraping, for example, a smooth surface produces many tiny irregular pertubations which are essentially a series of tiny impulses of random magnitudes. A regularly corrugated surface produces regularly spaced impulses which correspond to a periodic impulse train excitation. An irregularly shaped coarse surface produces a combination of irregular small random-valued impulses along with larger impulses. We may characterize the latter kind of action as a chaotic time series.

4. Rubbing - high-friction continuant

When two surfaces are sufficiently smooth to produce high-friction couplings between them then we witness another type of force interaction, that of rubbing. Consider, for example, rubbing a finger around the lip of a wine glass. The friction between the lip and finger causes a sequence of start-stop pertubations that, at the right wavelengths, build up to a self-sustained resonating oscillation in the acoustic shell of the glass. The lip must vibrate to create boundary conditions for the modes of vibration of the rest of the shell. The motion of the lip due to these boundary conditions becomes a factor in the slipping of the finger against the friction of the glass. Because the motion of the glass lip is periodic after suitable buildup, the resulting friction-based driving function is also periodic at the same rate thus providing a coupling of two harmonic oscillators sharing a common mode of vibration.

It is interesting to note that the motion described above is basically no different from that of the action of a violin bow on a string or even a saw, or that of windscreen wipers rubbing against the glass surface of the windscreen. High-friction forces play an important role in the world of mechanical vibrating systems and account for a large number of sound types in addition to those mentioned above such as squeaky brakes and footsteps on a basket-ball court.

5. Jet Stream - low pressure differential wave front

A jet stream force interaction is a continuous flow of air, perturbed by some mechanism, to produce an alternating force in a resonating chamber. This is the basic mechanism of flute and organ pipes. Each type of pipe interacts in a different way with the jet-stream function, all wind instruments and many natural phenomena such as the howling wind, are described by resonant responses to jet-stream excitations.

The nature of jet streams can be periodic due to a mechanism such as a reed, or quasi-periodic due to turbulence at an edge or surface. An example of a turbulent jet-stream excitation force is that of a flute or recorder. The chaotic nature of the jet stream produces spectral flux in the structure of the resulting sound.

6. Plosion - high pressure differential wave front

A plosion is a form of rapid pressure gradient experienced as a wavefront. Plosions are created by rapid accelerations in air volumes due to a large pressure differential. An example of a plosion is a

balloon burst. The pressure difference between the contained volume of air and the surrounding air creates a potential which is held in equilibrium by the restoring force of the elastic balloon shell. If this shell ruptures, or is punctured, then the potential energy is rapidly deployed into restoring equilibrium in pressure between the contained volume of air at the atmospheric pressure conditions. This results in a shock wave that is sent through the medium. There are many classes of sound for which plosions are a component. These include breaking and cracking sounds because of the sudden release of energy associated with fracture.

2.1.13 Higher-Order Force Interactions

Each of the above force interaction types can be grouped into higher-level interaction behaviors. Consider, for example, the action of dragging a rod across a corrugated surface at a constant velocity. The motion is roughly described by a series of regularly-spaced impulses. Assuming linearity, each of the impulses is essentially an independent impulse which can be considered separately from the others by the law of superposition of linear systems. This interaction then is a form of higher-level structure since it can be constructed out of the combination of a number of lower-level force interactions. The advantage of treating them separately is that their structural relationships are revealed, which is not the case if they are considered as a single sequence. This type of higher-order interaction is called an iterated-impulse and the corrugated-scraping action is a quasi-periodic, roughly-constant amplitude impulse train.

Another class of iterated impulse actions are those of bouncing. Here the distance between iterations is determined by a time-constant which can be expressed in one of several ways. One way of relating this behavior to physical invariants is by bounding the time constant of exponentially-iterated interactions by a constant of elasticity. The resulting iterative behavior is parameterized in terms of an elastic object bouncing action. By decay of both the impulse amplitudes and inter-impulse delay times we can characterize the interactions due to a bouncing, lossy, elastic system. A system whose iterations act in other ways, such as delay times getting longer, or amplitude increasing, must have an external driving force component. It has been shown, by Warren and Verbrugge (1988), that subjects are able to estimate the elasticity of an object from presentation of auditory bounce stimuli. This implies that higher-order structural interactions play an important role in the perception of natural sound events.

Aside from deterministic iterations there are several classes of stochastic iterated sequences that are important for characterizing higher-order structure in physical interactions. The first of these is Poisson shot noise. A Poisson sequence is characterized by a single parameter that determines the expectation of an impulse at a given point in the sequence. The expectation is expressed in terms of a delay from the last generated impulse. The basic form of a Poisson sequence is given by the probability that a number of points $n = k$ occur in an interval of length $t = t_1 - t_2$ and is a random variable of the following form:

$$P\{n(t_1, t_2)\} = \frac{e^{-\lambda k}(\lambda t)^k}{k!} \qquad [24]$$

where the quantity λt is the Poisson parameter and characterizes both the mean and variance of the probability distribution function. Poisson sequences describe, to a very good approximation, scattering behaviors due to plosions.

Many natural sounds exhibit a form of Poisson excitation and a decaying Poisson parameter during the course of the event. This decay in activity roughly corresponds to the decay in impulsive interactions during the course of breaking, smashing and spilling events. The spacing of the impulses is indeterminate but their mean spacing decays exponentially during the course of the event sequence.

For irregularly-textured granular surfaces, the higher-order structure of impulse interactions is one of chaotically-spaced impulses embedded in a sequence of smooth noise due to the contrast in frictions across the surface. Thus the act of dragging a rod across an irregularly-textured surface has a noisy higher-level structure. Different types of noise characterization apply depending on the nature of the texture. For example, a Poisson variable is most characteristic of certain forms of texture due to physical deformation such as crumpling, a Gaussian may best describe a smoothly distributed texture such as sand-paper but a chaotic variable due to a fractal-dimension dynamical state variable may be most characteristic of textures such as those of ceramic surfaces.

These characterizations of higher-order interactions are somewhat speculative. But by consideration of the underlying dynamics of physical systems and the structure of interactions we may find just cause for our adoption of stochastic sequence descriptions of higher-order transformation structures. In all these cases, the forces produce excitations that are high-level with respect to low-level resonance structures. Later in this chapter we will adopt a framework for representing both higher-level structures and lower-level components and transformations for the purposes of formally representing sound structures not only as vibrating systems, but also representation as structured event sequences.

2.1.14 Materials

Many broad effects on vibrating systems are produced by the nature of the materials from which the system is constructed. Perhaps the most general of the effects is that produced under the forces described by Hooke's law. That is, when part of a solid body undergoes a displacement of some nature, about an equilibrium position, the potential forces due to displacement are linearly proportional to it. Thus a constant relates the displacement to a restoring potential. This constant is known as the spring constant and generalizes, in the case of rigid materials, to the quantity of stress/strain called the Young's modulus.

1. Restoring Force

The Young's modulus, then, describes the stiffness or the yielding nature of a material and its potential for restoring its configuration under pertubations that do not take the material beyond its elastic limit. As we shall see, there are many types of restoring force under different forms of displacement stress in a material; such as shear modulus and group modulus. These different mea-

sures of a material's elasticity provide the terms for deriving the equations of motion under various actions such as torsional, longitudinal and transverse displacements.

One of the main effects of a change in stiffness of a material is that the propagation speed of waves in the medium is affected. This is due to the changes in the temporal action of the restoring force under different elastic values. Transverse waves have frequency-dependent propagation speeds thus contributing to an element of temporal dispersion in the spectrum. This is not the case, however, for compressional and torsional waves; those waves created by shears and torsional stresses on a material. The dispersion property is non-uniform when the surface of the material is made to bend thus offering different constraints to different frequencies. A torsional shear, for example, is not a bend in shape, it is a rotational stress. This is best seen for a rod. A steel rod can be vibrated in three basic ways, longitudinally, torsionally and transversely. The first two displacements do not affect the shape of the rod thus they do not affect the round-trip uniformity of wave propagation. The latter effects deformations of the basic shape properties of the rod thus creating different path constraints for differing frequencies, this leads to a dispersive spectral characteristic.

In general materials with a high Young's modulus vibrate more rapidly and decay faster than materials with a low Young's modulus. Thus systems of a particular size and shape will exhibit similar modal characteristics but they will be affected in both their dispersion and fundamental period characteristics by the material stiffness depending on the manner of oscillation within the material.

2. Density

The density of a material determines the inertia, and subsequently the momentum, of the particles in the vibrating system. Increased density means that each mass-spring element has a greater mass per unit volume. This in turn affects the speed of oscillation of the vibrating system, greater mass implies greater period thus lower frequencies of the modes of vibration.

3. Internal Damping

The Young's modulus gives a measure of stress over strain per unit area such that the unit of measurement is $\frac{N}{m^2}$. But there is a time component to elastic behavior caused by an increase in strain after some characteristic time interval τ. The second elastic expansion that this causes is a property of the specific material, it can range anywhere from milliseconds to seconds. In viscoelastic materials this elongation increases slowly but without limit, see (Fletcher and Rossing 1991; French 1971).

In order to represent the property of second elastic strain the Young's modulus is represented as a complex quantity:

$$E = E_1 + iE_2, \qquad [25]$$

the imaginary component represents the second elastic response. The relaxation formula exhibits a peak at the relaxation frequency $\omega = \frac{1}{\tau}$. In the most general case, E_1 and E_2 are frequency depen-

dent. As a function of frequency the most general expression for the decay time for internal damping is:

$$\tau_2 = \frac{1}{\pi f}\frac{E_1}{E_2} \quad [26]$$

where f is a modal frequency. For some materials, such as gut or nylon strings the effect of internal damping can be very strong. For metals such as steel the effect is negligible. These effects of internal damping must be included in physical object-modeling strategies if the resulting sound is to be perceptually matched to the modeled materials.

4. Homogeneity

Many of the properties that we have so far discussed have been assumed to apply uniformly in all dimensions of a vibrating system. Materials that exhibit roughly uniform behavior are called *isotropic*. Whilst this is a good assumption for most materials there are common materials whose mechanical characteristics are different along different dimensions. An example is wood, which is an *orthotropic* material. Wood has different elastic properties in each orthogonal dimension. Hence the elastic modulus for wood is expressed as three elastic moduli of the form:

$$\frac{v_{ij}}{E_i} = \frac{v_{ji}}{E_j}, \qquad i,j \in \{X, Y, Z\} \quad [27]$$

where X, Y and Z are the orthogonal dimensions, E_i and E_j are the elastic moduli of which there are three and v_{ij} and v_{ji} are the six Poisson ratios. The equations of motion for vibrating systems are easily modified by substituting the orthotropic independent values of E and v for their isotropic counterparts. For plates the effect of this transform is to produce two different propagation speeds in the different dimensions:

$$f_{mn} = 0.453h\left[c_x\left(\frac{m+1}{L_x}\right)^2 + c_y\left(\frac{n+1}{L_y}\right)^2\right], \quad [28]$$

where

$$c_x = \sqrt{\frac{E_x}{\rho(1-v_x v_y)}}, \quad [29]$$

and

$$c_y = \sqrt{\frac{E_y}{\rho(1-v_x v_y)}}. \quad [30]$$

Thus isotropic materials exhibit uniform dispersion and wave speed propagation in all orthogonal dimensions and orthotropic materials do not.

5. Material Transformations

If we consider two edge-supported rectangular plates, one made from glass the other from copper, and then proceed to write the equations of motion for each we obtain the formula of Equation 8, for longitudinal waves. Recall that a large part of the expression in the equation represents linear dimensions and we have already seen the effects of their scaling. There are also terms relating to the physical properties of the constituent materials such as Young's modulus and density. These properties are all collected into a single term c_L which represents the propagation speed of a longitudinal wave within the solid medium.Of course other wave types are possible depending on the nature of the interaction force; such as torsional waves, flexual waves, transversal, etc. Recalling the equation for each of these wave types:

$$c_L = \sqrt{\frac{E(1-v)}{\rho(1+v)(1-2v)}} \quad [31]$$

is the speed for longitudinal waves where E is the Young's modulus, v is Poission's ratio, and ρ is the density.

$$c_\tau = \sqrt{\frac{GK_\tau}{\rho I}} \quad [32]$$

is the speed of torsional waves where G is the shear modulus, K_τ is the torsional stiffness factor and ρI is the polar moment of intertia per unit length. Torsional waves in a bar are non-dispersive, so they have a wave velocity that is independent of frequency. In many materials the shear modulus is related to the Young's modulus and Poission's ratio by the equation:

$$G = \frac{E}{2(1+v)}. \quad [33]$$

The equation for longitudinal waves is given by:

$$c_L = \sqrt{\frac{E(1-v)}{\rho(1+v)(1-2v)}} \quad [34]$$

Thus we see that transformations of materials in a modeled sound result, primarily, in transformations of propagation speeds within the material. This, in turn, affects both the frequencies of the modes of vibrations, and the time-constant of damping. The former effect is easy to infer from the change in propagation time, the latter effect occurs due to a difference in the periodic rate of damping at boundary conditions caused by the change in speed of the wave.

2.1.15 Topology and Configuration

1. Medium - Solid, Liquid, Gas

It has been noted by the results of several studies into the perception of environmental sounds that the perceptual characteristics of sounds generated by different media, Solid, Liquid and Gas, have distinct properties and thus rarely get confused. Whereas sounds generated through the same medium have a possibility of confusion, (Gibson 1966; Gaver 1993; VanDerveer 1979).

Rayleigh (1894) notes that several important general properties of vibrations in solids do not generalize to vibrations in liquids or gasses. These differences are primarily in the nature of restoring forces and boundary conditions. In order for a medium to exhibit excitatory behavior it must have the ability to convert kinetic energy to potential energy thus create a restoring potential about an equilibrium, which Rayleigh and others called the virtual velocities of the elements under kinetic displacement. In general, for small displacements and pertubations in solids, Hooke's law determines that the restoring forces are linearly proportional to the displacement. For liquids and gasses, however, the restoring forces operate in much different ways, and the dispersion waves operate in a different manner. Thus the medium of vibration is a very strong characteristic of a physical vibrating system; solids, liquids and gasses have distinctly different properties thus giving distinctly different vibratory characteristics under force displacements.

2. Surface Topology (Rigid and Semi-Rigid Structures)

The surface topology corresponds to the shape and nature of a rigid or semi-rigid mechanical system. Many acoustical studies have been carried out on the nature of the vibratory mechanics of surfaces and structures of different rigid and semi rigid arrangements. Examples are strings, membranes, plates, shells and tubes. The topological structure of a surface determines, to a large degree, the natural modes of vibration of a system, parameterized by the size and material make-up of the system. The parameters affect the transformational components of the equations, but there are several physical properties that are left unchanged between these transformations. The differences in the forms of physical equations due to surface topology are rather complicated due to the different boundary conditions and methods of support. For the purposes of our investigation into acoustical invariants we offer a very general ontology of the physical nature of these forms.

The simplest mechanical vibrating systems from a physical point of view are single particle-spring systems. The study of all other physical topologies for vibrating systems is in terms of these elemental units, and their couplings within a material. Elemental particle-spring systems can be combined in one dimension to form strings and wires, they can be expressed in two-dimensions to form membranes and plates, adding a third dimension create shells, cavities, tubes and other volumetric topologies.The significance of topology from the point of view of affordance has already been discussed previously, but we here re-iterate that a volumetric cavity affords resonance due to jet-streams and plosions in a manner independent of its ability to carry torsional, longitudinal and transverse waves around the shell. Thus the affordance of topology is manifold (*sic*) with respect to its ecological acoustical properties.

3. Size

We have already seen one of the primary effects of size change of a system on its vibrational modes in the form of the general law of similarity, see Section 2.1.6. The fundamental mode of vibration is affected by changes in size, as are the other partials, in a uniform manner.

4. Topological Containments

As mentioned above, topological containments are volumes which form cavities which can contain air or some other transmitting medium such as a liquid or another solid. Consider a wine glass, for example, which has both the acoustic shell property, by way of being smooth in curvature and therefore roughly spherical, and the resonant cavity property. These properties tell us that the glass is capable of longitudinal, torsional and transverse vibrations as well a excitation due to air jets and plosions entering the open end. A near-closed cavity has the spring of air property of the Helmholtz resonator and thus begins to change its characteristic to that of a mass-spring system as outlined in Section 2.1.4.

5. Topological Dicontinuouty

Discontinuouties in topological surfaces often provide additional constraints which govern motion of the surface. for example, it was discovered by Rayleigh that addition of mass to a part of a system would affect the speed of wave propagation by increasing the period, thus reducing the frequency. Modes for which the mass occurs close to a node, a point where there is no motion in the mode, are not affected by the addition of mass thus the speed of modal vibrations is a function of mass layout on the vibrating surface. The converse is also true, that subtracting mass affects a mode is exactly the opposite manner.

Holes are to be considered important in the special case that the topology forms a cavity, for holes provide an impedance boundary to forces external to a spring of air system such as that of the Helmholtz resonator. The effect of this is to reduce the restoring forces at the surface of the shell in an air pocket or cavity which are at a maximum when the internal pressure of the cavity is great with respect to the restoring forces of the shell medium and thus affects the frequency of vibration of the system. An example of this is a tennis ball. When the restoring force of air pressure dominates the terms then the air-spring system dominates the terms of the resulting sound. Now when we consider a glass ball the restoring forces are dominated by the stiffness of the glass, thus the high-frequency glass vibratory modes dominate the resulting sound. Thus the relationship between the topological volume, the surface material stiffness, and the impedance characteristics of a contained air volume all contribute to the sound of topological volumes in complimentary, and predictable ways.

These observations are based on a displacement force acting at the boundary of the shell on either side of the volume surface. The same is true, however, of forces due to air pressure waves travelling inside the volume. If the material is not stiff enough to reflect the pressure waves back to the opening, then no internal oscillation will ensue. Thus plosive and jet-stream excitation forces will produce oscillatory behavior in a resonant cavity depending upon the stiffness of the surrounding walls.

6. Support

The method of support for an object creates additional boundaries in the surface of vibration. For example, a plate supported at its edges produces a very different style of vibration from one which is clamped at the edges. Another example is that of the support of bars and rods. Bars supported at their edges tend to damp out lower frequencies of vibration because the support mass grossly affects the lower modes of vibration. Higher modes have more nodal points thus a single point of support slows down wave propagation at the non-nodal points but it does not critically damp the higher modes. An example of this is the hanging of windchimes. Those supported at their ends have inharmonic and rapidly-decaying partials, those supported at a node in their fundamental mode of vibration have longer-lasting harmonically rich oscillations, Gaver (1993).

2.1.16 The Representational Richness of Affordance Structures

Consider the glass container shown in Figure 6. Physical equations generally describe several types of motion for the physical structure, namely those of transverse pressure wave propagation within the cylindrical tube, as well as torsional, longitudinal and transverse waves travelling within the shell structure. As an example of the complexity of affordance, consider the case where the tube is closed at one end; then the affordance structure of the object becomes very complex. For example, the object affords filling due to its capacity to contain solids, liquids and even gasses. This has a direct affect upon the sounds which the system is capable of generating. Due to the shape properties of the tube, the structure affords rolling, and standing upright as a support structure. Due to its nature as a solid the structure affords hitting, scraping and various other modes of impact excitation. If the object is made out of glass then its structure affords bouncing, due to elasticity of the materials, or if the elastic limit is surpassed under a heavy force the object affords breaking. Each of these states is affected by the conditions of each of the others. For example, the glass bottle may be filled for each of the actions described above and predictably simple consequences result; except in the case of breaking, where the destruction of the container structure results in a spilling of the liquid as well as the particulate scattering of the container itself, see Figure 6.

This example serves to illustrate some of the complexity in affordance structure of objects. The ecological acoustics view discussed by (Gibson 1966; Gaver 1983; Warren and Verbrugge 1988)

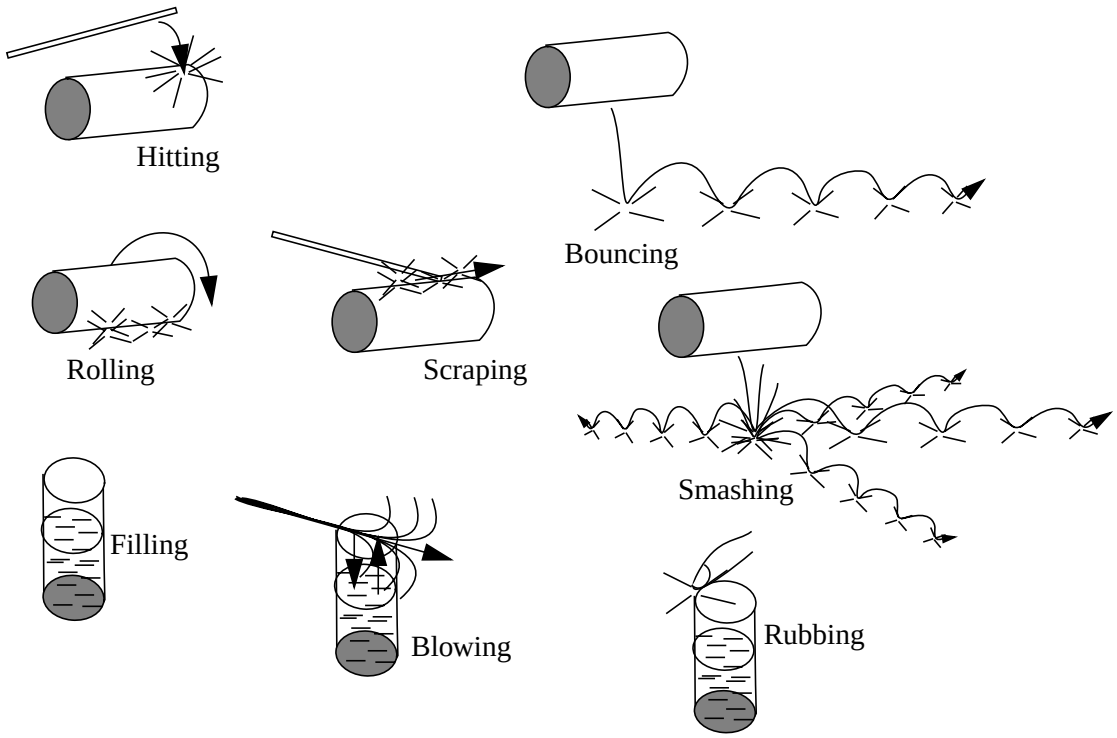

FIGURE 6. The many affordance structures of a glass container. Each of these actions produces a different sound, but the underlying physical object remains the same until fractured.

and others suggests that such affordances are directly perceivable. This view is still contentious so we shall not pursue it here. But it is perhaps useful to note at this juncture that the affordance structure is multiplex and we have no trouble perceptually interpreting these changes under vastly differing circumstances. We should ask the question what are the physical invariances of such affordances? It should be possible, if the direct pickup of information in the environment is governed by laws of invariance in physical systems, to find a set of governing principles which may contribute to the perception of each of the affordance structures outlined above. It is this richness in affordance structure that potentially makes the concept of modeling broad-invariant structure as opposed to detailed physical structure a compelling method for synthesis of realistic sound effects.

Of course, the affordances described for the glass container are not specifically auditory affordances. That is, they specify an underlying physical state which affects all the sensory data that can be perceived from the object. This implies that representation of affordance structure of an object is not domain or medium specific. It is, rather, a general property of the underlying nature of the object and should thus be manifest in these general terms. Consider for example a virtual environ-

ment which is represented, not by visual appearances or auditory behaviors or haptic data, but rather is represented in terms of affordances. Represented in this manner, it is possible to derive the correct visual, auditory and haptic responses for objects in the environment without having to store all the data specific to each mode.

Along with the affordances of containers due to their topology, materials and density we can also describe the affordances of other classes of physical layout. Consider for example the layout of physical membranes and plates (which are both forms of planar surface). The equations governing the propagation of longitudinal waves for these types of layouts have been described above. But there are many possible modes of excitation that could give rise to one of a number of wave propagation structures. If we consider the surface represented by a stretched membrane, such as that of a drum, we can see how this can act both as a surface, a spring board and a container of sorts.

As an application of affordance structures, we propose that objects in a virtual environment could be represented using an affordance graph structure. This graph structure could be queried for inferring many properties, such as possible sound interactions, visual states and properties of interactions with other objects. The consequence of such representation schemes being used as data structures instead of modally specific structures is that entire virtual environment could be rendered without the need for explicit modeling of sensory modes.

2.1.17 The Trace of Physical Symmetries in Auditory Energy Distributions

Upon looking at time-frequency distributions (TFDs) of various sounds we should expect to see the discernible trace of changes in underlying physical structure between slightly differing mechanical systems. In the signal domain, under certain well-defined conditions, we can develop techniques for transforming and manipulating the various aspects of a signal such as the fundamental frequency of excitation, the overall pattern of modal vibrations, the response time of modal vibrations and many others. Some examples of such transformations have already been given, the timbre-warping transformations and gender transformations for speech synthesis in Section 2.1.8 and Section 2.1.9.

These methods suggest that if a signal component can be derived to represent one of the physical properties of an underlying system, then we can transform it according to the principles of physical invariance in order to control the perception of physical object properties. The hypothesis is that by altering signals generated by physical systems in physically plausible ways we will alter the perception of physical objects in a predictable manner. Such manipulations are the goal of the current thesis. We intend to identify invariant components and transformations for real acoustic signals and use these invariants to generate physically meaningful transformations in order to create novel sound structures.

2.1.18 A Theory of Acoustic Information based on Ecological Perception

We have seen in this section that our concern has been with broad generalizations of physical acoustic systems from the perspective of ecological perception, and have shown that much of the structure of the underlying physical equations of nature that are described by Newton's laws, is preserved under broad classes of change, such as size, materials, shape, support and force interactions. Indeed it is the view of ecological perception that the counterpoint of changing and unchanging components of physical structures between different events is precisely what the perceptual systems are most sensitive to. By this view, the information necessary to specify a physical event exists in the energy distribution of an auditory signal in the form of broad classes of invariant properties and transformational structures. Furthermore, it is deemed that the identification of these components and their relationships *specifies* the underlying physical events. Thus the information for acoustic systems lies not in the abstract consideration of atomic, perceptual features in an auditory time-frequency distribution, but perhaps is better characterized as the components of the time-frequency distribution that correlate with physical invariants in the sounding world. This view is suggested by perceptual researchers in the field of ecological perception; (Gibson 1966; Gaver 1993; Warren and Verbrugge 1994; Mace 1977).

Gibson (1966) notes that the ear's evolutionary development is considered as an extension of the staciocyst in small sea-dwelling creatures. The staciocyst is the tiny organ responsible for the vestibular sense, which in small animals is often as simple as determining orientation in the vertical plane due to the effect of gravity on a small mass held inside a sack with sensors attached to the outside. The sensors directly code for the motion of the animal as well as its orientation without the need for higher-level brain functions. These properties are fundamental to the physical world in which the creature lives thus the evolution of such a mechanism perhaps became biologically advantageous as a part of a self-governed propulsion system. It is not, then, so contentious a view that a part of the auditory system may be involved with the direct extraction of physically meaningful data from an auditory signal by the same kinds of mechanisms that are generally considered to code for acoustic environments, such as binaural path delays between the ears and suppression of early echoes occurring after a direct stimulus in a reflective environment. If there are such mechanisms of direct sensitivity to physical invariants in the sounding world in the ear/brain system then the primary information for sound source-event understanding resides in this representation. As a theoretical view of acoustic information, the focus is moved away from the low-level mechanics of the auditory physiology of the ear as providing the primary perceptual cues, toward the physical facts of the vibratory environment as the primary conveyors of ecologically significant acoustical information. As Mace (1977) eloquently articulated as a summary of Gibson's view of perception: "...ask not what's inside your head but what your head is inside of."

2.2 Auditory Group Theory

In this section we develop a mathematical definition of the basic elements of a structured-audio representation using a formal framework for describing invariant structures within sounds. We derive our concepts from the theory of groups which has also been used for delimiting structurally significant elements in the ecological theory of vision, (Gibson 1966; Warren and Shaw 1985). Our treatment of Group theory is not directly related to these visual theories, which rely primarily on groups of affine transformations, but the basic spirit of the treatment is seen to have something in common. Our group concepts are formed out of observations of regularities across several different domains of audio phenomena; namely environmental audio, music and speech.

These groups are defined in terms of sets of elementary signals, represented in the complex domain, transformed by groups of symmetry-preserving transforms. As argued above, it is considered that these elementary sequences and operations correspond to invariants in the perceived structure of auditory objects, i.e. objects that are considered elementary and that cannot be decomposed further without losing their structural semantics.

It is considered that these groups constitute a powerful means of expressing various types of auditory objects and that the uses for such a representation extend beyond the scope of the current work. It is also considered that the elementary sequences and operators defined in this section are to be considered a subset of the available possibilities for sequences and transformations, but that this subset is representative of a large range of auditory objects.

2.2.1 Formal Definition of Group-Theoretic Invariants

We have discussed at some length the merits of adopting a view of perception in which a counterpoint of persistence and change specifies events. We must now take a step further in this direction and consider what, exactly, we mean by persistence and change. It is not enough to merely state that something is persistent or changing, we must propose a formal framework within which we can identify such structures in sound events.

We proceed in this section with the view that a style of change and a mode of persistence are fundamental to the physical characteristic of an observable system, and that any trace of the system in terms of sensory stimuli must reflect various aspects of these underlying characteristics in terms of information which is available to the perceptual system for decoding. Furthermore, a stronger view will be adopted; that this information is sufficient in order to specify a good deal of information about the nature of an event without recourse to inference or memory structures. We do, however, caution that the role of inference and memory is undisputable in many cases; for example, in the case of complex mixtures of events since the structure of the stimulus is corrupted by the interactions of the various elements and the limitations of the sensory decoding apparatus. Another example is the symbol grounding mechanism which allows one to name a structure that has a persistence across a number of events. With this caution in mind we propose a methodology for the definition of invariants in auditory perception. Invariants which specify the structure of events to be decoded by the auditory system.

Invariant structure is specified by an operation which affects a change in the state of a system such that it gives rise to a detectable change in sound structure but which leaves some definable component unchanged. This operation, then, specifies both a changing and an unchanging component of the sound under its action. Formally, let us denote a sound object waveform by the variable $\mathbf{W}$ and a transformation T_E which preserves a part of $\mathbf{W}$ that we shall call $\mathbf{S}$ and alters a complimentary part of $\mathbf{W}$ that we shall call $\mathbf{E}$. We say that T_E is a *symmetry-preserving* operation with respect to $\mathbf{S}$ and a *symmetry-breaking* operation with respect to $\mathbf{E}$. To put this operation in more concise form we can define the relationship via the following implication:

$$T_E\{\mathbf{W}\} \Rightarrow T_E\{\mathbf{E}\} \times \mathbf{S}, \qquad [35]$$

that is, the transformation T_E of a sound object $\mathbf{W}$ implies that some component of $\mathbf{W}$ called $\mathbf{E}$ is changed and some component of $\mathbf{W}$ called $\mathbf{S}$ is left unchanged. Furthermore, the relationship on the right-hand side of the expression is defined in terms of a product, so we are assuming that, in some as-yet undefined domain of representation, $\mathbf{S}$ and $\mathbf{E}$ are factors of $\mathbf{W}$. We now define a second operation on $\mathbf{W}$, denoted by T_S, which performs the complimentary operation with respect to the components $\mathbf{E}$ and $\mathbf{S}$. Thus T_S alters $\mathbf{S}$ and leaves $\mathbf{E}$ unchanged. We shall define the relationships of this operation in terms of its actions on the elements of $\mathbf{W}$ by the implication:

$$T_S\{\mathbf{W}\} \Rightarrow \mathbf{E} \times T_S\{\mathbf{S}\}, \qquad [36]$$

and we interpret this relation in the same manner as Equation 35.

The purpose of these relations for sound object description will become clear presently. We have defined two components of a sound object, each of which is left unchanged by some operation and altered by another operation. The component that remains unchanged under a transformation we shall call the *structural invariant* of that operation and the component that is altered we shall call he *transformational invariant* of that operation. Now, operations which preserve the symmetry of a component are called *symmetric* and operations which destroy the symmetry of some component are called *anti-symmetric*, hence each of the transformations is a symmetry-preserving operation with respect to its structural invariant and a symmetry-breaking operation with respect to its transformational invariant. In addition, each operation's symmetry properties are inverted with respect to each other's structural and transformational invariants. To clarify, if $\mathbf{S}$ is the structural invariant of the operation T_E and the transformational invariant of the operation T_S, then conversely, $\mathbf{E}$ is the structural invariant of the operation T_S and the transformational invariant of the operation T_E. We express the relationship between T_E and T_S as a pair of dual *anti-symmetric* operations with respect to a pair of *dual-symmetric* invariants.

We now propose that invariants $\mathbf{E}$ and $\mathbf{S}$ belong to a group, in the strict mathematical sense of a group, where the operations T_E and T_S are dual anti-symmetric subgroups of the overlying group. It is only in the identification of elements of persistence and in the identification of dual subgroup

operations that produce styles of change that we can expect to uncover the structural nature of sound objects. We propose, then, that *auditory group theory* is concerned with the identification of structures with dual symmetric and anti-symmetric group properties in the sounding world, and that the counterpoint of symmetric invariants and anti-symmetric operations defines the information that is relevant to the perception of auditory events from the point of view of an observer. We recall that the fundamental hypothesis of ecological perception is that "information exists as invariant aspects of these patterns and changes in the energy distribution." Warren and Shaw (1985). Quite simply, we interpret this to imply that, under a wide range of conditions, the signal emitted by an event presents the necessary information to determine that such an event indeed happened; however this must be interpreted as being plausible only in the absence of destructive effects imposed by sensory transduction machinery such as masking effects, and only when the energy distribution affords interpretation as an un-corrupted event whole, i.e. such that no occlusion or partial cancellation of the event has destroyed the energy distribution. We shall assume that such principles of well-formdness of an energy distribution hold with respect to an underlying event in the physical world.

2.2.2 Representation of Auditory Group Invariants

So how do we identify and represent auditory invariants? We now define a method using local Lie groups in order to represent transforms with desirable symmetry properties. These transforms are represented by functions called Lagrangians, we use the Lagrangian to obtain locally-linear partial differential equations. The solution of these equations enables us to determine the form of a function with the specified symmetry properties in terms of the Lagrangian partial derivatives. For a derivation of the representation of local Lie group transformations see Appendix 1. The following section follows Moon (1996).

2.2.3 The Local Lie Group Invariance Theorem

The functional of a transformed variable x' is represented by the integral equation:

$$J(x') = \int_a^b L\left(t', x(t'), \frac{d}{dt'}x'(t')\right)dt' \qquad [37]$$

This functional is invariant under a local Lie group T_ε only when the following relation holds up to first-order terms in ε:

$$J(x') = J(x) + o(\varepsilon) \qquad [38]$$

By solving the derivative of Equation 37 for the Lagrangian functional, $J(x)$, we arrive at a theorem which states that $J(x)$ is invariant under the local Lie group T_ε with generators τ, ξ and η if and only if:

$$L_t\tau + L_x\xi + L_{\dot{x}}\eta + L\dot{\tau} = 0. \quad [39]$$

This notation describes, in terms of partial derivatives, those parts of the Lagrangian which are affected by the local Lie group transform and those parts remain unaffected or *invariant* under the transform. We refer to this definition of invariance in the following sections.

In order to investigate the invariance properties of various classes of transformation we write down the transformation group for the desired Lagrangians. We then obtain the generators of this group by infinitesimal representation of the functionals. The infinitesimal representation specifies the form of a quasi-linear partial differential equation which is solved by integration in order to specify the form of the Lagrangian. The general method of solution for quasi-linear partial differential equations using the integral surface method is given in Appendix I. For a detailed account of the application of this method see (Bluman and Cole 1974; Moon 1995).

In the following sections we start with an analysis of several very simple transforms operating in the time-amplitude domain. These can be thought of as transformations of one-dimensional signals which specify specific forms of invariance. We will then generalize these results to the problem of specifying invariant functional for time-frequency transforms, which will lead us to an analysis of various types of structured audio representation.

2.2.4 Time-Shift Invariance

The transformation group for the time-shift operations is:

$$T_\varepsilon : t' = t + \varepsilon, \qquad x' = x \quad [40]$$

this specification is in the form of the global representation of a local Lie group, this gives the generators: $\tau = 1$ and $\xi = 0$. Recalling the Lagrangian invariance condition:

$$L_t\tau + L_x\xi + L_{\dot{x}}\eta + L\dot{\tau} = 0. \quad [41]$$

which for the given generators, ξ, η and $\dot{\tau}$ all go to zero, and thus reduces to the following form:

$$L_t = 0. \quad [42]$$

specifies that the Lagrangian partial L_t is invariant under the local Lie group T_ε so the Lagrangian exhibits no dependence on time. The characteristic equation of this group is given by the following partial differential equation:

$$\frac{dt}{1} = \frac{dx}{0} = \frac{d\dot{x}}{0} = \frac{dL}{L} \quad [43]$$

noting that expressions such as $\frac{dt}{1}$ are shorthand for $\frac{dt}{t} = 1$ following Bluman and Cole (1974). Using this notation it is seen that the differential equation is only dependent upon the ratio:

$$\frac{\mathrm{dt}}{1} = \frac{\mathrm{dL}}{\mathrm{L}} \quad \textbf{[44]}$$

This characteristic specifies the form of functions with the time-shift invariance group property:

$$L(t, x, \dot{x}) = f(x, \dot{x}) \ . \quad \textbf{[45]}$$

which states that for an arbitrary function, $f(x, \dot{x})$, the form only depends on the variables x and $\dot{x}$. In this case the solution to the Lagrangian PDE is trivial since it is dependent upon one term L_t.

2.2.5 Amplitude-Scale Invariance

The transformation group for amplitude-scale transformations is:

$$\mathrm{T}_\varepsilon : t' = t, \qquad x' = (1 + \varepsilon)x \quad \textbf{[46]}$$

the generators of which are $\tau = 0$ and $\xi = x$. The Lagrangian invariance condition gives:

$$L_x x + L_{\dot{x}} \dot{x} = 0 \quad \textbf{[47]}$$

which specifies the following partial differential equation:

$$\frac{dx}{x} = \frac{d\dot{x}}{\dot{x}} = \frac{dt}{0} = \frac{dL}{0} \quad \textbf{[48]}$$

By this characteristic equation the form of the Lagrangian is dependent only on the first two ratios:

$\frac{dx}{x} = \frac{d\dot{x}}{\dot{x}}$ and a constant t.

$$L(t, x, \dot{x}) = f\left(\frac{x}{\dot{x}}, t\right) \quad \textbf{[49]}$$

A specific example of this form of invariance is given by the function:

$$L(t, x, \dot{x}) = \frac{1}{\left(1 + \frac{\dot{x}}{x}\right)} \quad \textbf{[50]}$$

which is invariant to amplitude scale changes.

2.2.6 Time-Scale Invariance

The transformation group for a time-scale shift may be written as:

$$T_\varepsilon : t' = (1+\varepsilon)t, \qquad x' = x \tag{51}$$

for which the generators are $\tau = t$ and $\xi = 0$. The Lagrangian invariance condition specifies the form:

$$tL_t - \dot{x}L_{\dot{x}} = -L. \tag{52}$$

which has the characteristic system:

$$\frac{dt}{t} = \frac{dx}{0} = -\frac{d\dot{x}}{\dot{x}} = \frac{dL}{0}. \tag{53}$$

The general form of the Lagrangian satisfying these invariance conditions is given by:

$$L(t, x, \dot{x}) = \frac{1}{t} f(x, t\dot{x}) \tag{54}$$

2.2.7 Frequency-Shift Invariance

In order for a transformation to exhibit frequency-shift invariance using the exponential form for a frequency shift operator the transformation group is:

$$T_\varepsilon : t' = t, \qquad x' = xe^{j\varepsilon t} \tag{55}$$

which defines a local Lie group for which the generators are $\tau = 0$ and $\xi = jtx$. The Lagrangian invariance condition establishes that:

$$L_x tx + L_{\dot{x}}(x + t\dot{x}) = 0 \tag{56}$$

which has the characteristic system:

$$\frac{dx}{\dot{x}} = t\frac{dx}{x} = t\frac{d\dot{x}}{\dot{x}} \tag{57}$$

recognizing the fact that $dt = 0$ this PDE can be re-written as:

$$\frac{dx}{x} = \frac{d\dot{x}}{\dot{x}} \tag{58}$$

which gives the general form of the Lagrangian as:

$$L(t, x, \dot{x}) = f\left(\frac{x}{\dot{x}}, t\right) \quad [59]$$

Recall that this form of invariance looks somewhat like amplitude-scale invariance described above. However, the generators for the frequency-shift group are local generators and do not specify the global representation. For example:

Let $x(t) = \cos t$. Then from Equation 59 we have $L(t, x, \dot{x}) = -\cos t / \sin t$, $x'(t') = e^{j\varepsilon t}\cos t$ and $\dot{x}'(t') = e^{j\varepsilon t}(-\sin t + j\varepsilon\cos t)$. Thus the relation expressed in Equation 58 only holds in the limit of the local Lie group parameter as $\varepsilon \to 0$. Because the transform is non linear, the generators of this group only specify the infinitesimal representation for the transform. This suggests that functions which use the exponential frequency-shift operator only have the invariance property over a vector field in the range of small ε.

2.2.8 Frequency-Shift Invariance Alternate Form

So far we have derived invariance properties by considering a signal in the time-amplitude plane. We can also consider the signal in the time-frequency plane and solve for the Lagrangian functional in the same manner as above. Consider a signal $E(t, \omega)$, we can write the functional for a frequency transform as:

$$T_\varepsilon : t' = t, \qquad \omega' = \omega + \varepsilon \quad [60]$$

The infinitesimal form of the functionals, resulting from a Taylor series expansion about $\varepsilon = 0$, gives the generators $\tau = 0$ and $\xi = 1$. This leads to the following invariance condition:

$$L_\omega = 0 \quad [61]$$

for which the resulting Lagrangian can be written simply as:

$$L(t, \omega, \dot{\omega}) = f(t, \dot{\omega}) \quad [62]$$

This form states, simply, that a frequency-shift invariance specified in the time-frequency plane changes with respect to time and frequency derivatives but not with respect to the frequency value.

2.2.9 Summary of Invariant Components of Common Signal Transforms

Table 2 shows a list of commonly-used transformations with their corresponding invariance struc-

TABLE 2. Summary of Local Lie Group Transforms for Structured Audio Algorithms

Signal Transform	Amplitude Invariance	Time Invariance	Frequency Invariance	Phase Invariance
T_ι Amplitude Shift	no	yes	yes	yes
T_α Amplitude Scale	no	yes	yes	yes
T_δ Time Shift	yes	no	yes	yes
T_τ Time Scale	yes	no	no	no
T_π Time-only Stretch (local time and phase shift)	yes	no	yes	no
T_ω Frequency Scale	yes	no	no	no
T_Ω Frequency-only Shift	no	yes	no	no
T_ϕ Phase Shift	no	no	yes	no

ture. These elementary signal transformations are used to specify the form of structured audio algorithms in the next section. For each of the transforms we can determine which parts of the signal are invariant and which parts are transformed. For example, we note that time-scale operations alter the frequency content of a signal, but time-stretch operations do not. As we shall see, this is because time-stretch operations seek to preserve the local amplitude/frequency structure of a sound extending it in a local region of the signal by shifting with respect to a frame-rate.

These transforms are an important set of descriptors for structured audio operations due to the fact that they all have the group property. We have already seen that for any Lie group transform the operations are associative:

$$T_{U_{\varepsilon 1}} T_{U_{\varepsilon 2}} = T_{U_{\varepsilon 2}} T_{U_{\varepsilon 1}} = T_{U_{\varepsilon 1 + \varepsilon 2}}, \qquad [63]$$

this property is useful since it determines that it does not matter which parameter is applied first since the parameter of the product of two transforms is the sum of the parameters. The second important property is that of closure:

$$T_{U_{\varepsilon 1}} T_{U_{\varepsilon 2}} = T_{U_{\varepsilon 3}}, \qquad [64]$$

this property determines that the result of applying two transforms of the same type always results in a third transform of the same type therefore having the same symmetry properties. The third important property of auditory group transforms is the property of invertibility:

$$T_{U_{\varepsilon 1}} T_{U_{-\varepsilon 1}} = T_{U_0} = I \quad [65]$$

this property is extremely important because it determines that every transformation has a corresponding complimentary transformation whose parameter is simply the negative of the transformation parameter. This property specifies the form of the identity transform which, as we shall see, is also associative. These properties make the auditory groups normal subgroups of the overlying group of signal transformations. What this means from the perspective of signal processing is that the transformations are linear. The result of combining two transformations is associative:

$$T_{U_{\varepsilon 1}} T_{V_{\varepsilon 2}} = T_{V_{\varepsilon 2}} T_{U_{\varepsilon 1}} \quad [66]$$

and the result of combining multiple transformations with an inverse transformation produces the relation:

$$T_{U_{\varepsilon 1}} T_{V_{\varepsilon 2}} T_{U_{-\varepsilon 1}} = T_{V_{\varepsilon 2}} T_{U_0} = T_{V_{\varepsilon 2}}. \quad [67]$$

In the following section, knowledge of the symmetry properties of auditory group transforms is an important component to analyzing structured audio transform algorithms for producing a specified form of invariance.

2.2.10 Structured Audio Algorithm Analysis

In this section we apply the methods outlined above to the analysis of several different classes of audio transform. These methods enable us to define, in a formal manner, what a structured audio transform is and how it affects the invariants of a sound. Recall that the form of a structured audio transform was defined in terms of two separate transforms giving a composite transform T. Using the notational devices developed above for local Lie groups we can express the structured audio relation in the form:

$$T\{\mathbf{W}\} = T_{U_{\varepsilon 1}}\{\mathbf{E}\} T_{V_{\varepsilon 2}}\{\mathbf{S}\} \quad [68]$$

where $T_{U_{\varepsilon 1}}$ and $T_{V_{\varepsilon 2}}$ are two separate transforms that belong to the one-parameter family of local Lie group transformations described above, where $\mathbf{E}$ and $\mathbf{S}$ are separate components of the signal which combine by their product to form $\mathbf{W}$. In the discussion that follows we take $\mathbf{E}$ to represent an excitation signal, such as a glottal pulse in speech or the force-interaction of a scrape for natural sounds, and the $\mathbf{S}$ component represents resonant structures in the sound such as formant structures in musical instrument and speech sounds, or vibratory modes in natural acoustic responses. As we shall see, this division of a sound into excitation and resonance structures allows a conceptual framework for understanding the effects of various classes of auditory transform upon a sig-

nal. The form of transformations of $\mathbf{E}$ and $\mathbf{S}$ is not always considered linear but, as we shall see, the effect of each local Lie group transform is to produce changes predominantly in one or the other component.

We seek to characterize the following audio transforms in terms of their dual transformation structures. Therefore we present each of the audio transforms in the following section in the context of the Lie group transformations on the individual $\mathbf{E}$ and $\mathbf{S}$ components.

2.2.11 Classes of Structured Audio Transform

2.2.12 The Tape Transform (An Unstructured Audio Transform)

The first example of an audio transform that we consider is the tape transform. The tape transform is an operation on a waveform of the following form:

$$T_{tape}\{\mathbf{W}\} = T_{\omega_{\varepsilon 1}}\{\mathbf{E}\}T_{\omega_{\varepsilon 1}}\{\mathbf{S}\} = T_{\omega_{\varepsilon 1}}\{\mathbf{ES}\} \quad [69]$$

where $T_{\omega_{\varepsilon 1}}$ is the local Lie group for a frequency-shift transform. The tape transform produces modifications of a waveform analogous to speeding up or slowing down a tape recorder during playback. The transform collapses the underlying $\mathbf{E}$ and $\mathbf{S}$ components because the local Lie group is the same for both components with the same parameter $\varepsilon 1$. Thus, by the linearity of the transform, the relation in Equation 69 is obtained. The transform produces shifts in both frequency and time-scale of the underlying signal components.

The effect of this transform is most easily understood for speech signals, for which the fundamental pitch of the speech is not only altered, but the formant structures are also shifted thus producing the “munchkin” effect that is associated with such transforms. This transform is equivalent to that of band-limited re-sampling of a waveform for pitch-shift effects. It is of limited use as a structured audio transform because the frequency-scale transform also produces a corresponding time-scale transform thus failing to separate the spectral content of the waveform from the temporal content. Therefore we refer to the tape transform as an *un*structured audio transform.

The desired representation of structure can be achieved in two ways by the auditory group representation. The first is to alter the transformation so that it affects the time-scale of a sound without affecting the frequency scale. The second approach is to separate the excitation and spectral components from the sound and control them independently. We present examples of both these approaches in the following sections.

2.2.13 Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) has been used extensively for audio processing. It is defined as a time-varying DFT; (Allen 1977; Allen and Rabiner 1977). The STFT is not a structured audio transform in itself, but it forms the basis of many audio transforms that have been used

to attempt structured control over sound, thus we consider it briefly here as a background to subsequent analyses. The STFT analysis equation is:

$$X[l, k] = \sum_{n=0}^{N-1} w[n]x[n + lH]e^{-j\omega_k n}, \qquad l = 0, 1\ldots \qquad \textbf{[70]}$$

where $w[n]$ is a real window than is chosen to minimize the main-lobe and side-lobe effects of applying a rectangular window to a signal, H is the hop size of the window which is chosen such that $H < N$ so that the resulting analysis frames overlap in time by $N - H$ samples, and $\omega_k = \frac{2\pi k}{N}$.

The signal reconstruction equation for the STFT is:

$$x[n + lH] = \frac{1}{N}\sum_{m=0}^{N-1} X[l, k]w[m]e^{j\omega_k m}, \qquad l = 0, 1\ldots \qquad \textbf{[71]}$$

which produces overlapping signals which are summed to produce the final result. The utility of the short-time Fourier transform lies in the time-varying spectrum representation. For each time-frame, the DFT component of the STFT samples an underlying continuous Fourier transform at equally-spaced intervals on the unit circle, this results in a spectrum whose lowest-frequency component is $\frac{2\pi}{N}$ which we refer to as the *analysis frequency* ω_0. An analysis frequency value is chosen such that $\omega_0 > 20\text{Hz}$ which is the lowest threshold of frequency perception.

The STFT representation of a signal does little to characterize the content of the signal. Such characterization is necessary in order to perform spectral modifications for re-purposing and control. The STFT is also limited by the linear-frequency spacing. It is well-known that the auditory system performs an approximately logarithmic analysis. However, for the purposes of sound modeling the limitations of linear frequency spacing are not in the manner of information loss, rather the frequency representation is effectively redundant, with respect to human hearing, with an oversampling in the spectrum at higher frequencies. Therefore as long as the analysis frequency is chosen such that the spacing of Fourier components is less than a critical bandwidth in all regions of interest for a given sound then the STFT represents all of the important information for sound characterization.

2.2.14 The Phase Vocoder

A structured audio transform that uses the STFT as a front-end is the phase vocoder, (Moorer 1978; Portnoff 1981; Dolson 1986). The basic operation of a phase vocoder is to provide an estimate of the magnitude and phase for all analyzed frequencies at each frame of the time-varying

analysis signal. The frequency and phase components are simply derived from the STFT as the polar form of each $X[l, k]$:

$$X[l, k] = |X[l, k]|e^{-j\angle X[l, k]} \quad [72]$$

The phase vocoder is used to affect independent control over the temporal and spectral aspects of a sound. Temporal modifications are applied by effectively changing the hop size of the STFT re-synthesis equation, thus compressing or expanding the time-varying structure of the signal without altering the spectral content. This type of transform is called a time-stretch transform. In order to affect a time-stretch a shift in the reconstruction hop-size εH is introduced such that the effective new hop size is $H + \varepsilon H$. This lays down the original overlapping analysis frames at a new spacing but in order to match the phases at the frame boundaries, to avoid periodic discontinuouties, an equivalent shift in the phase of each component must be introduced:

$$x[n + l(H + \varepsilon H)] = \frac{1}{N}\sum_{m=0}^{N-1} |X[l, k]|e^{-j(\angle X[l, k] + \varepsilon\angle X[l, k])}w[m]e^{j\omega_k m}, \quad [73]$$

Using the complex exponential form of a sinusoid, the effect of this time-transform on a single sinusoidal component is:

$$A_l\cos\{\omega_l(n + l(H + \varepsilon H)) + (\phi_l + \varepsilon\phi_l)\} = \frac{A_l}{2}\{e^{j\omega_l n}e^{j\phi_l} + e^{-j\omega_l n}e^{-j\phi_l}\}e^{-j\omega_l lH}e^{-j\omega_l \varepsilon H}e^{j\varepsilon\phi_l} \quad [74]$$

where A_l is the amplitude of the sinusoid at frame l, ω_l is the frequency and ϕ_l is the phase at frame l. The last three terms in the equation correspond to a linear-phase shift for the frame, a linear-phase delay increment for the time-expansion and an additive phase increment for matching the phase of the sinusoid.

From the point of view of auditory group transforms the time-expanding phase vocoder transform is of the form:

$$T_{pvoc1}\{\mathbf{W}\} = T_{\pi_{\varepsilon 1}}\{\mathbf{E}\}T_{\pi_{\varepsilon 1}}\{\mathbf{S}\} \quad [75]$$

where $T_{\pi_{\varepsilon 1}}$ is the time-stretch transform which produces expansions/contractions in time without producing frequency shifts, (this contrasts with the form of time-scale invariance discussed above which produces frequency scaling as well as time scaling). The time-expansion is essentially a global transform that leaves the local structure of the signal intact. By the form of Equation 74 we see that the time-stretch essentially produces a time-shift and a reverse phase shift in each sinusoidal component for each frame of the inverse short-time Fourier transform. Furthermore, this shift is linear across the components thus producing constant group delay and phase shift effects. This suggests that the transform $T_{\pi_{\varepsilon 1}}$ is really a time-localized version of the time-shift invariance and

phase-shift invariance local Lie group transforms discussed previously, and its localized effect is produced by considering each STFT frame as an independent signal.

From the form of Equation 75 we see that the phase vocoder time stretch does not separate the **E** and **S** components of the waveform, since the transformation is exactly the same for both. Thus the effect of this transform is to modify the temporal flow of both the excitation structure and formant structures in the waveform without modifying their spectral content. This is useful, for example, for speech waveforms. Time-stretch modifications of speech produce the required speeding up or slowing down without the munchkin effect of the tape transform.

The phase vocoder is also useful for producing frequency shifts without altering the time structure of a sound. This is achieved by a frequency-scale transform followed by a time-stretch transform to undo the time-scale effects of the frequency transform. Thus the phase vocoder frequency transform is a composition of two auditory group transforms:

$$T_{pvoc2}\{\mathbf{W}\} = T_{\Omega_{\varepsilon 1}}\{\mathbf{ES}\} = T_{\omega_{\varepsilon 1}}\{T_{\pi_{\varepsilon 1}}\{\mathbf{ES}\}\}, \qquad [76]$$

where $T_{\Omega_{\varepsilon 1}}$ is the frequency-only scale transformation which is a composition of two local Lie groups: $T_{\omega_{\varepsilon 1}}$, the frequency-shift transform producing a shift $\varepsilon 1$ in the spectrum as well as a time-scaling by $-\varepsilon 1$, and $T_{\pi_{\varepsilon 1}}$ is a time-stretch that produces the inverse time-scale of $T_{\omega_{\varepsilon 1}}$ without altering the frequency structure.

For all its benefits as a structured audio transform the phase vocoder has several limitations. One problem is that the frequency-shift transform does not separate the **E** and **S** components of the waveform. This means that frequency shifts of both components are produced simultaneously, the net effect of which is that the fundamental pitch of a sound cannot be altered independently of the formant structure. Or, more generally from a natural sound perspective, the excitation structure cannot be de-coupled from the resonance structure of a sound. In order to address this problem the broad spectral envelope **S** of the sound is sometimes estimated, deconvolved and re-applied to the spectrum after frequency shifting. With this de-coupling the phase-vocoder frequency transform becomes:

$$T_{pvoc3}\{\mathbf{W}\} = T_{\omega_{\varepsilon 1}}\{T_{\pi_{-\varepsilon 1}}\{\mathbf{E}\}\}\mathbf{S} \qquad [77]$$

This transform fulfills the requirements of a full structured audio transform since it de-couples the underlying excitation and formant components. Thus we call the transform in Equation 76 a semi-structured transform since it does not decouple the components, but it does separate the local time structure from the global transformation structure.

The second problem is that the time-stretch algorithm assumes the local signal structure between successive frames is similar. This means that the sinusoidal components of the underlying spectrum must be slowly varying with respect to the analysis frame rate, (the hop size of the analysis). The source of this constraint is in the implicit phase modeling, which assumes that the spectrum is

deterministic. For stochastic signals the problem of phase reconstruction is different and therefore must be addressed separately.

2.2.15 Dual Spectrum Transformations (SMS, LPC)

A relatively common approach to modeling time-varying spectra of complex sound events is to decompose the STFT into a smaller number of time-varying sinusoidal components. This approach was adopted by McAulay and Quatieri (1986) and Serra (1990a, 1990b) for the purposes of speech and musical instrument modeling. Such decompositions rest on the supposition that the underlying signal comprises time-varying sinusoidal elements which change slowly with respect to the frame rate. This smoothness constraint is applied to peak tracking in the time-frequency distribution (TFD) and yields the time-varying parameters.

Serra (1990) used an STFT representation for a front-end to the spectral modeling synthesis (SMS) system. The analysis proceeds by first matching the spectrum as closely as possible with a set of time-varying sinusoidal (deterministic) components. The detection and tracking of sinusoidal parameters follows a set of heuristics which are designed to perform well with harmonic and quasi-harmonic sounds, see Table 3 and Table 4.

These heuristics are designed to cope with possibly non-harmonic partial structures in a sound as well as the possible termination of earlier partials and the possible onset of new partials during the course of a sound. They appear to work reasonably well for sounds in which the sinusoidal components are slowly varying with respect to the frame rate of the time-frequency distribution (TFD), see Figure 7. However, such sinusoidal modeling techniques break down when the spectrum comprises noise structures. Or if the sinusoidal structures are rapidly changing as is the case with the phase vocoder discussed above.

TABLE 3. Spectral Modeling Synthesis Peak Detection Heuristics

Peak Detection Parameters	Range	Description
Low Freq	0 - 10kHz	lowest partial frequency
High Freq	0.01 - 22.050 kHz	highest partial frequency
Magnitude Threshold	0.3 = 0dB	threshold for peak selection

TABLE 4. Spectral Modeling Synthesis Peak Continuation Heuristics

Peak Continuation Parameters	Range	Description
Max Frequency Deviation	0 - 100%	% frequency shift to closest peak in the next STFT frame
Peak Frequency Contribution	0 - 100%	% contribution of current peak frequency to the next frame's peak.
Fundamental Contribution	0 - 100%	% contribution of fundamental estimate to next frame
Number of Partials	0 - N	The number of partials to track throughout the TFD.

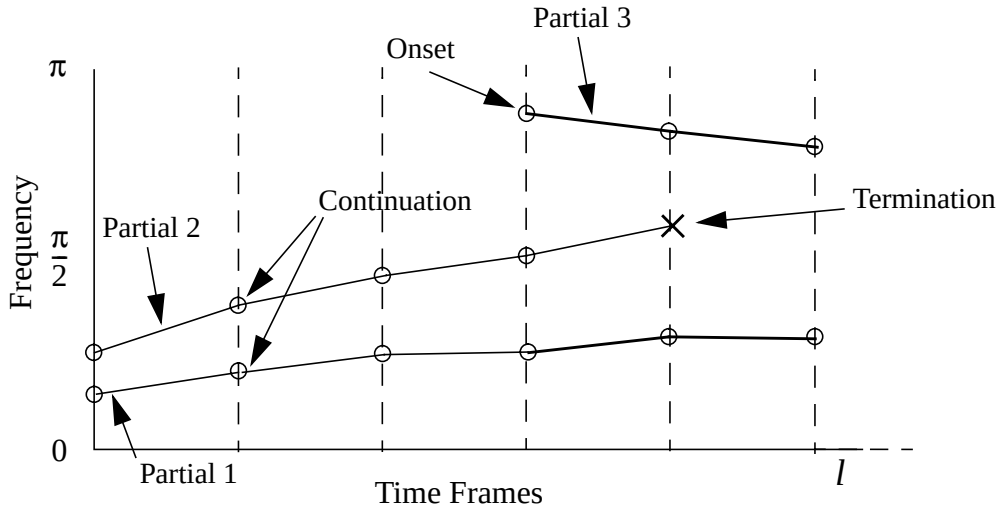

FIGURE 7. Sinusoidal Peak Tracking of a Time-Frequency Distribution

Serra's system employed a second decomposition in order to address the problem of incomplete spectral matching using sinusoidal tracking techniques. Using a technique analogous to residual estimation in LPC analysis the deterministic component (matched by the sinusoids) is subtracted from the TFD thus yielding a residual spectrum. This residual spectrum is then fitted by a series of broad-band spectral estimators which attempt to model the stochastic structure of the sound. We call this type of modeling dual-spectrum modeling where the duality is expressed as a heuristic partitioning of the TFD into deterministic and stochastic structures.

Structured audio control of dual-spectrum representations proceeds in much the same manner as the phase vocoder. The equations governing time-stretch and frequency-shift transforms for the sinusoidal components are exactly the same as those described for the phase vocoder above. The equations governing time-stretch of the stochastic component are, however, different. The difference lies in the phase reconstruction component of the transformation. Whereas for the phase vocoder time-stretch requires alterations by the stretch factor in the phase of each component, for stochastic modeling this term is replaced with a random phase term. Thus the time-stretch transform for the stochastic component of a dual-spectrum representation is:

$$\mathrm{x}[n + l(H + \varepsilon\mathrm{H})] = \frac{1}{N} \sum_{m=0}^{N-1} |\mathrm{X}[l, k]| \mathrm{e}^{-\mathrm{j}\varphi} \mathrm{w}[m] e^{j\omega_k m}, \quad \textbf{[78]}$$

where φ is a uniform random phase distribution. Whilst this technique is useful for analyzing the sounds of speech, musical instruments and limited classes of natural sound, it does not often characterize the content of the sound in structured manner. The heuristics for assigning sinusoidal components to the TFD do not distinguish between excitation structures and resonance structures, thus they mix the two in an unknown manner. In addition, the resulting residual spectrum used for stochastic approximation is a mixture of the noise components of the excitation and formant structures in a sound. A structured representation should articulate both excitation structures and formant structures as independently controllable elements and we conclude that dual spectrum representation does not perform such a decomposition.

Dual spectrum representations generally identify broad-band quasi-stationary and narrow-band quasi-stationary components within a signal. Although useful for modeling musical instruments and speech sounds such a decomposition does not go far enough in its characterization ability for the purposes of modeling the larger class of natural sounds.

2.2.16 Cepstral Transforms

The cepstrum, (Bogert et al. 1963; Oppenheim and Schafer 1989) is an extremely useful representation for structured audio. Under certain constraints it is able to produce the separation of the excitation and formant structures of a sound. As with the other transforms discussed above, the general constraints are that there is only one source at a time in each region of the signal. Unlike the dual-spectrum representations, the cepstrum explicitly models the product of the excitation and formant structures rather than resorting to heuristic guess work. The cepstral lifter is a logarithmic function of the Fourier transform that produces a signal that represents separation of wide-band and narrow-band spectral components. The cepstral representation was employed by Stockham et al. (1975) to effect a signal separation of the singing voice of Enrico Caruso from noisy recordings with orchestral accompaniment. The extracted voice of Caruso was also used by Charles Dodge for his piece *Any Resemblance is Purely Coincidental.*

The complex cepstrum is defined by:

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi}[\log|X(e^{jw})| + j\angle X(e^{jw})]e^{j\omega n}d\omega \quad [79]$$

that is, the inverse Fourier transform of the complex logarithm of the Fourier transform of a sequence. The cepstral signal represents wide-band and narrow-band components as non-overlapping additive different regions, thus it is possible to separate each component using a cepstral "lifter". In the application of Stockham et al. (1975) the problem was applied to blind deconvolution of the singing voice from an orchestral background. A signal-of-interest smoothness constraint had to be utilized in order to successfully extract the voice, the orchestral background was more time-varying in nature thus collapsed to a noisy signal compared to the smoothness constraints. Transforms based on a cepstral decomposition have the following form:

$$T_{cep}\{\mathbf{W}\} = T_{U_{\varepsilon 1}}\{\mathbf{E}\}T_{V_{\varepsilon 2}}\{\mathbf{S}\} \quad [80]$$

which is the general form of a structured audio transform affecting the excitation and formant components of the signal separately.

Limitations with cepstral decompositions are that the signal is constrained to contain only two convolutive components plus noise. For the purposes of natural sound modeling this model is not general enough.

2.2.17 Multi-Spectrum Time-Frequency Decompositions

The form of structured transform that we seek for representing natural sound structures must be able to cope with a multiplicity of *a-priori* unknown signal types some of which are convolutive and others of which are additive in the spectrum of a sound. Many of the natural sounds that we seek to characterize in the next chapter have a multiplicity of noisy spectral components, each of which is considered to be statistically independent. In order to represent such sounds for the purposes of structured control we seek a generalized structured audio representation which is not subject to the same limitations as the signal models described above.

The form of the structured representation that we seek is:

$$T_{general}\{\mathbf{W}\} = \sum_{i=1}^{\rho} T_{U^{(i)}_{\varepsilon 1_i}}\{\mathbf{E}_i\}T_{V^{(i)}_{\varepsilon 2_i}}\{\mathbf{S}_i\} \quad [81]$$

that is, we seek a transformation structure for natural sounds in which an arbitrary number ρ of spectral components are represented, and transformed independently by the local Lie groups $T_{U^{(i)}_{\varepsilon 1_i}}$ and $T_{V^{(i)}_{\varepsilon 2_i}}$. Such a decomposition, ambitious as it is, comprises a general structured-audio transform. A representation such as this is needed for characterizing complex sound events such as smashing and multiple bouncing objects. In the next chapter we present methods for extracting the independent components of a time-frequency distribution using statistical basis methods. These methods are not subject to heuristic spectral modeling or monophonic signals. Rather they seek to

characterize the signal space in terms of statistically independent features in the spectrum. This leads to a representation that is characterized by Equation 81.

2.2.18 Auditory Group Modeling of Physical Properties

A successful decomposition of a signal into the structured form of Equation 81 is the first part of our methodology and is the subject of Chapter III. The second part involves the meaningful transformation of these structures in order to create physically-plausible re-purposing of the extracted sound features. For example, we would like to extract the features of a glass-smash sound and re-use them for creating other smashing sounds, perhaps of a glass of a different size, or a different material such as pottery. In this section we relate the use of auditory group transforms to the form of physical invariants discussed earlier in this chapter.

TABLE 5. Summary of Audio Transforms and Corresponding Physical Property Transforms

Auditory Group Transform	Corresponding Physical Property Transforms
T_α Amplitude scale	Source-event type, source-event force
T_δ Time shift	Scatterings, iterations, other higher-level structures.
T_π Time-only stretch	Source-object materials (increase/decrease in damping). Event spreading, faster/slower.
T_ω Frequency shift / T_τ Time Scale	Source-object size/scale, shape and size of cavities, fundamental period of driving functions. Higher-order structures such as scatterings. Liquids. Speech. Chirping.
T_Ω Frequency-only shift	Event size shifting, preserves event time structure.
T_{f_0} Lowpass filter	Force and type of source event interaction

The audio transform of amplitude scale T_α of an independent component of a natural sound can serve to indicate an increase or decrease in the magnitude of the force interaction with a source object, or set of objects. For example, one component of hitting a ball harder with a bat is the positive change in amplitude of the signal. This view is very simplistic however, for there are other components of the sound that are affected by greater-magnitude force interactions such as the bandwidth of the excitation signal. Therefore we include a transform for filtering of an excitation signal T_{f_0} in order to represent this variable bandwidth behavior.

2.3 Summary of Approach

2.3.1 A Note on Proper and Improper Symmetry

Symmetry is a fundamental measure of similarity. We can characterize symmetry as the repetition of sub-structure within a larger structure. However, the common interpretation of symmetry may lead us to a mis-representation of what symmetry means for the purposes of similarity formalisms. Let us therefore be explicit. We denote by *improper symmetry* those forms of similarity that are generated by literal reflections of an object or group across an axis of symmetry. This type of symmetry is particular to a class of structures that perhaps cannot be physically realized. We denote by *proper symmetry*, henceforth to be called just symmetry, a form of persistence in an underlying object or group which is only defined under particular styles of change. Our reasons for adopting this interpretation will become clearer throughout the course of this section. For now, let us suffice in recognizing this definition for the purposes of summarizing our approach.

2.3.2 1. The Princple of Underlying Symmetry / Regularity

The first principle of our approach rests on the proposition that, for all the apparent complexity in natural sound-generating systems, there are a set of simple underlying symmetries which reduce the space of sound possibilities to a smaller set of fundamental elements with transformations defined upon them. These units can be represented as well-defined symmetry-preserving transformations of basic equations and we claim that they are capable of characterizing much of the inherent structure in sound. As we have seen in this chapter, physical laws exhibit properties of invariance under well-defined styles of transformation; a simple example of this is that of change in the size of a sound-generating system, by strict uniform re-scaling of all linear dimensions, the result of which is the same relative modes of vibration as the reference system but an altered fundamental mode which reflects a change in the absolute structure of the modal vibrations. Thus the governing equations remain essentially the same, but a simple and predictable change in sound structure is *specified* by the said transformation. We can recognize other symmetries; such as the substitution of different materials and changes in topological structure of sound-generating systems. By recognizing, and strictly defining, such invariants we claim: *we can reduce the process of description of sound-generating systems to that of symmetry transforms on a relatively small set of representative elementary physical systems*.

2.3.3 2. The Principle of Invariants Under Transformation

Our second principle is that of the representability of the physical elements of sound-generating systems in the domain of signals and systems with laws of invariance being represented by a well-defined set of symmetry transforms operating upon the signal-domain elements. If these signal transforms exhibit invariance properties that correspond with those of physical systems then we can say that the signal/system symmetry transform model represents the underlying symmetry structure of physical objects and events.

One could argue that the nature of such a representation is arbitrary and any mathematical system can be said to represent an underlying physical structure, but the form of our thesis is that the symmetrical properties of the physics can be formalized, and the symmetrical properties of the signal-level representation can be formalized, in such a way that a strict mathematical relationship

between the two can be maintained. Thus the two modes of representation, physical systems and signal-domain representations, can be said to be related. The second principle can be summarize as follows: *signal and system symmetry transform methods can be representative of the physical properties of sound structures in so far as they reflect similar symmetry properties.*

2.3.4 3. The Principle of Recoverability of Similarity Structure

The third principle on which this thesis is constructed is that of the identifiably of invariant elements and symmetry transformations within natural sound signals. Following the arguments of invariance in physical laws we propose that the trace of underlying physical symmetries in sound events is discernible in the domain of signal representation. Our approach, then, is to analyze sounds to obtain their similarity structure by the extraction of signals and transformations by recognizing symmetry within the signal; this serves to identify characteristic features of a signal. Thus: *the extracted symmetry structure is representative of the underlying physical structure in a sound event*.

2.3.5 4. The Principle of Representation Based on Control of Invariant Features

The fourth principle is that of the affordance of control in the representations outlined in the previous two sections. That is, we consider the representational elements along with their transformational operations to be a *structured representation* of the underlying physical event structure of sound events, and that this structure is non arbitrary in terms of natural symmetries and is thus a psychophysically relevant parameterization of a sound. Thus, modifications of the representational elements and/or their transformational structure is meaningful in terms of underlying physical event structures. Furthermore, since we have chosen our transformational operators to be representative of symmetries in physical laws, we can postulate that alterations of the said structured representation, along the paths of transformation that are well-defined within the system, are meaningful in terms of the underlying domain of physical sound-event structure. So we consider that: *our representation is controllable in physically-meaningful ways, which is generally not the case with the canonical set of general-purpose audio representation schemes*.

2.3.6 5. The Principle that Perception Uses the Above Representational Form

The final, and perhaps most important, principle on which this thesis is based is that of the connection between an underlying physical representation in a signal and the perceptibility of that representation under the influence of structural changes as outlined in principle 4. That is, since there is a strong relationship between signal representation schemes and underlying symmetries in the physical properties of sound events, the manipulation of representational structure affords predictable changes in the perception of structure in underlying physical events. This is precisely because the ear/brain system is directly sensitive to the underlying symmetries in the laws of nature. In short, certain shifts in representation structure afford the *perception* of a shift in underlying physical structure; the representation structure is designed to exhibit the same types of invariance and transformation properties as the underlying physical structure and, furthermore, we hope to achieve this without the need for explicit physical modeling of sound-generating systems by instead using transformations of invariant features.

In summary, then, the five principles outlined in this section form the basis of the thesis, each of these views can be substantiated by physical and perceptual evidence and the result is a formal representation structure for sound events that contains the necessary components for meaningful querying and re-structuring of acoustical information. Our thesis, then, claims a non-arbitrary representational structure for audio which can be used for many different purposes from analytical decomposition to synthetic rendering of sound objects.

2.4 Summary of Chapter

In the preceding pages we have given a broad view of the nature of sound objects. It is clear at this juncture that there is no simple method of analysis and description of the many possibilities exhibited by sounds, but it is also clear that there are many structural symmetries inherent in the phenomena that we can exploit as a basis for a broad theory of sound organization. We have developed the framework for such a theory under the title auditory group theory and have argued for the specific content of our methods. In the following chapters we demonstrate how the principles embodied in auditory group theory can be applied to the diverse problems of analysis/synthesis algorithm design, sound-structure analysis and sound-object synthesis.