Chapter I: Ecological Acoustics

1.1 Ecological Perception

The ecological approach to perception can be summarized as follows: much of what an organism needs to get from a stimulus, for the purposes of its ecological activities, can be obtained by direct sensitivity to invariant structures in the world in which it lives. That is, possibly complex stimuli may be considered as elemental from the perspective of an organism's perceptual apparatus and, furthermore, this perception may be unmediated by higher-level mechanisms such as memory and inference. This was the approach developed by Gibson and his followers for the field of visual perception, (Gibson 1966; Gibson 1979). In addition to the visual system, Gibson also considered the other senses including the auditory system from the point of view of direct pickup of invariants in the world. While there are contentious issues in Gibson's view, at least from the point of view of cognitive psychology, there are subtleties in the notion of direct perception that are often overlooked by a desire to understand perception as a product of higher brain functions. It is our belief that these ideas merit closer attention for consideration as a possible contributing factor in the auditory system and a justification of the ecological approach to understanding natural sounds is the subject of this chapter.

Research on models of auditory perception has, in the past, been concerned with the systematic grouping of low-level simple stimuli, or perceptual atoms which have been studied in the context of Gestalt and cognitive psychology, see for example (Bregman 1990; Brown 1992; Cooke 1991; Ellis 1996). These studies demonstrate several important results, for example the role of higher-level attentional mechanisms such as signal prediction for perceptual restoration of missing or occluded signal components (Ellis 1996), and the effects of proximity in time and frequency on the systematic grouping of auditory objects. Such groupings are said to form *auditory streams,* each of which is a perceptually separate component of the stimulus. This field of investigation, is called auditory scene analysis. Computational approaches to auditory scene analysis are concerned, then, with speculative enquiry into the nature of stream segregation from the point of view of low-level sensory stimuli.

The ecological approach, however, suggests that perception is not specified by the systematic integration of low-level simple stimuli, such as individual pixels in the retina or narrow-band frequency channels in the cochlea but that it is specified by directly perceivable, if complex, groups of features. Such features are manifest in a stimulus signal because they are caused by events in the world that exhibit certain symmetries. The general hypothesis is that the perceptual apparatus of an organism has evolved to be directly sensitive to the symmetries that occur in its natural environment and therefore its perceptual systems implement algorithms for the pickup of these features. These features are called *invariants*, (Gibson 1966; Shaw and Pittenger 1978). There has been a steady growth in the consideration of this view as characterizing aspects of the auditory system with several experiments having been conducted into the possible existence of invariants as well as speculations as to their signal properties (Gaver 1993, 1994; VanDerveer 1979; Warren and Verbrugge 1984; Wildes and Richards 1988). The general results of this body of work suggest that certain structures of sound events are lawfully and invariantly related to fundamental properties of physical systems and force interactions, and that human auditory perception may be directly sensitive to such structures. Whilst this body of literature has shed light on previously little understood aspects of natural sound perception and has suggested directions for future work, there has been no prevailing mathematical framework within which to articulate the findings in a systematic manner. In the next chapter we develop such a framework, based on group theory. Whereas computational auditory scene analysis is concerned with modeling low-level attentional mechanisms in audio perception, the approach of auditory group theory is to represent physical invariant symmetries of audio signals and to develop an algorithm that can extract these invariant components from recordings. In the remainder of this chapter we develop the background of the ecological approach to auditory perception.

1.1.1 Attensity and Affordance

The degree and scale of sensitivity of a particular organism to the acoustic environment depends on the appropriateness of the various sound signals for its survival. That is, an organism will be sensitive to properties of the acoustic environment that potentially affect its state of being; either positively or negatively. To give a visual example, the concept of a chair has very little to do with the perception of geometric visual primitives for the purposes of identifying an object that can be sat upon. Rather, a chair is considered to be an object in the environment that is capable of supporting the weight and sitting posture of the observer. Thus a desk can be used as a chair if it is the correct height and is stable enough, the fact that chairs tend to take on semi-regular forms has more cultural significance than perceptual significance. The ecological view suggests that the percept of affordance of sitting is unmediated by inference, it is a direct percept of the rigid body structure in an object that has evolved as a perceptual algorithm. Gibson's somewhat controversial hypothesis is that such percepts do not always need the interjection of cognitive functions of action and planning.

The appropriateness of an object or event for the ecology of an organism is called its *affordance structure*, (Gibson1966). The affordance structure of an event, then, is that which determines whether or not an organism should attend to it for a particular type of encounter. For example, the sound of an empty bottle specifies the affordance of filling (Gaver 1993). The perception of affor-

dance depends on the scale and restrictions of the environment in which an organism lives. The concept of affordance is an important one, it leads to the reason why different organisms attend to different sound properties, and may hold clues as to which types of sound-object structures are considered elemental from a human perspective.

A beautiful illustration of the concept of affordance and the effects of a change of scale is the 1996 french film *mikrocosmos* in which stunning footage of the microworlds of insects is scaled to human size. This is coupled with extremely clever, “scaled” sound effects; such as the sound of an irregular heavy thudding, the thudding turns out to be a bird eating up ants with its beak with a deathly precision. Thus the affordance structure of the thudding sound to the small insects of that microworld is potentially of great ecological significance. The ecological significance of the same sound at a human scale is, of course, not as great. The degree of importance that a particular sound structure holds for an organism is called its *attensity,* Shaw *et al.* (1974), and is proposed as the name for a measure of ecological significance of an object or event for an organism in a particular environment.

1.1.2 Complexity of Percept versus Complexity of Stimulus

Perceptual simplicity in a sound structure may have nothing to do with the simplicity of the stimulus from the point of view of analyzing the signal. On the contrary, there appears to be an inverse relationship between simplicity of stimulus and simplicity of perception. Consider, for example, the spectrograms of Figure 4and Figure 5. In the first figure the glass smash sound appears as a number of separate features; a low-frequency decaying noise component at the beginning plus a wide-band impulse, as well as numerous particles scattered in the time-frequency plane. It is not easy from the spectrogram to discern the structure of the sound, we do not “see” an obvious representation of smashing. Similarly, the coin bouncing sound of the second figure shows features of a wide-band impact spaced regularly and exponentially in time, an invariant of bouncing events, with a high-frequency ringing component, which is an invariant of small metallic objects. For all their signal complexity, these sounds present no confusion to our perceptual systems. We are unlikely to confuse the action of bouncing with the action of smashing, or the material property of metal with that of glass. It seems, then, that the more complex the structure of the stimulus the easier it is to discern its cause. That is, breaking and bouncing events specify their source actions by their overall structure and are not well represented by the micro-details of their time-frequency distributions.

The inverse relationship between percept complexity and stimulus complexity is articulated succinctly by Johansson with regard to the visual system, “... what is simple for the visual system is complex for our mathematics and what is mathematically simple is hard to deal with for the visual system”, cited in Jenkins (1985). Jenkins proposes that the same principle operates in the auditory domain.

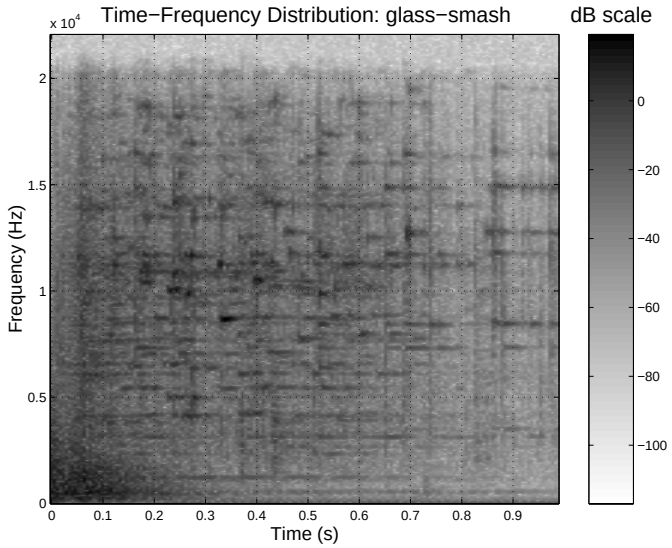

FIGURE 4. Spectrogram of the sound of a smashing glass. There are many components to this sound, such as low-frequency decaying impact noise and high-frequency particle scattering; but we perceive a single event: smashing.

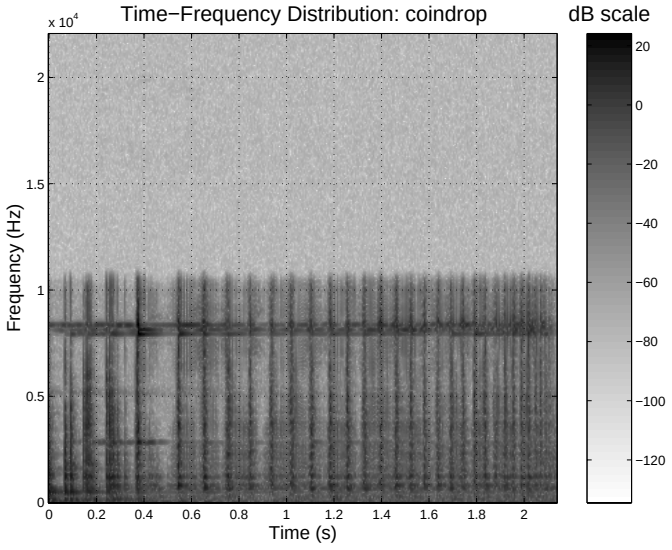

FIGURE 5. Spectrogram of a coin dropping sound. The features of this sound are a sequence of impacts that get closer in time. The metal rings after each bounce thus creating a constant high-frequency component.

The apparent paradox may be understood if we consider that in order for humans to extract meaningful ecological information from the world, in terms of objects, actions and their affordance structure, there must be a large degree of high-level information, residing within the structure of the signal. We do not need the context of seeing a coin drop in order to determine that a sound was indeed caused by a small metallic object dropping. The greater the quantity of structured information within a signal, the easier it is to identify the underlying causes of the sound; it is for this reason that the task of event identification becomes easier the richer a signal becomes. Consider that musical instrument categorization becomes easier when a range of stimuli are presented, as in a melody for example, rather than just a single note; or that the identification of an approaching friend by listening to their footsteps requires more than one footstep for recognition of their gate. The latter example illustrates an important concept of ecological acoustics, *everyday listening.*

1.1.3 Everyday Listening and Reduced Listening

The act of *everyday listening* is concerned with the extraction of as much information as is necessary to determine the underlying event of a sound, in this regard Gaver makes a phenomenological distinction between everyday listening and other types of listening: "*everyday listening*...is the experience of listening to events rather than sounds. Most of our experience of hearing the day-to-day world is one of everyday listening" (Gaver 1993).

We make a distinction between two broad classes of attention to sound; each considers different hierarchical levels of information a the sound structure. Gaver makes this distinction by considering the difference between *musical* listening and *everyday* listening: "the distinction... is between experiences, not sounds... it is possible to listen to any sound either in terms of its [inherent] attributes or in terms of those of the event that caused it." (Gaver 1993). Whilst we recognize that this is a useful distinction to make, the term *musical listening* may diminish the generality of the concept. For example, it is possible to listen to a piece of music both in terms of its sources and in terms of the abstract qualities of the sounds, this is also acknowledged by Gaver; what Gaver calls musical listening is a general property of our perceptual system that extends beyond the realm of traditional musical sounds. Thus we make a similar distinction, but in order to disambiguate the use of the term *musical* we follow Schaeffer and refer to the act of listening to inherent sound qualities, without regard to their causal identity, as *reduced listening;* (Smalley 1986; Schaeffer 1966). Thus we recognize a separation between *everyday* listening and *reduced* listening in much the same manner that Gaver proposes a phenomenological separation between *everyday* and *musical* listening.

Everyday listening is concerned with the relationships of sound structures to their underlying physical causes. We propose that the distinction between everyday listening and reduced listening is mainly in terms of the category assignment of the structural interpretation. More specifically, the inherent structure in the sound is precisely what we attend to at the reduced level of listening, and it is the relating of this inherent structure to the act of event recognition that we refer to as everyday listening. Thus everyday listening is not distinct from reduced listening, rather it is a higher-level listening experience due to the additional considerations it demands. Our premise is, then, that *inherent sound structure* is a necessary component of *source identification* and that it is the atten-

tion to sound structure that allows us to recognize various classes of complex sound events and to consider them as being similar.

1.1.4 Persistence and Change as Perceptual Units

In order to probe further into the ecological approach to auditory perception we now consider the concept of a sound *event*. For the purposes of modeling sound phenomena the notion of an event is very much contained in more general notion of change of an underlying physical state through time. But, which changes and what scale of time are appropriate to the identification of the constituents of a sound object?

The common insight that Gibson and Johansson brought to the understanding of visual event sequences was that spatio-temporal change in visual phenomena was the starting point of perception. Johansson described the importance of this insight with the following statement: “Change of excitation has been shown to be a *necessary* condition for visual perception”, Johannson (1958) cited in Warren and Shaw (1985). Whilst Johannson and Gibson worked primarily in the field of visual perception, the essential nature of change applies to the auditory sense also. The role of the perception of change in auditory stimuli was recognized by Risset and Mathews (1969) in their seminal study of the time-varying spectra of trumpet tones. Their work pointed out that the previously held notion of the primacy of steady-state components in a sound was invalid from the perspective of synthesizing realistic musical instrument sounds. Instead they considered a new model of sound in which the dynamic components of a spectrum are considered primary and the steady-state components were considered redundant for the purposes of instrument classification and tone-quality assessment, which is known generally as the attribute of *timbre* in a sound. This new model, which we shall call the *dynamic* model, was quickly adopted as the framework in which to study auditory phenomena and sound structure, and the framework lead to a number of revealing studies on the nature of timbre, (Plomp 1970; Grey 1975; Risset and Mathews 1979; Wessel 1979). This work represents the dominant view of sound structure in the psychological literature and it claims to represent the perceptually important dynamic structures that comprise auditory phenomena from the perspective of musical instrument sound structures.

However, it is not enough to recognize that change happens, hence there is structure. We must look a little more closely at what we specifically mean by change for auditory phenomena; for change implies that a variable exists by which it can be articulated. Shaw and Pittenger defined an event as, “a minimal change of some specified type wrought over an object or object-complex within a determinate region of space-time.”, Shaw and Pittenger (1978). Warren and Shaw argue that this view has profound ramifications for the investigation of perception, namely that “events are primary, and empty time and static space are derivative.”, Warren and Shaw (1985). The traditional view of objects, from Descartes to Locke, is precisely that they are static. But by the dictum of Shaw and Pittenger we are given an alternate view which is substantiated by twentieth-century physics; an object is stationary in so far as it *seems* stationary from the perspective of an observer. For the field of perception we interpret this as stationarity from an organisms’ *ecological* perspective. Thus what we mean by an event has to be related to the scale of measurement that we choose; a scale that is related to our moment to moment needs as organisms in the environment. It is

through these perceptual-phenomenological inquiries that Warren and Shaw offer up a definition of an event for the purposes of psychological investigation: “Most basically, then, events exhibit some form of *persistence* that we call an object or layout, and some *style of change* defined over it”, Warren and Shaw (1985).

An event, then, not only defines change, but change can only be defined in terms of some form of persistence. Far from being circular, this definition allows us to critically assess what we mean by an event and what we mean by change. This very simple notion leads to a remarkably powerful set of constraints on which to define what we mean by sound objects and sound structure.

1.1.5 Persistence and Change in Sound Structures

The nature of sound is inherently and necessarily temporal. Sound is a series of pressure variations in air which arrive at the ear continuously through time. Thus from a physical point of view sound is a measure of air pressure as a function of time at a particular point of observation. One view of persistence and change for sound phenomena then is that of air as a persistent medium and air pressure as a style of change. This, the physicists view of nature, is a remarkably useful representation for many types of investigation into the nature of sound phenomena, as we shall see later. However, it was postulated by Helmholtz in the nineteenth century that the sound of simple and complex tones could be considered not as a complex changing functions of time, but as relatively simple functions of time that could be understood in terms of *frequency.* Helmholtz’ studies on simple vibrating systems under the influence of driving forces of various kinds lead him to the supposition that the ear performed some kind of frequency transform, roughly analagous to the decomposition of a signal into separate ferequency components performed by a Fourier transform, Helmholtz (1954/1885). Helmholtz concluded that, from the perspective of the ear, a periodic change in air pressure at frequencies in the range of 20Hz-20kHz produces a percept of a *persistent* senstation. Furthermore, this sensation couldbe described by a superposition of simple sensations in the frequency domain, or Fourier domain, corresponding to a superposition of sinusoidal components in the time domain. Thus was born the classical model of hearing as developed by Helmholtz in his famous treatese on sound and the sense of audition: “*On the Sensations of Tone”*. , Helmholtz (1954/1885).

Throughout the nineteenth century, and well into the twentieth century, this form of Fourier persistence in a sound signal was considered to be the primary constituent of tone quality or *timbre*. Whilst Helmholtz himself noted that sounds generally had transitionary elements occuring at the onset, it was considered that the longer portion of a sound was steady state and that this element was representative of our perception. Therefore a sound was considered to be well approximated by an infinite Fourier decomposition since the steady-sate portion of a sound was considered primary from the point of view of perception. This model of hearing is now known as the *classical* model, Risset and Matthews (1977), and still resonates in the writings of auditory researchers to this day.

With the advent of computer analysis of sound using Fourier decomposition techniques and the synthesis of sound using various computer-based techniques it was quickly found that the Helm-

holtz' model was not sufficient for producing convincing sounds. Risset and Matthews developed time-varying analysis techniques that demonstrated the importance for capturing the change in Fourier components over short-duration frames (of roughly 20msecs), Risset and Mathews (1969). The changing spectral components produced a dynamic waveform that was perceived as satisfactory from the perspective of musical instrument sound synthesis.

However, even within the dynamic model of timbre, there is no accounting for higher-level temporal behaviour. Thus when we attempt to generalize the findings of Risset and Mathews it is difficult to account for the widely differing dynamic properties of many kinds of natural sounds. We now consider the example of Warren and Shaw, that an event is defined by some form of persistence and some style of change. From this view we proceed in extending and clarifying the findings of researchers such as Risset and Matthews on account of an ecological basis for perception. Ecological perception leads us to consider whether there are invariant components in natural sound spectra. If so, then we may be able to obtain some clues as to the nature of similarity between sounds generated by different physical systems; this similarity structure is determined by both the static and changing components of a sound.

It is the change of a persistent variable that gives rise to structure in an event; without the change there is no structure, and without the persistence it is impossible to define the change. Therefore any measurable quantity formed out of an event is the trace of an underlying structure of physical change articulated over the course of something that stays physically constant. So in the analyses of dynamic sound spectra we should expect to find that there is some component to the sound that is static and some component that is articulated thus defining the style of change. Hence it is not short-time change in a Fourier spectrum that specifies a sound, but it is also some form of underlying persistence, a persistence that exists even during transitory phases of a sound such as often occurs in the attack of a musical instrument note. In order to account for the persistent and changing components of an auditory signal we must look to the underlying physics of mechanical sound-generating systems, we shall explore this further in Section 2.4.3.

1.1.6 Hierarchical Structure in Sound Events

In addition to recognizing that event structure is delimited by the styles of change of persistent variables, the notion of higher-level structure can also be induced in much the same way. The point-light experiments described in Johannson (1973) point to a notion that styles of change operate hierarchically in visual phenomena, that is, the local transformation structure of a single point of light relates to the global perception of walking, dancing, running and gymnastics by higher-level styles of change across the low-level stimuli. The successful identification of these high-level behaviours by experimental subjects, in the absence of extra contextural cues such as body features, indicates that the style of motion across the points of light is sufficient to specify the change characteristic of the entire body structure. Warren and Shaw present a view that *change-specified structure* and *change-specified change* are the fundamental units of events and that these units form the basic elements of analysis for perceptual investigations Warren and Shaw (1985).

Extrapolating these findings to the domain of auditory stimuli we suggest that there is an element of change within the structure of a sound object beyond that recognized by Risset and Mathews, namely that of *higher-order temporal structure*. Whereas Risset and Mathews' research suggested that short-time changes in the Fourier spectrum reflected important features of the transient component of note onsets for musical instruments, they did not suggest that higher-order temporal structure may be important for perception of similarity between sounds. The similarity structure of many natural sounds is delineated not on the order of *Fourier* time or *short-time* changes but on the order of *high-level* changes within the structure of the sound. By high-level we are refering to the timing of onset components of sub-events due to an inherent multiplicity within the sound structure. An example is the perception of breaking events. Clearly a glass smash has the overall percept of a single sound object, yet within the structure there are a multiplicity of particles which exhibit a *massed* behaviour; this massed behaviour has characteristics that are common across many different breaking events suggesting that there is a similarity quality operating within the higher-order structure, i.e. beyond Fourier-time and short-time structure, of the sound events.

We break the similarity structure of a sound into three components, Fourier persistence, short-time change and high-level change. The persistent part of a sound can be measured in the manner of Fourier persistence because the cochlear mechanics of the ear, being sensitive to changes on the order of 50 msec and shorter, represents such micro-temporal change as an approximately static quality in log frequency space for rates of change in air pressure greater than 20 Hz, which is simply $\frac{1}{0.050}$, and represents the *frequency perception threshold* of the cochlear mechanism. We shall call components whose characeristics give rise to frequency perception *Fourier-time* components. Fourier-time components are static in their perception, but by the physical nature of their makeup we know they exhibit periodic change over a window of perception that lasts 50 msec for the lowest-frequency components and less than 50 msec for higher-frequency components.

Aside from Fourier-time changes operating above the frequency-perception threshold, changes occuring at rates less than 20 Hz, and continuous in terms of a function of the underlying Fourier-time components, are perceived as *short-time change* in the static frequency spectrum. These rates of change are below the freqeuency-perception threshold and therefore articulate *perceptual short-time*; short-time spectral changes are *perceived* as change whereas Fourier-time changes are perceptually static. Although very simple, it is very important to delineate these terms if we are to proceed in identifying persistence and change in sound structures. It makes no sense from a perceptual point of view, and perhaps even from a physical perspective, to treat these different styles and rates of change as part of the same phenomena when, from an ecological perspective, they are entirely different forms of information.

We could characterize the short-time style of change in a sound as change-specified structure. That is, the underlying Fourier-time components are specified under small changes which are perceived below the frequency-perception threshold. But what of changes in the style of change of Fourier components? Larger changes which are not perceived as small and continuous from the perspective of short-time perception. Warren and Shaw (1985) consider a form of structural specification which they call *change-specified change*. What this means is a style of change operating over the

short-time change structure in an event. We consider this category of change to delineate the higher-order structure of a sound object in the same terms as Johannson's point-light walker experiments demonstrated visual sensitivity to differences in styles of change of lights positioned at the joints of an articulated human motion sequence.

Thus in the sound examples of Figure 4 and Figure 5, the glass smash event is specified by the Fourier persistence that is characteristic of glass sounds (its spectral features), a short-time change structure that reflects impact and damping in each individual particle (short-time temporal features), and a high-level change structure that reflects the scattering of particles in the time-frequency plane (global time-frequency structure). The coin-drop sound specifies Fourier persistence due to the small metallic coin object, short-time change that reflects the individual impacts, and a high-level structure that represents the form of exponentially-decaying iterations which is characterisitc of bouncing sounds. This tri-partite decomposition of sounds is necessary for the description of natural sound events and it is not represented by previous analysis/synthesis models of sound.

From the ecological perspective of human auditory perception, sound objects reveal similarities in their affordance structures. That is, an underlying physical action is recognized by the mechanism of recognition of a style of persistene and change in a physical event. An example of this can be found in the sound-event identification studies of VanDerveer in which confusions between events such as "hammering" and "walking" suggest that both of these sound structures afford consideration as either event because both the similarity in the mechanical structure of the events, and hence the similarity structure of their corresponding sound objects, are closely matched. If the encounter were framed in the context of "woodwork" then observers may more readily be persuaded that the perceived action is indeed hammering, a similar priming could operate the other way in order to persuade the perception of walking.

This ambiguity in affordance of sound structure is precisely what enables a Foley artist to trick us into believing the footsteps and door slams that we hear in a movie; for these sounds are very rarely constructed from the physical events which they are made to represent. So we see that the concept of sound-stucture similarity, beyond that which has been of primary concern to psychologists studying timbre, has been used effectively for many years by artists and composers, but it is only recently that higher-level structure has started to become the focus of detailed scientific scrutiny, (Schubert 1974; VanDerveer 1979; Warren and Verbrugge 1988; Gaver 1993).

We conclude this section with a remark from Warren and Verbrugge, "sound in isolation permits accurate identification of classes of sound-producing events when the temporal structure of the sound is specific to the mechanical activity of the source", Warren and Verbrugge (1988) see also (Gibson 1966; Schubert 1974). Thus higher-order structure may be specific to classes of events such as hammering, walking, breaking and bouncing, and lower-order structure may not play the primary role which it has been assigned by the classical and dynamic view of sound structure. Such a shift in sound-structure characterization implies that we must be prepared to invert the prevailing theoretical view of perception as an integration of low-level perceptual atoms and consider

that at least part of the mechanism must be concerned with the identification of higher-order structure without regard to the specifics of the features in low-level components.

1.1.7 Illusions of Affordance: The Example of Foley

It is the very goal of sound-object modeling to deliver the necessary information by which an observer can infer an underlying object and action in relation to some task or senario. We consider the example of a film sound track. The on-screen action of a film is often balanced by a sound track that presents additional or complementary information to that of the visual display. The goal of the additional cues is often to enhance the sense of immersion in the scene and to provide information about the off-screen environment such as providing a cue for an action that cannot be seen.

Foley artists and sound designers regularly exploit physical invariance properties in order to create an illusion of a particular sound event to support the on-screen illusion of action. The technique is named after the radio and film pioneer Jack Foley who, as a Universal Studios technician in the 1950s, became known for his synchronized sound effects such as the reverberating footsteps of an actor moving down a hallway. Many of the effects are acheived using a small, but ingeneous, set of tools and objects that are capable of making many varieties of sound, such as small metal objects, trays full of gravel, bells, door knockers, and water pools, (Mott 1990). The remarkable fact is that entire radio shows or film soundtracks were *performed* live by a Foley artist and recorded in sync with the action in the case of film. Furthermore this was achieved with only a modest collection sound-generating objects.

An example of Foley sound is that of footsteps in a film. Each footstep is carefully cued to the action to convey extra information about the action of an actor. Fore example, a sudden slowing down or shuffling sound can imply a surprise action. Also, we are often presented with sound that are a little louder and lower in pitch than we might normally hear. But the manipulation of the sound in this manner affords the perception of something larger, and more dramatic than a realistic recording. Sound designers, who's job it is to create sounds using various computer and electronic synthesis and manipulation tools, often create enhanced effects for use in films. By manipulating the sounds in various ways they can often be given added dramatic effect which can add much in the way of tension and repose during the course of action in a film, Mott (1990). A Foley artist will substitute sounds for keys jangling, locks being opened, coins dropping, footsteps on gravel and wood, water dripping, and many other seemingly arbitrary sound events.

The reason for our senses suspending disbelief on account of sounds not generated by an accurate source is due to the affordance of the sound structure for the perception of the intended event. Small metallic plate objects can substitute for keys because they have all of the necessary features that afford being perceived as keys, i.e. metalic and small. Another example is that of footsteps which are generated by treading in a large box containing appropriate materials, such as gravel or sand, there are only a small number of such sounds that are required to create all the footsteps for a film. The lesson of Foley, then, is that it is only necessary to capture the essential components of a sound, those components that afford the specification of the required physical event and it is therefore not necessary to account for all the details of an event.

We can perhaps guage the success of such techniques if we consider that virtually none of the sounds we hear in a modern film are generated on the set. They are all placed there by hand in audio post production, a process that takes many weeks during the late stages of film making.

What if audio producers could be given a tool suite that could transform a set of “sound objects” in the many ways that they need? For example, a footstep model would generate footsteps parameterized by various desired properties such as gender, size, gate, shoe type, ground type. One application of the current study is to build sound models to assist the audio production process by offering controllable sound effects. The purpose of these models in sound effects and Foley audio production is to speed up production time and to offer ceative control over sound materials. However, potentially the most interesting use of such a system would be for generating sound effects for interactive media; an interactive sound modeling program could act as an automatic Foley server capable of generating plausible sound effects from descriptions of objects, events and actions. Such a system would rely on transforming various physical properties of sound features for producing the desired effects, or it may attempt to explicitly model all the underlying physical features of the interactive environment and render sounds from detailed physical descriptions. The evidence for not persuing the latter approach rests in the lesson of Foley. That is, we only need to match sounds in so far as their affordance structure matches that of a desired sound percept.

1.1.8 Studies in Environmental Audio Perception

In order to probe at understanding the perceptual structure and relationships of everyday sounds such as those generated by Foley artists, several researchers have investigated catagorical perception and similarity ratings of everyday sounds.

VanDerveer’s study on thirty common natural sounds in a free identification task suggested that listeners could identify the source events very accurately at a rate of about 95% correct recognition, VanDerveer (1979). The sounds included clapping, footsteps, jingling, and tearing paper. Those sounds for which there was a high degree of causal uncertainty were described in terms of their abstract structural qualities rather than a source event, this accounted for only a few of the sounds which could not be identified. VanDerveer also found that clustering in sorting tasks and confusion errors in free identification tasks tended to show the gouping of events by common temporal patterns. For example, the sound of hammering was confused with the sound of walking. Both of these sounds share a periodic impusive pattern. This effect suggests that similarity judgements may operate on the high-level structure of certain classes of sound. The higher-order structure in the sound may also provide significant information about an event, for example consider that a listener’s ability to detect whether footsteps are ascending or descending the stairs is likely a a product of higher-order structure in the sound, Gaver (1993).

Warren and Verbrugge suggest that the auditory system may in fact be designed to pick up information more readily from higher-order structure, such as changes in spectral layout and distribution of onset components within an event, than quasi-stable elements, Warren and Verbrugge (1988). They showed that listeners were able to distinguish between breaking and bouncing categories by re-arrangement of the structural components of a bounce sound to sound like that of

breaking. The point of interest is that the information necessary to categorize the sound as breaking was sufficiently represented by higher-order structural features rather than any re-arrangement in low-order features.

In a follow-up experiment, it was shown that presentation of a single bounce period was enough for listeners to judge the elasticity of a ball. This is remarkable since the physical equations dictate that observation of two periods is necessary in order to derive the elastic constant. This experiment suggests a similar result as the dynamic vector visual display experiments of Johansson, that we perceive the events by imposing constraints corresponding to what an ecologically reasonable interpretation of the underlying acoustical variables are. The ecologically probable limits act as guiding contraints for the perception of underlying physical events, Warren and Verbrugge (1988).

There have been a small number of studies in the production of environmental sound events by synthetic means. Gaver describes methods for creating, amongst other sounds, impacts and scraping sounds using simple sinusoidal and noise components. His methods are based on an ecological approach to understanding the important structural content of these sounds, Gaver (1994). The algorithms generate patterns of spectra using sinusoidal and filter-bank techniques which are controlled via the algorithm parameters, such as frequency and decay time of partials, to specify various kinds of object behaviours. The impression of size is controlled by shifts in spectral excitation and force is specified by amplitude of partials. Materials are suggested by the damp-time of spectral components. The results of Freed, on the perception of mallet hardness as a ratio of high-to-low frequency energy, suggest that perceived hardness of objects can be modeled directly in the same way, Freed (1990). Similar findings are those of Wildes and Richards (1988) who propose several invariant properties of the sounds of various materials, which could lead to a synthesis method for cuing material properties such as glass-ness or wood-ness.

There are several implications of the ecological view of perception for timbre which are indeed supported by the existing timbre literature. For example, Grey (1975) suggested that the topology of musical-instrument perceptual similarity spaces derived by multi-dimensinal rescaling was determined by the grouping of instrument sounds by intrument family. The cases in which musical instruments did not group by physical similarity were cases in which structural features of the sounds of different physical systems were similar. Grey found, for example, that the overblown flute tended to cluster with the violins, and that this was perhaps due to a similarity in the time-structure of the excitation component, whch is a bowed string in the case of a violin and a turbulent jet stream in the case of the flute, Grey (1975). These results suggest a physically based interpretation of the timbre space and that the abstract dimensions sought by timbre researchers are traces of the underlying physical properties of the physical systems. Thus it appears that timbral control by manipulation of the axes of a multi-dimensional timbre space may only be possible for very limited domains of sound. If any structural features are not in common between two sounds it makes no sense to traverse a timbre sapce between them, since the features of one cannot be systematically mapped onto the features of the other in a one-to-one fashion.that

Many studies in vision have suggested that kinematic experience of the real world plays a role in perception. They imply that perception is constrained by kinematics construed at an ecological

scale. For example, Warren and Verbrugge (1985) discuss experiments suggesting that subjects were able to estimate the elasticity of the ball by observing one period in either the auditory or visual domain. The modality of the information did not matter, the judgements were accurate for both. The overall sensativity of humans to such information is remarkable considering the complexity of vibrating systems and their interactions with acoustic environments.

1.1.9 Summary: Implications of Invariants for Structured Audio

The general contribution of ecological acoustics experiments has been to suggest the important role of high-level structure within sounds for the perception of events. This view contrasts with the prevailing local-structure explanations of the dominant theories of sound which have been primarily concerned with musical instrument timbre and vowel sound qualities. The findings of ecological acoustics lead us to seek high-level structure in sounds by way of recognizing structural invariants and their transformations within a sound structure. The examples of a glass smash and a coin drop illustrate that there are persistent components within the sound as well as change structures. The general framework that we adopt for our approach to structured audio representation is to specify sound structures in terms of the three structural hierarchical elements of Fourier-time persistence, short-time change and high-level change structures.

The goal of this thesis is to provide a set of working methodologies for representing the internal structure of natural sound events and re-purposing this structure for generating new sound events. If we can represent the said structural elements by algorithmic methods, and lift invariants out of sound-event recordings, then these elements form the basis of the perceptually meaningful structure description of a sound and constitute structured feature descriptions of sound events. This structured representation can then be used to re-purpose the invariant components by applying modifications that give rise to the perception of a specifiable change event.

In the next chapter we set out to delimit some of the invariants in physical acoustic systems, and to develop the mathematical framework by which such invariants and their transformations can be applied to the problem of structured audio representation.