Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio

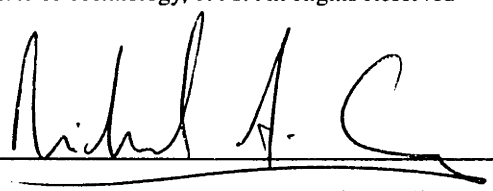
Michael Anthony Casey

B.A. (Hons),
University of East Anglia, 1990
Norwich, U.K.

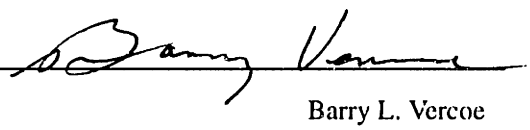
A.M.,
Dartmouth College, 1992

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy at the Massachusetts Institute of Technology

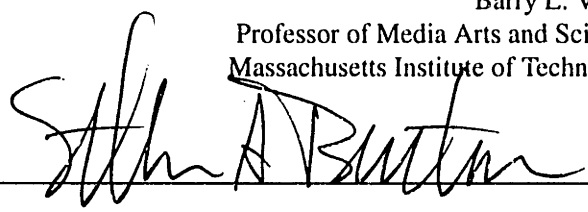
February, 1998

Author 

Program in Media Arts and Sciences
January 9th, 1998

Certified by 

Barry L. Vercoe
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Accepted by 

Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences
Massachusetts Institute of Technology

Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio

Michael Anthony Casey

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning on January 9, 1998, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy.

Abstract

To date there have been no audio signal representation methods capable of characterizing the everyday sounds that are used for sound effects in film, TV, video games and virtual environments. Examples of these sounds are footsteps, hammering, smashing and spilling. These environmental sounds are generally much harder to characterize than speech and music sounds because they often comprise multiple noisy and textured components, as well as higher-order structural components such as iterations and scatterings. In this thesis we present new methods for approaching the problem of automatically characterizing and extracting features from sound recordings for re-purposing and control in structured media applications.

We first present a novel method for representing sound structures called auditory group theory. Based on the theory of local Lie groups, auditory group theory defines symmetry-preserving transforms that produce alterations of independent features within a sound. By analysis of invariance properties in a range of acoustical systems we propose a set of time-freqeuency transforms that model underlying physical properties of sound objects such as material, size and shape.

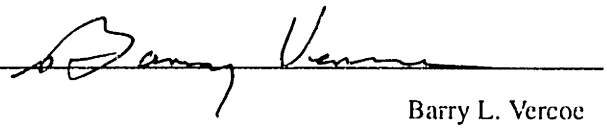
In order to extract features from recorded sounds we have developed new statistical techniques based on independent component analysis (ICA). Using a contrast function defined on cumulant expansions up to fourth order, the ICA transform generates an orthogonal rotation of the basis of a time-frequency distribution; the resulting basis components are as statistically independent as possible. The bases are used in conjunction with auditory group transforms to characterize the structure in sound effect recordings. These characteristic structures are used to specify new sounds with predictable, novel features.

For our results we have implemented auditory group models that are capable of synthesizing multiple sound behaviors from a small set of features. These models characterize event structures such as impacts, bounces, smashes and scraping as well as physical object properties such as material, size and shape. In addition to applications in video and film media, the methods presented herein are directly applicable to the problem of generating real-time sound effects in new media settings such as virtual environments and interactive games, as well as creating new sound synthesis methods for electronic music production and interactive music experiences.

Thesis Advisor: Professor Barry L. Vercoe
Professor of Media Arts and Sciences

This work was performed at the MIT Media Laboratory. Support for this work was provided in part by MERL - A Mitsubishi Electric Research Laboratory.

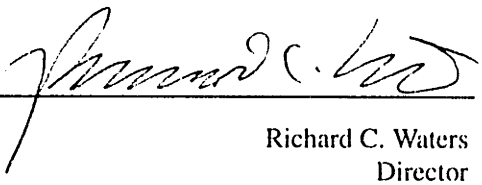
Doctoral Dissertation Committee

Thesis Advisor 

Barry L. Vercoe
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis Reader 

Whitman Richards
Professor of Cognitive Science
Massachusetts Institute of Technology

Thesis Reader 

Richard C. Waters
Director
MERL - A Mitsubishi Electric Research Laboratory

Acknowledgements

Many people have contributed to the development of these ideas and methods over the last few years. First and foremost I would like to thank my advisor, Barry Vercoe, for supporting this research and for providing an environment in which it has been possible to freely pursue independent avenues of thought.

I would also like to thank the other members of my doctoral committee, Whitman Richards and Richard Waters, both of whom have provided lucid and valuable insights into my work and have contributed to its development in numerous ways. During the final months of preparation I was fortunate to have conversations with many experts in the fields of signal processing, auditory perception and sound modeling; in particular I am grateful to Michelle Covell, Bill Freeman, William Gaver, Gaile Gordon, Marvin Minsky and Malcolm Slaney for helping me to strengthen my methods and arguments.

I firmly believe that the work environment makes the greater part of the learning experience and for this I would like to thank my colleagues in the Machine Listening Group at the MIT Media Lab. I have had three office mates, Jeff Bilmes, Eric Scheirer and Paris Smaragdis, all of whom have freely shared their time, opinions and expertise. Paris provided the necessary encouragement for me to take the bold step from using SVD decomposition techniques to formulating an ICA algorithm for time-frequency analysis; this was indeed a valuable contribution to my work. In addition to my office mates, during my five years at MIT the Machine Listening Group has comprised others to whom I wish to extend my gratitude: Dan Ellis, Jonathan Feldman, Bill Gardner, Youngmoo Kim, Adam Lindsay, Keith Martin, Joe Pompei and Nicolas Saint-Arnaud. Also thanks to Molly Bancroft, Betty Lou McClanahan and Connie Van Rheenen for valuable support at the Media Lab.

In addition to my doctoral advisor, I would like to acknowledge three other advisors; Cedric Carnell at Lutterworth Grammar School, Denis Smalley at the University of East Anglia and Jon Appleton at Dartmouth College for encouraging my path into higher education. Thanks are also due to Simon Atkinson, Ray Guillette and George Kachadorian for playing such an active role in the development of this work over the last several years.

Finally I would like to thank Amy van der Linde, my parents Connie and Tony, and my sister Angela for all their love and support during these last few years. For this, I dedicate my thesis to you.

Contents

Introduction

Structured Audio

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

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

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

Controllability

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

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

Scalability

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

Compactness

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

Ideas to be Investigated

Structured Audio Event Representation

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

Feature Extraction

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

Structured Re-purposing and Control

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

Applications for Structured Audio

Automatic Foley

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

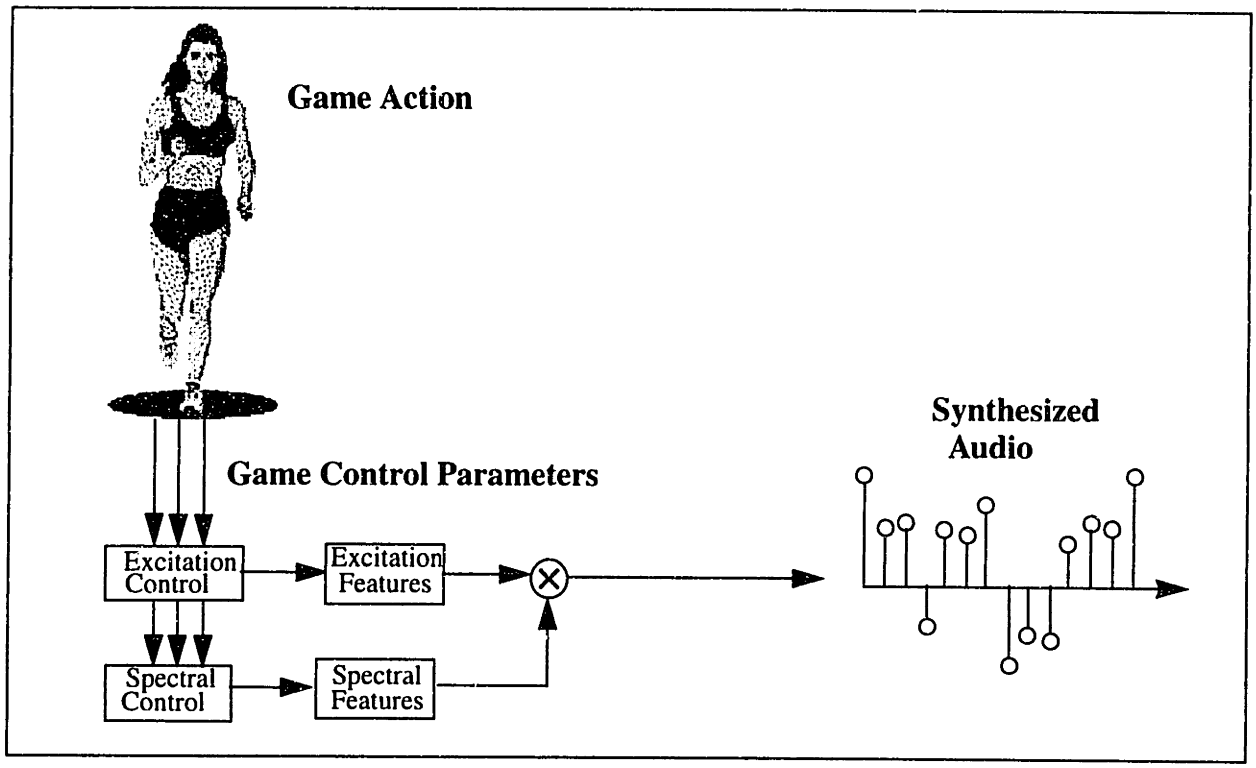


FIGURE 1. Automatic Foley Generation. The audio for an interactive game can be generated from a structured audio description of the materials and action parameters of a scene. This allows automatic synthesis of appropriate sounds such as the footsteps of a runner, which depend on the motion dynamics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

Producer and Sound Designer's Assistant

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

Low-Bandwidth Audio Representations

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

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

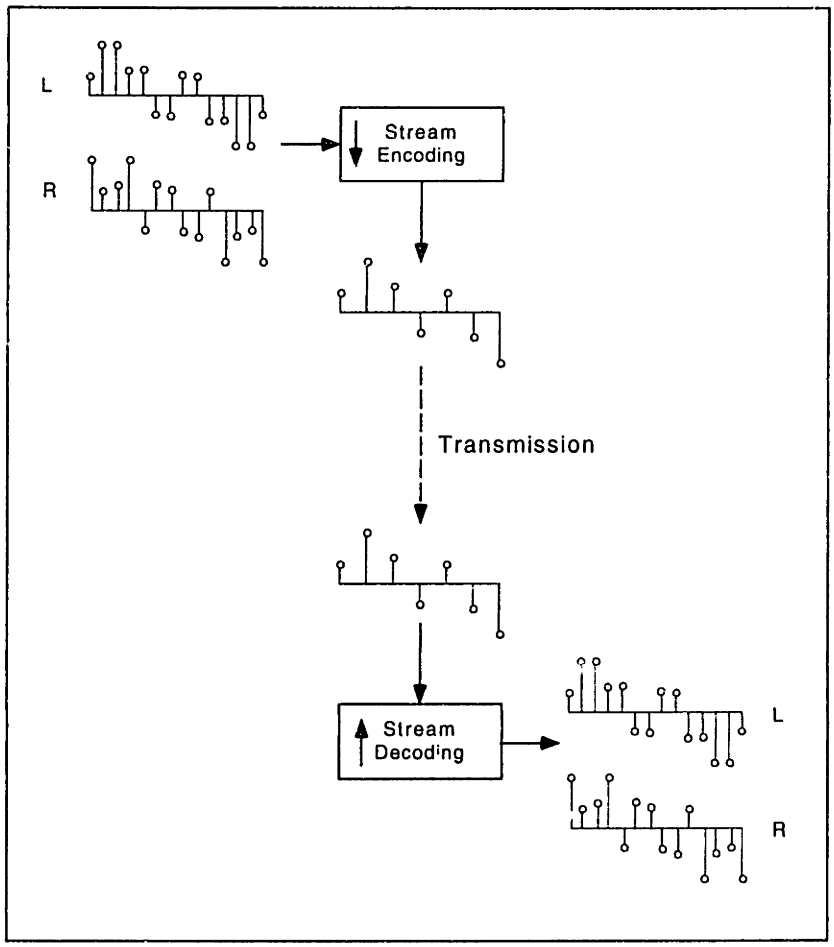


FIGURE 2. Streaming audio flow diagram. An audio source is compressed into a smaller representation using a stream encoder. Encoded streams must be decoded at the receiving end before being rendered.

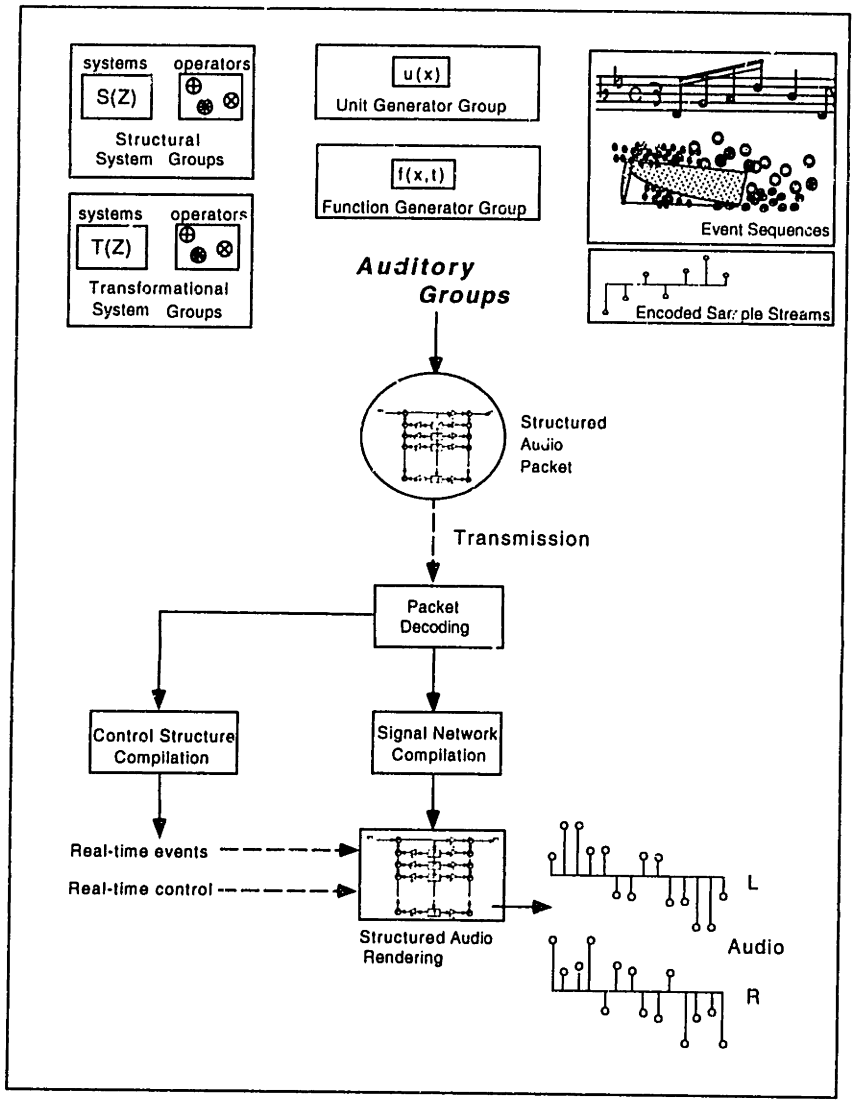


FIGURE 3. Structured audio flow diagram. Structured audio sound events are represented using a combination of elementary building blocks called auditory invariants which can be represented by mathematical groups for the purposes of formal description. It is this representation that is transmitted rather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

Auditory Invariants

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

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

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

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

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

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

Thesis Overview and Scope

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

Chapter 1: Ecological Acoustics

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

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

Chapter 2: Auditory Group Theory

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

Chapter 3: Statistical Basis Decomposition of Time-Frequency Distributions

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

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

Chapter 4: Structured Sound Effects using Auditory Group Transforms

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

Scope of Current Work and Results / Findings

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

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.

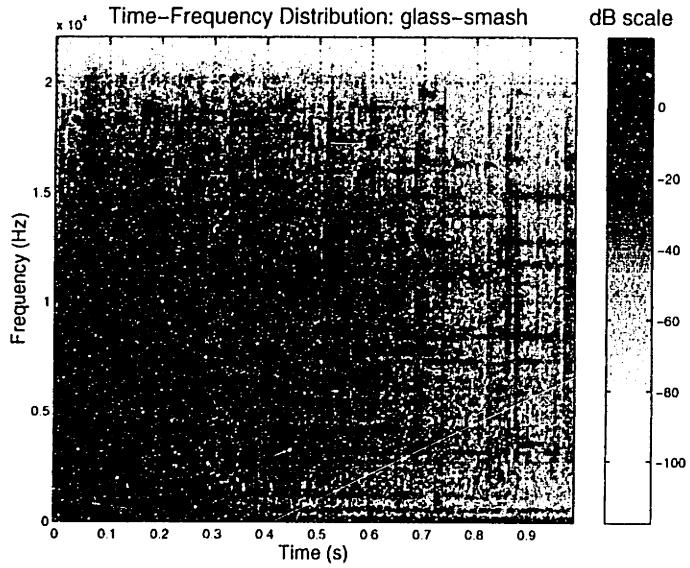


FIGURE 4. Spectrogram of the sound of a smashing glass. There are many components to this sound, such as low-frequency decaying impact noise and high-frequency particle scattering; but we perceive a single event: smashing.

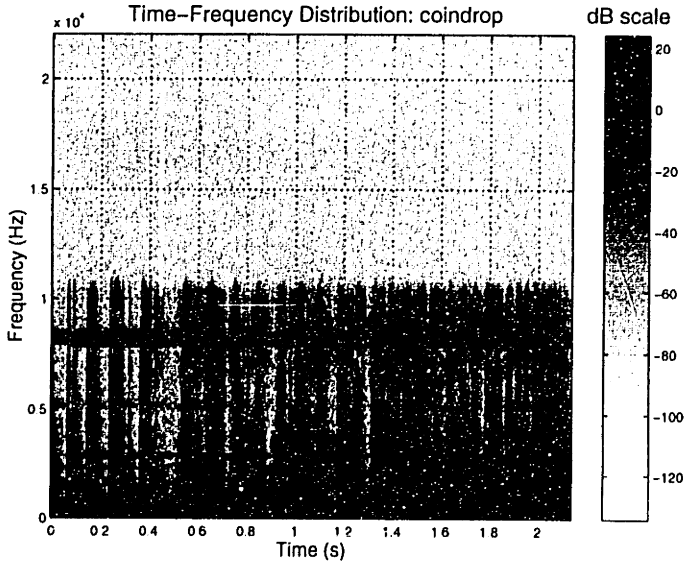


FIGURE 5. Spectrogram of a coin dropping sound. The features of this sound are a sequence of impacts that get closer in time. The metal rings after each bounce thus creating a constant high-frequency component.

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; Kisset 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.

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 $Z(x, y, t) = X(x)Y(y)T(t)$. Following Fletcher and Rossing (1991), the displacement amplitude is given by:

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

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:

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

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:

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

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:

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

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:

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

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 notes, "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:

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

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:

$$\mathbf{Ka} = \mathbf{k} \qquad [13]$$

where

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

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

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

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:

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

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:

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

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:

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

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

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

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

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

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:

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

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:

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

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:

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

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:

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

where X, Y and Z are the orthogonal dimensions, E_i and E_j are the elastic moduli of which there are three and ν_{ij} and ν_{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 ν for their isotropic counterparts. For plates the effect of this transform is to produce two different propagation speeds in the different dimensions:

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

where

$$c_x = \sqrt{\frac{E_x}{\rho(1-\nu_x\nu_y)}}, \qquad [29]$$

and

$$c_y = \sqrt{\frac{E_y}{\rho(1-\nu_x\nu_y)}}. \qquad [30]$$

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:

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

is the speed for longitudinal waves where E is the Young's modulus, ν 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:

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

The equation for longitudinal waves is given by:

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

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)

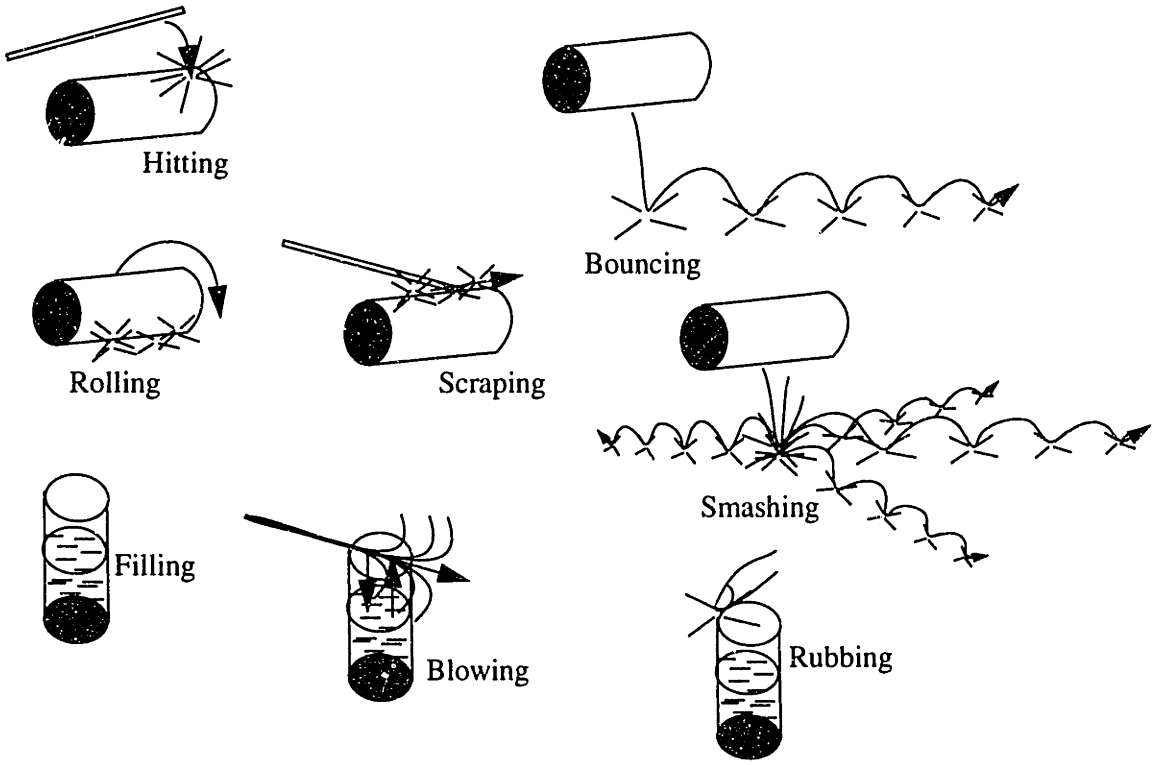


FIGURE 6. The many affordance structures of a glass container. Each of these actions produces a different sound, but the underlying physical object remains the same until fractured.

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 **W** and a transformation T_E which preserves a part of **W** that we shall call **S** and alters a complimentary part of **W** that we shall call **E**. We say that T_E is a *symmetry-preserving* operation with respect to **S** and a *symmetry-breaking* operation with respect to **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 **W** implies that some component of **W** called **E** is changed and some component of **W** called **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, **S** and **E** are factors of **W**. We now define a second operation on **W**, denoted by T_S, which performs the complimentary operation with respect to the components **E** and **S**. Thus T_S alters **S** and leaves **E** unchanged. We shall define the relationships of this operation in terms of its actions on the elements of **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 **S** is the structural invariant of the operation T_E and the transformational invariant of the operation T_S, then conversely, **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 **E** and **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:

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

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

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

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:

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

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

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

which specifies the following partial differential equation:

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

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.

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

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

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

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:

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

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:

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

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:

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

for which the resulting Lagrangian can be written simply as:

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

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 structure. 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.

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

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:

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

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:

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

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

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

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:

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

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:

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

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:

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

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:

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

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:

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

where $T_{\pi_{\varepsilon l}}$ 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 $T_{\pi_{\varepsilon l}}$ 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.

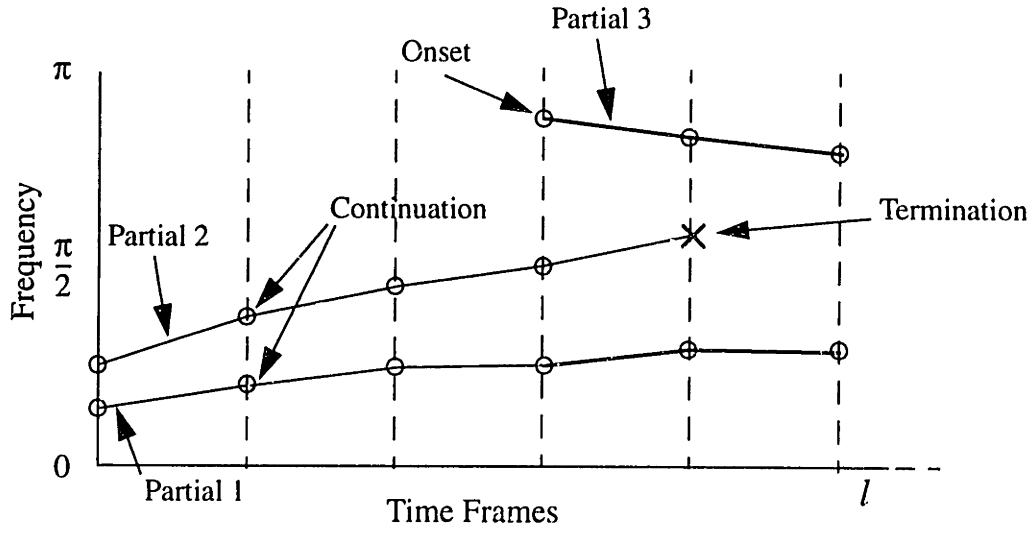


FIGURE 7. Sinusoidal Peak Tracking of a Time-Frequency Distribution

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:

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

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:

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

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:

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

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:

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

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.

Chapter III: Statistical Basis Decomposition of Time-Frequency Distributions

3.1 Introduction

In the previous chapters we outlined the need for a method for decomposing an input time-frequency distribution (TFD) into independently controllable features that can be used for re-synthesis. In this chapter we describe a suite of techniques, related to principal component analysis (PCA), that decompose a TFD into statistically independent features. As we shall show in this chapter, statistically-independent decomposition of a Gaussian distributed TFD is performed by a singular value decomposition (SVD). For non-Gaussian TFDs we develop an independent component analysis (ICA) algorithm.

We first introduce the concept of PCA and the necessary mathematical background. We then consider the computation of a robust PCA with SVD and develop the theory for SVD in Section 3.3.7. In Section 3.3.8 we give an example of the application of SVD to sound-structure modeling which demonstrates the potential merits of the technique. We then consider some important limitations of SVD in Section 3.3.15 which are due to the implicit dependence on second-order statistics only. In Section 3.3.16 we consider extensions to SVD to include higher-order statistical measures and, in Section 3.3.18, we consider an information-theoretic interpretation of PCA which provides the framework for developing a higher-order independent component analysis (ICA) algorithm for feature decomposition.

3.2 Time Frequency Distributions (TFDs)

As with any signal characterization scheme, there must be a front-end which decomposes the signal into low-level mathematical objects for further treatment. In this section we shall outline several representations which could be used for a front-end analysis, and we make our choice for further development based on several design criteria; i) efficiency of the transform, ii) data preservation and invertibility, iii) ease of implementation.

Most of the salient characteristics of audio signals exist in the short-time spectro-temporal domain. That is the domain of representation of a signal in which time-varying spectral features can be represented directly without need for further transformation. An example of such an analysis is the well-known short-time Fourier transform (STFT).

3.2.1 Desirable Properties of the STFT as a TFD

Although the short-time Fourier transform is limited in its characterization abilities it does have several very desirable properties. Firstly it can be implemented extremely efficiently using the fast Fourier transform (FFT). For most sound analysis applications an FFT-based analysis will run in real time on standard microcomputer hardware. Secondly, since the Fourier transform can be thought of as a linear operator, there are well-defined signal-processing operations which produce stable, invertible results that are easily implemented without the call for compensating machinery as is the case for many other TFD representations. Thus, for the purposes of sound modeling, we consider that the STFT is a reasonable time-frequency representation.

The main problem of interest with the STFT, as with most TFD representations, is in the redundancy of spectral information. The STFT with an appropriately selected analysis frequency errs on the side of inclusion rather than on the side of omittence of important information. Therefore it is with little loss in generality that we choose the STFT as a front-end frequency analysis method in the following sections. It should be emphasized, however, that all of the statistical basis reduction techniques presented can be applied to any TFD; we shall give examples of the application of statistical basis reduction methods to alternate TFDs in Chapter IV.

3.2.2 Short-Time Fourier Transform Magnitude

It was Ohm who first postulated in the early nineteenth century that the ear was, in general, phase deaf, Risset and Mathews (1969). Helmholtz validated Ohm's claim in psycho-acoustic experiments and noted that, in general, the phase of partials within a complex tone (of three or so sinusoids) had little or no effect upon the perceived result. Many criticisms of this view ensued based on faulting the mechanical acoustic equipment used, but Ohm's and Helmholtz' observations have been corroborated by many later psycho-acoustic studies, Cassirer (1944).

The implications of Ohm's acoustical law for sound analysis are that the representation of Fourier spectral components as complex-valued elements possessing both a magnitude and phase component in polar form is largely unnecessary, and that most of the relevant features in a sound are represented in the magnitude spectrum. This view is, of course, a gross simplification. There are many instances in which phase plays an extremely important role in the perception of sound stimuli. In fact, it was Helmholtz who noted that Ohm's law didn't hold for simple combinations of pure tones. However, for non-simple tones Ohm's law seems to be well supported by psycho-acoustic literature.

Consideration of Ohm's acoustical law has lead many researchers in the speech and musical analysis/synthesis community to simplify Fourier-based representations by using the magnitude-only

spectrum. In the case of the STFT this results in a TFD known as the short-time Fourier transform magnitude (STFTM), Griffin and Lim (1989). The downside in using the STFTM representation appears at the re-synthesis stage. Because the phases have been eliminated a phase-model must be estimated for a given STFTM, the phase must be constrained in such a manner as to produce the correct magnitude response under the operation of an inverse Fourier transform and Fourier transform pair. This property of a phase model is expressed by the following relation:

$$|\hat{\mathbf{Y}}| \approx \left|\mathrm{FT}\{\mathrm{FT}^{-1}\{|\mathbf{Y}|e^{-j\hat{\phi}}\}\}\right| \qquad [82]$$

where $|\mathbf{Y}|$ is a specified STFTM data matrix, $\hat{\phi}$ is a phase-model matrix and $|\hat{\mathbf{Y}}|$ is the approximated magnitude response matrix for the given magnitude specification and phase model. Since there is a discrepancy between $|\hat{\mathbf{Y}}|$ and $|\mathbf{Y}|$ for most values of $\hat{\phi}$ a least-squares iterative phase estimation technique is used to derive the phase model, Griffin and Lim (1984). We discuss this technique further in the next chapter.

Without loss of generality, then, we will use the STFTM representation in the examples given in this chapter. The algorithms are defined for complex-valued spectra but work on magnitude-only spectra without the need for modification.

3.2.3 Matrix Representation of TFDs

We represent an arbitrary TFD by a matrix $\mathbf{X}$ which we refer to as the data matrix. In the case of the STFT the data matrix is:

$$\mathbf{X}_{mn} = \mathrm{X}[l, k] \qquad [83]$$

where m and n are the row and column indices of a matrix $\mathbf{X}$. Thus the data matrix can be thought of as a two-dimensional plane with points (m, n). This interpretation of the data matrix will be useful when we discuss applications of auditory group transforms to TFDs.

The statistical basis reduction techniques discussed later in this chapter are sensitive to the orientation of the data matrix. The is due largely to the consideration of *variates* in vector form for which measures of a particular variable occupy the columns of a matrix. Thus a data matrix has the *variates* in the columns and the *observations* in the rows.

3.2.4 Spectral Orientation

As defined above the data matrix is in spectral orientation. That is, the variates are functions of the frequency variable $\omega_k = \frac{2\pi k}{N}$. There are N columns such that each column represents the complex spectral value of a signal at a particular frequency $\frac{2\pi n}{N}$ where n is the column index of the data matrix. Thus in spectral orientation the observations are the time-varying values of the spectrum at a particular frequency.

$$\mathbf{X} = \begin{bmatrix} X(1,0) & X(1,1) & \ldots & X(1,N-1) \\ X(2,0) & X(2,1) & \ldots & X(2,N-1) \\ \ldots & \ldots & \ldots & \ldots \\ X(M,0) & X(M,1) & \ldots & X(M,N-1) \end{bmatrix} \quad [84]$$

The corresponding covariance matrix is $N \times N$ and is defined by:

$$\Phi_{\mathbf{X}} = E[\mathbf{X}^T\mathbf{X}] - \mathbf{m}^T\mathbf{m} \quad [85]$$

where $\mathbf{m}$ is a vector of column means for the data matrix.

3.2.5 Temporal Orientation

An alternative method of representing a TFD using matrix notation is to orient the matrix temporally. The variates are functions of the time-frame variable l and the observations operate through frequency.

$$\mathbf{X} = \begin{bmatrix} X(1,0) & X(2,0) & \ldots & X(M,0) \\ X(1,1) & X(2,1) & \ldots & X(M,1) \\ \ldots & \ldots & \ldots & \ldots \\ X(1,N-1) & X(2,N-1) & \ldots & X(M,N-1) \end{bmatrix} \quad [86]$$

In spectral orientation the covariance matrix is $M \times M$. In general the choice of orientation of a data matrix is determined by the desirable characterization properties of any subsequent analysis. If the matrix is in temporal orientation then a covariant statistical analysis, one that relies upon the covariance matrix, will yield results that are sensitive to the time-frame variates. However, since for most sound analysis purposes $M \gg N$, the cost of computation of the covariance and subsequent decomposition can be prohibitively great, or at least many orders of magnitude greater than computing the covariance in spectral orientation, see Sandell and Martins (1995).

3.2.6 Vector Spaces and TFD Matrices

For a given TFD in spectral orientation the frequency variates span the column space and the observations span the row space of the data matrix. The row vector space is generally much larger than the column vector space in spectral orientation, and the converse is true of temporal orientation.

1. Column Space of a TFD

The column space $\Re(\mathbf{X})$ of an $m \times n$ TFD matrix is a subspace of the full m-dimensional space which is $\mathbf{R}^m$ in the case of a spectrally-oriented STFTM representation, that is the m-dimensional vector space spanning the field of reals. The dimension of the column space is of interest to us here. It is defined as the rank $\mathbf{r}$ of the matrix which is the number of linearly independent columns.

2. Row Space of a TFD

Conversely, the row space $\Re(\mathbf{X}^{\mathrm{T}})$ is a subspace of $\mathbf{R}^{n}$. Thus the co-ordinates represented by a set of observations can be thought of as a linear combination of the column vectors, which are the *basis*, and conversely the basis functions themselves can be thought of as a linear combination of observations. The dimension of the row space is the rank $\mathbf{r}$ which is also the number of linearly independent rows.

3. Null Space of a TFD

TFDs contain a good deal of redundancy. This redundancy manifests itself in the null space of the data matrix. For an $m \times n$ TFD matrix the null space is of dimension $n - \mathbf{r}$ and is spanned by a set of vectors which are a basis for the null space. For TFD data matrices the null space arises from the correlated behavior between the variates. The correlations between frequency bins in a spectrally-oriented TFD data matrix are expressed as linear dependencies in the vector spaces. Thus information about one of the correlated components is sufficient to specify the other components, therefore the remaining components are not well-defined in terms of a vector space for the TFD matrix. In many cases the dimensionality of the null-space of a TFD is, in fact, larger than the dimensionality of the column and row spaces, both of which are $\mathbf{r}$. From this observation we form a general hypothesis about the vector spaces of TFDs:

$$rank\{\aleph(\mathbf{X})\} \gg rank\{\Re(\mathbf{X})\} \qquad [87]$$

from which it follows that $n \gg 2\mathbf{r}$. Estimation of the rank of the TFD thus provides a measure of the degree of redundancy within a sound with respect to the chosen basis of the TFD.

3.2.7 Redundancy in TFDs

For any given frequency analysis technique, the chosen basis functions for projecting a signal into the time-frequency plane are extremely elemental. In the case of the STFT these basis functions are a set of complex exponentials linearly spaced in frequency. Each analysis bin of an STFT frame is thus a projection of the time-domain signal onto an orthogonal basis spanned by the said exponential functions. In the case of the continuous Fourier transform the basis is infinite thus defining a Hilbert space and the DFT (which is used by the STFT) effectively samples this space at discrete intervals. Such a basis is designed to span all possible complex-valued sequences representing each spectral component as a point in a high-dimensional space. Indeed Fourier's theorem states that *any* infinitely-long sequence can be decomposed into an infinite sum of complex exponentials. Thus each infinitesimal frequency component within a signal gets an independent descriptor.

Clearly natural sounds are not this complex. There is a good deal of redundancy in the signals. Much of the redundancy is due to the grouped nature of physical vibrations. That is, a set of frequencies generated by a system are generally related to a fundamental mode of vibration by some form of statistical dependence. We have seen this behavior in the form of the acoustical equations given in Chapter II. The inter-partial dependence of a sound spectrum may be a linear function, a non-linear function, resulting in harmonic, inharmonic or stochastic components, but in each case there is a non-zero joint probability between many of the marginal components defined for each

frequency of vibration. Such statistical dependence within a sound results in uniform motion of a set of points in the time-frequency plane of a TFD. The motion may be linear or non-linear but, nevertheless, the resulting spectrum is statistically dependent to some degree.

Redundancy is an important value for information in the signal since by eliminating it we are able to see what actually varies during the course of a sound and, by inspecting it, we see what stays essentially the same. In fact, the concept of redundancy has been the subject of some perceptual theories. For example, Barlow (1989) considers the concept of redundancy to be fundamental to learning and argues that it is redundancy that allows the brain to build up its "cognitive maps" or "working models" of the world.

Somewhat less ambitious is the claim that redundancy in the low-level projection of a sound onto a spectral basis is a necessary component to extracting meaningful features from the sound, or at least it is a good point of departure for investigating methods for characterizing the structure of natural sounds. This observation leads quite naturally to an information theoretic interpretation of the task of feature extraction and characterization of natural sound TFDs.

3.3 Statistical Basis Techniques for TFD Decomposition

3.3.1 Introduction

In view of the prevailing redundancy in TFDs we seek methods for identifying the null space and characterizing the row and column spaces in terms of a reduced set of basis vectors. The general hypothesis is that the reduced space will represent salient information in the TFD. A stronger hypothesis is that the redundancy-reduced basis may represent the perceptually most important information in the signal. These are the ideas to be investigated in this section.

3.3.2 Principal Component Analysis (PCA)

Principal component analysis was first proposed in 1933 by Hotelling in order to solve the problem of decorrelating the statistical dependency between variables in multi-variety statistical data derived from exam scores, Hotelling (1933). Since then, PCA has become a widely used tool in statistical analysis for the measurement of correlated data relationships between variables, but it has also found applications in signal processing and pattern recognition for which it is often referred to as the Karhunen-Loeve transform, Therrien (1989). The use of PCA in pattern recognition is born out of its ability to perform an optimal decomposition into a new basis determined by the second-order statistics of the observable data.

3.3.3 Previous Audio Research using PCA

The use of Principal Component Analysis for audio research can be traced back to Kramer and Mathews (1956) in which a PCA is used to encode a set of correlated signals. In the 1960s there was some interest in PCA as a method for finding salient components in speech signals, of particular note is the work of Yilmaz on a theoryof speech perception based on PCA, (Yilmaz 1967a,

1967b, 1968), and the application of PCA to vowel characterization, (Plomp *et al.* 1969; Klein *et al.* 1970; Zahorian and Rothenburg 1981). Yilmaz was concerned with the identification of invariants in speech, thus his work is perhaps the most relevant to the current work. PCA has also been applied in the processing of audio signals for pattern recognition applications by basis reduction of the Short-Time Fourier Transform (STFT), Beyerbach and Nawab (1991), and in modeling Head-Related Transfer Functions for binaural signal modeling, Kistler & Wightman (1992).

In addition to speech and acoustical encoding, PCA of musical instrument sounds has been researched quite extensively, (Stautner 1983; Stapleton and Bass 1988; Sandell and Martens 1996). The results for musical instrument modeling are reported to be of widely varying quality with little or no explaination of why some sounds are better characterized than others by a PCA. In the following sections we develop an argument that suggests some important limitations with PCA, and with its numerical implementation using SVD. This leads us to a new approach for decomposing time-frequency representations of sound into statistically salient components.

3.3.4 Definition of PCA

PCA has many different definitions but they all have several features in common. These can be summarized as follows:

PCA Theorem: *The k-th principal component of the input vector* $\mathbf{x}$ *is the normalized eigenvector* v_k *corresponding to the eigenvalue* λ_k *of the covariance matrix* Φ_x, *where the eigenvalues are ordered* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$.

where the covariance matrix is defined in Equation 85. A proof of this theorem may be found in Deco and Obradovic (1996). A PCA is, then, a linear transform matrix $\mathbf{U}$ operating on a TFD matrix $\mathbf{X}$ as follows:

$$\mathbf{Y} = \mathbf{XV} \qquad [88]$$

with $\mathbf{Y}$ representing the linearly-transformed TFD matrix. If the rows of the linear transformation matrix $\mathbf{V}^T$ are the eigenvectors of the covariance matrix Φ_x then it is said to perform a Karhunen-Loeve Transform of the input column space $\Re(\mathbf{X})$. In this case $\mathbf{V}$ is an orthonormal matrix and thus satisfies the following relations:

$$\mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I} \qquad [89]$$

and the relationship between the input and output covariance can be expressed as:

$$\Phi_y = \mathbf{V}^T\Phi_x\mathbf{V} = \mathbf{V}^T\mathbf{V}\Sigma = \Sigma \qquad [90]$$

where Σ is a diagonal matrix of eigenvalues which correspond to the variances of a set of independent Gaussian random variables which span the input space:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma_N^2 \end{bmatrix}. \qquad [91]$$

(For a derivation of diagonalization of the covariance matrix see Appendix II). Under this definition, the PCA essentially linearly decorrelates the output variates in $\mathbf{Y}$ such that each column is statistically independent to second order with respect to the other columns. Traditionally, in statistical texts, the matrix of eigenvectors $\mathbf{V}$ is referred to as the weights matrix and the linearly transformed matrix $\mathbf{Y}$ is referred to as the scores of the PCA. This nomenclature follows Hotelling's original formulation.

3.3.5 Joint Probability Density Functions and Marginal Factorization

We now assume a statistical interpretation of TFD data matrix variates. The probability density function (PDF) of each column of the input is defined as a marginal density in the joint probability density of the column space of the TFD. A definition of statistical independence is derived in the form of the relationship between the joint probability distribution of the columns of a TFD and the individual column marginal distributions. Specifically, the output columns are statistically independent if and only if:

$$p_{\mathbf{Y}}(\mathbf{Y}) = \prod_{i=1}^{N} p_{\mathbf{Y}_i}(\mathbf{Y}_i) \qquad [92]$$

that is, the output marginals are independent PDFs if and only if their joint density function can be expressed as a product of the marginals. In the case of Gaussian input densities, PCA decorrelates the input PDF to second order and thus exhibits the marginal factorization property described by Equation 92, see Comon (1994) and Deco and Obradovic (1996).

3.3.6 Dynamic Range, Scaling, Rank, Vector Spaces and PCA

There are several problems with PCA as defined above for the purposes of TFD decomposition. The first is that since PCA is defined as a diagonalization of the input covariance, the system loses sensitivity to lower magnitude components in favor of increasing the sensitivity of higher magnitude components. This is because the input covariance is essentially a *power* representation of the input variates. The result of transforming to a power representation is a loss in dynamic range due to finite word-length effects and numerical precision in floating-point implementations.

This relates to an issue on the usefulness of PCA in general. PCA depends on the scaling of the input coordinates. This is referred to in the literature as the "scaling problem". The problem manifests itself in the solution of the diagonalization using eigenvalues. The pivoting requires scaling of each row in order to yield a Gaussian elimination, the condition number of the TFD matrix deter-

mines the sensitivity of the data to scaling and whether or not the matrix is indeed singular to working precision.

PCA does not define a solution when the columns of the input matrix are linearly dependent. In this case the null space of the matrix is non empty. In fact, for TFDs we have already developed the hypothesis that the null space is in fact much larger than the row and column space of the data matrix, see Equation 87. Equivalently we can interpret the identification of the size of the null space as a rank estimation problem, we can see this in the relation defined in Equation 95. The PCA definition as diagonalization of the covariance does not explicitly provide a method for handling the null space of a matrix. This is because methods involving the identification of eigenvalues rely on full-row rank of the data covariance matrix. Therefore this form of PCA is of little practical use in implementing redundancy reduction techniques for TFDs. However, we shall refer to the canonical PCA form in our discussions in the following sections since the basic theoretical framework is somewhat similar for the null-space case as well as the case of non-Gaussian input distributions.

Another problem with the definition of PCA in the form outlined above is that the resulting basis spans only the column space of the input. Thus it does not generalize to the problem of identifying a basis for the row space. The covariance matrix is necessarily square which renders it invertible under the condition of full column rank. The covariance is also a symmetric matrix which is defined by the relation $\Phi_x = \Phi_x^T$, thus the row space and the column space of the input representation are collapsed to the same space, namely a power representation of the column space of the TFD. In performing a PCA using the covariance method we are thus discarding information about the space of row-wise observations in favor of characterizing the column-wise variates.

In order to address the problems of dynamic range and row/column space basis identification, we seek a representation which does not rely on the covariance method; rather, the sought method should directly decompose the input TFD into a separate basis for the row and column space of the TFD data matrix. We know that the rank of the row and column spaces is equal thus the null space will be the same from both points of view.

We now develop practical techniques for decorrelating input components of a TFD. These techniques are defined so as to address the problems of dynamic range, scaling, vector-space representation and matrix rank that we have discussed in this section.

3.3.7 The Singular Value Decomposition (SVD)

The singular value decomposition has become an important tool in statistical data analysis and signal processing. The existence of SVD was established by the Italian geometer Beltrami in 1873 which was only 20 years after the conception of a matrix as a multiple quantity by Cayley. As we shall see, the singular value decomposition is a well-defined generalization of the PCA that addresses many of the problems cited above.

A singular value decomposition of an $m \times n$ matrix $\mathbf{X}$ is any factorization of the form:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} \qquad [93]$$

where $\mathbf{U}$ is an $m \times m$ orthogonal matrix; i.e. $\mathbf{U}$ has orthonormal columns, $\mathbf{V}$ is an $n \times n$ orthogonal matrix and Σ is an $m \times n$ diagonal matrix of singular values with components $\sigma_{ij} = 0$ if $i \neq j$ and $\sigma_{ii} \geq 0$; (for convenience we refer to the i th singular value $\sigma_i = \sigma_{ii}$). Furthermore it can be shown that there exist non-unique matrices $\mathbf{U}$ and $\mathbf{V}$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0$. The columns of the orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ are called the left and right singular vectors respectively; an important property of $\mathbf{U}$ and $\mathbf{V}$ is that they mutually orthogonal.

We can see that the SVD is in fact closely related to the PCA. In fact the matrix product $\mathbf{U}\Sigma$ is analogous to the matrix $\mathbf{Y}$ defined for PCA:

$$\mathbf{Y} = \mathbf{XV} = \mathbf{U}\Sigma \qquad [94]$$

Because both the singular vectors defined for an SVD are square and have orthonormal columns their inverses are given by their transposes. Thus $\mathbf{V}^{-1} = \mathbf{V}^{\mathrm{T}}$. Now the relation in Equation 94 can be expressed $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$ which is the definition of an SVD.

The first major advantage of an SVD over a PCA is that of rank estimation and null-space identification. $\aleph\{\mathbf{X}\}$ can be identified for both the left and right singular vectors as the space spanned by vectors corresponding to the singular values for which $\sigma_j = 0$, whereas if $\sigma_j \neq 0$ then the corresponding singular vectors $\mathbf{U}_j$ and $\mathbf{V}_j$ are in the range of $\mathbf{X}$ which is spanned by the column space of the left and right singular vectors which, in turn, span the row space and column space of the data matrix $\mathbf{X}$.

The upshot of these observations is that we can construct a basis for each of the vector spaces of $\mathbf{X}$. Recalling the relation between the rank of the null space and the rank of the row and column spaces of a matrix:

$$rank\{\aleph(\mathbf{X})\} = N\text{–}rank\{\Re(\mathbf{X})\} \qquad [95]$$

the SVD provides a theoretically well-defined method for estimating the rank of the null space, specifically it is the number of zero-valued singular values. This in turn defines the rank of the data matrix $\mathbf{X}$.

The SVD defined thus has implicitly solved the problems inherent in the PCA definition. Firstly, the SVD decomposes a non-square matrix, thus it is possible to directly decompose the TFD representation in either spectral or temporal orientation without the need for a covariance matrix. Furthermore, assuming a full SVD, the decomposition of a transposed data matrix may be derived from the SVD of its complimentary representative by the relation:

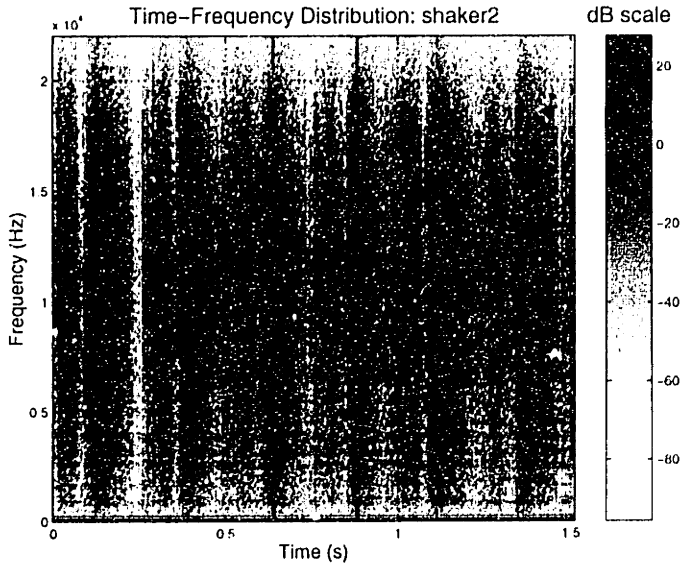


FIGURE 8. Short-time Fourier transform TFD of 1.5 seconds of a percussive shaker. The vertical dark regions are the shake articulations. Analysis parameters are N=1024, W=512, H=256. Sample rate is 44.1kHz.

$$\mathbf{X}^{\mathrm{T}} = \mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}} \qquad [96]$$

which follows from the relation $\Sigma^{\mathrm{T}} = \Sigma$. This means that the full SVD decomposition of a matrix in spectral orientation can be used to specify an SVD decomposition in temporal orientation and *vice-versa*. Thus the direct SVD decomposition keeps all the relevant information about the null, row and column spaces of a data matrix in a compact form.

Since the SVD decomposes a non-square matrix directly without the need for a covariance matrix, the resulting basis is not as susceptible to dynamic range problems as the PCA. Thus, components of a TFD that lye within the working precision of a particular implementation are not corrupted by squaring operations. Theoretically it is not in fact possible to invert a non-square matrix. Thus implementation of a SVD is a compromise between the theoretical definition and practically tractable forms. The machinery of compromise in the SVD is the psuedoinverse of a matrix.

3.3.8 Singular Value Decomposition of Time-Frequency Distributions

3.3.9 A Simple Example: Percussive Shaker

Figure 8 shows the STFTM TFD of a percussive shaker instrument being played in regular rhythm. The observable structure reveals wide-band articulatory components corresponding to the shakes and a horizontal stratification corresponding to the ringing of the metallic shell. What does not show clearly on the spectrogram is that the rhythm has a dual shake structure with an impact

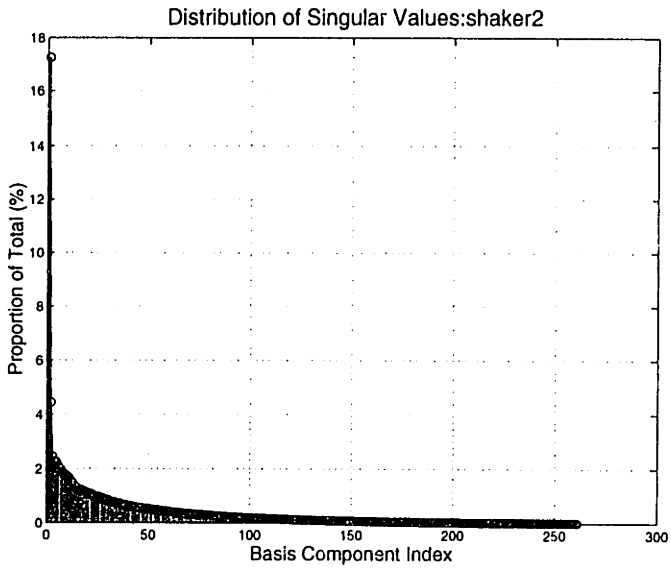


FIGURE 10. The singular values of an SVD of the percussive shaker sound. The first 19 component account for approximately 50% of the total variance in the signal.

occurring at both the up-shake and down-shake of the percussive action. This results in an anacrusis before each main shake. We would like a basis decomposition to reveal this elementary structure using very few components.

From an acoustical perspective the magnitude of the broad-band regions corresponds to the force of the shake. The shaker comprises many small particles which impact the surface of the shell creating a ramped impact and decay which has Gaussian characteristics.

3.3.10 Method

The shaker TFD was treated in spectral orientation which is the transpose of the spectrogram representation show in the figure. A full SVD decomposition was performed on the STFTM for 1.5 seconds of the sound.

3.3.11 Results

The singular values of the SVD are shown in Figure 10. The first three singular vectors decay rapidly from accounting for 17% of the total variance in the signal to accounting for approximately 2.4% of the total variance in the signal. Since the first three components have a much steeper decay than the rest of the components they are considered to hold the most important characteristic information in the TFD.

The first three left and right singular vectors are shown in Figure 11. The third singular vectors demonstrate the intended structure of an anacrusis followed by a down-beat for each shake. The left temporal vectors articulate this structure clearly. The right spectral vectors reveal the broad-band nature of the shake impacts. The remaining singular vectors account for the temporal pattern-

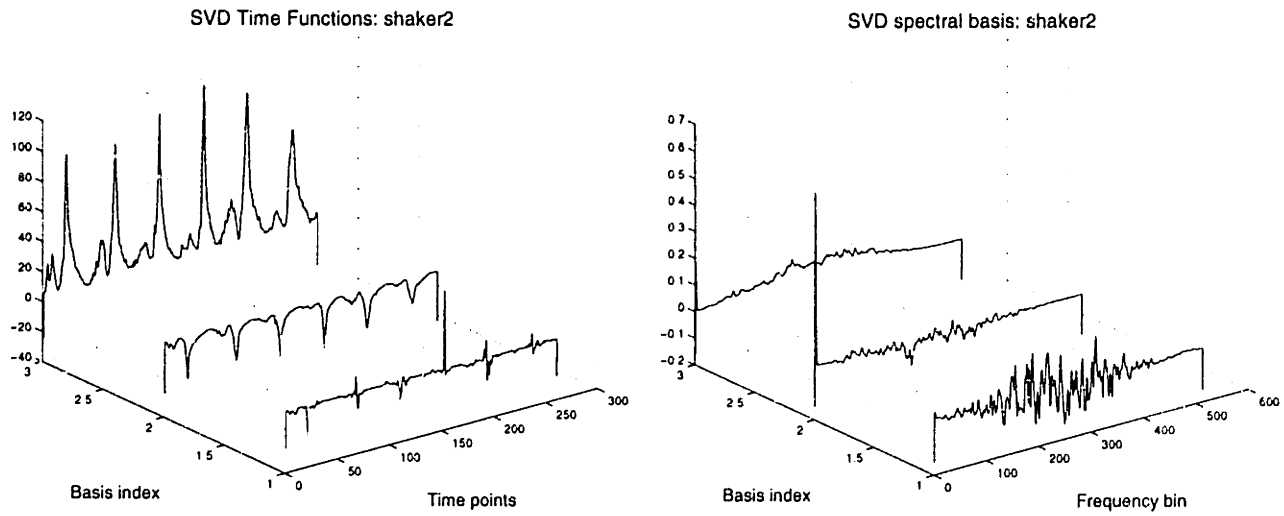


FIGURE 11. First three left and right singular vectors of the shaker sound. The left singular vectors correspond to a time-function for each of the right-singular vector spectral basis components. The outer-product of each pair of basis components forms an independent spectral presentation for that basis

ing and broad-band excitation of the particulate components of the shaker sound as well as the spectral structure of the metallic shell which is exhibited by the narrow-band spectral structure of the first right singular vector. From these results we conclude that the SVD has done a remarkably efficient job of representing the structure of the shaker sound.

3.3.12 A More Complicated Example: Glass Smash

Figure 12 shows 1.00 second of a glass smash sound. We can see from the figure that a number of discernible features are visible in this spectral representation; namely a low-frequency decaying impact noise component, a wide-band onset component and a series of high-frequency scattered particulate components which correspond to the broken glass shards. Ultimately, we would like a basis decomposition to represent these elements as separate basis functions with independent temporal characteristics.

From an ecological acoustics perspective, the bandwidth of the onset click, and the rate of decay of the low-frequency impact noise as well as the number of high-frequency particles serves to specify the nature of the event. In this case the glass-smash is relatively violent given the density of particles and the bandwidth and decay-times of the noise components.

From a signal perspective it is reasonable to treat this sound as a sum of independent noisy components since the individual particles corresponding to sharding are generated by numerous independent impacts. Each particie, however, contains formant structures as is indicated by the wide-band synchrony of onsets in the particulate scattering. This synchrony is manifest as correlations in the

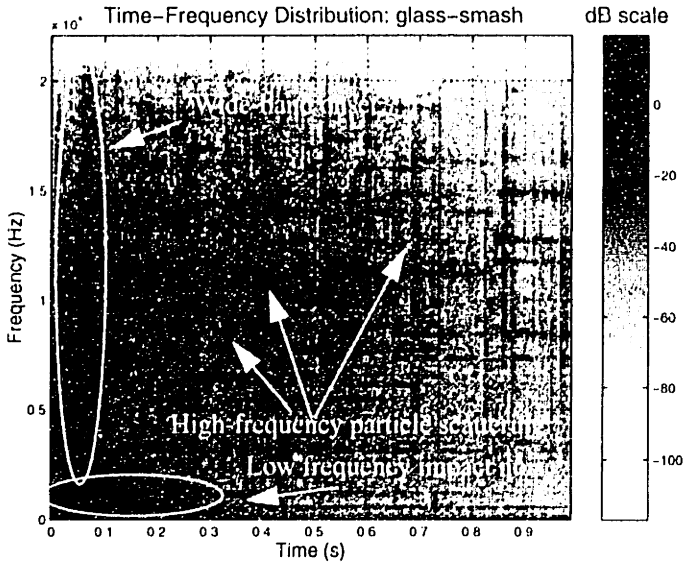


FIGURE 12. Short-time Fourier transform TFD of a glass-smash sound. Analysis parameters are N=1024, W=512, H=256. Sample rate is 44.1kHz.

underlying pdfs of the marginal distributions of the glass-smash TFD. From these observations it seems reasonable that an SVD might reveal quite a lot about the structure assuming that the nature of the statistical independence of spectral components is roughly Gaussian. For such a noisy sequence this assumption seems like a reasonable first-order approximation.

3.3.13 Method

The data matrix $\mathbf{X}$ is first decomposed using a full SVD as described in Section 3.3.7. This yields a set of orthonormal basis functions for both the row space and column space of $\mathbf{X}$ as well as a diagonal matrix Σ of singular values. In this example we chose to represent the matrix in spectral orientation which is essentially the transpose of the spectrogram orientation shown in Figure 12.

3.3.14 Results

The singular values for the glass smash sound are distributed across much of the basis thus suggesting a relatively noisy spectrum in terms of the distribution of Gaussian variances in the orthogonal basis space, see Figure 13. The left and right singular vectors of the spectrally-oriented SVD are given in Figure 8. The 5th left singular basis vector shows a pattern of decaying amplitude through time which corresponds to the low-pass spectral-basis component of the 5th right singular vector.

Other discernible features in the left singular vectors are the time patterns of the glass shards, bases 1-4, which are iterated with a decaying envelope through time. The narrow-band nature of the

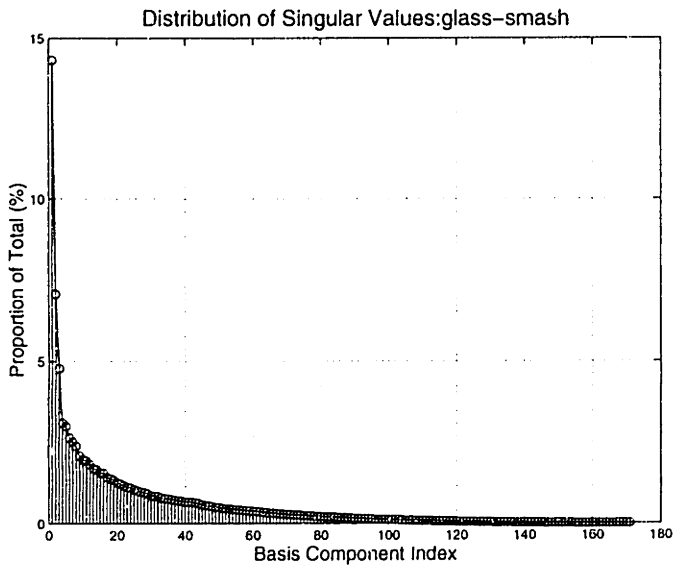


FIGURE 13. Distribution of singular values for the glass-smash sound. The first 14 singular values account for approximately 50% of the variance of the original signal.

peaks in the first 4 right singular vectors suggest high-Q filter characteristics which are due to the ringing of glass particles in the spectrum.

It was our goal in applying the SVD to the glass-smash sound to reveal elements of the complex structure in the noisy TFD shown in Figure 12. The coarse structure of the sound is indeed revealed by the decomposition but it does not appear that the signal has been characterized as successfully as the shaker example. We now discuss possible causes for inadequacy in a PCA of a TFD using the SVD.

3.3.15 Limitations of the Singular Value Decomposition

As we have discussed previously, an SVD decorrelates the input covariance by factoring the marginals of the second order statistics. This has the effect of rotating the basis space onto the directions that look most Gaussian. Whilst this assumption is valid for TFDs whose independent components comprise Gaussian-distributed magnitudes we conjecture that this assumption is too limiting for the case of most sounds. Speech and music sounds have been shown to have probability density functions which are non-Gaussian, therefore their PDFs are characterized by cumulants above second order, see [Bell&Sejnowski86] [Sejnowski88].

As an illustration of this point consider the scatter plot shown in Figure 14. The input distribution is a 2-dimensional uniform random variable which is evenly distributed in both dimensions. An SVD produces a basis which generates the basis rotation shown by the scatter plot in Figure 15. The SVD essentially creates a basis for the most Gaussian directions in the data without sensitivity to alternate distributions.

Thus we seek an alternate formulation of the SVD which is sensitive to higher-order statistical measures on the input data. We interpret this as necessitating a dependency on cumulants at higher than second order. The hypothesis is that such a decomposition will enable a more accurate statistically independent decomposition of data that is not Gaussian distributed.

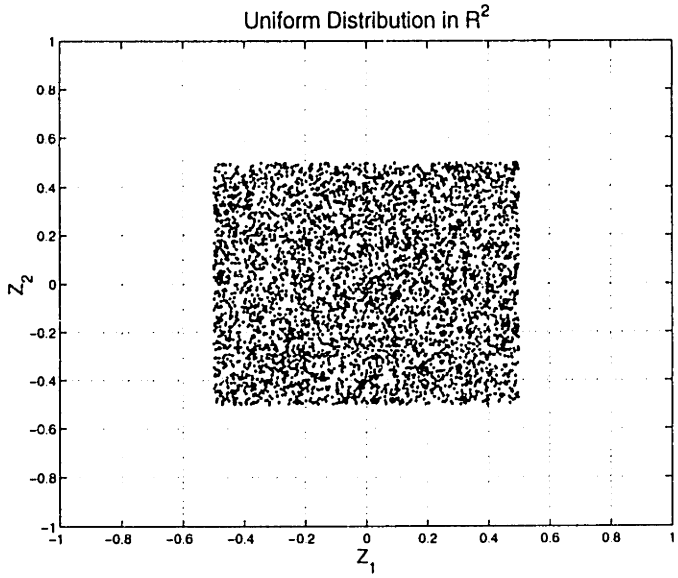


FIGURE 14. Scatter plot of a uniformly distributed random variable Z in 2-space.

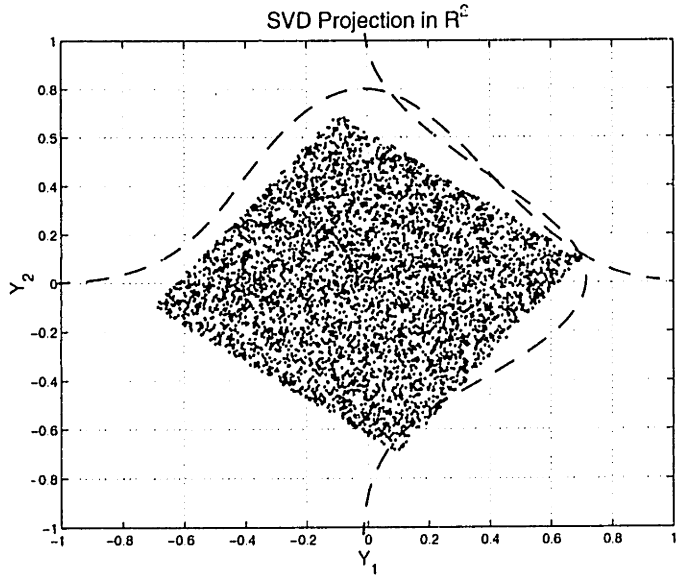


FIGURE 15. Scatter plot of SVD transformation of uniformly-distributed random variable Z. The SVD rotates the basis into the most Gaussian-like directions shown by the dashed lines. Clearly, this basis is not the best possible characterization of the input space.

3.3.16 Independent Component Analysis (ICA)

The concept of an ICA was first proposed in 1983 by Herault and Jutten who produced an iterative on-line algorithm, based on a neuro-mimetic architecture, for blind signal separation, see Jutten and Herault (1991). Their algorithmic solution to the problem of separating an unknown mixture of signals became the basis for a a number of different investigations into the application of statistical methods for identifying independent components within a data set. The blind source separation problem is related to the ICA problem by the need to identify statistically independent components within the data. For blind signal separation (BSS) the independent components correspond to *a-priori* unknown signals in a linear mixture, and for the ICA

Giannakis et al. (1989) used third-order cumulants to address the issue of identifiability of ICA. The resulting algorithm required an exhaustive search and is thus intractable for practical applications. Other mathematically-based techniques for identifying ICA were proposed by Lacoume and Ruiz (1989), who also used a cumulant-based method, and Gaeta and Lacoume (1990) proposed a maximum likely hood approach to the problem of blind identification of sources without prior knowledge.

An alternative method of investigating the existence of ICA was the method of Cardoso (1989) who considered the algebraic properties of fourth-order cumulants. Cardoso's algebraic methods involve diagonalization of cumulant tensors, the results of which are an ICA. Inouye and Matsui (1989) proposed a solution for the separation of two unknown sources and Comon (1989) proposed a solution for possibly more than two sources. These investigations form the mathematical foundations on which independent component analysis has continued to grow.

Recently many neural-network architectures have been proposed for solving the ICA problem. Using Comon's identification of information-theoretic quantities as a criteria for ICA Bell and Sejnowski (1996) proposed a neural network that used mutual information as a cost function. The resulting architectures were able to identify independently distributed components whose density functions were uni-modal. Bell's network maximizes the mutual information between the input and output of the neural network which has the effect of minimizing the redundancy. Amari *et al.* (1996) proposed a using a different gradient descent technique than Bell which they called the *natural gradient.* These, and many other, neural network-based architectures were proposed as partial solutions to the problem of blind signal separation.

Aside from the BSS problem of additive mixtures, several architectures have been proposed for addressing the problem of convolved mixtures of signals. Among these are architectures that employ feedback weights in their neural network architectures in order to account for convolution operations. A novel approach to the problem of convolutions of signals was proposed by Smaragdis (1997), in which the convolution problem is treated as a product of spectral components thus the architecture seeks to factor the spectral components into independent elements.

All of the techniques and architectures introduced above have been applied to the problem of separation of sources in one form or another. An alternate view of ICA is that it is closely related to PCA. This is the view that we take in this section. We develop the necessary mathematical background in order to derive an algorithm which is capable of producing a *basis* in which spectral components are lifted into independent distributions. Our methods are closely related to the algebraic methods of Comon and Cardoso and are seen as a higher-order statistical extension of the SVD.

3.3.17 The ICA Signal Model: Superposition of Outer-Product TFDs

For the purposes of feature extraction from a TFD using an ICA we must be explicit about our signal assumptions. Our first assumption is that the input TFD is composed of a number of *a-priori* unknown, statistically independent TFDs which are superposed to yield the observable input TFD. This assumption of superposition is represented as:

$$\mathbf{X} = \sum_{i=1}^{\rho} \chi_i + \sum_{j=1}^{\kappa} \Upsilon_j \qquad [97]$$

where χ_i are the latent independent TFDs of which there are ρ, and the Υ_j are an unknown set of noise TFDs of which there are κ. Observing that the superposition of TFDs is a linear operation in the time-frequency plane and under the assumption that the inverse TFD yields the corresponding latent superposition of signals then Equation 97 is interpreted as the frequency-domain representation of a blind signal separation problem. In this form the signal model defines the domain of signal compositions that we are operating under but it does nothing to define the form of the features that we might want to extract as characteristic components of the signals.

A second, stronger assumption is that each independent TFDs χ_i is uniquely composed from the outer product of an *invariant* basis function $\mathbf{y}_i$ and a corresponding *invariant* weighting function $\mathbf{v}_i$ such that:

$$\chi_i = \mathbf{y}_i \mathbf{v}_i^T. \qquad [98]$$

These functions are invariant because they are statistically stationary vectors which multiply, using the outer-product of two vectors, to form a TFD matrix. Under the assumption of the outer-product form of the TFD the vectors are defined to be stationary since there is no way to affect a time-varying transform.

This latter assumption seems on the surface quite limiting. After all many natural sounds are composed of non stationary spectral components which may shift in frequency during the course of the sound. However, recalling our framework from the previous chapter, the utility of a statistical basis decomposition comes not from the ability to fully characterize the transformational structure of a

sound, but it is in its ability to identify likely candidates to be treated as *invariants* for a sound structure. These invariants are to be subjected to further treatment in the next chapter in which we use them to identify the time-varying structure of a sound. We recall the observation of Warren and Shaw (1985) that structure must be defined as a form of persistence and a style of change. The statistical decomposition of TFDs provides much of this structure in the form of spectral invariants and temporal fluctuations, but time-varying frequency components are not represented by the techniques. We must define time-varying frequencies in terms of a form of persistence and it is this form that we seek to identify.

We conjecture that the time-varying components of a natural sound are constrained in their range between each time frame of a TFD, thus the change in an invariant is relatively small at each time frame. Recall from our discussion on auditory group theory that such small changes in an invariant component can be used to identify the form of a transformation. The basis techniques on which we rely for extraction of the invariant features are dependent upon the PDFs of the invariant components. Thus, under the assumption of small changes between frames, it is assumed that each PDF is stationary enough over a portion of the TFD that it is well represented by the ensemble statistics of the TFD.

However, the argument is a little more subtle than this. Since the statistical techniques presented herein are batch techniques, operating on the entire data matrix with no update rule, there is actually no dependence upon the order of the frames in a TFD. Thus we would get equivalent results if we presented the frames in a random order. So it is the time-average PDF of an independent spectral component that determines the form of an invariant. For example, if the component oscillates about a mean frequency such that the average density of the centre frequency is greater than the density of the peak frequency deviations then the distribution of the average PDF will be representative of the underlying invariant.

These arguments lead us to our third assumption for the ICA signal model: that the underlying invariant functions of the independent TFDs are distributed in time-frequency in such a way that their average PDF is, in fact, representative of an invariant component. In the case that they are not centered on a mean frequency value we observe that the statistics will yield a series of similar TFD basis components that differ by the nature of the underlying transformation. Since the basis decomposition techniques order the basis components by their singular values, i.e. their salience in the input TFD, we take the components that have larger singular values as being representative of the invariants that we seek. It is extremely unlikely that a single time-varying component will yield very high singular values for each of its mean spectra in the statistical decomposition. This leads us to assert that the decompositions are valid representatives of the underlying TFDs but care must be taken in interpreting the results.

By representing the basis components $\mathbf{y}_i$ and $\mathbf{v}_i$ as the columns of two matrices we arrive at an assumed signal model for the input TFD:

$$\mathbf{X} = \mathbf{Y}_\rho \mathbf{V}_\rho^{\mathrm{T}} + \Upsilon \qquad [99]$$

where $\Upsilon = \sum_{j=1}^{\kappa} \Upsilon_j$ is the summed noise matrix. The components $\mathbf{Y}_\rho$ and $\mathbf{V}_\rho$ both have ρ columns. Thus for an $m \times n$ input TFD $\mathbf{X}$, $\mathbf{Y}_\rho$ is an $m \times \rho$ matrix and $\mathbf{V}_\rho$ is an $n \times \rho$ matrix. We call this model a superposition of outer-product TFDs and it defines the structure of the features that we seek in a statistical basis decomposition of a given input TFD.

3.3.18 ICA: A Higher-Order SVD

For our formulation of ICA we start with an SVD of a TFD data matrix:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}. \qquad [100]$$

The statistically independent basis that we seek is an orthogonal set of vectors that span the column space $\Re(\mathbf{X})$, thus the SVD factorization is a good starting point since $\mathbf{V}^{\mathrm{T}}$ already spans this space, but under a rotation of basis on the assumption of Gaussian input statistics.

We would like the ICA to be an orthogonal basis like that of the SVD but we impose different constraints corresponding to maximization of higher-order cumulants in the PDFs of the data space. We define the matrix $\mathbf{Z} = \mathbf{V}^{\mathrm{T}}$ which is the matrix of random vectors whose PDFs we seek to factor. Now, the random vector $\hat{\mathbf{z}} \in \mathbf{Z}$ has statistically independent components if and only if:

$$p_{\mathbf{Z}}(\hat{\mathbf{z}}) = \prod_{i=1}^{N} p_{\mathbf{Z}_i}(\hat{\mathbf{z}}_i). \qquad [101]$$

where N is the dimensionality of $\hat{\mathbf{z}}$. Thus we seek to factor the joint probability density function of $\hat{\mathbf{z}}$ into the product of its marginal PDFs in order to achieve statistically independent basis components.

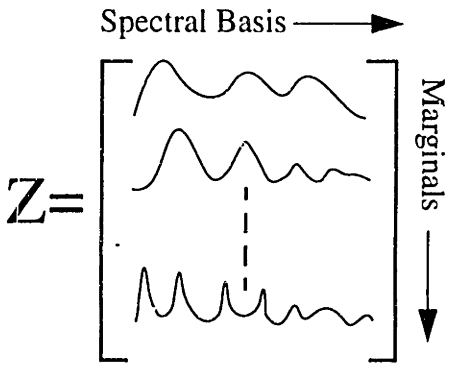


FIGURE 16. Illustration of the orientation of spectral basis components in the ICA decomposition of the variable Z. The marginals are the orthogonal complement of the basis functions.

As an illustration of the effect of the factorization let us consider $\mathbf{X}$ as a TFD in spectral orientation. The column space of $\mathbf{X}$ is occupied by the frequency-bins of the chosen time-frequency transform with each row of $\mathbf{X}$ an observation or time-slice of the TFD. The matrix $\mathbf{V}$ is a basis for the column space of $\mathbf{X}$ with each column of $\mathbf{V}$ corresponding to a spectral basis component. The matrix $\mathbf{Z}$, then, contains the spectral basis functions row-wise as shown in Figure 16. Now, the random vector $\hat{\mathbf{z}}$ has a joint PDF which operates down the columns of the matrix $\mathbf{Z}$. A successful factorization of the joint probability density function of $\mathbf{Z}$ will therefore result in statistically independent rows of $\mathbf{Z}$ which corresponds to a statistically-independent spectral basis for the spectrally-oriented TFD.

We could equally decide to represent the TFD in temporal orientation. The column space of $\mathbf{X}$ would thus have variates corresponding to the temporal amplitude functions of a TFD, each weighting an underlying spectral basis component which, in the case of temporal orientation, is represented by the left singular vectors which are the columns of $\mathbf{U}$. A successful factorization would result in a set of statistically-independent amplitude basis functions. The orientation that we choose may have a significant effect on the resulting characterization of a TFD. We explore issues of data matrix orientation later in this chapter.

For the defined matrix $\mathbf{Z}$, we can effect the factorization shown in Equation 101 by a rotation of the basis $\mathbf{V}$. This corresponds to rotating the rows of $\mathbf{Z}$ such that they point in characteristic directions that are as statistically independent as possible based on a criteria which we shall soon define. Thus the ICA can be thought of as an SVD with a linear transform performed by a new matrix $\mathbf{Q}$ such that:

$$\mathbf{Z}_{ICA} = \mathbf{Q}\mathbf{Z}_{SVD} = \mathbf{Q}\mathbf{V}^{T}. \qquad [102]$$

3.3.19 Information-Theoretic Criteria For ICA

Having defined the form of the ICA we now seek to define a criteria for statistical independence that will yield the factorization of Equation 101. In order to do this we must define a distance metric δ between the joint-probability density function $p_z(\hat{z})$ and the product of its marginals:

$$\delta\left(p_z(\hat{z}), \prod_{i=1}^{N} p_{z_i}(\hat{z}_i)\right) \qquad [103]$$

In statistics, the class of *f*-divergences provides a number of different measures on which to base such a metric. The Kullback-Leibler divergence is one such measure and is defined as:

$$\delta(p_x, p_z) = \int p_x(u)\left(\log\frac{p_x(u)}{p_z(u)}\right)du\,. \qquad [104]$$

Substituting Equation 103 into Equation 104 yields:

$$I(p_z) = \int p_z(\hat{z})\left(\log\frac{p_z(\hat{z})}{\prod_{i=1}^{N} p_{z_i}(\hat{z}_i)}\right)d\hat{z} \qquad [105]$$

where $I(p_z)$ is the average mutual information of the components of $\mathbf{z}$. The Kullback-Leibler divergence satisfies the relation:

$$\delta(p_x, p_z) \geq 0 \qquad [106]$$

with equality if and only if $p_x(u) = p_Z(u)$, Comon (1994). Thus, from Equation 103, the average mutual information between the marginals $\hat{z}_i$ becomes 0 if and only if they are independent, which implies that information of a marginal does not contribute to the information of any other marginal in the joint PDF of $\mathbf{z}$.

3.3.20 Estimation of the PDFs

Having defined a suitable criteria for ICA we must now tackle the problem of estimation of the PDF of $\mathbf{z}$ since the densities are not known. We do, however, have data from which to estimate the underlying PDFs.

The Edgeworth expansion of a density $\mathbf{z}$ about its best Gaussian approximate ϕ_z for zero-mean and unit variance is given by:

$$\frac{p_z(u)}{\phi_z(u)} = 1 + \frac{1}{3!}k_3h_3(u) + \frac{1}{4!}k_4h_4(u) + \frac{10}{6!}k_3^2h_6(u) + \frac{1}{5!}k_5h_5(u) + \frac{35}{7!}k_3k_4h_7(u) \quad [107]$$
$$+ \frac{280}{9!}k_3^3h_9(u) + \frac{1}{6!}k_6h_6(u) + \frac{56}{8!}k_3k_5h_8(u) + \frac{35}{8!}k_4^2h_8(u) + \frac{2100}{10!}k_3^2k_4h_{10}(u)$$
$$+ \frac{15400}{12!}k_3^4h_{12}(u) + o(m^{-2})$$

where k_i denotes the cumulant of order i of the scalar variable u and $h_i(u)$ is the Hermite polynomial of degree i defined by the recursion:

$$h_0(u) = 1, \quad h_1(u) = u \quad . \quad [108]$$
$$h_{k+1}(u) = uh_k(u) - \frac{\partial}{\partial u}h_k(u)$$

With a method for estimating the PDF from an ensemble of data we are able to proceed with parameterizing the linear transform $\mathbf{Q}$ so that the ICA basis vectors in $\mathbf{Z}$ satisfies our independence criteria as closely as possible.

3.3.21 Parameterization and Solution of the Unitary Transform Q

In order to obtain the rotation matrix $\mathbf{Q}$ a parameterization in terms of the Kullback-Leibler divergence on $\mathbf{z}$ is utilized.

With the solution of $\mathbf{Q}$ we arrive at a form for the ICA transform:

$$\mathbf{X} = \mathbf{U\Sigma Q^T QV^T} \quad [109]$$

since $\mathbf{Q}$ is unitary, the quantity $\mathbf{Q^TQ} = \mathbf{I}$, thus rotations of the basis components do not affect the reconstruction of $\mathbf{X}$.

3.3.22 Uniqueness Constraints

The formulation of ICA in this manner does not specify the basis uniquely. In fact, it expresses an equivalence class of decompositions for which there are infinitely many possibilities. In order to define a unique ICA additional constraints must be imposed on the form of Equation 109.

Firstly, the decomposition is invariant to permutations of the columns. Thus the same criteria for ordering of basis components as the SVD is utilized; namely that the basis components are permuted in decreasing order of their variances. We denote by $\mathbf{P}$ the permutation matrix that performs this ordering. Permutation matrices are always invertible and they have the property $\mathbf{P^TP} = \mathbf{I}$. The second criteria for uniqueness stems from the fact that statistics are invariant under scaling. That is, the PDF of a scaled random vector is the same as the unscaled vector's PDF. A scaling is chosen such that the columns of $\mathbf{V}$ have unit norm. We denote by Λ the invertible diagonal matrix of scaling coefficients. Finally an ICA is invariant to sign changes in the basis components. The unique-

ness constraint is chosen such that the sign of the largest modulus is positive. We denote by $\mathbf{D}$ the diagonal matrix comprising values from [1,-1] which performs this sign change. As with the other uniqueness constraint matrices, the sign-change matrix is trivially invertible.

These uniqueness constraints give the final form of the ICA:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{Q}^T\mathbf{P}^T\Lambda^{-1}\mathbf{D}\mathbf{D}^T\Lambda\mathbf{P}\mathbf{Q}\mathbf{V}^T \quad [110]$$

with

$$\mathbf{Z} = \mathbf{D}^T\Lambda\mathbf{P}\mathbf{Q}\mathbf{V}^T \quad [111]$$

and

$$\mathbf{Y} = \mathbf{U}\Sigma\mathbf{Q}^T\mathbf{P}^T\Lambda^{-1}\mathbf{D}. \quad [112]$$

Since $\mathbf{X}$ and $\mathbf{Z}$ are both given we can compute the left basis by a projection of the data against the right basis vectors:

$$\mathbf{Y} = \mathbf{X}\mathbf{Z}^T = \mathbf{V}\mathbf{Q}^T\mathbf{P}^T\Lambda^T\mathbf{D}. \quad [113]$$

The outputs covariance Φ_Y is diagonalized by the unitary transform $\mathbf{Q}$ but, unlike the SVD, this diagonalization is based on the contrast defined by fourth-order cumulants.

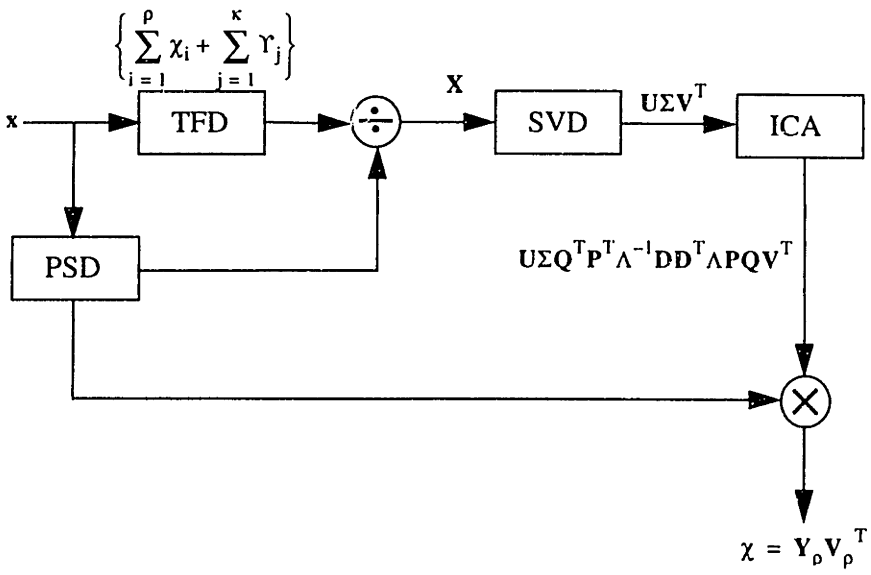


FIGURE 17. Signal flow diagram of independent component analysis of a time-frequency distribution. The input is a signal x whose latent variables χ_i we seek. An optional power-spectral density normalization is applied followed by the SVD and ICA transforms. The output is a set of ρ basis vectors which span the signal space of the TFD of x.

3.4 Independent Component Analysis of Time-Frequency Distributions

So far we have discussed the mechanics of the ICA and SVD algorithms in relative isolation from their application to sound analysis. We are now in a position to discuss the general application of independent component analysis to time-frequency distributions. In this section we investigate methods for utilizing an ICA for identifying features in the TFD of a signal. The feature extraction problem involves the estimation of many unknown variables in the signal. As we shall see, using the signal assumptions defined above in conjunction with a careful application of an ICA we are able to obtain results that appear to meet our demands to a remarkably high degree.

3.4.1 Method

The method proceeds as follows. An input signal x is assumed to contain ρ independent components that are combined under the signal model of Equation 97. All of ρ, χ_i and Υ_j are assumed unknown *a-priori*, Figure 17.

A time-frequency transform produces a TFD which expresses the signal model in the time-frequency plane. For many natural sounds, the power of the signal in increasing energy bands may decrease rapidly due to the predominance of low-frequency energy. Thus, from the point of view of statistical measures on the data, the variates are scaled by the average spectrum of the TFD. In order to compensate for the power-loss effect at high frequencies, and to *sphere* the data to a reasonable scale in all variates, an optional power-spectral density estimation and normalization step is incorporated.

The power spectral density of a TFD is calculated using Welch's averaged periodogram method. The data sequence $x[n]$ is divided into segments of length L with a windowing function $w[n]$ applied to each segment. The periodogram segments form a separate TFD which can be estimated from the analysis TFD in the case of an STFT. In the most general case, however, this may not be possible so we represent PSD normalization as a separate path in the signal flow diagram of Figure 17.

The periodogram or the lth segment is defined as:

$$I_l(\omega) = \frac{1}{LU}\left|X_l(e^{jw})\right|^2 \qquad [114]$$

where L is the length of a segment and U is a constant that removes bias in the spectral estimate and $X_l(e^{jw})$ is a short-time Fourier transform frame as described previously. The average periodogram for a signal $x[n]$ is then the time-average of these periodogram frames. If there are K frames in the periodogram then the average periodogram is:

$$\bar{I}_l(\omega) = \frac{1}{K}\sum_{l=0}^{K-1} I_l(\omega). \qquad [115]$$

Thus the average periodogram provides an estimate of the power-spectral density (PSD) of $x[n]$. Assuming that the PSD is nowhere equal to zero we can perform the normalization of the TFD by division in the frequency domain as indicated in the figure. Once the data matrix $\mathbf{X}$ is obtained an SVD is performed which yields a factorization of the data matrix into left and right basis vectors and a matrix of corresponding singular values, see Equation 93.

From the singular values Σ, the rank ρ of the TFD can be estimated. In order to do this we pick a criteria $\Psi \in [0...1]$ that specifies the amount of total variance in the TFD that we wish to account for in the resulting basis. In the case of data compaction applications Ψ is chosen relatively high, typically around 0.95, so that the reconstruction from the reduced basis results in as little loss as possible. However, for the purposes of feature extraction we can choose Ψ much lower since we seek to characterize the primary features in the data space rather than account for all the variance. Typical values for Ψ were determined empirically to be in the range $0.2 \le \Psi \le 0.5$. Given this variance criteria, estimation of the rank ρ of $\mathbf{X}$ is achieved by solving the following inequality for ρ:

$$\frac{1}{\text{trace}(\Sigma^2)} \sum_{i=1}^{\rho} \Sigma^2(i,i) \geq \Psi. \qquad [116]$$

This estimate of the rank of the data matrix provides a good approximation of the number of statistically independent components in the TFD. Thus the following ICA problem can be reduced from the problem of generating a full set of independent columns in the basis space to that of generating exactly ρ independent columns. Since the singular vectors of the SVD are sorted according to their singular values in decreasing order of importance, we choose the first ρ columns of $\mathbf{V}$ for the estimation and solution of the ICA.

Thus the first ρ right singular vectors of the SVD are used to obtain a basis with an ICA, the vectors are transposed and stored in a matrix $\mathbf{Z}$ which is the observation matrix for the ICA decomposition. An iterative procedure is employed which first estimates the cumulants for each pair of rows $(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ in $\mathbf{Z}$; of which there are $\rho\frac{(\rho-1)}{2}$ pairs. From these cumulants the angle α that minimizes the average mutual information $I(p_z)$, defined in Equation 105, is calculated such that the unitary transform $\mathbf{Q}^{(i,j)}$ performs a rotation about the angle α in the orthogonal plane of $(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$. It can be shown that a set of planar rotations, derived from estimates of α, that maximize the pairwise independence (i.e. minimize the average mutual information) of the rows of $\mathbf{Z}$ are a sufficient criteria for independence. That is, pair-wise independence specifies global independence. For a proof of this conjecture see Comon (1994). After each iteration, $\mathbf{Z}$ is updated by applying the unitary transform $\mathbf{Z} = \mathbf{Q}^{(i,j)}\mathbf{Z}$. The iterations continue on the set of row pairs in $\mathbf{Z}$ until the estimated angles α become very small or until the number of iterations k has reached $1 + \sqrt{\rho}$.

After these operations have been performed, $\mathbf{Z}$ contains a set of ρ basis components in the rows which are as statistically independent as possible given the contrast criteria of maximization of fourth-order cumulants in $\mathbf{Z}$. As discussed previously, these components are not unique for the statistics are invariant to scaling, ordering and sign changes in the moduli of the vector norms. Applying uniqueness constraints we first compute the norm of the columns of $\mathbf{V} = \mathbf{Z}^T$ which specify the entries of the diagonal scaling matrix Λ. In order to solve for the ordering the entries of Λ are sorted in decreasing order; this specifies the permutation matrix $\mathbf{P}$ whose rows generate the said ordering of entries in Λ. Finally a diagonal matrix of entries with unit modulus an possibly different signs is constructed such that the entry of the largest modulus in each column of $\mathbf{Z}$ is positive real; this specifies the matrix $\mathbf{D}$.

With the full specification of the ICA in hand we can compute a basis for the row-space of $\mathbf{X}$ using the relation:

$$\mathbf{Y}_\rho = \mathbf{X}\mathbf{Z}_\rho^T = \mathbf{V}_\rho\mathbf{Q}^T\mathbf{P}^T\Lambda^T\mathbf{D} \qquad [117]$$

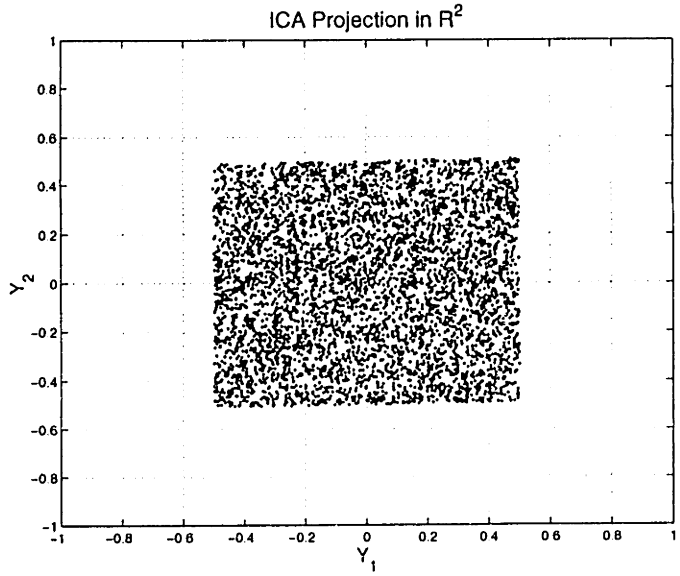


FIGURE 18. Scatter plot of the output of an ICA for the input of an arbitrary linear transformation of a bi-variate uniform distribution. The plot shows that the PDFs have been accurately characterized by the ICA since they have been separated correctly (compare with Figure 15 on page 97).

this operation is equivalent to performing the unitary transform and uniqueness operations on the first ρ left singular vectors of the preceding SVD scaled by their singular values:

$$\mathbf{Y}_\rho = \mathbf{U}_\rho \Sigma_\rho \mathbf{Q}^T \mathbf{P}^T \Lambda^{-1} \mathbf{D}. \qquad [118]$$

With these bases in place we are able to specify the form of the latent independent TFDs which form the independent features of the original TFD:

$$\chi = \mathbf{Y}_\rho \mathbf{V}_\rho^T \qquad [119]$$

thus each column $\mathbf{Y}_j$ and $\mathbf{V}_j$ specifies a basis vector pair for an independent TFD χ_i, and the independent χ_i's sum to form the signal TFD of $\mathbf{X}$, which is an approximation $\hat{\mathbf{X}}$. The residual spectrum $\mathbf{X} - \hat{\mathbf{X}}$ specifies the near-uncorrelated noise components of $\mathbf{X}$ which is also obtainable by an ICA transform of the SVD basis components that were not used in the identification of χ:

$$\Upsilon = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{Y}_{M-\rho} \mathbf{V}_{N-\rho}^T. \qquad [120]$$

As a final comment on the ICA before we demonstrate its application to sound analysis we again consider the bi-variate uniform distribution of Figure 14. Recall that the SVD basis did not ade-

quately characterize the PDFs due to its Gaussian basis criteria, see Figure 15; an ICA transform of the SVD basis produces the basis rotation shown in Figure 18, which demonstrates that the ICA is capable of characterizing non-Gaussian distributions. In fact, the bi-variate uniform distribution is one of the most difficult joint-PDFs for an ICA to characterize and it bodes well for the algebraic approach that we were able to factor this example correctly, see (Bell and Sejnowski 1995; Amari *et al.* 1996).

3.5 Examples of Independent Component Analysis of TFDs

We now give some examples of the application of ICA to analysis of noisy and textured natural sounds. These sounds have traditionally been very difficult to characterize with sinusoidal-based analysis techniques such as the dual-spectrum representations considered earlier in this chapter. ICA characterizations are not limited to noisy spectra, however. A harmonic sound will also have a discernible PDF which can be separated from other components. In fact, PCA techniques have been successfully applied to harmonic spectra in previous research as outlined previously; see, for example Bayerbach and Nawab (1991). These studies have demonstrated the applicability of PCA techniques to sinusoidal tracking applications. In the following examples, therefore, we focus on the much harder problem of analysis and characterization of sounds with very noisy TFDs.

3.5.1 Example 1: Bonfire sound

1. Method

The first example is that of a bonfire; the spectrogram is shown in Figure 19. The discernible features in this sound are a low-pass continuous Gaussian noise component and a number of intermittent wide-band crackles. We would like the ICA to treat these as separable features of the TFD. An ICA analysis was applied to the bonfire sound with no PSD normalization since there was no frequency band in which energy was disproportionately high compared with the other bands.

2. Results

The singular values of the SVD of the bonfire sound are shown in Figure 20. There is a steep roll-off in the first three singular values followed by a steady exponential decay for the remainder of the components. The first three singular values account for 40% of the total variance in the bonfire signal. This is a very high quantity for just three components, so we would like to investigate the parts of the sound that they represent.

Figure 21 and Figure 22 show the SVD and ICA basis vectors for the bonfire sound respectively. The left singular vectors of the SVD decomposition correspond to amplitude functions through time of the TFD. These characterize the row-space of the TFD in spectral orientation. Each of the three components shown seem to exhibit both the intermittent crackling properties as well as the Gaussian noise sequence properties described above. However, they are not well separated into statistically-independent components. An inspection of the right SVD singular vectors similarly shows that the wide-band and low-pass components are mixed between the basis vectors. Thus we conclude that the SVD has not characterized the bonfire sound satisfactorily.

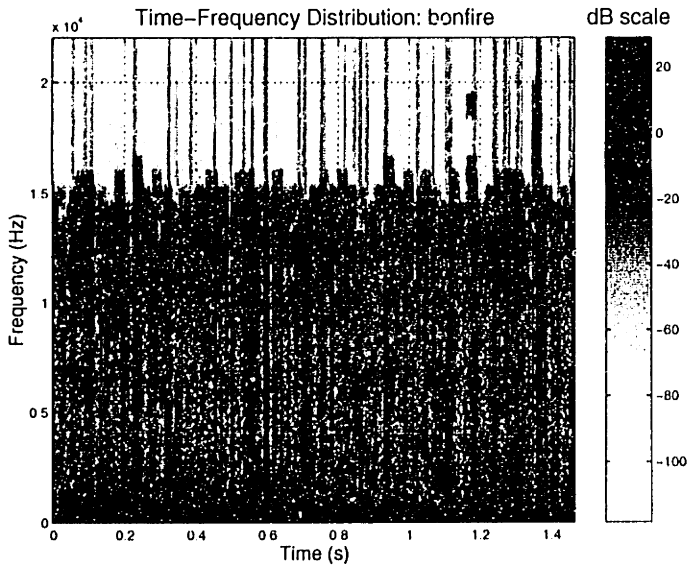


FIGURE 19. STFT spectrogram of bonfire sound. The sound contains intermittent wide-band click elements as well as low-pass and wide-band continuous noise elements.

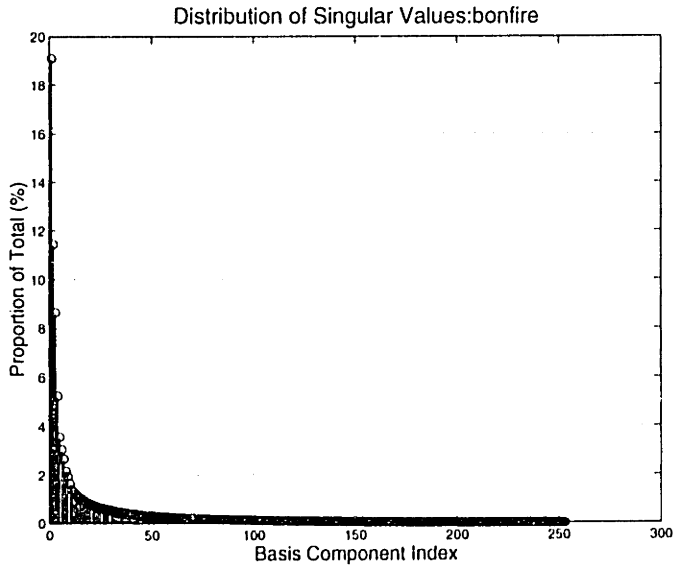


FIGURE 20. Singular values of bonfire sound. The first three singular values account for 40% of the total variance in the data. This implies that they are good candidates for features.

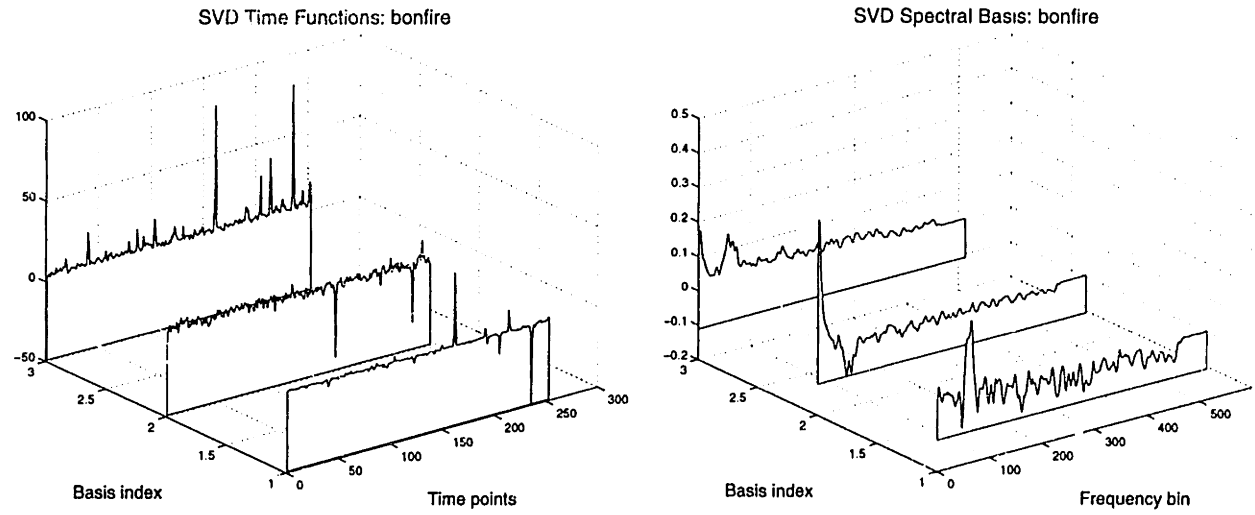


FIGURE 21. SVD basis vectors of a bonfire sound. The left singular vectors seem to mix both the continuous noise elements as well as the erratic impulses. The right singular vectors exhibit a similar mixing of spectral features with notches in some spectra occurring at the peaks of others.

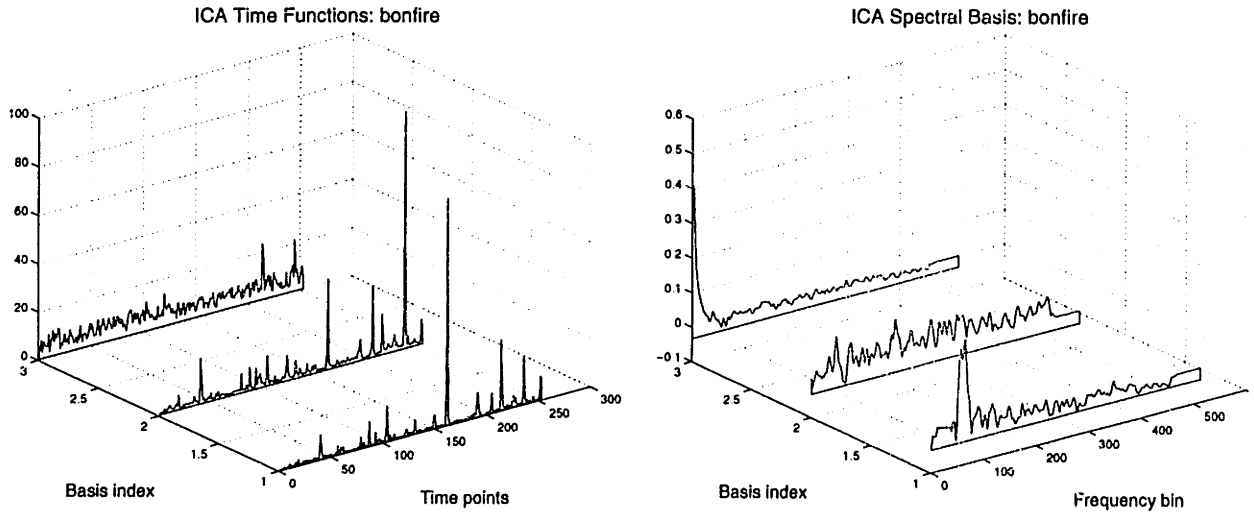


FIGURE 22. ICA basis vectors of the bonfire sound. The left singular vectors exhibit the desired characteristics of erratic impulses and continuous noise densities as independent components. The right singular vectors also exhibit independence with low-pass and wide-band components clearly distinguished.

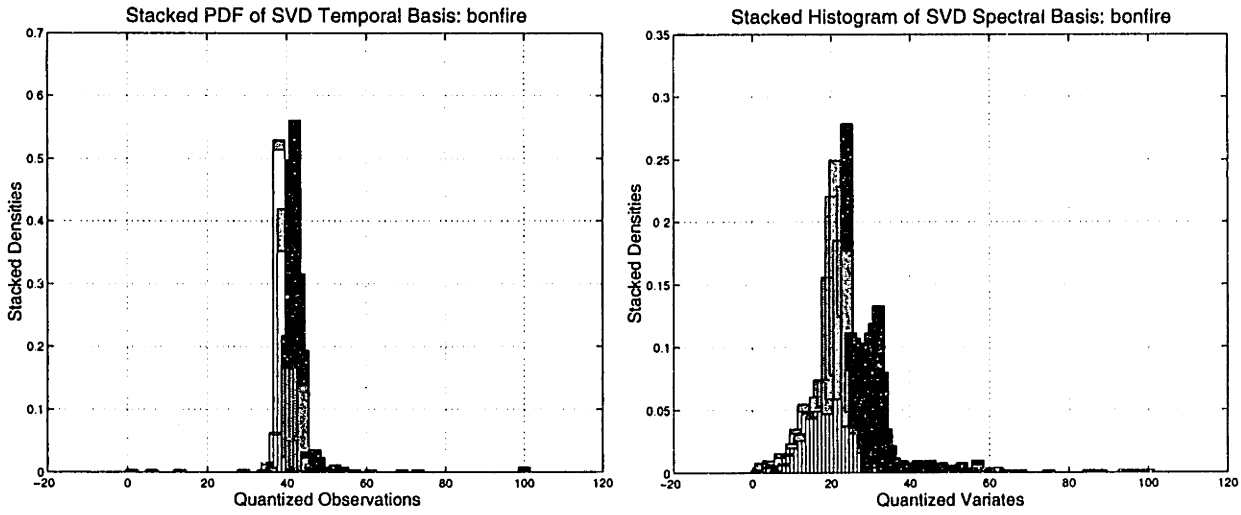


FIGURE 23. Stacked histograms of bonfire sound SVD basis vectors. The left singular vectors exhibit a clustering around the quantized 40. They are roughly evenly distributed each side implying an even distribution. The right singular vectors appear somewhat Gaussian and are centered at quantized values around 20 and 30.

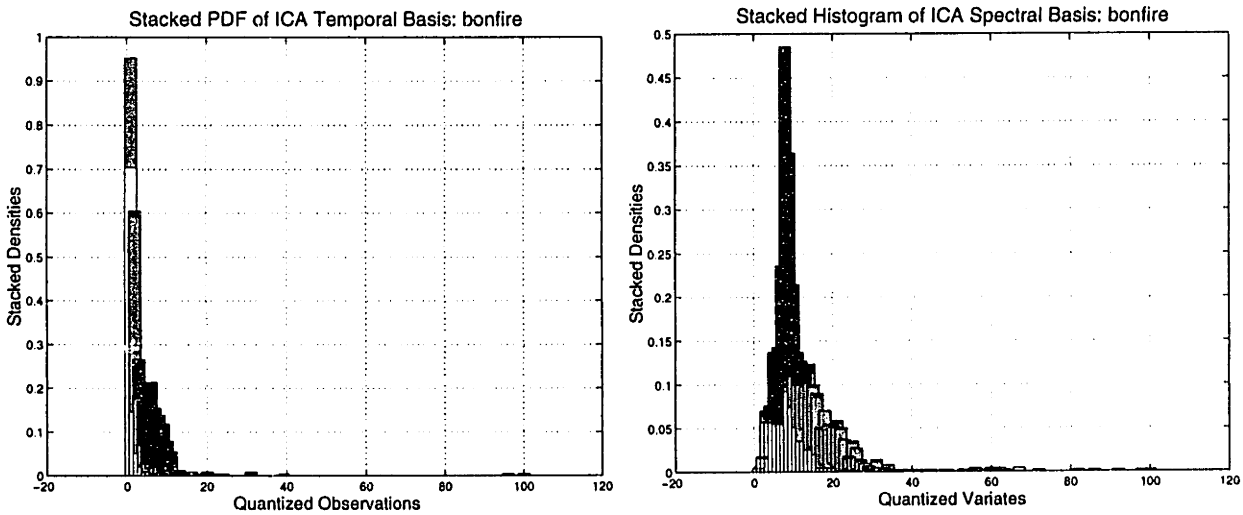


FIGURE 24. Stacked histograms of bonfire ICA basis vectors. The left singular vectors are skewed to the right-hand side resembling an exponential distribution. The right singular vectors are also skewed to the right. Several of the components are extremely peaky, suggesting a high kurtosis.

An inspection of the ICA basis functions, however, reveals a more promising characterization. The three left singular vectors show a clear distinction between continuous noise and intermittent crackling elements; the first two basis functions corresponding to the amplitude functions of the intermittent crackling and the third component corresponding to the amplitude function of the continuous noise component. Similarly, inspection of the right singular ICA vectors shows a clear separation between wide-band and low-pass components. The first two right basis functions correspond with the first two left functions and represent wide-band spectral components; namely, the spectral envelopes of the crackling components. The third right ICA basis vector shows a low-pass component, it corresponds with the continuous noise amplitude function of the third left basis vector and is the spectral envelope of the continuous noise component.

Figure 23 and Figure 24 show stacked histograms of the values of each set of basis vectors. The values were quantized into 150 bins and the histograms of each vector are stacked, in turn, on top of the histogram of its predecessor in order that they can be more easily inspected. The main difference between the SVD and the ICA histograms is that the SVD components are spread roughly evenly around a mean value, thus approximating the components with Gaussian-like PDFs. The ICA histograms, on the other hand, are skewed to one side, thus exhibiting underlying PDFs that are approximately exponential. The kurtosis (measure of fourth-order cumulants) of the PDFs of the ICA distribution is much higher than that of the SVD distribution, suggesting that the contrast function based on fourth-order cumulants is a successful method for contrasting a set of basis vectors against a joint-Gaussian PDF which has no cumulants above second order. Thus, if there exists higher-order cumulants in the estimated PDF of a given basis, the SVD will only find the best Gaussian approximation which, in the case of high kurtosis or skew in the actual PDF, will not be adequate for characterizing the basis components of a TFD.

Table 6 shows the values of the fourth-order cumulants for each of the right basis vectors for both SVD and ICA decompositions. The value of fourth-order cumulants is called *kurtosis* and it is a measure of the peakiness of a PDF. A kurtosis of 0 is a Gaussian distribution, with positive kurtosis being more peaky than a Gaussian and negative kurtosis being flatter than a Gaussian. For example, exponential distributions have a height positive kurtosis and uniform distributions have a low negative kurtosis. The sign of a distribution's kurtosis is called its modality. The table, then, shows that the kurtosis of the first and second components is greater than Gaussian for both SVD and ICA decompositions. However, the ICA has maximized the kurtosis to a much higher degree than the SVD suggesting that the Gaussian criteria does not point the basis vectors in the directions of

TABLE 6. Fourth-Order Contrast Values for SVD and ICA Right Basis Vectors of Bonfire

Basis Component	SVD Kurtosis	ICA Kurtosis
1	32.593	51.729533
2	4.736	18.102202
3	1.748	-0.640642
Contrast from Gaussian	1087.648	3004.044

greatest cumulants. The third component is somewhat Gaussian in both cases, suggesting that the third spectral component is in fact Gaussian. The contrast value at the bottom of the tables is the sum of squares of the kurtosis values and is used as a measure of deviation from normality (Gaussian-ness) for a PDF. Clearly the ICA has produced a greater deviation from normality, which suggests that higher-order PDFs exist in the signal. From the above results it is clear that ICA has done a better job of characterizing the features of the input TFD than SVD, thus we conclude that the ICA is a useful new technique for characterizing sound features.

3.5.2 Example 2: Coin dropping and bouncing sound

1. Method

The second example is that of a coin dropping and bouncing. The STFTM spectrogram of this sound is shown in Figure 26. Clearly discernible are the individual bounces which get closer through the course of the TFD, they exhibit a wide-band spectral excitation up to the Nyquist frequency, which in this case is 11.025 kHz because the original sound was sampled at only 22.050 kHz as compared with 44.1kHz for the other examples. This re-sampling does not affect our analysis since it just means that the spectrum is band-limited and it will serve as a good test of our methods for the restricted spectrum case. Since the coin is made out of a metallic material there is a high-frequency ringing component. This high-frequency component is observed to be continuous throughout the sound. We see this component centered at about 8kHz. Also discernible are a number of low and mid-frequency ringing components which are triggered by bounces, but do not continue for long. These components are metallic ring components with a rapid decay corresponding to internal damping in the coin and damping due to coupling with the surface of impact. We analyzed this sound with a 44.1kHz STFT in the same manner as the previous sounds, with no PSD normalization.

2. Results

The singular values of the coin drop sound are shown in Figure 25. The first three components account for approximately 33% of the total variance in the original sound. Again, as with the bonfire sound, the singular values decay rapidly at first followed by a steady exponential decay in the higher components.

Figure 27 and Figure 28 show the SVD and ICA basis vectors for the coin drop sound respectively. The left SVD basis vectors show some structural characterization. The individual bounces are relatively well delimited, but there is ambiguity across the functions. The right singular vectors have failed to separate the high-frequency, wide-band and low-mid components discussed above. Both the high-frequency components and the low-frequency components are mixed across the basis set. In contrast, we see that the left ICA basis vectors delineate three temporal behaviors: a decaying iterated amplitude sequence, a steady amplitude sequence with a single large spike and lastly we see the iterated impact amplitude functions for the bounce sequence. Similarly, the right basis vectors of the ICA show a clear separation of the TFD spectrum into three different spectral behaviors. The first is a low-pass component, corresponding with the rapidly decaying impact sequence of the first temporal vector which perhaps represents the initial high-velocity impacts of the coin with the impact surface, the second shows a predominance of mid and high-frequency spectral components corresponding with the ringing mentioned above and which is corroborated by the

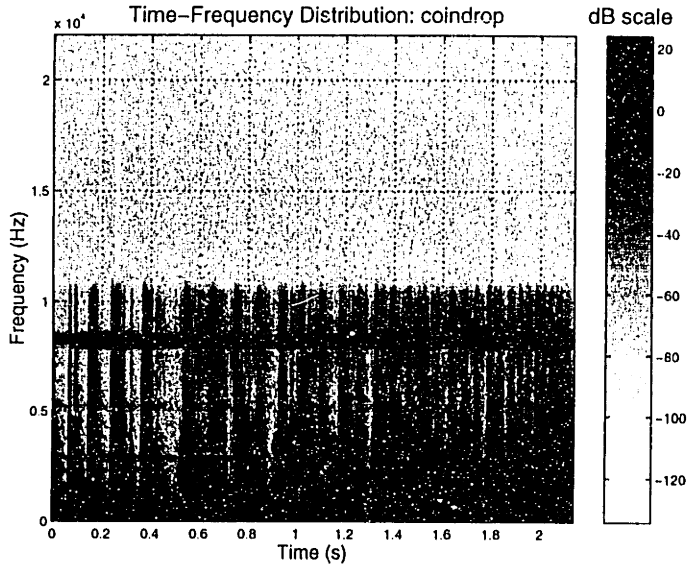


FIGURE 26. STFT spectrogram of coin drop sound. The features of interest in this TFD are the distinct high-frequency ringing component, the exponentially-decaying wide-band impact sequence (vertical striations) and low and mid-range ringing components.

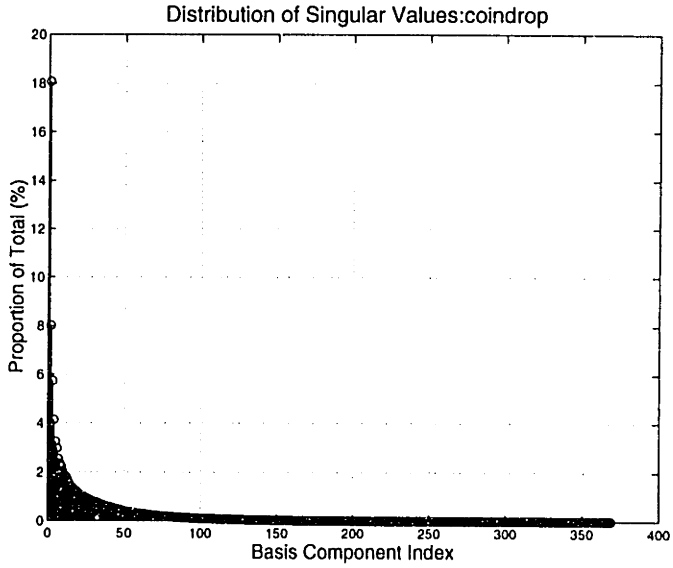


FIGURE 25. Singular values of coin drop sound. The first three components account for 33% of the origin variance in the TFD. The remaining components drop off gradually and exponentially.

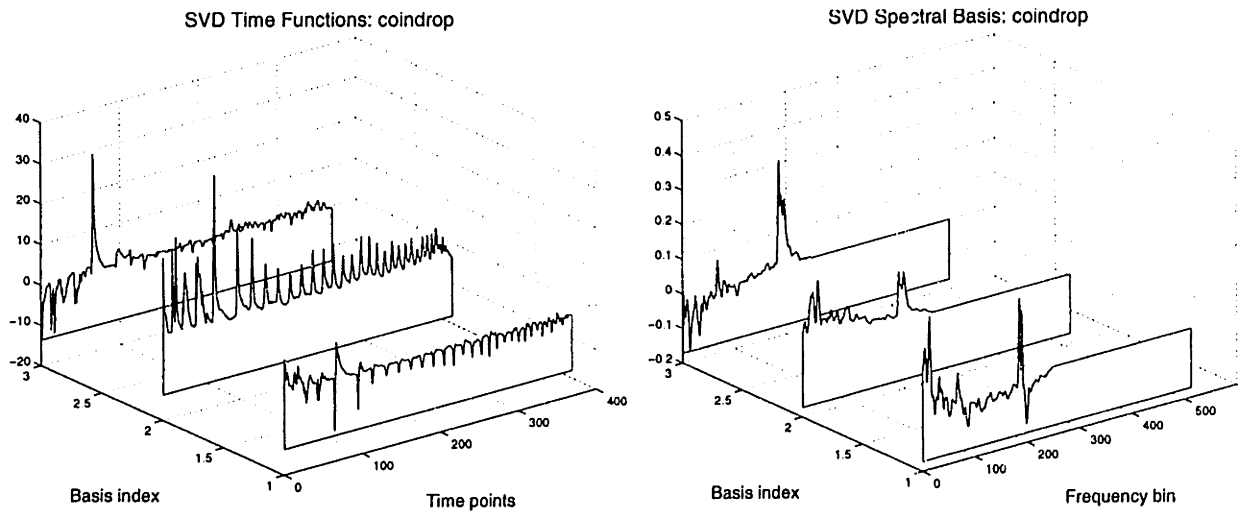


FIGURE 27. SVD basis vectors for the coin drop sound. The left singular vectors capture the structure of the coin bounces but there is some ambiguity. The right singular vectors fail to separate the high-pass, wide-band and low-pass components of the sound.

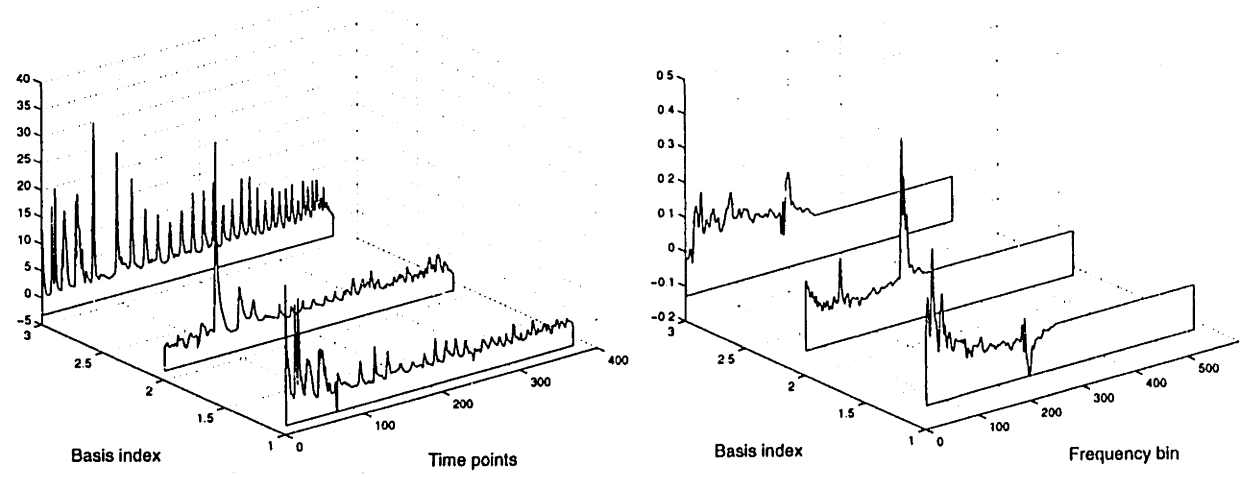


FIGURE 28. ICA basis vectors for the coin drop sound. The left singular vectors appear to reveal much of the temporal structure of the sound with the iterated behavior and the decaying envelope clearly delimited. The right singular vectors show a clear distinction between the low-frequency, high-frequency and wide-band components that we sought to characterize.

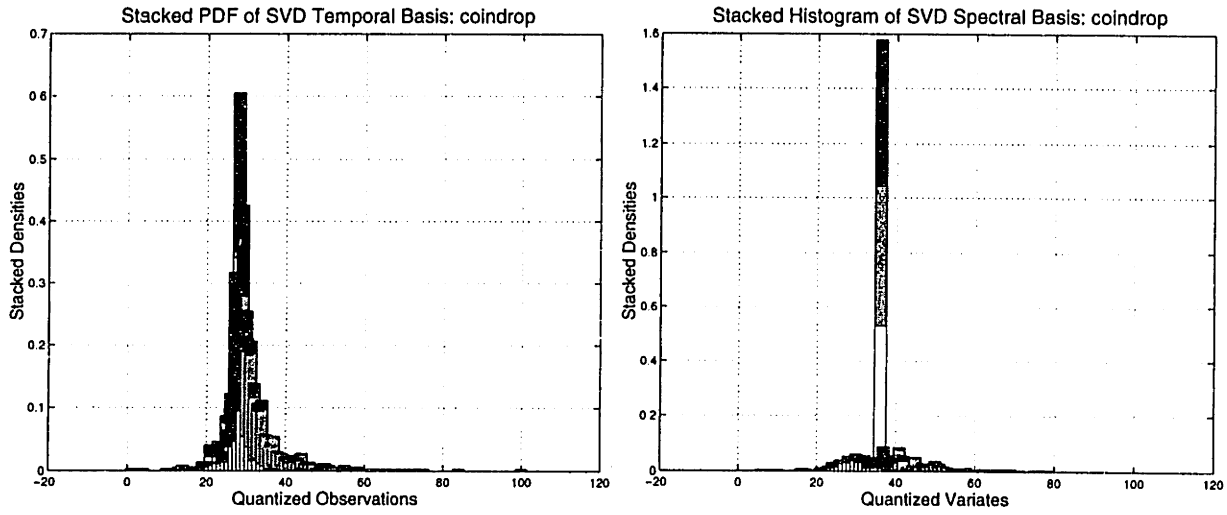


FIGURE 30. Stacked histograms of coin drop SVD basis functions.

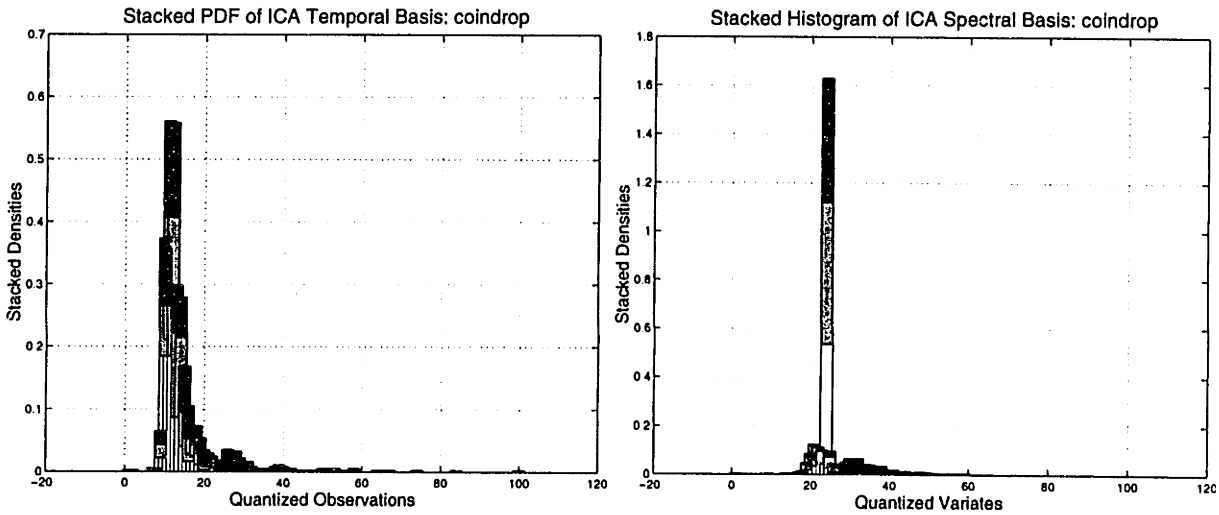


FIGURE 29. Stacked histograms of coin drop ICA basis functions.

continuous nature of the second ICA left amplitude function but for a single spike, and finally the third component exhibits the wide-band spectral content of the impacts, which again is supported by inspection of the left ICA amplitude functions. As with the previous example, these results suggest that the ICA has performed in a superior manner with respect to the characterization of features in the input TFD.

Figure 30 and Figure 29 show stacked histogram approximations of the PDFs of the coin drop sound for the SVD and ICA left and right vectors respectively. As with the last example, we can see that the SVD basis components are centered around a mean value and that the ICA values appear more skewed, suggesting the presence of higher-order cumulants in the underlying PDFs.

Table 7 shows the kurtosis values for the SVD and ICA right basis vectors respectively. As with the previous example, the ICA consistently maximizes the higher-kurtosis values whilst marginally changing lower values. This suggests that there is a single Gaussian component in the spectrum and that the other two components are non-Gaussian with relatively high kurtosis values. The contrast measure of the ICA compared with the SVD indicates that the ICA has a greater departure from normalcy and thus has performed better at characterizing the higher-order statistics of the input TFD.

TABLE 7. Fourth-Order Contrast Values for SVD and ICA Right Basis Vectors of Coin Drop

Basis Component	SVD Kurtosis	ICA Kurtosis
1	3.404	1.930
2	14.913	37.532
3	7.646	15.931
Contrast from Gaussian	292.476	1666.224

3.5.3 Example 3. Glass Smash Revisited

1. Method

Earlier in this chapter we gave an example of the application of SVD to the characterization of a glass smash sound. In this example we revisit the glass smash sound with an ICA analysis. The spectrogram of the glass smash sound is shown in Figure 8. In this example we not that the glass smash sound has a lot of energy in the low-frequency bands for the first few milliseconds of the sound due to the force of the impact. Since this energy is scaled disproportionately with respect to the energy of the subsequent glass particles shards we used a power-spectrum normalization of the input TFD as shown in the signal flow diagram of Figure 17.

2. Results

Figure 31 and Figure 32 show the singular values of the glass smash sound for non-normalized and PSD-normalized TFDs respectively. The subtle difference between the two is that the first component of the non-normalized TFD is much larger than the first in the PSD-normalized TFD. As we

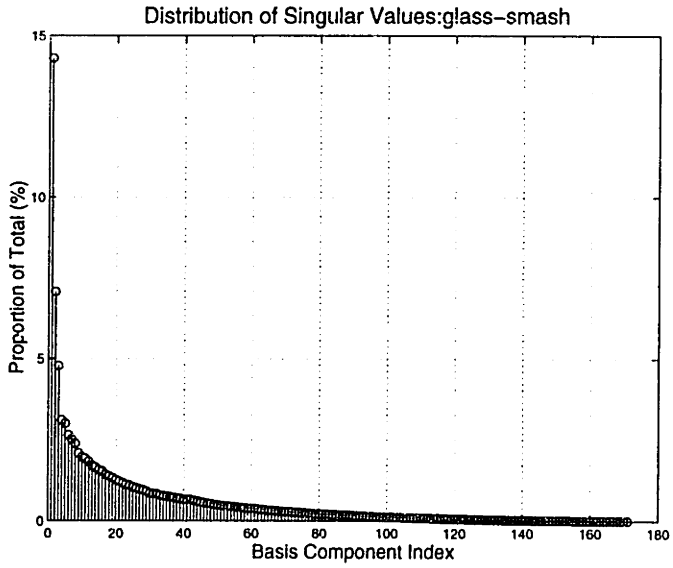


FIGURE 32. Singular values of glass smash PSD-normalized TFD decomposition.

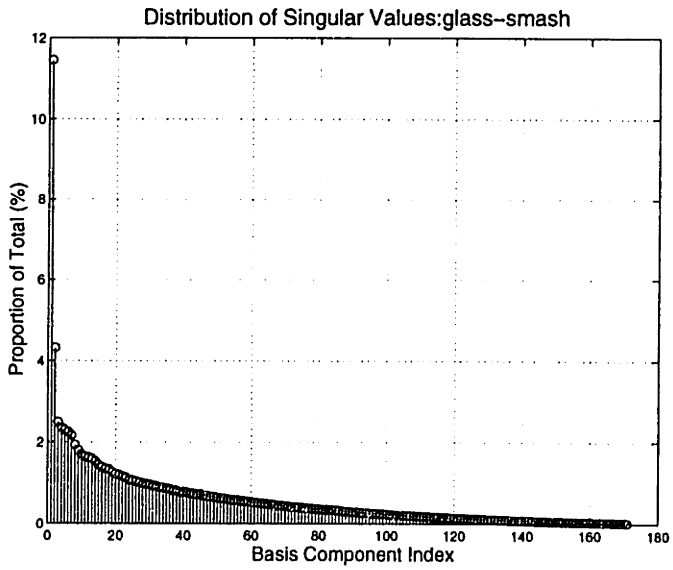


FIGURE 31. Singular values of glass smash non-normalized TFD decomposition.

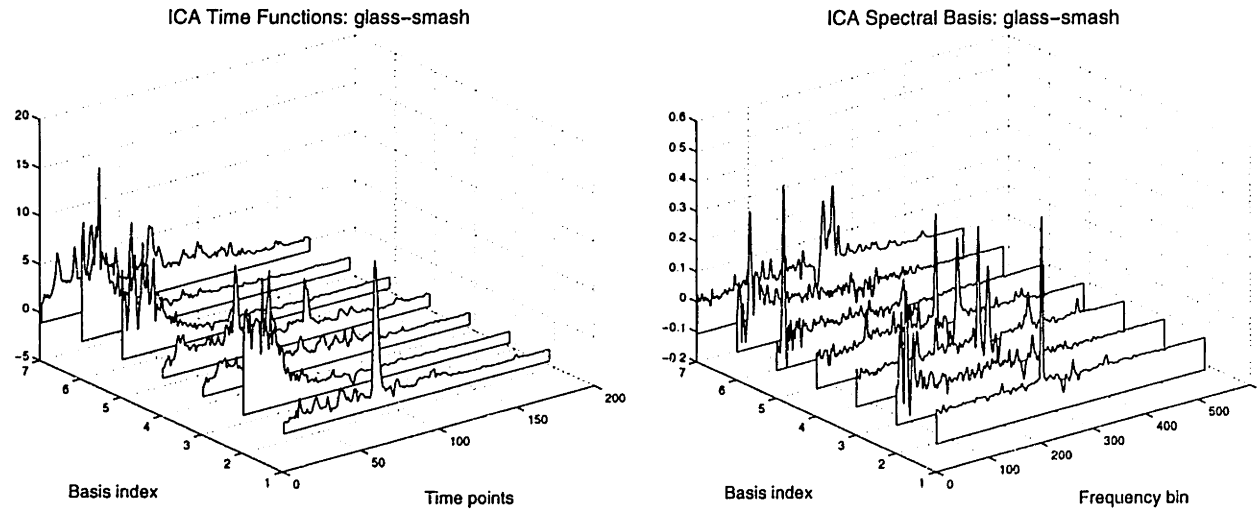


FIGURE 33. ICA basis functions for the non-normalized glass smash sound.

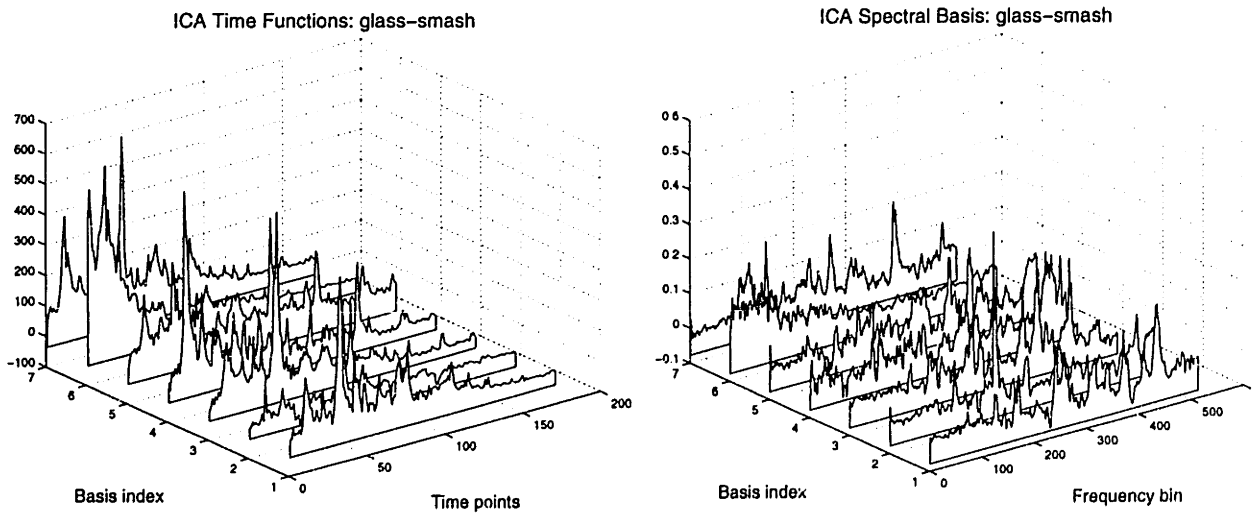


FIGURE 34. ICA basis functions for the PSD-normalized glass-smash sound.

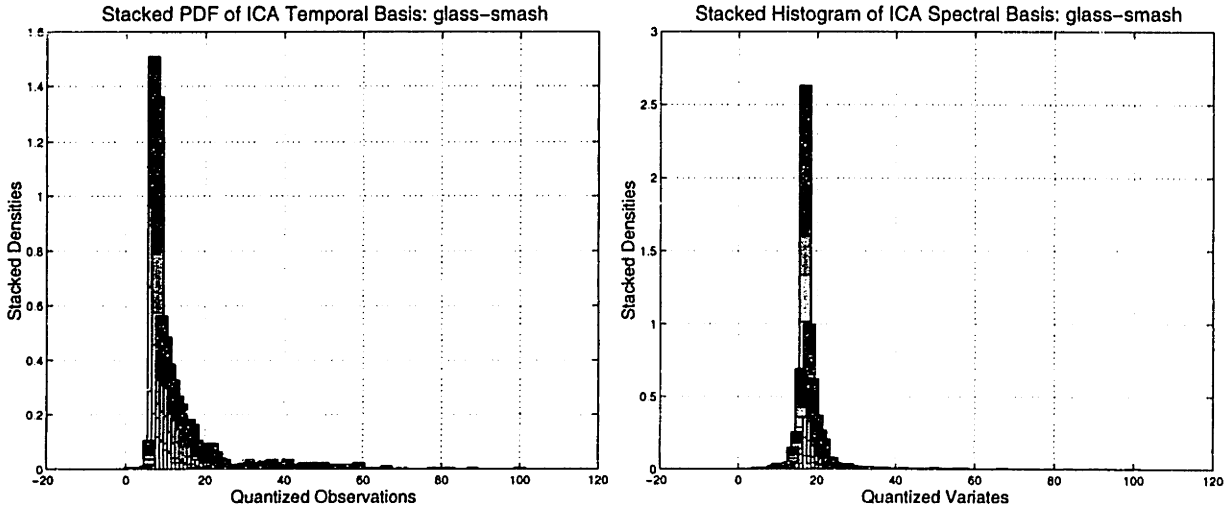


FIGURE 35. Stacked histograms of ICA basis functions for the non-normalized glass smash sound.

shall see, this first component corresponds with the low-frequency impact noise and the PSD normalization has had the effect of reducing its dominance for the decomposition.

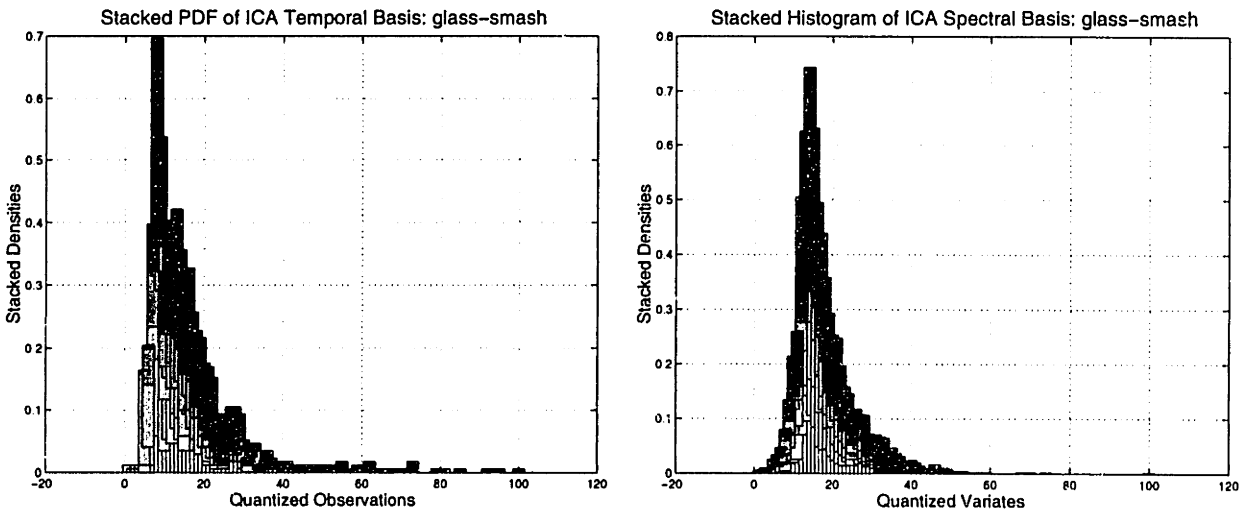


FIGURE 36. Stacked histograms of ICA basis functions for the PSD-normalized glass smash sound.

shall see, this component corresponds with the low-frequency impact noise. The effect is to enhance the presence of otherwise diminished components.

Figure 33 and Figure 34 show the basis vectors both the non-normalized and PSD-normalized input TFD respectively. The non-normalized functions show no less than three low-frequency impact components, suggesting an over-representation of this portion of the sound. The PSD-normalized vectors, however, show a more even spread of features, with only one feature corresponding to the low-frequency impact component. Amongst the other distinguishable components in the PSD-normalized basis are the presence of high-frequency narrow-band components which are not represented by the non-normalized spectrum. These components correspond with individual glass particles that are scattered throughout the spectrum. Thus they all exhibit high-Q narrow-band components, but with different fundamental frequencies. These observations are corroborated by inspection of the left-amplitude functions which show decaying-iterated time functions for these spectral components. From these results we conclude that the PSD-normalized spectrum has helped to significantly reduce the bias toward the low-frequency components in the input TFD. We suggest that PSD normalization may be a necessary tool for feature extraction from impact and explosion sounds due to the disproportionate representation of low frequency components in their TFDs.

Table 8 shows the values of the kurtosis for SVD and ICA decompositions of the PSD-normalized glass smash TFD. The ICA consistently maximizes the kurtosis over the SVD decomposition thus suggesting the presence of higher-order cumulants in the underlying PDFs. The contrast measure from normalcy suggests that the ICA has characterized the higher-order spectral structure of the glass-smash sound to a much greater degree than the SVD. The resulting basis vectors are therefore statistically independent to a much greater degree than the SVD basis vectors suggesting that ICA is again a better choice of decomposition for the input TFD.

TABLE 8. Fourth-Order Contrast Values for SVD and ICA Right Basis Vectors of Glass Smash

Basis Component	SVD Kurtosis	ICA Kurtosis
1	-1.827016	2.825701
2	0.024450	1.715675
3	1.463005	6.736605
4	1.218856	4.580318
5	2.932699	3.636371
6	1.359961	2.056869
7	7.047597	35.665819
Contrast from Gaussian	67.083415	1366.793833

3.6 Summary

In this chapter we have introduced the general notion of statistical basis decomposition of an arbitrary time-frequency distribution as a means for extracting features from sounds whose spectral and temporal properties are *a-priori* unknown. We developed a framework for investigating these techniques by considering principal component analysis. We have shown that PCA is not generally suitable for the decomposition of sound TFDs due to its reliance on a covariance representation which has the effect of decreasing the dynamic range of the input TFD with respect to the numerical accuracy of a given implementation. PCA was also shown not to characterize the different vector sub-spaces of a TFD; namely, the row space, the column space and the null space. Since, by our definition of a signal model, an input TFD is not necessarily assumed to be of full rank we sought a solution that would perform the said characterization.

The singular value decomposition was introduced as a method which directly decomposes a non-square matrix. This enables decomposition of a TFD without the need for covariance representation. This has the advantage of increasing the dynamic range of the decomposition over the standard PCA techniques. In addition to the advantage of rectangular decomposition, the SVD also provides a characterization of the vector spaces outlined above. This enables us to identify features for the row-space and column space of a TFD as well as enabling us to ignore the null-space components. Furthermore, this decomposition enables us to estimate the rank of the input TFD which provides a reasonable estimate of the number of statistically independent components in the input signal. However, we demonstrated that an SVD is limited by its assumption of Gaussian input statistics and showed that, for many sounds, such an assumption is not valid.

In order to address these problems we discussed a higher-order statistics extension of the SVD called independent component analysis. An ICA was shown to decompose the TFD of several complicated natural sounds in a manner that showed better feature characteristics than the corresponding SVD decomposition. The source of the better performance was the ability of the ICA to maximize cumulants at higher order than was possible using an SVD. Finally we showed that, in some cases, it is necessary to pre-process a TFD in order to remove bias toward high-energy low-frequency components. This was the case with a violent impact sound, glass smashing, and we demonstrated the superior performance of a power-spectral-density normalized TFD input over the standard non-normalized input. The ICA was also shown to outperform the SVD in this case.

In the larger context of our thesis, the techniques presented herein serve to identify features in a TFD. As discussed in this chapter, these features are considered to correspond directly to the structural invariants in an auditory group theory representation of a sound. Thus these methods are seen as a part of an auditory group analysis scheme. Whilst these components appear to be good candidates for sound-structure characterization, they do not offer a good characterization of time-varying components in a TFD. In the next chapter we consider how the statistical bases extracted from a TFD can be used to characterize time-varying structures within a sound in order to construct controllable models for sound-effect synthesis.

Chapter IV: Structured Sound Effects using Auditory Group Transforms

4.1 Introduction

In this chapter we develop signal processing techniques for implementing real-time, controllable sound-effects models. Since the domain of investigation of this thesis is the representation and synthesis of environmental sounds we have had to develop new signal representation methods in order to characterize the structure inherent in these sounds. The algorithms presented in this chapter are based on developing both the auditory group theory representation of sound structure and the use of statistical basis functions as structured audio source material. We start this chapter with a consideration of direct synthesis from the basis functions using inverse Fourier transform methods. Issues of reconstruction of time-frequency distributions from reduced basis representations will be covered

4.2 Resynthesis of Independent Auditory Invariants from Statistical Bases

In the previous chapter we investigated the problem of extracting statistically independent features from a time-frequency distribution under a signal model of superposition of outer-product independent TFDs. In this section we investigate the problem of reconstructing the independent TFDs from their statistical bases. These independent signals form the basic material for subsequent structured audio representation and synthesis.

4.2.1 Spectrum Reconstruction from Basis Components

Recall from the previous chapter the signal model of independent outer-product TFDs:

$$\chi = \mathbf{Y}_\rho \mathbf{V}_\rho^{\,T} = \sum_{i=1}^{\rho} \mathbf{Y}_i \mathbf{V}_i^{\,T} \qquad [121]$$

where χ is the spectrum reconstruction of the signal space of the analyzed sound. $\mathbf{Y}_i$ and $\mathbf{V}_i$ are the independent left and right basis vectors of each χ_i in χ of which there are ρ corresponding to

the estimated rank of the analyzed TFD. This equation specifies an algorithm for independent component resynthesis.

First let us investigate the reconstruction of a full-spectrum signal TFD from a basis set. Both the SVD and ICA basis components span the same subspace of a TFD. We can see this by investigating the form of the ICA analysis equation:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{Q}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\Lambda^{-1}\mathbf{D}\mathbf{D}^{\mathrm{T}}\Lambda\mathbf{P}\mathbf{Q}\mathbf{V}^{\mathrm{T}} = \mathbf{Y}\mathbf{Z}. \qquad [122]$$

This equation describes the factorization of a TFD into a left basis set $\mathbf{Y}$ and a right basis set $\mathbf{Z}$ that span the full space of $\mathbf{X}$. But let us now consider the unitary transform $\mathbf{Q}$. The unitary transform produces orthogonal rotations of the basis vectors of an SVD: $\mathbf{U}$ and $\mathbf{V}$. This means that each plane of the SVD basis functions is mapped into the same plane, but it is rotated about the origin. Thus each plane of the SVD basis spans exactly the same Euclidean subspace as the resulting ICA basis under the unitary transform $\mathbf{Q}$. Furthermore, it is evident that the relation $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}$ holds because $\mathbf{Q}$ is unitary. The upshot of this is that the reduced ensemble of ρ ICA basis vectors spans exactly the same subspace of $\mathbf{X}$ as the reduced ensemble of ρ SVD basis vectors, but the rotation of the basis within that subspace is different between the two, with ICA basis vectors pointing in directions that maximize the fourth-order cumulants of the underlying PDFs as described in the previous chapter.

Now let us consider the uniqueness constraints $\mathbf{P}$, Λ and $\mathbf{D}$. $\mathbf{P}$ is a permutation matrix which is an orthonormal matrix and thus has the property $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$ which is the identity matrix. Λ is a diagonal scaling matrix and is used in conjunction with its trivial inverse (inversion of the diagonal entries) in Equation 122 which produces the relation $\Lambda^{-1}\Lambda = \mathbf{I}$. Finally, $\mathbf{D}$ is a matrix of real diagonal entries with unit modulus thus $\mathbf{D}\mathbf{D}^{\mathrm{T}} = \mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{I}$. Thus, for a reduced basis decomposition of ρ basis vectors, Equation 122 reduces to the following form:

$$\hat{\mathbf{X}}_\rho = \mathbf{U}_\rho\Sigma_\rho\mathbf{I}_\mathbf{Q}\mathbf{I}_\mathbf{P}\mathbf{I}_\Lambda\mathbf{I}_\mathbf{D}\mathbf{V}_\rho^{\mathrm{T}} = \mathbf{U}_\rho\Sigma\mathbf{V}_\rho^{\mathrm{T}} \qquad [123]$$

where $\mathbf{I}_\mathbf{Q}$ etc. indicates an identity transform due to a pair of complimentary matrices. This relation serves as a proof that the subspace spanned by the SVD basis is exactly the same as the subspace spanned by the ICA basis because they reconstruct exactly the same signal.

The utility of this proof is that, for a full spectrum reconstruction of the composite signal TFD χ, the resynthesized TFD is exactly the same for both the SVD and ICA cases. Thus there is no distinction between the two methods for the purposes of data compaction in the formulation of ICA that we developed in the previous chapter. However, since the *independent* basis components point in different directions through the same Euclidean subspace, there are quantitative differences between the two sets of individual bases. It is only the ensemble subspace that remains the same under the ICA transform. This is a desirable property of the ICA since it is an invertible represen-

tation of a TFD with respect to its SVD factorization. These points serve as a strong argument for using an algebraic form of an ICA rather than adopting *ad hoc* learning algorithm techniques.

If now consider the sum of independent TFDs formulation:

$$\chi = \mathbf{Y}_1\mathbf{V}_1^{T} + \mathbf{Y}_2\mathbf{V}_2^{T} + \ldots + \mathbf{Y}_\rho\mathbf{V}_\rho^{T} \qquad [124]$$

we have shown that the effect of the ICA is to produce a different set of independent summation terms than an SVD, one that characterizes each independent component of χ in a more satisfactory manner than an SVD basis but which also preserves the vector space spanned by the SVD.

4.2.2 Example 1: Coin Drop Independent Component Reconstruction

As an illustration of these points we consider the spectrum reconstruction of the coin drop sound whose full spectrum is shown in Figure 26 in ChapterIII. To illustrate the first point, that the ICA and SVD span the same subspace of $\hat{\mathbf{X}}_\rho$, the 3-component reconstruction of the coin drop sound is shown in Figure 37 and Figure 38 for an SVD and an ICA basis respectively. These spectrograms are exactly the same thus empirically corroborating the proof of subspace equivalence.

We can see that the TFD reconstruction has captured most of the salient detail in the original TFD with only 3 components; the original non-factored TFD had 257 components ($\frac{N}{2}+1$ due to symmetry of the Fourier spectrum of a real sequence). The white patches in the TFD represent areas where the reconstruction produced negative values. Since this TFD is an STFTM each point in the TFD is a magnitude which, by its definition, cannot be negative for a complex spectrum. These negative values arise because a good deal of the original TFD space has been cropped by the projection onto a 3-dimensional subspace thus additive compensating spectral magnitudes have been eliminated. The negative values are clipped at a threshold of -100dB in order to create a sensible magnitude spectrum reconstruction. Comparison of the white patches with the original TFD reveals that these areas were of low magnitude in the original TFD. Only a very small portion of the reconstructed TFD needs to be clipped in the case of full independent-component signal spectrum reconstruction.

We now investigate the resynthesis of each independent component χ_i using both the SVD and ICA factorizations. Figure 40 and Figure 39 show a reconstruction of the first independent component of both the SVD and ICA factorizations respectively. This reconstruction can be viewed as the first summation term in Equation 124. What is immediately noticeable is that the SVD factorization produces an independent component TFD that seems to oscillate about two complimentary spectral bases producing a mesh-grid type pattern. This behavior stems from the tendency of an SVD factorization to contain negating components as well as additive components in each basis vector, which is due to the fact that the components are not really statistically independent. In contrast, a consideration of the ICA independent component shows clearly a behavior that matches the low and mid-frequency ringing components of the original coin drop TFD.

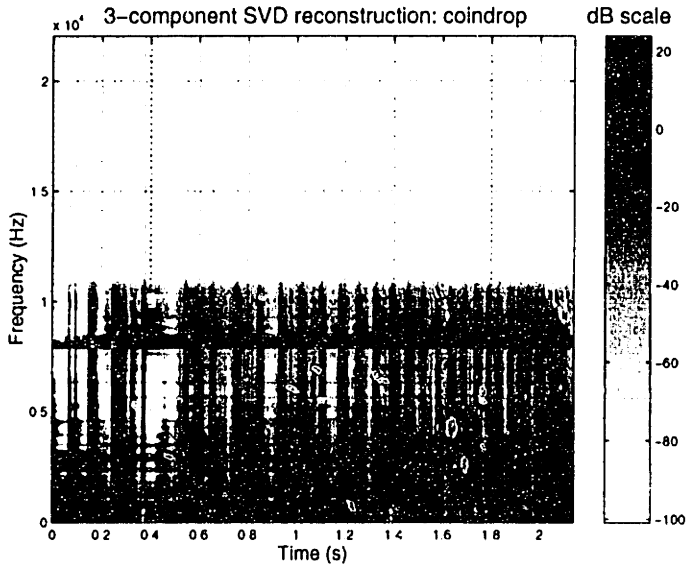


FIGURE 37. SVD 3 basis-component reconstruction of coin drop TFD.

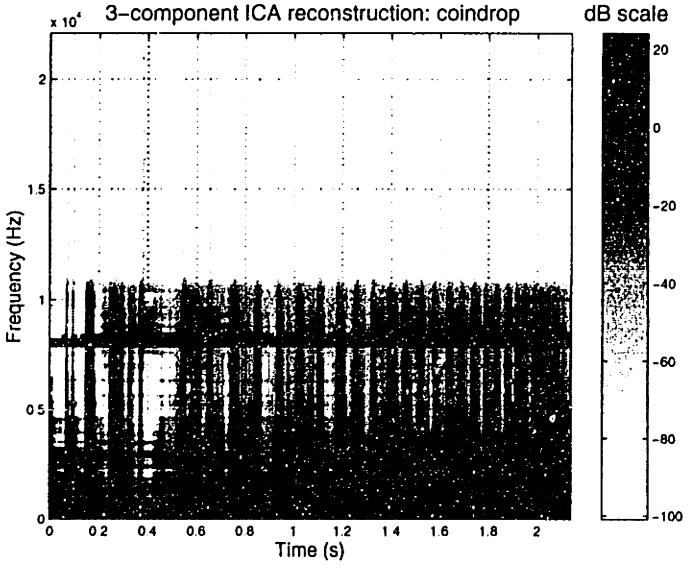


FIGURE 38. ICA 3 basis-component reconstruction of full coin-drop TFD.

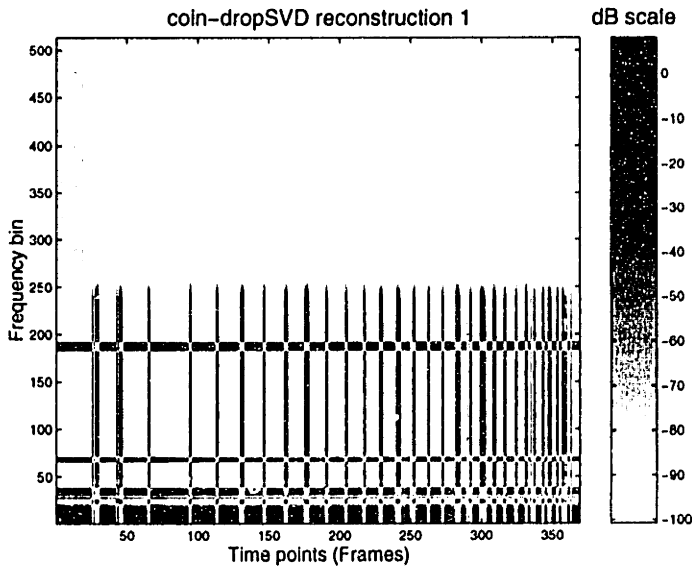


FIGURE 40. SVD reconstruction of first coin drop independent TFD.

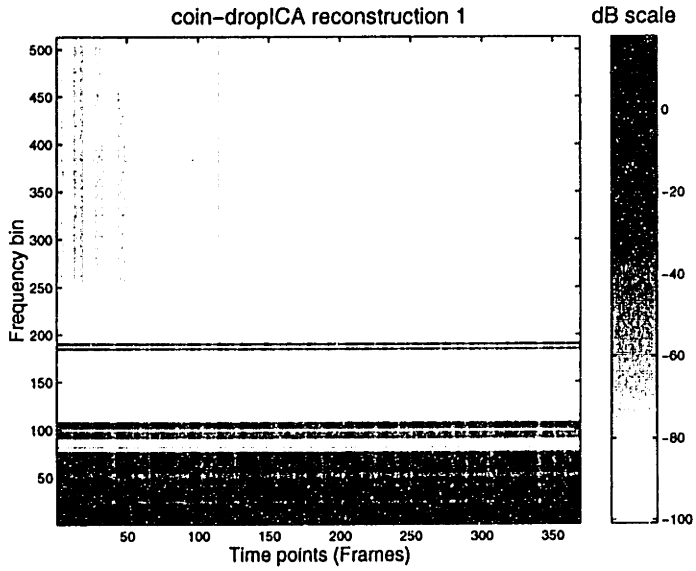


FIGURE 39. ICA reconstruction of first coin drop independent TFD.

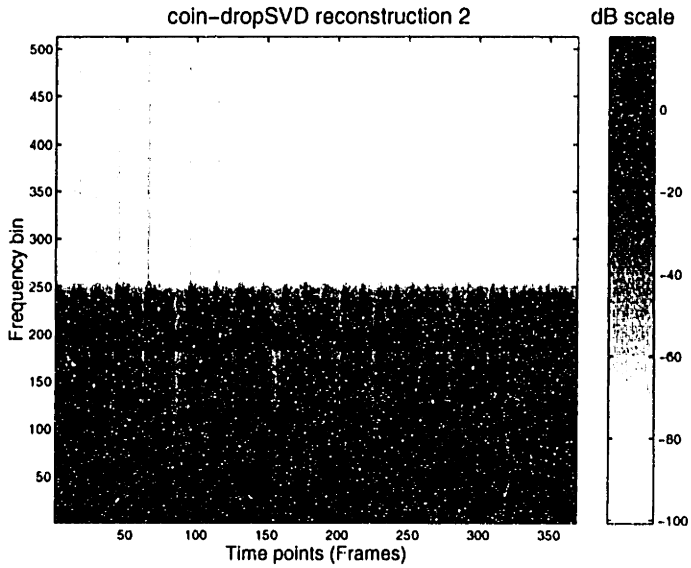


FIGURE 41. SVD reconstruction of second coin drop independent TFD.

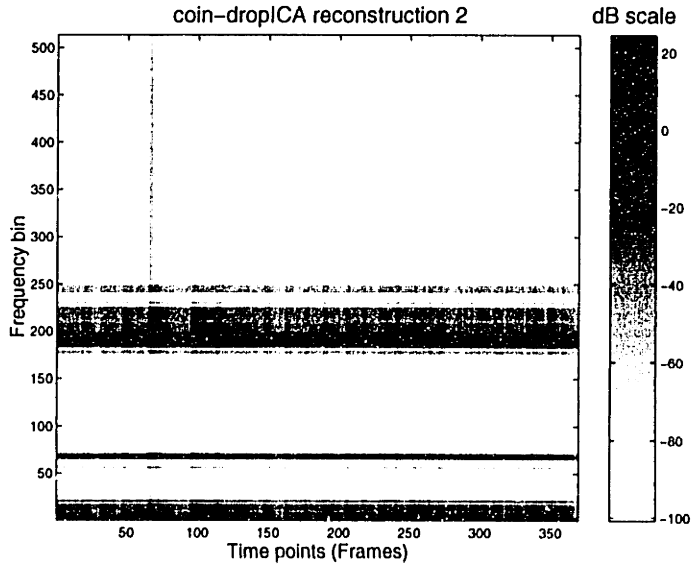


FIGURE 42. ICA reconstruction of second coin drop independent TFD.

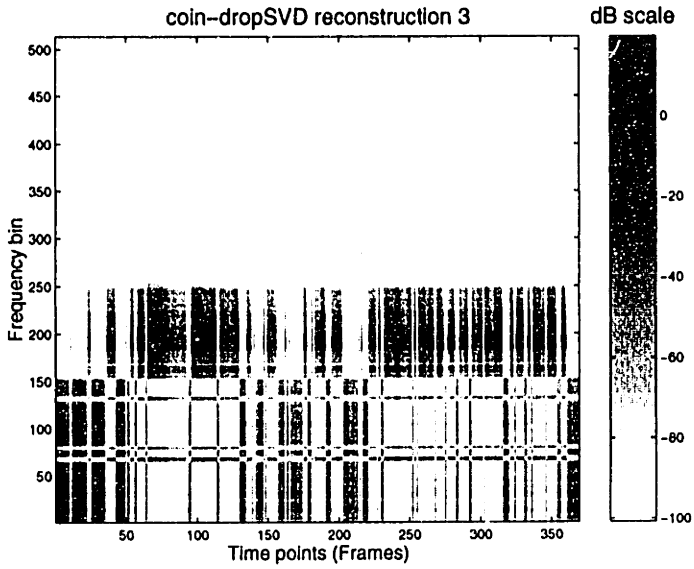


FIGURE 44. SVD reconstruction of third coin drop independent TFD.

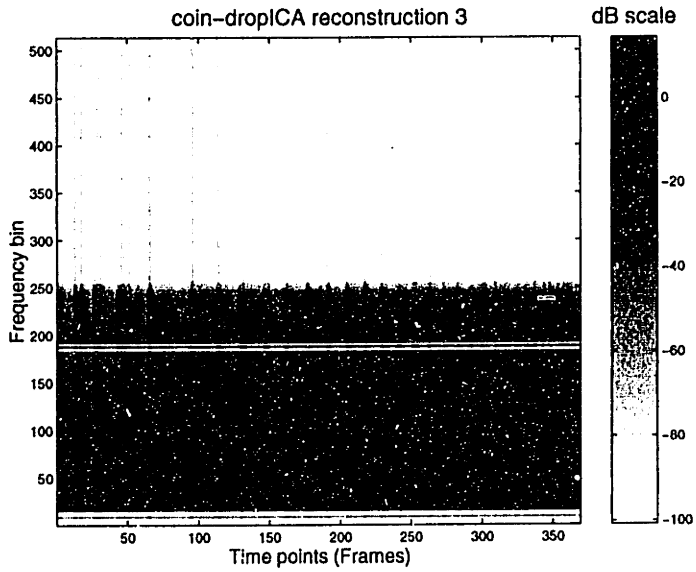


FIGURE 46. ICA reconstruction of third coin drop independent TFD.

Figure 41 and Figure 42 show TFD reconstructions of the second independent component for the SVD and ICA respectively. In this example the SVD component exhibits more stable behavior than in the previous example, but it has not successfully decorrelated the high-frequency ring component from the wide-band bounce-impacts. We can see this by the presence of both horizontal and vertical structures in the spectrum. The second ICA component has produced complimentary ringing components to those found in the first ICA component. These seem to correspond to mainly to high-frequency modes in the coin with some low and mid-frequency ringing also present.

Figure 44 and Figure 46 show TFD reconstructions of the third independent component. Again, the SVD has resorted to the mesh-grid pattern seen in the first SVD independent component. The third ICA component, however, reveals a clear vertical structure corresponding to the impacts of the coin on a surface with very little horizontal structure present in the TFD. This independence of this structure is due to the fact that the wide-band signal is an instantaneous excitation function for the coin, thus it is itself independent of the ringing from a physical point of view. It is the cause of the ringing, thus we see that the ICA has successfully decomposed the coin drop sound into independent components that make sense from a physical perspective as a source/filter model. This example has shown how we can reconstruct a set of two-dimensional TFDs from pairs of one-dimensional vectors. The vectors represent the most important structure of the TFD in terms of its orthogonal time-varying amplitude and time-static spectral components.

4.2.3 Example 2: Bonfire Sound

In this example we discontinue discussion of the SVD since we have sufficiently covered its limitations from the point of view of TFD feature extraction as well as TFD independent component resynthesis. The full spectrum TFD of the bonfire sound is shown in Figure 19 in ChapterIII. The 3-component reconstruction for the bonfire sound is here shown in Figure 47. The signal TFD reconstruction shows that the basis vectors have successfully captured both the wide-band erratic crackling as well as the continuous low-pass and wide-band noise densities of the sound.

Inspecting the independent TFD resynthesis we see from Figure 48 that the first independent component has captured some of the wide-band crackling components of the sound, with the dark-gray regions representing energy approximately 20-30dB above the light-gray regions. The second independent component, shown in Figure 50, shows a much more characteristic wide-band intermittent crackling component than the first. We note that the first component contains energy at lower frequencies, at around the 60th frequency bin, than the second which is perhaps the source of the independence. The third component, shown in Figure 49, clearly shows a horizontal continuous noise structure which contains none of the intermittent crackling thus demonstrating that the ICA has successfully separated the continuous noise component from the intermittent wide-band component.

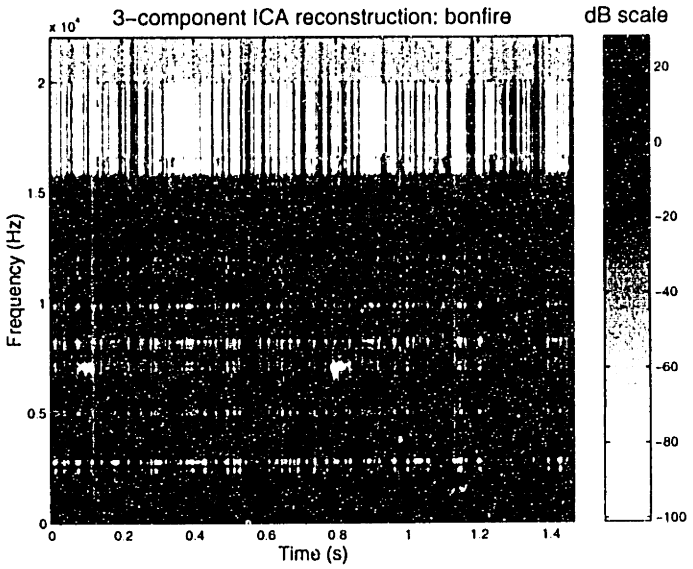


FIGURE 47. ICA 3-component reconstruction for the bonfire sound. (See Figure 19 on page 111 for full-spectrum TFD).

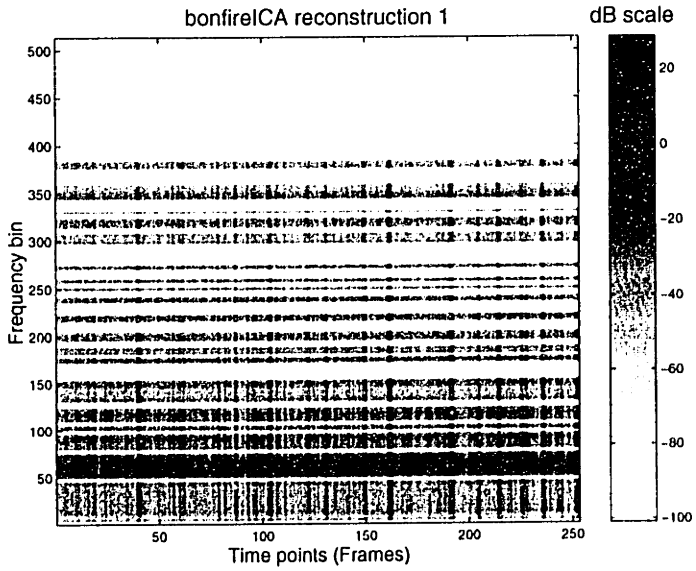


FIGURE 48. First Independent TFD reconstruction from ICA basis.

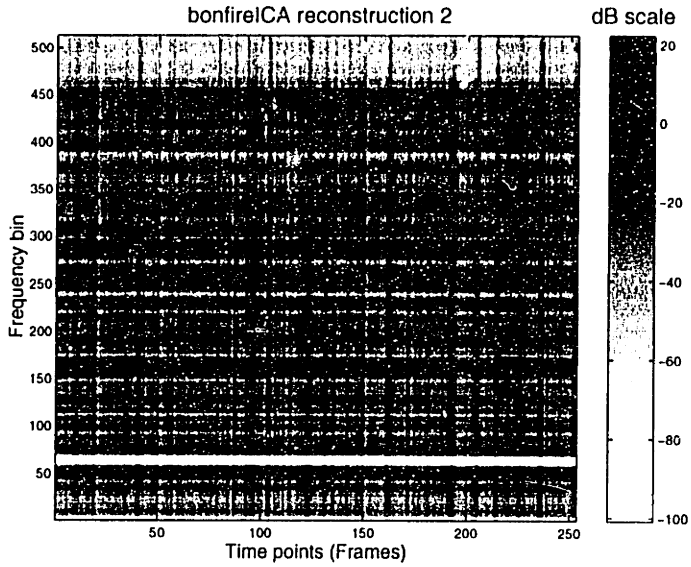


FIGURE 50. Second Independent TFD reconstruction from ICA basis.

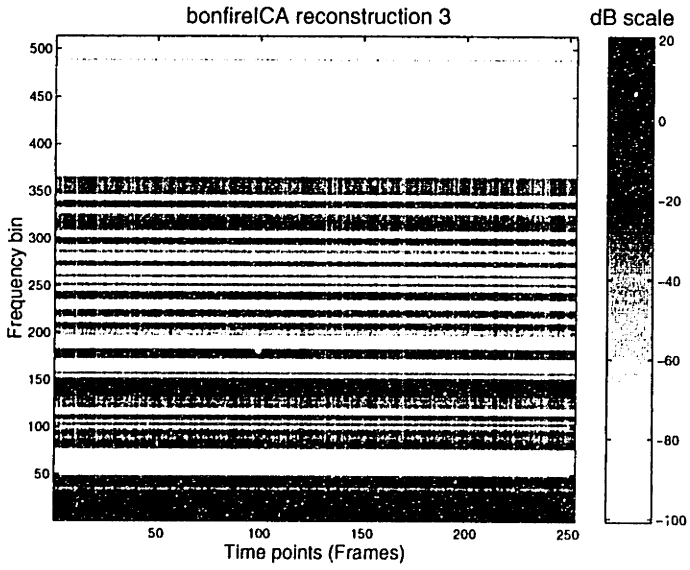


FIGURE 49. Third Independent TFD reconstruction from ICA basis.

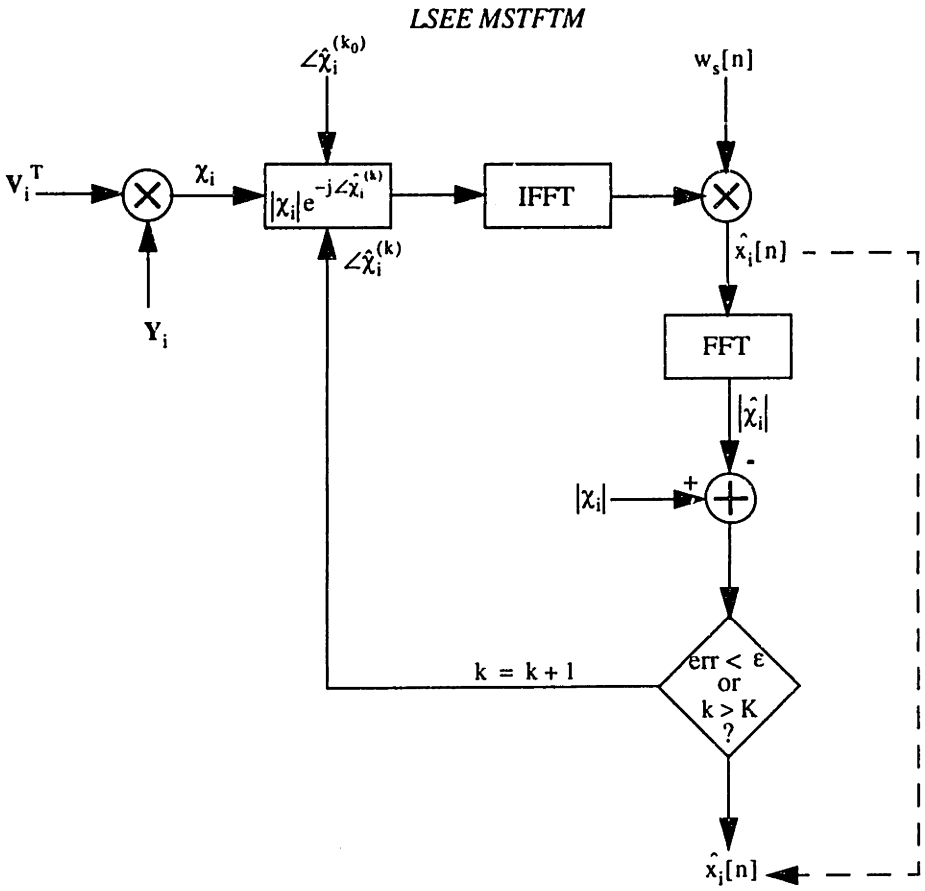


FIGURE 51. Independent component signal reconstruction algorithm. Least-Squares Error Estimation Modified Short-Time Fourier Transform Magnitude, based on Griffin and Lim (1984).

4.2.4 Signal Resynthesis from Independent Component Spectrum Reconstruction

Having obtained a set of independent magnitude spectrum reconstructions the problem at hand is how to estimate a signal for the independent component. Recall that we are assuming a magnitude-only spectrum representation for the TFD. We further assume, in this section, that the TFD can be transformed into an STFTM representation under some mapping. This is the approach used, for example, in Slaney et al. (1996) in which a correlogram representation is transformed into an STFTM representation for the purposes of correlogram inversion. Such a framework is quite general and enables us to proceed with little loss of generality in the methods.

Figure 51 shows a general algorithm for estimating a signal from a modified short-time Fourier transform magnitude representation. The STFTM TFD is constructed using the outer-product of the independent component vectors as described previously. Following the algorithm of Griffin and Lim (1984), which is also the algorithm used by Slaney et al. (1996), we seek to estimate a phase spectrum for the TFD such that the inverse transform yields a signal whose forward transform produces a magnitude TFD that minimizes the error with respect to the specified independent component in the least-squares sense.

To understand the problem consider that an arbitrary initial phase spectrum $\angle\hat{\chi}_i^{(k_0)}$ combined with the specified TFD magnitude $|\chi_i|$ in general is not a valid STFT since there may be no sequence whose STFT is given by $|\chi_i|e^{-j\angle\hat{\chi}_i^{(k_0)}}$, see Griffin and Lim (1984). The LSEE MSTFTM algorithm shown in Figure 51 attempts to estimate a sequence whose STFTM $|\hat{\chi}_i|$ is closest to the specified $|\chi_i|$ in the least squared error sense. By expressing the distance between the estimated and specified spectrum as a distance measure:

$$\delta\left\{\hat{x}_i[n], |\chi_i|e^{-j\angle\hat{\chi}_i}\right\} = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\chi_i(\omega) - \hat{\chi}_i(\omega)\right|^2 d\omega \qquad [125]$$

and applying Parseval's relation:

$$\delta\left\{\hat{x}_i[n], |\chi_i|e^{-j\angle\hat{\chi}_i}\right\} = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \left|x_i[mH-n] - \hat{x}_i[mH-n]\right|^2 \qquad [126]$$

a quadratic solution for $\hat{x}_i[n]$ is found the form of which is an overlap-add synthesis procedure with an error-minimizing windowing function $w_s[n]$:

$$x_i[n] = \sum_{m=-\infty}^{\infty} w_s[mH-n]\hat{x}_i[mH-n] \qquad [127]$$

$$w_s[n] = \left(\frac{2\sum_{m=0}^{\left(\frac{L}{H}\right)-1} \frac{H}{L}}{\sqrt{4a^2+2b^2}}\left[a + b\cos\left(\frac{2\pi n}{L} + \frac{\pi}{L}\right)\right] \right) \qquad [128]$$

where $a = 0.54$ and $b = -0.46$ are the Hamming window coefficients and H and L are the STFT hop size and window lengths respectively. Equation 128 holds only if $L = 4H$.

The procedure is iterative, estimating a signal then transforming it back to the Fourier domain for the next iteration and converging to an optimal solution (in the quadratic sense) over successive iterations. A solution is considered adequate if the total magnitude error across the entire TFD satisfies $\left\{\hat{x}_i[n], |\chi_i| e^{-j\angle\hat{\chi}_i}\right\} < \varepsilon$, with ε chosen as an arbitrarily small number.

The LSEE MSTFTM algorithm can be helped by the incorporation of different initial conditions for the phase estimates. For harmonic, or quasi-periodic sounds an initial linear-phase condition for the independent component TFD provided good results. For noise-based spectra initial phase conditions of a uniformly distributed random sequence produced good results. Using an appropriate choice for initial phase conditions in the MSTFTM algorithm we found that the solution converges satisfactorily after approximately 25 iterations.

4.3 Auditory Group Re-synthesis

4.3.1 Signal Modification using the LSEE MSTFTM

The independent component re-synthesis algorithm described above recovers a signal from a magnitude STFT so it can be used for signal modifications of an independent component feature. We take the final spectrum estimate of the MSTFTM algorithm as the spectrum for modified re-synthesis. The form of signal modifications for the MSTFTM follow closely the form of the phase vocoder which was previously discussed in Chapter II. The main auditory groups corresponding to phase vocoder transforms are: T_π and T_Ω which are the time-only and frequency-only transforms corresponding to resynthesis hop-size alterations and frequency-scale transforms with compensating time-only transforms as discussed in Chapter II.

Using these transforms, it is possible to implement a real-time synthesis algorithm based on the inverse FFT. The algorithm performs much in the same way as the phase vocoder implementing independent time-stretch and frequency-scale operations on the estimated LSEE MSTFTM TFD discussed in the previous section, see Figure 47. Because the transformations are associative, the order of the component transformation is not important.

Whilst it is possible to use the auditory group transformed IFFT as a real-time algorithm for independent control over sound features the implementation is somewhat expensive, even given the relative efficiency of the FFT. So rather than focusing upon an FFT-based implementation we turn our attention to more efficient techniques.

4.3.2 Efficient Structures for Feature-Based Synthesis

One way of improving on the ISTFT re-synthesis method is to construct an equivalent, more efficient, model in the time domain. The basis for an efficient filter-based implementation for independent component resynthesis is that the TFD of each independent component comprises essentially a single filter that is modulated in amplitude for each time frame. This amplitude modulation, together with the phase estimates provided by the MSTFTM, constitute the total information necessary to reconstruct the signal. Therefore it is possible for us to design filters for each independent component right spectral basis vector and drive it with filters designed from the left amplitude basis vectors of an independent component TFD.

There are two general approaches to the problem of filter design for independent component modeling: finite impulse response (FIR) and infinite impulse response (IIR) modeling. We start with a consideration of FIR modeling because of its close relationship to the inverse Fourier transform method of re-synthesis discussed above.

4.3.3 FIR Modeling

Figure 54, Figure 52, and Figure 55 show FIR filters for the three features of the bonfire ICA analysis. The FIR filters are obtained by a zero-phase inverse Fourier transform:

$$s_i[n] = \frac{1}{N}\sum_{k=0}^{N-1} V_i[k]e^{j\frac{2\pi k}{N}n} \qquad [129]$$

where $V_i[k]$ is a single DFT vector obtained from the Fourier magnitude values in $\mathbf{V}_i$, the i-th column of the right basis vectors. The Z-transform of $s_i[n]$ is $S_i(Z)$ and we shall refer to it shortly.

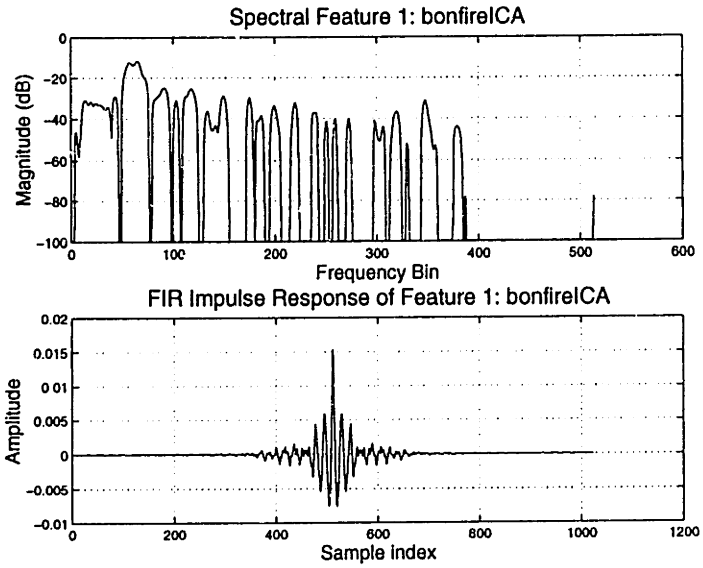


FIGURE 54. Bonfire sound: linear-phase FIR filter for spectral independent basis component 1.

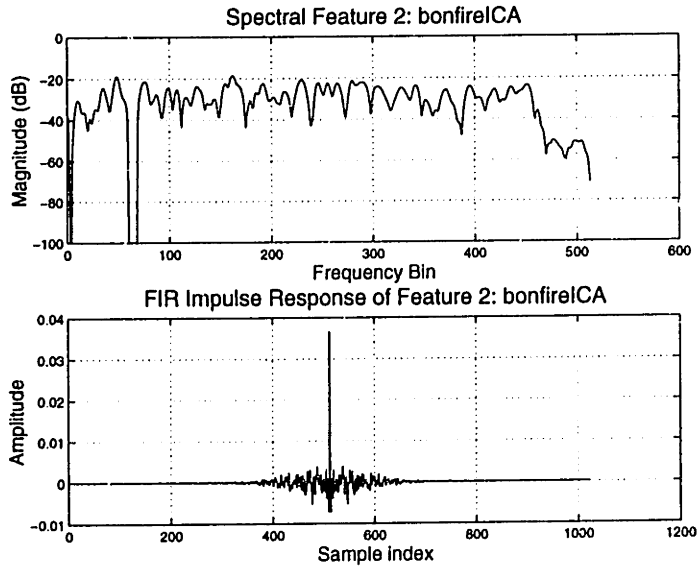


FIGURE 52. Bonfire sound: linear-phase FIR filter for spectral independent basis component 2.

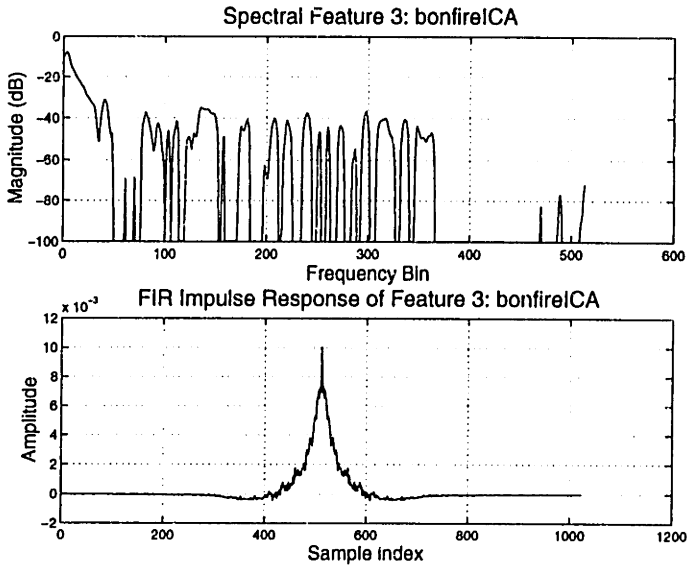


FIGURE 55. Bonfire sound: linear-phase FIR filter for spectral independent basis component 2.

These linear-phase FIR filters are used in conjunction with a matching set of IIR filters that are estimated from the left independent component basis vectors of the TFD. These IIR filters are low-order approximations of the time-amplitude basis functions and operate at the *frame rate* of the source STFT. In the examples that follow all of the time-amplitude IIR models of left basis vectors are 8th-order auto-regressive models obtained by an auto-covariance linear-predictive coding analysis (LPC). This type of analysis was discussed previously in Chapter II so we shall not explain it here.

The LPC analysis yields a set of filter coefficients for each independent component time function as well as an excitation signal. We can represent the form of this system as:

$$E_i(Z) = \frac{B_i(Z)}{A_i(Z)}, \qquad [130]$$

where $B_i(Z)$ are the zeros of the system and represent an excitation source signal for the amplitude function, and $A_i(Z)$ is an 8-th order prediction filter for the time-response of the amplitude function. These IIR models generate an impulse train spaced at the STFT frame rate. Time-stretch control is produced by altering the spacing of the impulse train which corresponds to a shift in the hop-size for the underlying STFT representation, this corresponds to the auditory group transform T_π which produces time-only alterations of a signal. Frequency-shift control is produced in the

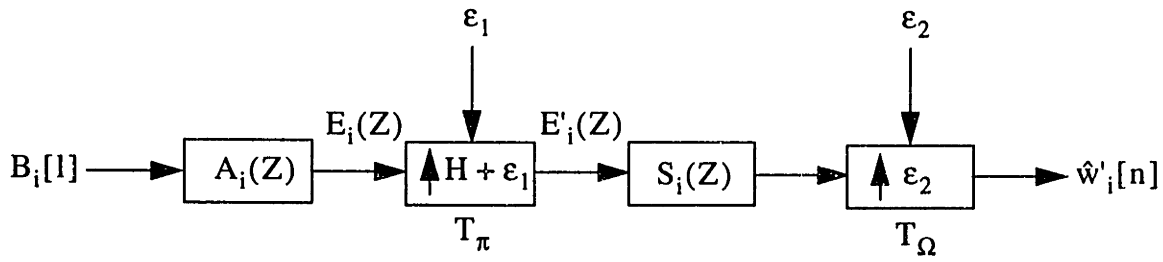


FIGURE 56. Implementation of structured independent component re-synthesis using a linear-phase FIR model $S_i(Z)$. The system function of the amplitude signal $E_i(Z)$ is specified at the Fourier transform frame rate. It is interpolated to the required hop size by an auditory group transform T_π which alters the global time structure of the sound. On the right, the transform T_π produces shifts in frequency by a factor ε_2.

same manner as the phase vocoder using the T_Ω frequency-only transform. The system flow diagram for FIR modeling of independent components is given in Figure 56. The input, on the left, is a source-excitation signal $B_i[l]$ expressed at the STFT frame rate, it's time response is generated by $A_i(Z)$ which is the prediction filter for amplitude functions. The time-scale-only auditory group transform produces shifts in the spacing of the amplitude functions which comprise a variably-spaced amplitude-varying impulse train $E'_i(Z)$. These frame-rate impulses are convolved with the FIR impulse response of the spectral basis component $S_i(Z)$ which is transformed by the frequency-shift-only auditory group transform. The result of these filtering operations is the synthesis of an independent component by separate control over the time-amplitude basis functions and the frequency basis functions of the underlying TFD.

By way of example, consider the IIR time-function models and excitation sequences in Figure 57-Figure 60. These signals are the systems-level implementation of the left ICA basis functions, and they are used in conjunction with the right-basis function FIR models described above. The first two figures show the IIR impulse response of the amplitude prediction filters. The figures show that the third independent component has a longer time response than the first. The third corresponds to a continuous noise component in the sound and the first corresponds to crackling components in the bonfire sound.

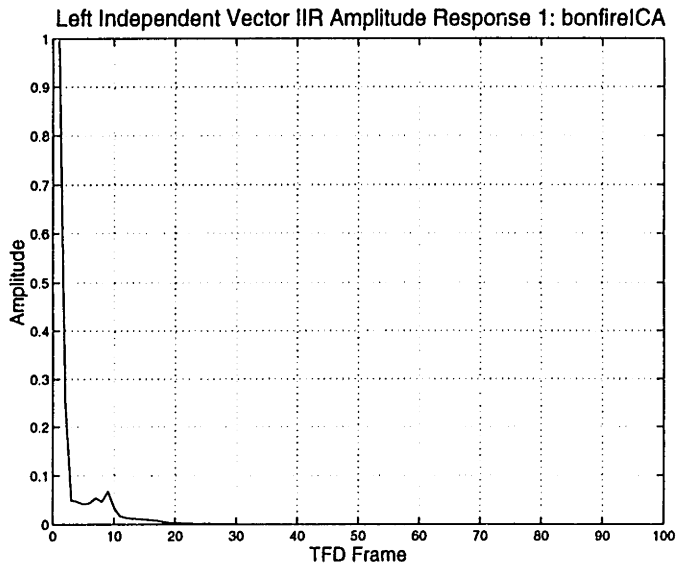


FIGURE 57. Impulse response of the first left independent vector IIR model of the bonfire sound.

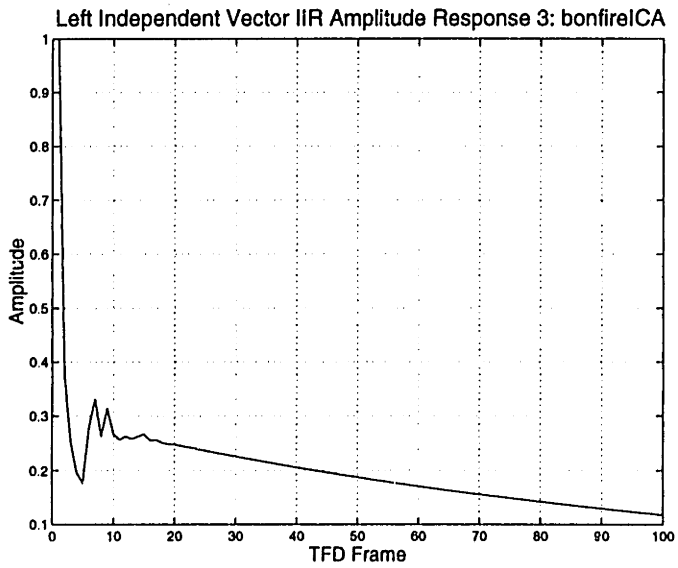


FIGURE 58. Impulse response of the third left independent vector IIR model of the bonfire sound.

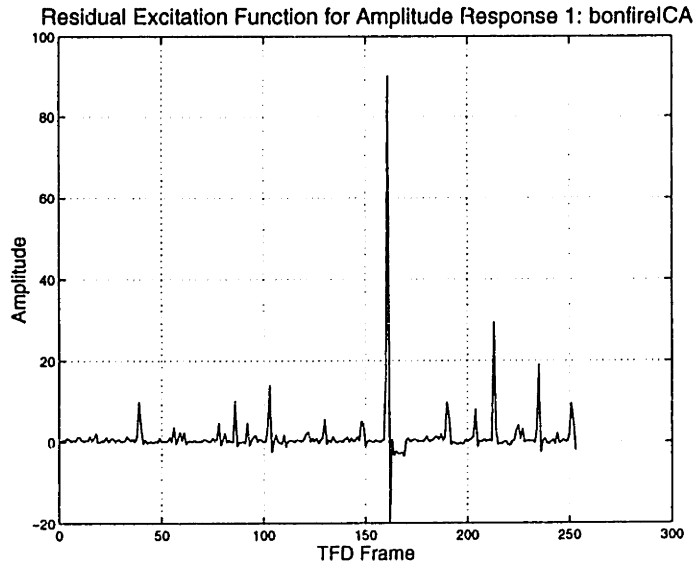


FIGURE 59. Excitation signal for first independent component amplitude function of the bonfire sound.

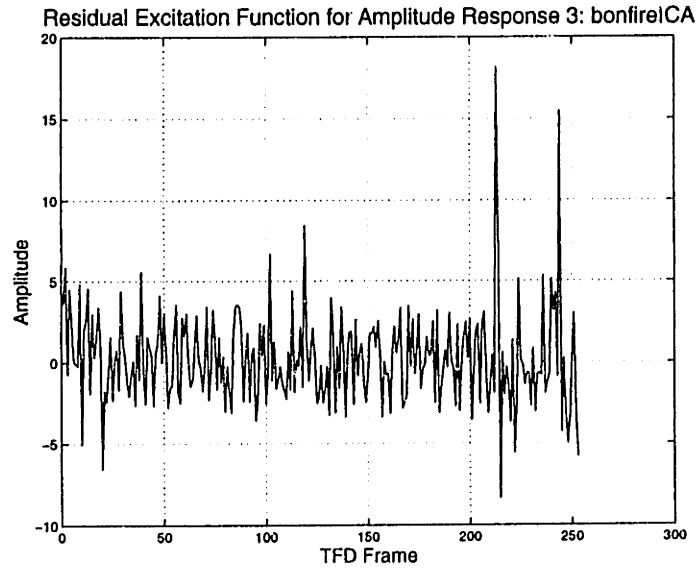


FIGURE 60. Excitation signal for third independent component amplitude function of the bonfire sound.

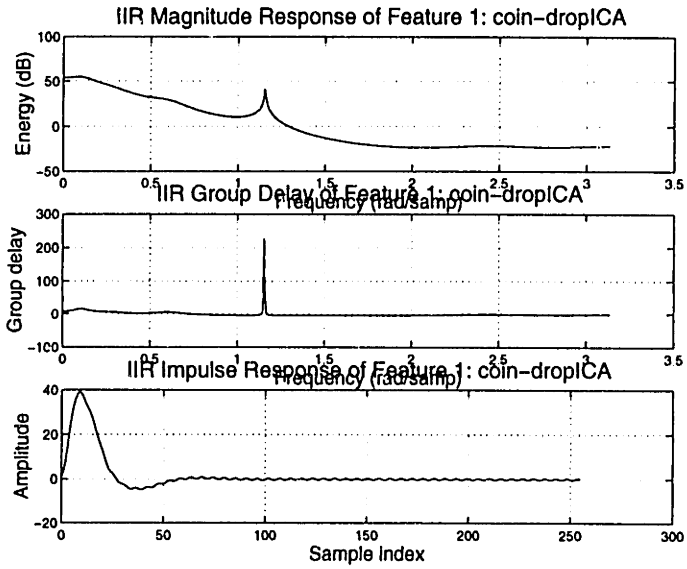


FIGURE 61. IIR Frequency Response, Group Delay and Impulse Response for the first independent component of the coin drop sound.

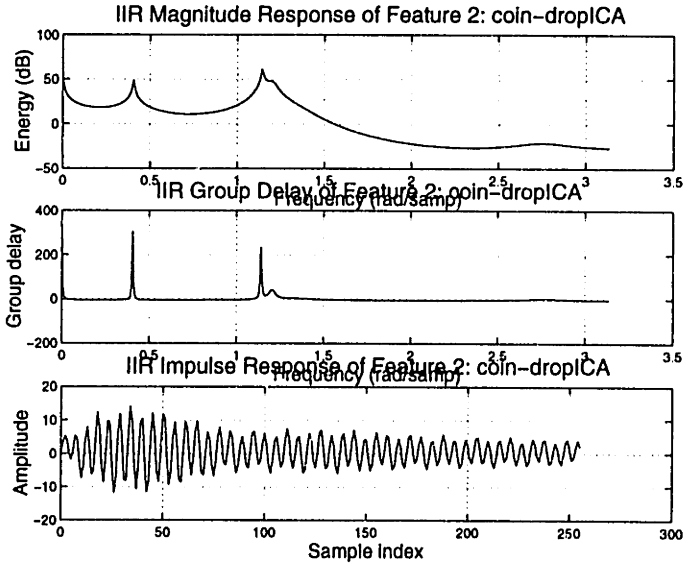


FIGURE 62. IIR Frequency Response, Group Delay and Impulse Response for the second independent component of the coin drop sound.

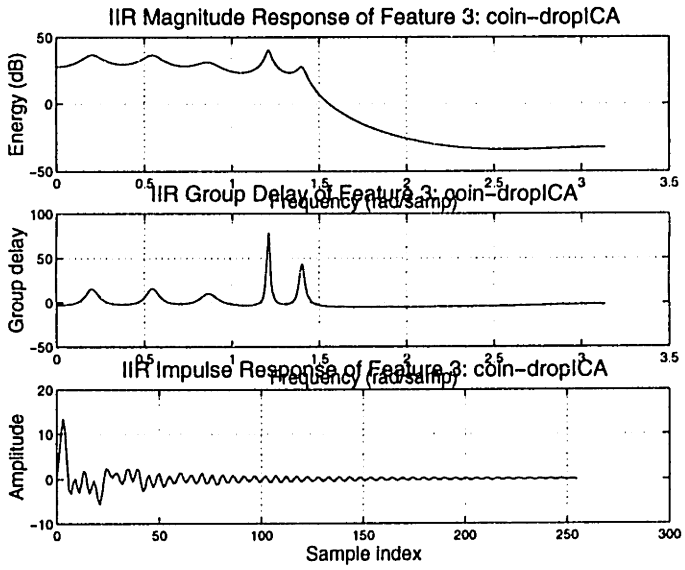


FIGURE 64. IIR Frequency Response, Group Delay and Impulse Response for the second independent component of the coin drop sound.

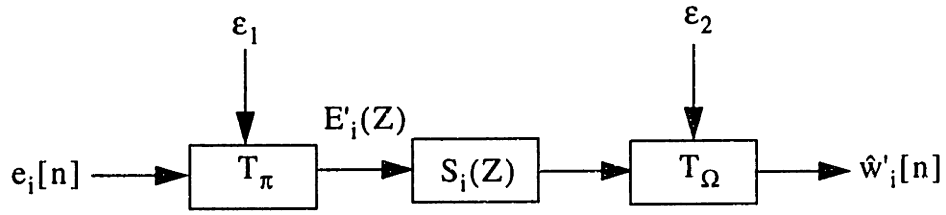


FIGURE 65. Implementation of structured independent component re-synthesis using an IIR system structure.

The second two figures show the excitation signals corresponding to the amplitude functions. We see a clear difference between the two illustrated components. The first shows an erratic impulsive behavior which is characteristic of crackling, and the second shows a continuous noise component that is characteristic of the bonfire sound.

4.3.4 IIR Modeling

The main problem with the FIR modeling technique is that it relies on a frame-rate much as the inverse short-time Fourier transform. In order to move away from the dependence on a frame rate we now consider the resynthesis of auditory invariants using IIR models. The method for IIR modeling of independent components starts with the MSTFTM signal for an independent component

$\hat{x}'_i[n]$. This signal is subjected to an LPC analysis using a higher-order than the LPC analysis used for FIR time-function modeling. The examples that we discuss in this section were obtained using a 12-th order LPC analysis. The LPC analysis yields a system function for each independent component as follows:

$$\chi_i(Z) = \frac{E_i(Z)}{S_i(Z)} \qquad [131]$$

where $\chi_i(Z)$ is the independent component sequence represented by its Z-transform, $E_i(Z)$ is an excitation sequence and $S_i(Z)$ is a spectral structure. This decomposition, for each of the ρ independent components gives us a convenient manner for extracting the components of a structured audio transform. Recall that the most general form of the structured audio transform was:

$$T_{structured}\{W\} = \sum_{i=1}^{\rho} T_{U^{(i)}_{\varepsilon 1_i}}\{E_i\} T_{V^{(i)}_{\varepsilon 2_i}}\{S_i\}. \qquad [132]$$

the explicit separation of the signals $E_i(Z)$ and $S_i(Z)$ in Equation 131 gives us the final form of our analysis. Not only are we able to extract a number of independent feature signals from a TFD, but we now also have a deconvolution of the excitation and spectral-structure components of each independent component. This allows us to implement transforms of the type specified by Equation 132, which are well-formed structured audio transforms. This is the re-synthesis framework that we adopt for structured-audio re-purposing and control of sound effects. The system flow diagram for structured audio resynthesis using IIR models is shown in Figure 65.

By way of example for the IIR independent component resynthesis method consider Figure 62 - Figure 64. These figures show the frequency response and impulse response for each of the three independent components of the coin drop sound. The first component is generally low-pass with a narrow-band spike component, from the impulse response we determine that this component is heavily damped and very lowpass. The second component has a longer time response which corresponds to the ringing of the coin, this is also manifest as high-Q regions in the frequency response. The third component is wide band and corresponds to the impact component of the coin bounces. These figures demonstrate that the IIR models capture the features of each independent component quite well. Thus the efficient form of independent component resynthesis does a good job of characterizing the structure of the original TFD.

4.3.5 Characterization of Excitation functions

From the examples given above, we propose that the excitation functions can be generalized into four broad classes of behavior: impacts, iterations, continuous noise and scatterings. Each of these types of excitation function can be approximated with a parameterized unit generator function. Table 9 lists four common excitation functions that we use for modeling a wide range sound feature behaviours.

TABLE 9. Excitation Function Generators and their Modeling Uses

Function	Signal Variable	Modeling Applications
Iteration	I[n]	Periodic, bouncing
Gaussian	G[n]	Scraping, blowing
Poission	P[n]	Scattering, impact, spilling
Chaotic	C[n]	Turbulence, jitter

For example, the unit generator function of iterations is an impulse train with exponentially spaced impulses. By setting the time constant of the exponential to decay we create an excitation signal suitable for modeling bouncing events. A constant spacing between impulses is useful for generating excitation signals for periodic events such as hammering and footsteps.

4.4 Auditory Group Synthesis Models

Having established methods for approximating the spectral features of an ICA analysis with IIR filters and characterizing the excitation structures using the generator functions shown in Table 9, we now are in a position to describe the architecture of general-purpose sound synthesis techniques that use features extracted from recordings in order to generate novel audio content.

Figure 66 shows the general form of an auditory group model. The $T_{e_i}\{E_i(Z)\}$ elements represent excitation signals and their auditory group transforms, and the $T_{s_i}\{S_i(Z)\}$ elements are the spectral structures and their auditory group transforms. The signal processing network defined in this way essentially implements Equation 132.

Assuming that the spectral structures are approximated using IIR filter models as described above we can implement the auditory group transforms of spectral structures using pole manipulation algorithms. In Chapter 2 we described the invariance properties of size changes and changes in the Young's modulus of materials by inspecting the physical equations governing the acoustics. Both of these transformations of a spectral feature can be implemented efficiently by appropriate manipulation of the roots of the denominator polynomial of each $S_i(Z)$.

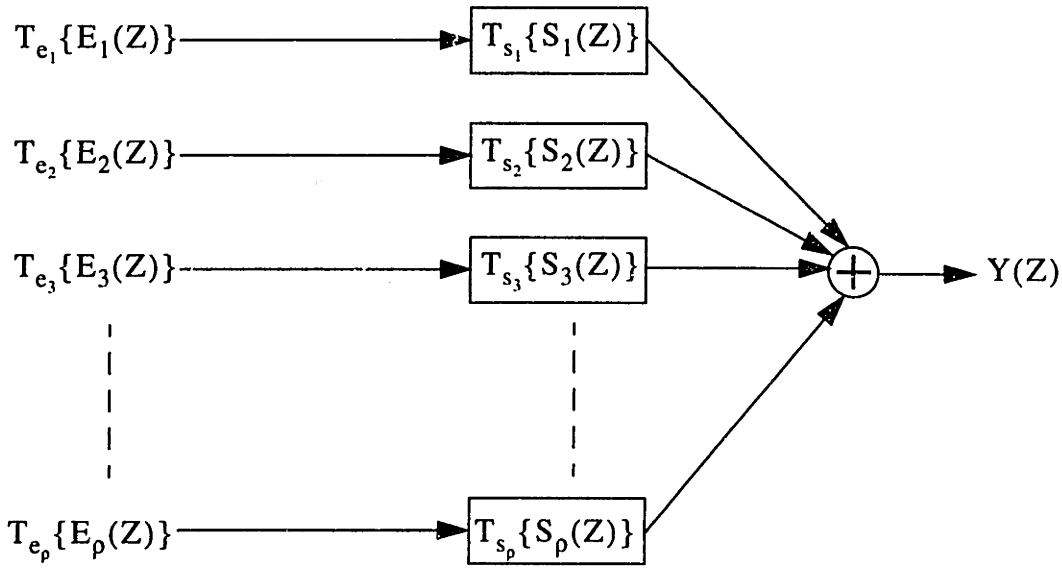


FIGURE 66. Schematic diagram of DSP implementation of multi-feature synthesis model with auditory group transforms T_e and T_s for each excitation signal and spectral structure respectively.

4.5 Future Directions

The system described thus far is capable of synthesizing a wide variety of sound types from single source/filter models to superpositions of source/filter models with time-varying behaviours. There are however some inherent limitations in the system as implemented thus far.

4.5.1 Orthogonality of ICA Transform

The first limitation of the current system is that the ICA transform is generated by an orthogonal rotation of the basis set from an SVD. Common (1994) points out that an ICA, if it exists for a given signal, is not constrained to be an orthogonal transform. Thus it should be possible to develop an algorithm that generates a non-orthogonal basis for an ICA. It is our expectation that improved separation performance for independent signal components will result from relaxing the constraint of basis orthogonality.

4.5.2 Weyl Correspondence and Transformational Invariant Tracking

Whilst we have shown that auditory group transforms can be used to specify physically-meaningful changes in an invariant, we have not fully demonstrated the possibility of tracking group transforms across a sound. For example, the sound of chirping birds would require a time-varying analysis that could separate the invariants from the frequency sweep transform generated by the chirp.

A number of recent developments in group-theoretic signal processing may be of importance to the correspondence and tracking problem. In particular, the time-frequency strip filters of Weisburn and Shenoy (1996) show promise as a method for tracking chirps in the time-frequency plane. Time-frequency strip filters are implemented as a special case of a *Weyl* filter, which relates the Weyl Correspondence to time-frequency analysis and thus provides group-theoretic origins for tracking transformational invariants.

4.5.3 On-Line Basis Estimation

A further requirement for successful tracking of invariants under transformation is that the basis estimation should utilize an on-line algorithm. The ICA algorithm that we developed in Chapter III serves our purposes well as long as the statistics of the input matrix are approximately constant across the matrix. For sounds with rapidly varying components, such as birds chirping, we require that the basis components be re-estimated for each frame of input, based on previous input. There are many such algorithms described in the ICA literature, for example Amari *et al.* (1996). A combination of on-line basis estimation and invariant tracking using time-frequency strip filters will allow greater accuracy in the analysis and characterization of complex everyday sound events.

4.6 Summary

In this chapter we first demonstrated that independent component basis vectors are better features of a sound than the corresponding singular value decomposition feature set, even though the two sets of basis vectors span exactly the same subspace of a time-frequency distribution. The independent components are used to reconstruct independent TFDs for each component of a sound. We gave methods for estimating a signal from the independent component TFDs based on an iterative procedure for minimizing phase errors.

These independent component signals can be used to further simplify the sound characterization by estimation of a set of filters using either FIR or IIR modeling techniques. The IIR modeling techniques were shown to be simpler in form than the FIR techniques but they are a little more computationally expensive. This extra expense, however, is eliminated when we consider the problem of phase modeling for FIR-based re-synthesis thus suggesting that IIR synthesis is a better model for implementing efficient resynthesis of independent components. The IIR filter model explicitly break each independent component signal into an excitation function and a spectral structure thus the combination of ICA analysis and IIR modeling results in a multi-source blind deconvolution of the latent statistical components in a TFD. This signal model is more ambitious

than most that are commonly used for audio signal processing algorithms and proves extremely powerful for audio re-purposing and control. The structure of the IIR resynthesis scheme was shown to be analogous to that of a well-formed auditory group transform thus satisfying all the conditions of a structured audio transform. To date there have been no audio signal processing methods capable of multi-source blind deconvolution and this technique may prove useful for application areas other than those presented.

In order to control sounds for structured re-purposing we presented a small collection of excitation modeling functions whose signals are representative of a wide range of natural sound behaviors. It was shown that a combination of these functions can be used to generate many different sound instances from a single input matrix representation. This excitation matrix is subject to control by auditory group transforms for generating novel features in sound structures. We also discussed transformation techniques for spectral-structure components that are used for modeling physical object properties such as size and materials.

The goal of this thesis was, at the outset, to find a method for representing and controlling natural sounds by their structured content. In this chapter we have demonstrated techniques for synthesizing and controlling the independent features of a sound in the desired manner. Our conclusion then, is that the said methodologies for representing and synthesizing natural sounds comprise a good working collection of tools with which to carry out the desired transforms.

Appendix I: Local Lie Group Representations

1.1 Definition of Invariants

Let us consider the problem by abalysing a system transform with specified invariance properties. We seek a system that is described by a Lagrangian functional:

$$J(x) = \int_a^b L(t, x, \dot{x}) \ , \qquad [133]$$

where L is the Lagrangian which is integrated in order to obtain the output function $J(x)$. It should be noticed that, along with the specifiable variables t and x, the Lagrangian also specifies a function in terms of the derivative signal $\dot{x}$.

We define a transformation, T as a mapping $T : R^2 \rightarrow R^2$, which maps a point (t, x) into a point (t', x') by:

$$T : t' = \phi(t, x), \qquad x' = \psi(t, x) \qquad [134]$$

where $\phi(t, x)$ and $\psi(t, x)$ are specified transformation functions. Composition of transformations is represented by:

$$S : t' = \gamma(t, x), \qquad x' = \omega(t, x) \qquad [135]$$

then

$$ST : t'' = \gamma(t', x'), \qquad x'' = \omega(t', x') \qquad [136]$$

and by substitution from Equation 135 we arrive at the composition functional:

$$ST : t'' = \gamma(\phi(t, x), \omega(t, x)), \qquad x'' = \omega(\phi(t, x), \omega(t, x)) \ . \qquad [137]$$

In order for a transformation to belong to the group it must have a corresponding inverse transform within some bounded region of the transform:

$$T^{-1} : t = \Phi(t', x'), \qquad x = \Psi(t', x') \quad . \tag{138}$$

Another necessary component of a group is the identity element which is represented by the compositions $T^{-1}T$ and TT^{-1}. We represent the identity element with the notation T_0.

Auditory group transformations are dependent upon a real parameter ε, for all ε in an open interval $|\varepsilon| < \varepsilon_0$. We can define this family one parameter transformations in the following manner:

$$T_\varepsilon : t' = \phi(t, x, \varepsilon), \qquad x' = \psi(t, x, \varepsilon) \quad . \tag{139}$$

As well as the existence of inverse and identity transforms for the functionals we also want the transforms to exhibit the *local closure property.* This property states that for ε_1 and ε_2 sufficiently small, there exists an ε_3 such that:

$$T_{\varepsilon_1} T_{\varepsilon_2} = T_{\varepsilon_3} \quad . \tag{140}$$

A one-parameter family of transformations that satisfies the above requirements of inverse, identity and closure is called a *local Lie group*, [Logan87] [Moon95] [Gilmore74] [Bourbaki75].

1.2 Transformations of points

Consider the transform:

$$T_\varepsilon : t' = t\cos\varepsilon - x\sin\varepsilon, \qquad x' = t\sin\varepsilon + x\cos\varepsilon \tag{141}$$

The composition of these functions has the form $T_{\varepsilon_1} T_{\varepsilon_2} = T_{\varepsilon_1 + \varepsilon_2}$, thus the parameters sum under composition. The identity transform is $T_0 = I$ and $T_\varepsilon^{-1} = T_{-\varepsilon}$. By these four properties the above transformation forms a local Lie group, the *rotation group.*

A Taylor series expansion of Equation 141 about $\varepsilon = 0$ yeilds the following form for T_ε :

$$t' = \phi(t, x, 0) + \phi_\varepsilon(t, x, 0)\varepsilon + \frac{1}{2}\phi_{\varepsilon\varepsilon}(t, x, 0)\varepsilon^2 + \ldots \qquad [142]$$

$$= t + \tau(t, x)\varepsilon + o(\varepsilon)$$

$$x' = \psi(t, x, 0) + \psi_\varepsilon(t, x, 0)\varepsilon + \frac{1}{2}\psi_{\varepsilon\varepsilon}(t, x, 0)\varepsilon^2 + \ldots$$

$$= x + \xi(t, x)\varepsilon + o(\varepsilon)$$

where $o(\varepsilon) \to 0$ faster than $\varepsilon \to 0$. The function subscripts ϕ_ε and ψ_ε denote a partial derivative, e.g. $\phi_x = \frac{\partial \phi}{\partial x}$. This representation constitutes a global representation of the local Lie group T_ε. The quantities τ and ξ are the *generators* of T_ε and they are defined as the partial derivatives of the component functions of the transformation:

$$\tau(t, x) = \phi_\varepsilon(t, x, 0), \qquad \xi(t, x) = \psi_\varepsilon(t, x, 0) \quad . \qquad [143]$$

The generators are also used to obtain an *infinitessimal representation,* or local representation, for the local Lie group obtained for small ε:

$$t' = t + \varepsilon\tau(t, x) + o(\varepsilon), \qquad [144]$$

$$x' = x + \varepsilon\xi(t, x) + o(\varepsilon)$$

For linear transforms the infinitessimal representation can be used to specify the global representation. This is not true for non-linear transformations for which the generators represent a local-linear transform of the vector field of the transformation for a small region of ε.

The rotation group transform, given by Equation 141, is a linear transform. We define the properties of a linear transform in Section ??? where we considered the class of normal subgroups. We can obtain the generators for the rotation group by solving the partial derivatives for the component functions, ϕ_ε and ψ_ε, evaluating at $\varepsilon = 0$ gives:

$$\tau(t, x) = \frac{\partial}{\partial\varepsilon}\phi(t, x, \varepsilon)\Big|_{\varepsilon = 0} = -x \qquad [145]$$

and

$$\xi(t, x) = \frac{\partial}{\partial\varepsilon}\psi(t, x, \varepsilon)\Big|_{\varepsilon = 0} = t \qquad [146]$$

substituting into the infinitessimal representation, Equation 144, we obtain the local representation of the rotation group transform:

$$t' = t - \varepsilon x + o(\varepsilon), \qquad [147]$$

$$x' = x + \varepsilon t + o(\varepsilon)$$

the generators of which are $\tau(t, x) = x$ and $\xi(t, x) = t$ thus specifying the global representation:

$$t' = t - x\varepsilon + o(\varepsilon), \qquad [148]$$

$$x' = x + t\varepsilon + o(\varepsilon)$$

We shall see the importance of the rotation group later when we consider a general class of transforms that operates in the time-frequency plane as time shifts and time-frequency rotations. This class of transforms will be used in the following chapters to characterize particular classes of sound structure invariant.

1.3 Transformations of functions

We now consider the effects of transforming a function by a local Lie group. Let $x = h(t)$, where $h(t) \in C$, the set of complex numbers. Under the local Lie group transformation T_ε produces a mapping from $x = h(t)$ to $x' = h'(t')$. We find the form of h' by noting that T_ε maps t to $t' = \phi(t, h(t), \varepsilon)$. Now, for ε sufficiently small there exists an inverse mapping of $\phi(t, h(t), \varepsilon)$, denoted by K, such that $t = K(t', \varepsilon)$. Then under T_ε:

$$x' = \psi(t, h(t), \varepsilon) = \psi(K(t', \varepsilon), h(K(t', \varepsilon)), \varepsilon) \equiv h'(t') \quad , \qquad [149]$$

where r t' and x' are the transformations of the Lagrangian $L(t, x, \dot{x})$. In order to fully specify the behaviour of the function under transformation we must also determine how derivatives of functions are transformed under T_ε:

By the chain rule:

$$\frac{d}{dt'}h' = \psi_t \frac{dt}{dt} + \psi_x \frac{dh}{dt}\frac{dt}{dt'} = (\psi_t + \psi_x \dot{h})\frac{dt}{dt'} \qquad [150]$$

where ψ_t and ψ_x are evaluated at $(K(t', \varepsilon), h(K(t', \varepsilon)), \varepsilon)$. Using this derivative we arrive at an extended group of transforms which represent the effect of the transform T_ε upon the Lagrangian $L(t, x, \dot{x})$:

$$t' = \phi(t, x, \varepsilon), x' = \psi(t, x, \varepsilon), \dot{x}' = \frac{\psi_t(t, x, \varepsilon) + \psi_x(t, x, \varepsilon)\dot{x}}{\phi_t(t, x, \varepsilon) + \phi_x(t, x, \varepsilon)\dot{x}} \quad [151]$$

The generators for the derivative element of the extended group is derived by the global representation method, described above, and it is given by:

$$\dot{x}' = \dot{x} + (\dot{\xi} - \dot{x}\dot{\tau})\varepsilon + o(\varepsilon) \quad [152]$$

the linear generator η is thus given by:

$$\eta = \dot{\xi} - \dot{x}\dot{\tau} \quad [153]$$

which extends the generators previously given in Equation 144.

Appendix II: Derivation of Principal Component Analysis

2.1 Eigenvectors of the Covariance Matrix Derivation

2.1.1 Principal Component Feature Extraction

Consider a random vector $\mathbf{x} \in \Re^n$ for which a basis $\mathbf{u}$ is to be determined such that an approximation of $\mathbf{x}$ can be obtained by a linear combination of m orthogonal basis vectors:

$$\hat{\mathbf{x}} = y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + \ldots + y_m\mathbf{u}_m \qquad [154]$$

where $m < n$ and y_j is the weighting coefficient of basis vector $\mathbf{u}_j$ formed by taking the inner product of $\mathbf{x}$ with $\mathbf{u}_j$:

$$y_j = \mathbf{x}^T\mathbf{u}_j \,. \qquad [155]$$

By forming a $k \times n$ observation matrix $\mathbf{X}$, the rows of which are the observations $\mathbf{x}_k^T$, we can express Equation 154 and Equation 155 respectively in the following form:

$$\hat{\mathbf{X}} = \mathbf{Y}\mathbf{U}_m^T \qquad [156]$$

$$\mathbf{Y}_m = \mathbf{X}\mathbf{U}_m \qquad [157]$$

where $\mathbf{U}_m$ is an $n \times m$ matrix whose columns are an uncorrelated basis for $\mathbf{X}$, and $\mathbf{Y}_m$ is a $k \times m$ matrix whose columns are the coefficients for each column vector in $\mathbf{U}_m$. By the orthog-

onality of $\mathbf{U}_m$ and by the imposition of an additional constraint that the columns of $\mathbf{U}_m$ have unit norm, i.e. $\|\mathbf{U}_j\| = 1$, it follows that:

$$\mathbf{U}_m^T \mathbf{U}_m = \mathbf{I}_m \qquad [158]$$

where $\mathbf{I}_m$ is a $m \times m$ diagonal matrix with unit entries. We refer to Equation 156 as the PCA-feature re-synthesis equation and Equation 157 as the PCA-feature projection equation where the columns of $\mathbf{Y}_m$ correspond to the estimated features and the columns of $\mathbf{U}_m$ are the projection vectors which perform the linear transformation of the input to the new uncorrelated basis.

The problem, then, for deriving a PCA is to obtain the matrix $\mathbf{U}_m$ such that the residual error in approximating $\mathbf{x}$ with $\hat{\mathbf{x}}$ is minimized:

$$\varepsilon = \mathbf{x} - \hat{\mathbf{x}} = \sum_{j=m+1}^{n} y_j \mathbf{u}_j \qquad [159]$$

that is, the expansion of all the unused features results in a minimal signal which is the residual error ε. A suitable criteria is the minimization of the expectation of the mean-square residual error:

$$\xi = E[|\varepsilon|^2] = E[|\mathbf{x} - \hat{\mathbf{x}}|^2] \qquad [160]$$

where the expectation of an arbitrary function of $\mathbf{x}$, say $\psi(\mathbf{x})$, is defined as the element-wise operation:

$$E[\psi(\mathbf{x})] = \int_{-\infty}^{\infty} \psi(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \qquad [161]$$

where $p_{\mathbf{x}}(\mathbf{x})$ is the probability density function of the ran m variable $\mathbf{x}$. Since the expectation operator is linear and due to the condition that the columns of $\mathbf{U}_m$ are orthonormal it follows that:

$$\zeta = E[\varepsilon^T \varepsilon] = E\left[\left(\sum_{i=m+1}^{n} y_i \mathbf{u}_i^T\right)\left(\sum_{j=m+1}^{n} y_j \mathbf{u}_j\right)\right] = \sum_{j=m+1}^{n} E[y_j^2] \quad [162]$$

from which it follows that:

$$E[y_j^2] = E[(\mathbf{u}_j^T \mathbf{x})(\mathbf{x}^T \mathbf{u}_j)] = \mathbf{u}_j^T E[\mathbf{x}\mathbf{x}^T]\mathbf{u}_j = \mathbf{u}_j^T \mathbf{R}\mathbf{u}_j \quad [163]$$

where $\mathbf{R}$ is the correlation matrix for $\mathbf{x}$. Now by substitution of Equation 163 into Equation 162 we arrive at the quantity to be minimized:

$$\zeta = \sum_{j=m+1}^{n} \mathbf{u}_j^T \mathbf{R}\mathbf{u}_j \quad [164]$$

We can express the minimization as a multi-variate differential equation by using a set of Lagrangian multipliers λ_j and setting the derivative of the expectation of the mean-square error with respect to the basis components $\mathbf{u}_j$ to zero. This is a necessary condition for the minimum and gives the characteristic equation:

$$\frac{\partial}{\partial \mathbf{u}_j}\zeta' = 2(\mathbf{R}\mathbf{u}_j - \lambda j \mathbf{u}_j) = 0, j = m+1, \ldots, n \quad . \quad [165]$$

It is well known that the solutions to this equation constitute the eigenvectors of the correlation matrix $\mathbf{R}$. It is also worth noting that the correlation matrix is related to the covariance matrix by the following expression:

$$\mathbf{Q}_x = \mathbf{R}_x - \mathbf{m}\mathbf{m}^T \quad . \quad [166]$$

Thus, for zero-mean or *centered* data, the problem is equivalent to finding the eigenvectors of the covariance matrix $\mathbf{Q}_x$. Since the columns of $\mathbf{U}$ are now determined to be the eigenvectors, we can re-express the residual error as the sum of the eigenvalues of the unused portion of the basis:

$$\zeta = \sum_{j=m+1}^{n} \lambda_j \quad [167]$$

and the solution to the minimization reduces to ordering the basis vectors $\mathbf{u}_j$ such that the columns with the smallest eigenvalues occur in the unused portion of the basis which also implies that the m columns of $\mathbf{U}_m$ which *are* used for reconstruction should comprise the eigenvectors with the m largest eigenvalues.

Now that we have arrived at the form of the solution for optimal orthonormal basis reconstruction (in the square-error sense) we must find a general form for representing the solution. Since the eigenvalues form the diagonal elements of the covariance of the transformed data with all other elements equal to zero we can express the solution to the eigenvalue decomposition as a diagonalization of the input covariance matrix $\mathbf{Q}_x$.

Bibliography

Abbas, H.M. and Fahmy, M.M. (1993). "Neural model for Karhunen-Loeve transform with application to adaptive image compression", *IEE Proceedings*-I, Vol 140, No. 2, April, pp. 135-143.

Abe, M. and Ando, S. (1995). "Nonlinear time-frequency domain operators for decomposing sounds into loudness, pitch and timbre". *International Conference on Acoustics, Speech, and Signal Processing*, pp 1368–71 vol.2., *IEEE*; New York, NY, USA.

Agajan, S. and Egiazarjan, K. (1987). "A new class of transformations in the theory of processing discrete signals", volume M. Tud. Akad. Szam.tech. Autom. Kut. Intez. (Hungary). *Problems of Computer Science.*

Allen, J. B. (1977). "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Time Fourier Transform.", *IEEE Acoustics, Speech and Signal Processing*, 25 (3), pp. 235-238.

Allen, J. B., & Rabiner, L. R. (1977). "A Unified Approach to Short-Time Fourier Analysis and Synthesis", *Proceedings of the IEEE*, 65 (11), pp. 1558 - 1564.

Amari, S., Cichocki, A., and Yang, H. H. (1996). "A New Learning Algorithm for Blind Signal Separation", In *Advances in Neural Information Processing Systems 8*, Editors D. Touretzky, M. Mozer, and M. Hasselmo, MIT Press, Cambridge MA.

Anderson, N. and Karasalo, I. (1975). "On computing bounds for the least singular value of a triangular matrix". *BIT*, 15:1–4.

Appleton, J.H. (1989). *21st-century musical instruments : hardware and software.* City University of New York, Brooklyn, New York.

Association, A.S. (1960). *Acoustical Terminology.* American Standards Institute, New York.

Aware Inc. (1993). *Speed of Sound Megadisk CD-ROM#1: Sound Effects.* Computer CD-ROM.

Balzano, G. (1980). "The group-theoretic description of 12-fold and microtonal pitch systems". *Computer Music Journal*, 4(4):66–84.

Baraniuk, R. (1995). "Marginals vs. covariance in joint distribution theory". *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1021–4 vol.2. , IEEE; New York, NY, USA.

Baraniuk, R. (1996). "Covariant time-frequency representations through unitary equivalence". *IEEE Signal Processing Letters*, 3(3):79–81.

Barlow, H. (1989). "Unsupervised Learning", *Neural Computation*, 1, pp. 295-311.

Barlow, J.L., Yoon, P.A., and Zha, H. (1996). "An algorithm and a stability theory for downdating the ULV decomposition". *BIT*, 36:14–40.

Barlow, J.L., Zha, H., and Yoon, P.A. (1993). "Stable chasing algorithms for modifying complete and partial singular value decompositions". Tech. Report CSE-93-19, Department of Computer Science, The Pennsylvania State University, State College, PA.

Barner, K. and Arce, G. (1997). "Design of permutation order statistic filters through group colorings". *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 44(7):531–48.

Barth, W., Martin, R.S., and Wilkinson, J.H. (1967). "Calculation of the eigenvalues of a symmetric tridiagonal matrix by the method of bisection". *Numer. Math.*, 9:386–393.

Bell A.J. & Sejnowski T.J. (1995). "An information-maximization approach to blind separation and blind deconvolution", *Neural Computation*, 7, 1129-1159.

Bell A.J. & Sejnowski T.J. (1995). "Fast blind separation based on information theory", in *Proc. Intern. Symp. on Nonlinear Theory and Applications*, vol. 1, 43-47, Las Vegas.

Bell A.J. & Sejnowski T.J. (1996). "Learning the higher-order structure of a natural sound", *Network: Computation in Neural Systems*.

Berman, S. and Grushko, I. (1983). "The theory of discrete signal processing". *Problemy Peredachi Informatsii*, 19(4):43–9.

Berry, M.W. (1992a). "A Fortran-m 77 software library for the sparse singular value decomposition". Tech. Report CS-92-159, University of Tennessee, Knoxville, TN.

Berry, M.W. (1992b). "Large scale sparse singular value computations". *Internat. J. Supercomp. Appl.*, 6:13–49.

Berry, M.W. (1993). "SVDPACKC: Version m 1.0 user's guide". Tech. Report CS-93-194, University of Tennessee, Knoxville, TN.

Berry, M.W. and Auerbach, R.L. (1994). "A block Lanczos SVD method with adaptive reorthogonalization". In Brown, J.D., Chu, M.T., Ellison, D.C., and Plemmons, R.J., editors, *Proceedings of the Cornelius Lanczos International Centenary Conference, Raleigh, NC, Dec. 1993*, pages 329–331. SIAM, Philadelphia.

Berry, M.W. and Golub, G.H. (1991). "Estimating the largest singular values of large sparse matrices via modified moments". *Numer. Algorithms*, 1:363–374.

Beyerbach, D. & Nawab, H. (1991). "Principal Components Analysis of the Short-Time Fourier Transform", in *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, "Digital Signal Processing", pp. 1725 - 1728.

Björck, Å. and Bowie, C. (1971). "An iterative algorithm for computing the best estimate of an orthogonal matrix". *SIAM J. Numer. Anal.*, 8:358–364.

Bluman, G. W., & Cole, J. D. (1974). *Applied Mathematical Sciences: No 13, Similarity Methods for Differential Equations.* New York: Springer-Verlag.

Bogert, B. P., Healy, M. J., and Tukey, J. W. (1963). "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Saphe Cracking." *Proceedings of the Symposium on Time Series Analysis*, M. Rosenblatt, (Ed.), New York: Wiley.

Boulanger, R. (1985). *The Transformation of Speech Sounds into Music using Spectral Intersection Synthesis.* PhD

thesis, UCSD, CARL.

Bregman, A. (1990). *Auditory Scene Analysis.* MIT Press, Cambridge, Mass.

Brown, G. (1992). *Computational auditory scene analysis: a representational approach.* PhD thesis, University of Sheffield.

Bunch, J.R. and Nielsen, C.P. (1978). "Updating the singular value decomposition". *Numer. Math.*, 31:111–129.

Businger, P. (1970). "Updating a singular value decomposition". *BIT*, 10:376–385.

Cardoso, J.-F. (1989). "Blind identification of independent components with higher-order statistics". In *Proc. Workshop on Higher-Order Spect. Anal., Vail, Colorado*, pages 157–160.

Cardoso, J.-F. (1990). "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem". In *Proc. ICASSP*, pages 2655–2658.

Cardoso, J.-F. and Comon, P. (1990). "Tensor based independent component analysis", In *Proc. EUSIPCO.*

Cardoso, J. F. (1995). "The equivariant approach to source separation", In *Proc. NOLTA*, pages 55-60.

Cardoso, J. F. (1995). "A tetradic decomposition of 4th-order tensors: application to the source separation problem", In M. Moonen and B. de Moor, editors, *Algorithms, architectures and applications*, volume III of *SVD and signal processing*, pages 375-382. Elsevier.

Casey, M. (1993). "Non-linear estimation of audio synthesis control parameters", In *Proceedings of the Internation Computer Music Conference*, Tokyo. ICMA.

Casey, M. (1994). "Understanding musical sound with forward models and physical models", *Connection Science*, 6(2&3):355–371.

Casey, M. (1996). "Multi-model estimation as a basis for computational timbre understanding", In *International Conference on Music Perception and Cognition*, Montreal.

Casey, M. & Smaragdis, P. (1996). "NetSound", *Proceedings of the International Computer Music Conference*, ICMA, Hong Kong.

Cassirer, E. (1944). "The Concept of Group and the Theory of Perception." *Philosophy and Phenomenological Research.* Vol. V (1), pp. 1-35.

Chan, T.F. (1982a). "Algorithm m 581: An improved algorithm for computing the singular value decomposition". *ACM Trans. Math. Software*, 8:84–88.

Chan, T.F. (1982b). "An improved algorithm for computing the singular value decomposition". *ACM Trans. Math. Software*, 8:72–83.

Chan, T.F. and Hansen, P.C. (1990). "Computing truncated SVD least squares solutions by rank revealing QR factorizations". *SIAM J. Sci. Statist. Comput.*, 11:519–530.

Chandrasekaran, S. and Ipsen, I. C.F. (1992). "Analysis of a QR algorithm for computing singular values". Tech. Report YALEU/DCS/RR-917, Yale University, New Haven, CT.

Chandrasekaran, S. and Ipsen, I. C.F. (1994). "Backward errors for eigenvalue and singular value decompositions". *Numer. Math.*, 68:215–223.

Charlier, J., Vanbegin, M., and Van Dooren, P. (1988). "On efficient implementations of Kogbetliantz's algorithm for computing the singular value decomposition". *Numer. Math.*, 52:279–300.

Charnley, T. and Perrin, R. (1978). "Studies with an eccentric bell". *Journal of Sound and Vibration*, 58(4):517–25.

Chemillier, M. (1987). "Free monoid and music". ii. *Informatique Theorique et Applications*, 21(4):379– 417.

Comon, P. (1992). "MA identification using fourth order cumulants". *Signal Processing, Eurasip*, 26(3):381–388.

Comon, P. (1994). "Independent Component Analysis, a new concept?", *Signal Processing, Elsevier*, 36(3):287–314. Special issue on Higher-Order Statistics.

Comon, P., Jutten, C., and Herault, J. (1991). "Separation of sources, part II: Problems statement". *Signal Processing*, 24(1):11–20.

Cooke, M. (1991). *Modeling auditory processing and organization*. PhD thesis, University of Sheffield.

Cosi, P., DePoli, G., and Lauzzana, G. (1994). "Timbre classification by nn and auditory modeling". *Proceedings of International Conference on Artificial Neural Networks*, volume 2 vol. xvi+xiii+1482, pages 925–8 vol.2. , Springer-Verlag; Berlin, Germany.

Courtot, F. (1991). "Representation and induction of musical structures for computer assisted composition". *European Working Session on Learning Proceedings*. Springer-Verlag; Berlin, Germany.

Cremer, L. (1984). *The Physics of the Violin*. Cambridge, MA: MIT Press.

Crummer, G., Walton, J., Wayman, J., Hantz, E., and Frisina, R. (1994). "Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch". *Journal of the Acoustical Society of America*, 95(5).

Cullum, J.K., Willoughby, R.A., and Lake, M. (1983). "A Lanczos algorithm for computing singular values and vectors of large matrices". *SIAM J. Sci. Statist. Comput.*, 4:197–215.

Culver, C. (1956). *musical acoustics*. McGraw Hill, New York.

Deco, G., and Obradovic, D. (1996). *An Information-Theoretic Approach to Neural Computing*. New York: Springer-Verlag.

Delprat, N. and Kronland-Martinet, R. (1990). "Parameters estimation for nonlinear resynthesis methods with the help of a time-frequency analysis of natural sounds". *Proceedings. Sound Control;* Personal Assistance; Yamaha-Kemble Music(UK); et al, ICMC; Glasgow, UK.

Demmel, J. and Kahan, W. (1990). "Accurate singular values of bidiagonal matrices". *SIAM J. Sci. Statist. Comput.*, 11:873–912.

Depalle, P., Garcia, G., and Rodet, X. (1993). "Tracking of partials for additive sound synthesis using hidden markov models". *Proceedings of ICASSP '93*, pages 225–8 vol.1., IEEE; New York, NY, USA.

Depalle, P., Garcia, G., and Rodet, X. (1995). "The recreation of a castrato voice, farinelli's voice". *Proceedings of 1995 Workshop on Applications of Single Processing to Audio and Accoustics*, IEEE; New York, NY, USA.

DeWitt, L. and Crowder, R. (1987). "Tonal fusion of consonant musical intervals: The oomph in stumpf". *Perception and Psychophysics*, 41:73–84.

Dolson, M. (1986). "The phase vocoder: A tutorial". *Computer Music Journal*, 10(4).

Dongarra, J., Bunch, J.R., Moler, C.B., and Stewart, G.W. (1979). *LINPACK Users' Guide*. SIAM, Philadelphia.

Drmavc, Z. (1997). "Implementation of Jacobi rotations for accurate singular value computation in floating point arithmetic". *SIAM J. Sci. Comput.*, 18.

Eisenstat, S.C. and Ipsen, I.C.F. (1993). "Relative perturbation techniques for singular value problems". Tech. Report YALEU/DCS/RR-942, Yale University, New Haven, CT.

Ellis, D. (1995). "Underconstrained stochastic representations for top-down computational auditory scene analysis". *Proceedings of 1995 Workshop on Applications of Single Processing to Audio and Accoustics*, Number 284., IEEE; New York, NY, USA.

Ellis, D. (1996). *Prediction-Driven Computation Auditory Scene Analysis*. PhD thesis, MIT.

Feichtinger, H., Strohmer, T., and Christensen, O. (1995). "Group theoretical approach to gabor analysis". *Optical Engineering*, 34(6):1697–704.

Feiten, B. and Gunzel, S. (1994). "Automatic indexing of a sound database using self-organizing neural nets". *Computer Music Journal*, 18(3):53–65.

Fernando, K.V. (1989). "Linear convergence of the row cyclic Jacobi and Kogbetliantz methods". *Numer. Math.*, 56:73–91.

Fernando, K.V. and Parlett, B.N. (1994). "Accurate singular values and differential qd algorithms". *Numer. Math.*, 67:191–229.

Flanagan, J. and Rabiner, L. (1966). "Phase vocoder". *Bell System Technical Journal*, 45:1493–1509.

Fletcher, N. H., & Rossing, T. D. (1991). *The Physics of Musical Instruments*. New York: Springer-Verlag.

Freed, D. J. (1990). "Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events." *Journal of the Acoustical Society of America*, 87(1), 311-322.

Freeman, W. T., and Tenenbaum, J. B. (1997). "Learning bilinear models for two-factor problems in vision" , *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97)*, Puerto Rico, U. S. A., June.

French, A. P. (1971). *Vibrations and waves*. New York: Norton.

French, A. P. (1975). *Newtonian mechanics*. New York:Norton.

Gaeta, M., & Lacoume, J. L. (1990). "Source separation without *a-priori* knowledge: The maximum likelihood solution", in Torres, Masgrau and Lagunas, (Eds.), *Proc. EUSIPCO Conf.*, Barcelona, Elsevier.

Ganesan, K., Marlot, M., and Mehta, P. (1986). "An efficient algorithm for combining vector quantization and stochastic modeling for speaker-independent speech recognition". , *Inst. Electron. & Commun. Eng. Japan; Acoust. Soc. Japan*, pp. 1069–71 vol.2. IEEE; New York, NY, USA.

Gaver, W. W. (1988). *Everyday Listening and Auditory Icons*. Ph.D. Dissertation, University of California, San Diego.

Gaver, W. W. (1993). "What in the World Do We Hear? An Ecological Approach to Auditory Source Perception." *Ecol. Psych.* (5)1

Gaver, W. W. (1994). "Using and Creating Auditory Icons." In *Auditory Display:Sonification, Audification, and Auditory Interfaces*, edited by G. Kramer. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVIII. Reading, MA: Addison Wesley.

George, E.B. and Smith, M.J. (1992). "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones". *Journal of the Audio Engineering Society*, 40(6):497–516.

Gerth, J. (1993). "Identification of sounds with multiple timbres". *Proceedings of 37th Annual Meeting on the Human Factors and Ergonomics Society, Human Factors & Ergonomics Soc.* vol.1. ; Santa Monica, CA, USA.

Giannakis, G., Inouye, Y., and Mendel, J. M. (1989). "Cumulant-based identification of multichannel moving average models", *IEEE Automatic Control*, Vol. 34, July, pp. 783-787.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems.* Boston:Houghton Mifflin.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston:Houghton Mifflin.

Golub, G.H. (1968). "Least squares, singular values and matrix approximations". *Aplikace Matematiky*, 13:44–51.

Golub, G.H. and Kahan, W. (1965). "Calculating the singular values and pseudo-inverse of a matrix". *SIAM J. Numer. Anal. Ser. B*, 2:205–224.

Golub, G.H. and Luk, F.T. (1977). "Singular value decomposition: Applications and computations". In *Trans. 22nd Conference of Army Mathematicians, ARO Report 77-1*, pages 577–605.

Golub, G.H., Luk, F.T., and Overton, M.L. (1981). "A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix". *ACM Trans. Math. Software*, 7:149–169.

Golub, G.H. and Reinsch, C. (1970). "Singular value decomposition and least squares solution". *Numer. Math.*, 14:403–420.

Golub, G.H., Solna, K., and Van Dooren, P. (1995). "A QR-like SVD algorithm for a product/quotient of several matrices". In Moonen, M. and De Moor, B., editors, *SVD and Signal Processing, III: Algorithms, Architectures and Applications*, pages 139–147. Elsevier Science B.V., Amsterdam.

Golub, G.H. and Van Loan, C.F. (1979). "Total least squares". In Gasser, T. and Rosenblatt, M., editors, *Smoothing Techniques for Curve Estimation*, pages 69–76. Springer-Verlag, New York.

Golub, G.H. and Van Loan, C.F. (1980). "An analysis of the total least squares problem". *SIAM J. Numer. Anal.*, 17:883–893.

Grey, J. (1975). *An Exploration of Musical Timbre.* PhD thesis, Stanford University Psychology Department.

Grey, J. (1977). "Multidimensional perceptual scaling of musical timbres". *Journal of the Acoustical Society of America*, 61(5):1270–7.

Grey, J. and Moorer, J. (1977). "Perceptual evaluations of synthesized musical instrument tones". *Journal of the Acoustical Society of America*, 62(2):454–62.

Grey, J. (1978). "Timbre discrimination in musical patterns". *Journal of the Acoustical Society of America*, 64(2):467–72.

Grey, J. and Gordon, J. (1978). "Perceptual effects of spectral modifications on musical timbres". *Journal of the Acoustical Society of America*, 63(5):1493–500.

Griffin, D., and Lim, J. S. (1984). "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Tansactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, pp. 236-243.

Gu, M. and Eisenstat, S.C. (1992). "A divide-and-conquer algorithm for the bidiagonal SVD". Tech. Report YALEU/DCS/RR-933, Department of Computer Science, Yale University, New Haven, CT.

Gu, M. and Eisenstat, S.C. (1993). "A stable and fast algorithm for updating the singular value decomposition". Tech. Report YALEU/DCS/RR-966, Department of Computer Science, Yale University, New Haven, CT.

Guillemain, P. and Kronland-Martinet, R. (1996). "Characterization of acoustic signals through continuous linear time-frequency representations". *Proceedings of the IEEE*, 84(4):561–85.

Hansen, P.C. (1987). "The truncated SVD as a method for regularization". *BIT*, 27:534–553.

Hansen, P.C. (1990a). "Relations between SVD and GSVD of discrete regularization problems in standard and general form". *Linear Algebra Appl.*, 141:165–176.

Hansen, P.C. (1990b). "Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank". *SIAM J. Sci. Statist. Comput.*, 11:503–518.

Helmholtz, H. L. F. (1954). *On the sensations of tone as a psychological basis for the theory of music.* (A. J. Ellis, Trans.) New York: Dover. (Original work published 1885).

Hotelling, H. (1933). "Analysis of a Complex of Statistical Variables in Principal Components", *Journal of Educational Psychology,* Vol. 24, pp. 417-441.

Howard, S. and Sirianunpiboon, S. (1996). "Wavelet coefficients, quadrature mirror filters and the group so(4)". *Proceedings of Third International Symposium on Time- Frequency and Time-Scale Analysis* (TFTS-96), IEEE; New York, NY, USA.

Huron, D. (1991). "Tonal consonance versus tonal fusion in polyphonic sonorities". *Music Perception*, 9(2):135–154.

Irino, T. and Patterson, R. (1994). "A theory of asymmetric intensity enhancement around acoustic transients". *Proceedings of 1994 International Conference on Spoken Language Processing, Acoustical Soc. Japan*; Tokyo, Japan.

Iverson, P. and Krumhansl, C. (1993). "Isolating the dynamic attributes of musical timbre". *Journal of the Acoustical Society of America*, 94(5):2595–603.

Jenkins, J. J. (1985). "Acoustic information for objects, places, and events." In W. H. Warren & R. E. Shaw (Eds.), *Persistence and change: Proceedings of the First International Conference on Event Perception.* pp 115-138, Hillsdale, NJ: Lawrence Erlbaum Associates.

Johansson, G. (1958). "Rigidity, stability and motion in perceptual space. *Acta Psychology, 14,* 359-370.

Johansson, G. (1973). "Visual perception of biological motion and a model for its analysis." *Perception and Psychophysics, 14,* 201-211.

Jordan, M. and Rumelhart, D. (1992). "Forward models: Supervised learning with a distal teacher". *Cognitive Science*, 16.

Jutten, C. & Herault, J. (1991). "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture", *Signal Processing,* Vol. 24, No. 1, pp. 1-10.

Karplus, K., and Strong, A. (1983). "Digital Synthesis of Plucked String and Drum Timbres". *Computer Music Journal.* 2(7):43-55.

Kanetkar, S. and Wagh, M. "Group character tables in discrete transform theory". *Journal of Computer and System Sciences*, 19(3):211–21.

Kistler, D., & Wightman, F. L. (1992). "A Model of Head-Related Transfer Functions Based on Principal Components", *Journal of the Acoustical Society of America,* Vol. 91, pp. 1637 - 1647.

Klein et al (1970). "Vowel Spectra, Vowel Spaces and Vowel Indentification", *Journal of the Acoustical Society of America*, 48, pg999-1009.

Kramer, H.P. and Mathews, M.V. (1956). "A Linear Coding for Transmitting a Set of Correlated Signals", *IRE Transactions Information Theory*, IT-2, 41-46.

Kruskal, J. (1964a). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". *Psychometrika*, 29.

Kruskal, J. (1964b). "Nonmetric multidimesional scaling: A numerical method". *Psychometrika*, 29.

Lacoume, J. L., & Ruiz, P. (1989). "Extraction of independent components from correlated inputs, A solution based on cumulants", *Proceedings of the Workshop on Higher-Order Spectral Analysis*, Vail, Colerado, June, pp. 146 - 151.

Lansky, P. and Steiglitz, K. (1981). "Synthesis of timbral families by warped linear prediction". *Computer Music Journal*, 5(3).

Laroche, J. and Meillier, J.-L. (1993). "A simplified source/filter model for percussive sounds". *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE; New York, NY, USA.

Laroche, J. and Meillier, J.-L. (1994). "Multichannel excitation/filter modeling of percussive sounds with application to the piano". *IEEE Transactions on Speech and Audio Processing*, 2(2):329–44.

Laughlin, R., Truax, B., and Funt, B. (1990). "Synthesis of acoustic timbres using principa! component analysis". *Proceedings ICMC*; Glasgow, UK.

Lee, M., Freed, A., and Wessel, D. (1992). 'Neural networks for simultaneous classification and parameter estimation in musical instrument control". In *Int. Soc. Opt. Eng.*. Adaptive and Learning Systems.

Lee, M. and Wessel, D. (1992). "Connectionist models for real-time control of synthesis and compositional algorithms". *Proceedings of the International Computer Music Conferece.*

Legitimus, D. and Schwab, L. (1990). "Natural underwater sounds identification by the use of neural networks and linear techniques". In *Proceedings of the International Neural Network Conference*, volume 2 vol. xlii+1098, pages 123–6 vol.1. Thomsom; SUN; British Comput. Soc.; et al, Kluwer; Dordrecht, Netherlands.

Legitimus, D. and Schwab, L. (1991). "Experimental comparison between neural networks and classical techniques of classification applied to natural underwater transients identification". *IEEE*; New York, NY, USA.

Lenz, R. (1989). "Group-theoretical model of feature extraction". *Journal of the Optical Society of America A (Optics and Image Science)*, 6(6):827–34.

Liu, Y. and Popplestone, R. (1994). "A group theoretic formalization of surface contact". *International Journal of Robotics Research*, 13(2):148–61.

Loeliger, H.-A. (1991). "Signal sets matched to groups". *IEEE Transactions on Information Theory*, 37(6):1675–82.

Mace, W. M. (1977). "James Gibson's strategy for perceiving: Ask not what's inside your head, but what your head is inside of." In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. E. Shaw & Bransford (Eds.), pp. 43-65. Hillsdale, NJ: Erlbaum.

Makhoul, J. (1975). "Linear prediction: A tutorial review". *Proceedings of the IEEE*, 63:561–580.

Mathews, M. (1969). *The technology of computer music*. MIT Press, Cambridge, Mass.

McAulay, R. and Quatieri, T. (1986). "Speech analysis/synthesis based on a sinusoidal representation". *IEEE Tr. ASSP.*

McIntyre, M. E., Schumacher, R. T., & Woodhouse, J. (1983). "On the Oscillations of Musical Instruments." *Journal of the Acoustical Society of America*, 74:1325-1345.

Mellinger, D. (1991). *Event formation in musical sound*. PhD thesis, Stanford University.

Miller, D. (1926). *The Science of Musical Sounds*. Mac Millan, New York.

Moon, T. (1996). "Similarity methods in signal processing". *IEEE Transactions on Signal Processing*, 44(4):827–33.

Moorer, J.A. (1978). "The use of the phase vocoder in computer music". *Journal of the Audio Engineering Society*, 24(9):717–727.

Mott, R. L. (1990). *Sound effects: Radio, TV, and film.* London: Focal Press.

Naparst, H. (1991). "Dense target signal processing". *IEEE Transactions on Information Theory*, 37(2):317–27.

Ney, H. (1990). "The use of a one-stage dynamic programming algorithm for connected word recognition". In Weibel, A. and Lee, K.-F., editors, *Readings in Speech Recognition*. Morgan Kaufmann Publishers.

Oja, E. (1982). "A Simplified Neuron Model as a Principal Component Analyzer", *J. Math. Biology*, 1, 267.

Oja, E., Karhunen, J., Wang, L., and Vigario, R. (1995). "Principal and independent components in neural networks - recent developments". *Proc. VII Italian Workshop on Neural Nets WIRN'95*, May 18 - 20, Vietri sul Mare, Italy.

Oja, E. (1995). *The nonlinear PCA learning rule and signal separation - mathematical analysis.* Helsinki University of Technology, Laboratory of Computer and Information Science, Report A26.

Oppenheim, A.V. (1989). *Dicrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, New Jersey.

Palazzo, R., J., Interlando, J., and deAlmeida, C. (1994). "Construction of signal sets matched to abelian and non-abelian groups". *IEEE International Symposium on Information Theory*, IEEE; New York, NY, USA.

Paul, J., Kilgore, E., and Klinger, A. (1988). "New algorithms for automated symmetry recognition". *SPIE - Int. Soc. Opt. Eng.* (USA). Intelligent Robots and Computer Vision. Sixth in a Series.

Pentland, A. and Turk, M. (1991). "Eigenfaces for recognition". *Journal of Cognitive Neuroscience*, 3(1).

Plomp, Pols and van de Geer (1967). "Dimensional Analysis of Vowel Spectra", *Journal of the Acoustical Society of Amer*
ica, 41(3), p707-712.

Plomp, R. (1970). "Timbre as a multidimensional attribute of complex tones". In Plomp, R. and Smoorenburn, G.G., editors, *Frequency Analysis and Periodicity Detection in Hearing*. A.W. Sijthoff, Leiden.

Portnoff, M. R. (1981). "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis", *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-29(3), pp. 374-390.

Puckette, M. and Brown, J. (1992). "An efficient algorithm for the calculation of a constant-q transform". *Acoustical Society of America.*

Pyt'ev, Y. (1971). "Signal preprocessor algorithm for recognition systems with similarity generalization". *Kibernetika*, 7(2):23–31.

Quatieri, T. and McAulay, R. (1989). "Phase coherence in speech reconstruction for enhancement and coding applications". In *ICASSP-89: 1989 International Conference on Acoustics, Speech and Signal Processing*, volume 4 vol. 2833, pages 207–10 vol.1. IEEE, IEEE; New York, NY, USA.

Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, New Jersey.

Rayleigh, Lord (1894). *The Theory of Sound.* Vol. 1, New York: Macmillan. (Reprinted by Dover, New York, 1945.)

Richardson, E. (1954). "The transient tones of wind instruments". *Journal of the Acoustical Society of America*, 26:960–962.

Richman, M., Parks, T., and Shenoy, R. (1995). "Discrete-time, discrete-frequency time-frequency representations".

International Conference on Acoustics, Speech, and Signal Processing, IEEE; New York, NY, USA.

Richman, M., Parks, T., and Shenoy, R. (1996). "Features of a discrete wigner distribution". *IEEE Digital Signal Processing Workshop Proceedings*, IEEE; New York, NY, USA.

Risset, J. (1966). *Computer study of trumpet tones*. Bell Labs Technical Report.

Risset, J. (1971). "Paradoxes de hauteur". In *Proceedings of the 7th international congress of Acoustics*, Budapest.

Risset, J. and Mathews, M. (1969). "Analysis of musical instrument tones". *Physics Today*, 22(2).

Risset, J. and Wessel, D. (1982). "Exploration of timbre by analysis and synthesis". In Deutsch, D., editor, *The Psychology of Music*, pages 26–58. Academic Press, New York.

Roads, C. and Strawn, J. (1987). *Foundations of Computer Music*. MIT Press, Cambridge, Mass.

Rockmore, D. (1995). "Fast fourier transforms for wreath products". *Applied and Computational Harmonic Analysis*, 2(3):279–92.

Rodet, X. (1996). "Recent developments in computer sound analysis and synthesis". *Computer Music Journal*, 20(1):57–61.

Rosch, E. (1975). "Cognitive reference points". *Cognitive Psychology*, 7:532–547.

Runeson, S. (1977). "On the possibility of smart perceptual mechanisms". *Scandanavian Journal of Psychology, 18*, pp. 172 - 179.

Saint-Arnaud, N. (1995a). *Classification of Sound Textures*. MIT Media Laboratory Masters Thesis, Cambridge, MA.

Saint-Arnaud, N. (1995b). "Sound texture resynthesis". In *Proceedings of the International Joint Conference on Artificial Intelligence*, Montreal. IJCAI.

Sandell, G. and Martens, W. (1995). "Perceptual evaluation of principal-component-based synthesis of musical timbres". *Journal of the Audio Engineering Society*, 43(12):1013–28.

Sayeed, A. and Jones, D. (1996). "Equivalence of generalized joint signal representations of arbitrary variables". *IEEE Transactions on Signal Processing*, 44(12):2959–70.

Schaeffer, P. (1966) *Traite des objects musicaux*. Paris: Seuil.

Schubert, E. D. (1974). "The role of auditory perception in language processing". In *Reading, Perception and Language*. D. D. Duane & M. B. Rawson (Eds.), pp. 97-130, Baltimore: York Press.

Serra, X. and Smith, J. (1990a). "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition". *Computer Music Journal*, 14(4):12–24.

Serra, X. and Smith, J. (1990b). "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition". In *Proceedings of the Fifth European Signal Processing Conference*. Elsevier; Amsterdam, Netherlands.

Setayeshi, S. and El-Hawary, F. (1994). "Neural network based signal prediction and parameter estimation for underwater layered media systems identification". *Proceedings of Canadian Conference on Electrical and Computer Engineering*, IEEE; New York, NY, USA.

Settel, Z. and Lippe, C. (1995). "Real-time musical applications using frequency domain signal processing". In *Proceedings of 1995 Workshop on Applications of Single Processing to Audio and Accoustics*, number 284. IEEE; New York, NY, USA.

Shaw, R. E., McIntyre, M., & Mace, W. M. (1974). "The role of symmetry in event perception." In MacLeod, R. B. & Pick, H. (Eds.), *Perception: Essays in honour of James J. Gibson.* Ithica: Cornell University Press.

Shaw, R. E., & Pittenger, J. B. (1978). "Perceiving change." In H. Pick and E. Slatzman (Eds.), *Modes of Perceiving and Processing Information.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Slaney, M., Covell, M., and Lassiter, B. (1996). "Automatic audio morphing". In *International Conference on Acoustics, Speech and Singnal Processing*, Atlanta, May.

Slawson, W. (1968). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America. 43,* 87-101.

Smalley, D. (1986). "Spectromorphology and Structuring Process". In Emmerson, S. (Ed.). *The Language of Electroacoustic Music.* London: Macmillan.

Smaragdis, P. (1997). *Information Theoretic Approaches to Source Separation*, Masters Thesis, MAS Department, Massachusetts Institute of Technology.

Smaragdis, P. (1997). "Efficient Blind Separation of Convolved Sound Mixtures",*IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.* New Paltz NY, October 1997.

Smaragdis, P. (1998). "Blind Separation of Convolved Mixtures in the Frequency Domain". *International Workshop on Independence & Artificial Neural Networks* University of La Laguna, Tenerife, Spain, February 9 - 10, 1998.

Smith, J. O. (1990). "Efficient Yet Accurate Models for Strings and Air Columns Using Sparse Lumping of Distributed Losses and Dispersion." In *Proceedings of the Colloquium on Physical Modeling.*

Smith, , J. O. (1992). "Physical modeling using digital waveguides". *Computer Music Journal*, 16:74–87.

Snell, J. (1983). "Lucasfilm audio signal processor and music instrument". In *IEEE Electronics Conventions; LA.*

Stankovic, R. and Stankovic, M. (1994). "Group theoretic models of linear systems: a common look at continuous, discrete and digital systems". *Proceedings of Third International Conference on Systems Integration*, IEEE Comput. Soc. Press; Los Alamitos, CA, USA.

Stapleton, J. C., & Bass, S. (1988). "Synthesis of Music Tones Based on the Karhunen-Loeve Transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 36, pp. 305-319.

Stautner, J. P. (1983). "Analysis and Synthesis of Music Using the Auditory Transform", Master's Thesis, MIT EECS Department, Cambridge, MA.

Stockham, T. G., Cannon, T. M., and Ingebretsen, R. B. (1975). "Blind Deconvolution Through Digital Signal Processing". *Proceedings of the IEEE,* Vol. 63, pp. 678-692.

Tellman, E., Haken, L., and Holloway, B. (1995). "Timbre morphing of sounds with unequal numbers of features". *Journal of the Audio Engineering Society,* 43(9):678–89.

Tenenbaum, J. B., and Freeman, W. T. (1997). "Separating Style and Content", in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan and T. Petsche, (Eds)., Morgan Kaufmann, San Mateo.

Tenney, J. (1965). The physical correlates of timbre. *Gravesaner Blaetter,* 26:106–109.

Therrien, C.W. (1989). *Decision Estimation and Classification.* John Wiley & Sons, New York, NY.

Therrien, C., Cristi, R., and Allison, D. (1994). "Methods for acoustic data synthesis". *Proceedings of IEEE 6th Digital Signal Processing Workshop*, IEEE; New York, NY, USA.

Thiele, C. and Villemoes, L. (1996). "A fast algorithm for adapted time-frequency tilings". *Applied and Computational Harmonic Analysis*, 3(2):91–9.

Toiviainen, P., Kaipainen, M., and Louhivuori, J. (1995). "Musical timbre: similarity ratings correlate with computational feature space distances". *Journal of New Music Research*, 24(3):282–98.

VanDerveer, N. J. (1979). "Ecological acoustics: Human perception of environmental sounds." *Dissertation Abstracts International*, 40, 4543B. (University Microfilms No. 80-04,002)

Vishnevetskii, A. (1990). "Fast group-theoretical transform (signal convolution)". *Problemy Peredachi Informatsii*, 26(1):104–7.

Wang, K. and Shamma, S. (1995). "Auditory analysis of spectro-temporal information in acoustic signals". *IEEE Engineering in Medicine and Biology Magazine*, 14(2):186–94.

Warren, W., & Shaw, R. E., (1985). "Events and encounters as units of analysis for ecological psychology." In W. H. Warren & R. E. Shaw (Eds.), *Persistence and change: Proceedings of the First International Conference on Event Perception.* (pp 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Warren, W. and Verbrugge, R. (1988). "Auditory perception of breaking and bouncing events". In Richards, W., editor, *Natrual Computation*. MIT Press, Cambridge, Mass.

Waters, R., Anderson, D., Barrus, J., Brogan, D., Casey, M., McKeown, S., Nitta, T., Sterns, I., and Yerazunis, W. (1997). "Diamond park and spline: Social virtual reality with 3d animation, spoken interaction, and runtime extendability". *Presence*.

Weisburn, B. and Shenoy, R. (1996). "Time-frequency strip filters". *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, IEEE; New York, NY, USA.

Weiss, L. (1996). "Time-varying system characterization for wideband input signals". *Signal Processing*, 55(3):295–304.

Wessel, D. (1973). "Psychoacoustics and music". *Bulletin of the Computer Arts Society*, 30:1–2.

Wessel, D. (1979). "Timbre space as a musical control structure". *Computer Music Journal*.

Wildes, R., & Richards, W. (1988). "Recovering material properties from sound." In W. Richards (Ed.), *Natural Computation* (pp. 356-363). Cambridge, MA: MIT Press.

Winham, G., & Steiglitz, K. (1970). "Input Generators for Digital Sound Synthesis", *Journal of the Acoustical Society of America* 47 2:ii, pp. 665-666.

Woodard, J. (1992). "Modeling and classification of natural sounds by product code hidden markov models". *IEEE Transactions on Signal Processing*, 40(7):1833–5.

Yilmaz, H. (1967a). "Perceptual invariance and the psychophysical law", *Perception and Psychophysics*, Vol. 2(11), 533-538.

Yilmaz, H. (1967b). "A Theory of Speech Perception", *Bulletin of Mathematical Biophysics*, Vol. 29, 793-824.

Yilmaz, H. (1968). "A Theory of Speech Perception II", *Bulletin of Mathematical Biophysics*, Vol. 30, 455-479.

Zahorian S.A. and Rothenburg, M. (1981). "Principal Components analysis for low-redundancy encoding of speech spectra", *Journal of the Acoustical Society of America*, 69(3), 832-845.