PERCEIVING THE SOURCES OF ENVIRONMENTAL SOUNDS WITH A VARYING NUMBER OF SPECTRAL CHANNELS

by

VALERIY SHAFIRO

A dissertation submitted to the Graduate Faculty in Speech and Hearing Sciences in partial fulfillment of the requirments for the degree of Doctor of Philosophy, the City University of New York

2004

This manuscript has been read and accepted for the Graduate Faculty in Speech and Hearing Sciences in satisfaction of the dissertation requirements for the degree of Doctor of Philosophy.

11/11/03
Date

Winifred Strange, Ph. D.
Chair of Examining Committee

11/11/03
Date

Martin Gitterman, Ph. D.
Executive Officer

Supervisory Committee:

James J. Jenkins, Ph. D.

Arlene Neuman, Ph. D.

Glenis Long, Ph. D.

Outside reader:

David Rosenthal, Ph. D.

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

PERCEIVING THE SOURCES OF ENVIRONMENTAL SOUNDS WITH A VARYING NUMBER OF SPECTRAL CHANNELS

by

Valeriy Shafiro

Adviser: Professor Winifred Strange

Nonspeech environmental sounds provide the listener with valuable information about sound-producing objects and events in the immediate environment. Avoiding a collision with an unseen car, responding to a ringing doorbell, or enjoying a bird song are only a few examples of the potential benefits of accurate object and event identification. However, relatively little is known about the cognitive factors and acoustic parameters that influence the perception of the sources of environmental sounds. A major difficulty associated with environmental sound research is the great variety in the types of sound sources in the environment, and the lack of a clear taxonomy of environmental sounds. In an attempt to provide an empirical basis for environmental sound classification and explore the acoustic parameters important in environmental sound perception, the present research investigated the effects of systematically varying the spectral resolution of familiar environmental sounds on listeners' ability to identify the corresponding sound sources.

Normally hearing adult listeners were asked to identify the sources of 60 familiar environmental sounds processed through a vocoder simulation of a cochlear implant with

varying numbers of frequency channels. Using a Latin square design, listeners heard 10 different sounds in each of six channel conditions (2, 4, 8, 16, 24, 32) followed by all 60 undistorted sounds. For each sound, listeners selected one of 60 response options that best described the source of the sound. Results indicate that identification performance continuously improved with an increasing number of frequency channels and reached 81% correct with 32 channels. However, for some sounds, identification accuracy declined when the number of channels was 16 or higher due, perhaps, to the spectral asynchrony introduced by unequal filter group delays across channels. Thus, overall identification accuracy did not substantially improve beyond 16 channels. In general, broadband temporally patterned sounds (e.g., 'helicopter', 'clapping', 'typing') tended to require fewer channels to be identified correctly than sounds with time-varying narrow band resonances (e.g., 'baby crying', 'doorbell', 'rooster'). These findings, compared with those of other investigators, suggest that, without training, the accurate perception of environmental sounds may require more spectral channels than that of speech sounds.

ACKNOWLEDGMENTS

One piece of wisdom that I learned as a graduate student is that research, and especially behavioral research, is inherently a highly collaborative undertaking. A twisted path from an initial idea to a manuscript describing an original experimental work entails many important decisions that affect research quality. Further, it requires a great deal of resources such as participants, equipment, finances, expert knowledge, and time commitments. Goodwill and cooperation of many people and institutions are required for a research project to come to fruition, even in a relative sense of satisfying an academic requirement. During my dissertation work, as during all preceding stages of my graduate training, I received help and support from many people. The full extent of their contributions to my work I am only now beginning to realize.

First of all, I would like to thank the members of my dissertation committee Drs. Winifred Strange, James Jenkins, Arlene Neuman and Glenis Long. It was Winifred Strange who initially channeled my somewhat chaotic thinking about the perception of environmental sounds into the form of a manageable project. Countless spontaneous and arranged discussions with Jim Jenkins kept me motivated throughout my work, and significantly broadened my perspective on the field of environmental sound research. Arlene Neuman has taught me most of what I know about the thorny issues of instrumentation during my tenure in her laboratory. She also provided valuable guidance on numerous aspects of this work ranging from hardware choices and calibration to data analysis. My design and signal processing decisions benefited considerably from the advice of Glenis Long, who was always willing to share her expertise on these matters, and kept me from losing track of the auditory periphery.

As a behavioral researcher, I encountered several challenges stemming from the signal processing part of the methodology I employed. To overcome these challenges, I relied heavily on the advice of three engineering experts. I would like to thank Philip Loizou for providing his algorithms for generating stimuli, and for his suggestions about signal processing alternatives. Patrick Nye has been generous with his time, and instrumental in his assistance during my graduate school years in enabling me to understand the fundamentals of signal processing and Matlab programming. I am grateful to Jim Kates for his willingness to share his knowledge over emails without ever meeting me in person.

In my frustrated search for good quality recordings of environmental sounds that served as the basis for my stimuli, I was helped by the advice of Gordon Hempton, DBA the Sound Tracker. I thank the post-production staff of CUNY-TV for making their CD sound effects libraries available to me.

I am indebted to Erika Levy, Kanae Nishi, Miwako Hisagi, and Yana Gilichinskaya as well as other members of the Speech Acoustics and Perception Laboratory for their wit, personal warmth and humor, which made innumerable boring and mundane tasks seem exciting and worth doing. I would like to thank the student, faculty and technical staff of the Department of Speech and Hearing Sciences for their valuable suggestions and assistance over the years.

Special thanks go to Dr. Lawrence Raphael for sharing his unique and unparalleled ability to discuss various complex topics of life and science in the form of anecdotes, metaphors, and abbreviations; to Dr. David Rosenthal for numerous illuminating discussions of issues in cognitive science; to Gary Chant for always keeping

his cables and adapters in place; and to Linda Ashour and Loretta Walker for their help with numerous administrative tasks involved in my research and graduate training.

My research interests and dreams might never have materialized if it were not for the unyielding support of my parents, Leonid and Galina, my wife Tatyana, and my brother Alex. I also have to acknowledge the patience of my son Samuel who conscientiously let me finish collecting data and major writing before delighting us with his appearance in this world.

Finally, I would like to thank the NIH-NIDCD for the financial support of my research [F31 DC006109].

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 3

CHAPTER 4

CHAPTER 1

INTRODUCTION

1.1 Research aims

This research was designed to accomplish several practical and theoretical goals. The practical goals of the project were the following. First, it aimed to establish the minimal number of frequency channels required for the identification of the sources of a large number of familiar environmental sounds. The majority of currently available cochlear implants process acoustic input signals by dividing them into a number of frequency channels and stimulating the listener's auditory nerve according to the dynamic intensity information (i.e., amplitude envelope) derived from the individual channels. The number of channels used during the signal-processing stage differs among specific implant models. To date, most cochlear implant research has focused on the number of frequency channels needed for optimal speech perception. Nevertheless, for a cochlear implant to be efficient in supplying the user with information about sound-producing objects and events in the environment, it is important to know the number of channels needed for the identification of sources of a large number of familiar environmental sounds that are encountered by listeners in daily life. The question of the required number of frequency channels has been vigorously researched for speech perception using simulated cochlear implant models (Loizou, Dorman & Tu, 1999; Shannon, Zeng & Wygonski, 1998; Dorman, Loizou & Rainey, 1997; Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995), but only marginally for environmental sounds (Gygi, 2001). Therefore, one aim of this project was to provide this basic and important information.

The second aim of this research was to determine the range of variability among individual environmental sounds in the number of channels required to identify corresponding sound sources. Previous research (Gygi, 2001) indicated that environmental sounds differ substantially in regard to the amount of spectral detail required for accurate source identification. For example, in filtering studies, it has been shown that 'siren' or 'cough' sounds could be accurately identified even when spectral information below 8000 Hz was removed. On the other hand, identification of 'thunder' and 'waves' sounds severely deteriorated when the energy below 300 Hz was removed, even though energy at higher frequencies was left intact. Similarly, Gygi (2001) found that a 'helicopter' sound was highly identifiable when the amplitude envelope of only a single channel was used to modulate white noise, but a 'ringing phone' sound and a 'sneeze' sound were identified quite poorly even with six frequency channels. Gygi's design, however, included only one channel and six channel conditions. Increasing the number of different channel conditions would make it possible to examine differences in the number of channels required to identify the sources of individual environmental sounds more closely. This information may be useful in making predictions about the identifiability of individual environmental sounds through cochlear implants with varying numbers of channels. In essence, it would provide a basis for establishing which sound sources are easy and which are difficult for cochlear implant users to identify, given the number of channels implemented in particular implants.

A third aim of the project was based on the expectation that individual environmental sounds could be consistently differentiated based on the number of channels required for source identification. This expectation was based on Gygi's

findings with one and six channels described above and pilot data collected by the author. In the pilot study, source identification of 15 environmental sounds processed through five different number-of-channels conditions was tested with two listeners in a closed-set response format. Preliminary results indicated that both listeners could identify the sounds of a 'helicopter' and 'breaking glass' with two channels, the sound of a 'coughing child' with eight channels, and the sound of a 'cat meowing' with sixteen channels. While the small number of sounds and listeners tested in the pilot prevented any conclusive statements about the data, it did indicate a strong possibility that individual environmental sounds could be grouped based on the number of channels required for source identification. In that case, it might be possible to identify acoustic parameters shared by all sounds characterized by the same number of channels, and ones that distinguish sounds between different number-of-channels groups. These data, in turn, could serve as an empirical foundation of a future model designed to predict the required number of channels for novel environmental sounds based on their acoustic structure.

The theoretical aims of the project, while different in scope, were closely related to its practical aims. A major obstacle to the general understanding of the perception of the sources of environmental sounds has been the absence of a comprehensive empirically-based taxonomy in which individual environmental sounds could be classified according to their perceptual similarities and differences. Although there are many possible ways in which a classification scheme can be constructed, there are considerable drawbacks associated with each. For instance, a classification can be based on the listening context in which sounds are heard (e.g., household sounds, office sounds, street sounds, etc.). Or, one may choose to classify sounds based on their relevance to the

listener's well being (e.g., danger sounds, relaxing sounds, distracting sounds, etc.). However, as suggested by Gaver (1993), these are not mutually exclusive categories, and do not provide strict constraints on the identity of member sounds. A sound in one category can easily be listed as a member in another. For example, the sound of a ringing telephone can be encountered in both home and office, just as the sound of an alarm can both represent a danger and be distracting. As an alternative, Gaver offered a way of classifying sounds based on the physical attributes of their sources. Indeed, the accurate perception of the source characteristics of a specific sound should be much more useful in generating a description of a particular environment, or determining the relevance of a sound to the well being of the listener, than the other way around. That is, knowledge of an environment can be obtained by identifying the sound sources in that environment. Knowing the general label for an environmental context, however, does not specify what individual sound sources are present in it at a given time.

Overall, Gaver's source-based taxonomy provides an attractive alternative to common context-based descriptions of environmental sounds. Unfortunately, it is grounded more in analytical reasoning than perceptual data. In fact, studies investigating listeners' perception of large selections of environmental sounds using such techniques as principal component analysis (Ballas, 1993), and multidimensional scaling (Gygi, 2001) have consistently found that the clustering of sounds along psychological dimensions only remotely resembles Gaver's grouping of sounds. Thus, while Gaver's approach to the perception of environmental sounds may provide an effective analytical tool for exploring the relationship between the source and the acoustic signal, it only partially reflects the dimensions along which listeners actually perceive the sources of

environmental sounds. On the other hand, a classification of environmental sounds based on the quantifiable amount of spectral detail (i.e., the number of frequency channels) required for the accurate identification of the underlying sources should provide a solid empirical footing for the perceptual grouping of individual environmental sounds. Therefore, one theoretical goal of the present study was to explore the possibility of classifying environmental sounds based on the number of channels required for source identification. This taxonomy, by its nature, would highlight acoustic parameters important for source identification, and could be used in the future for systematically exploring the relationships between the source and the acoustic signal, on the one hand, and the signal and the listener, on the other. Given such a taxonomy, further questions can be asked about why source identification of certain sounds requires a different number of channels than other sounds.

Another theoretical aim of the project was to investigate how the perceptual dimensions used by listeners in determining the sources of environmental sounds change as a function of changing spectral resolution, represented by different number-of-channels conditions. Multidimensional scaling solutions obtained on listeners' identification responses separately for each number-of-channels conditions could indicate changes in clustering of sounds in perceptual space. It was expected that sounds whose sources could not be identified at a given number-of-channels condition would occupy a different position in the perceptual space than in conditions when their sources could be identified correctly. This is because in each case such sounds would be likely to be evaluated along different perceptual dimensions. This expectation was supported by previous findings that listeners tended to describe environmental sounds in terms of the

sources responsible for or associated with their production. The sensory qualities of environmental sounds were reported only when listeners were asked about them specifically, or when listeners could not identify the sources based on the sound they heard (Vanderveer, 1979; Gaver, 1988; Ballas, 1993; Truax, 2001). Later, information about the clustering of sounds in perceptual space could be combined with the results of acoustic analysis performed on the sounds in each condition. This would reveal which acoustic parameters of environmental sounds are most salient perceptually in each condition, and how the salient acoustic parameters change in response to changes in spectral resolution of environmental sounds.

1.2 Research background and scope

Historically, research in sound-based perception and cognition has concentrated in several large topic areas. Much has been learned about the human perception of speech and the relationship between the acoustic signals and phonetic percepts. Text-to-speech synthesizers and speech recognition systems are, perhaps, the most widely known technological offspring of this line of research. Research on music perception has led to the development of electronic musical instruments and a new kind of electro-acoustic music. Psychoacoustic experiments with acoustically simple stimuli have provided invaluable knowledge about the psychophysical structure of human hearing, sound source localization in space, and physical constraints on hearing and perceptual organization of auditory stimuli. This information has provided a basis for the design of efficient signal compression strategies, development of virtual three-dimensional audio, and various types of electronic devices for aural communications. While the practical applications of

the knowledge accumulated in these research areas are quite numerous, and each of these areas can be further divided into a variety of smaller topics, these areas of concentration generally define the subject boundaries of the auditory perception research of the last century. However, some topics in auditory perception have not been studied to the extent warranted by their importance in human daily living or their theoretical relevance. One such area is the perception of sources of environmental sounds by humans.

Listeners' tendency to perceive environmental sounds primarily in terms of their sources, and to report on their sensory qualities only when sources cannot be easily identified (Vanderveer, 1979; Gaver, 1993; Ballas, 1993; Truax, 2001) is hardly surprising, given that sources are substantially more relevant to listeners' well being than the sounds they produce. Indeed, a listener's chances for survival upon hearing a gun shot, an explosion, or a lion roar would be drastically undermined if he or she would contemplate the sensory qualities of the sound first (e.g., the sound's pitch or loudness) before seeking cover. Thus, the information about the sources of environmental sounds rather than the sounds themselves has primary ecological value for the listener. From the listener's point of view, the main function of the acoustic signal of an environmental sound is to convey information about its source.

Nonelectric mechanical sound sources (e.g., breaking glass, footsteps, etc.) and the sounds they produce are linked together by a set of nonarbitrary lawful systematic relationships in which the physical properties of and dynamic changes in the sound source are causally related to the resulting acoustic signal. While much remains unknown about the exact nature of these relationships, the success of sound synthesis based on the numerically specified physical characteristics and behavior of the source

confirms that such relationships do exist, and further suggests that information about the source may be available in the corresponding acoustic signal (Gaver, 1993; Carello, Wagman, & Turvey, 2003). Indeed, research has demonstrated that human listeners are able to obtain remarkably detailed information about the sources of environmental sounds from the acoustic signal alone. For instance, it was shown that listeners could identify the configuration of two palms during clapping from the sound of the corresponding claps (Repp, 1987). Also, listeners can determine the length of a wooden rod from the sound it makes when it is dropped on a floor (Carello, Anderson, Kunkler-Peck, 1998). Similarly, listeners can estimate the hardness of a mallet used to strike a cooking pan from the resulting sound (Freed, 1989), or judge the fullness of a vessel from the sound of pouring water into it (Cabe & Pittenger, 2000). However, the sheer multitude of sounds that convey potentially useful information about objects and events in the environment, constitutes a major difficulty for constructing a general theoretical framework of environmental sound perception.

One way of overcoming this difficulty can be seen in attempts to determine a representative set of environmental sounds that supply the listener with information about the environment (Ballas, 1993; Marcell, Borella, Greene, Kerr & Rogers, 2000). Indeed, while the number of environmental sounds that may potentially be present in a listener's environment is virtually infinite, there is a finite subset of such sounds that the listener is exposed to in day-to-day life. Hence, measures of the familiarity of sounds (Marcell et al., 2000), or their frequency of occurrence (Ballas, 1993) can provide a basis for establishing a limited subset of representative sounds well known by a large population of listeners. This approach offers a practical advantage for narrowing the scope of

environmental sounds from all possible sounds to those that listeners actually hear often and whose sources they can easily identify. Its drawback, however, is that it may preclude the researcher from generalizing his/her findings to environmental sounds not included in the subset.

Another indication of a close relationship between the perceptual characteristics of environmental sounds and those of their sources can be seen in the findings of a similarity rating study that compared the perceptual space for environmental sounds with that of their sources. Gygi (2001) describes a study in which listeners made similarity judgments on (a) pairs of acoustic tokens of different environmental sounds, (b) pairs of the same sounds imagined by the listener presented only with the sounds' labels, (c) pairs of imagined source events corresponding to the sounds investigated in (a) and (b). Multidimensional scaling solutions derived from the three kinds of similarity data revealed very similar clustering of heard and imagined sounds and their imagined sources in a two-dimensional space across all three solutions. That is, overall, the stimuli tended to be positioned around the same coordinates regardless of whether the listeners judged the similarity of acoustic stimuli, the similarity in the imagined sounds, or the similarity in the source events underlying the sounds. This finding provides additional evidence for a perceptual transparency of environmental sounds and their sources.

In sum, it is more productive to emphasize the perception of sources of environmental sounds rather than their sensory or abstract semantic characteristics (e.g., dark – light, sharp – dull) when attempting to understand what information about the source these sounds convey to the listener, and how this information is being conveyed. This approach is supported by (1) listeners' overall tendency to describe sounds they hear

in the environment in terms of their sources, (2) much greater ecological relevance of sound sources than corresponding sounds to the listener's well being, (3) listeners' ability to obtain detailed information about many of the source characteristics and attributes from the sounds they produce, and (4) similarities in the perceptual space of environmental sounds and that of their sources.

In general, an understanding of the perception of sources of environmental sounds, i.e. the mechanism whereby the listener obtains information about the identity and behavior of one or more distal sound-producing objects and events in the environment, offers a variety of theoretical and practical rewards. Theoretically, it offers a novel approach to the age-old problems of how distal objects in the environment are perceived by the organism. Or, to put it in the classical language of Gestalt psychology: "How do things appear?" and "Why do things appear the way they do?" To date, the majority of general object perception models have been based mainly on research in vision (Jenkins, 1985; Marcell et al., 2000). Thus, studies of the auditory perception of the identity and behavior of objects in an environment through sound provides a new way of testing the predictions of such models. In addition, there are other long-standing theoretical issues in psychology that can be explored afresh in the auditory modality. The questions of whole-part relationships, knowledge representation, perceptual learning, and the units or objects of perception are as relevant for auditory as they are for visual perception. Moreover, as noted by Gygi (2001), environmental sounds have a unique place in auditory perception research because their perception is different from speech in some important respects but is similar to speech perception in others. Specifically, unlike physically simple stimuli such as pure tones and white noise, commonly used in

psychoacoustic research, environmental sounds provide the listener with meaningful information about their sources. It can, thus, be assumed that perception of the sources of environmental sounds requires a higher level of central processing than does the perception of simple stimuli. Therefore, examining similarities and differences between environmental sound and speech perception, on the one hand, and environmental sounds and simple acoustic stimuli, on the other, may lead to novel insights about central auditory processing of physically complex and ecologically relevant sounds.

An understanding of the perception of the sources of environmental sounds also has a number of practical applications. Perhaps the most beneficial application of this knowledge can be gained in the design of sensory aids such as cochlear implants. The perceptual handicap of deafness is not limited to speech, but also includes all other sounds that animate the world of hearing persons. In fact, some prelingually deafened individuals seek cochlear implantation in adulthood without much hope for developing speech, but in order to become more aware of various sound sources in their environment. Even a brief review of personal testimonials posted in various Internet based forums for cochlear implant users can give a sense of how much appreciation newly implanted cochlear implant users have for being able to hear environmental sounds. Given the important ecological function played by the sources of environmental sounds in the daily life of an individual, it becomes imperative to insure that cochlear implants improve not only speech perception but also the perception of environmental sounds for hearing-impaired persons.

In order to achieve this goal, it is important to answer the following three interrelated questions. (1) How are the sources of environmental sounds specified by the acoustic

structure of environmental sounds? (2) How do cochlear implants alter the acoustic structure generated by the sources of environmental sounds? (3) How do changes in environmental sounds' acoustics resulting from cochlear implant processing affect the perception of sound sources? Answers to the first and the third questions may help to determine the acoustic information that is sufficient and necessary for the accurate perception of the sources of environmental sounds. This knowledge, coupled with the answer to the second question, may help to determine whether this information is available to the listener after the signal processing performed by a cochlear implant, and indicate ways to improve implant design in order to supply the required information. Finally, a future investigation may be able to reveal whether and how listeners' ability to identify the sources of environmental sound from spectrally impoverished acoustic input can be improved through training.

CHAPTER 2

PREVIOUS ENVIRONMENTAL SOUND PERCEPTION RESEARCH

Systematic interest in the perception of environmental sounds and their sources by humans is a relatively recent development in psychology and hearing science. Nevertheless, there are several theoretical approaches and experimental paradigms that have been used or can be used in this area of research. There appears to be relatively little overall consensus in how different researchers understand the mechanism of environmental sound perception, or how to best define the questions that need to be asked about it. While many may agree with Bates' (2000) conjecture that "properly defining the problem is ninety percent of its solution," just how to define the problem properly seems highly debatable. This state of affairs is, perhaps, a necessary step in determining the experimental approach that would definitively clarify the mechanism of environmental sound perception. Until then, however, one may choose to concentrate on finding and exploring those questions that can lead to the definition of the right experimental approach.

2.1 Theoretical approaches

Theoretical approaches to the perception of environmental sounds have a number of historic underpinnings, and are largely rooted in the traditional schools of thought in psychology. Although many perceptual models attempt to incorporate elements of more than one approach, as their creators may see fit (Martin, 1999; McAdams, 1993; Gygi, 2001), there are nevertheless clear conceptual distinctions in how the perception of environmental sounds can be understood from different theoretical view points. Thus, it

may add clarity to the present discussion if the principal differences among main theoretical approaches are made explicit at the outset.

Two different meta-theoretical approaches have been applied to research in environmental sound perception: the ecological approach, and the information processing approach. Although these approaches are associated with a number of specific models, there are overarching principal differences in how perception is viewed in each case.

According to the ecological view, also known as the direct realist view, perception is direct (Gibson, 1979; Carello et al., 2003; Fowler, 1990). That is, the perceptual system is said to be tuned-in to the ecologically relevant properties of distal objects and events as specified in the stimulus array. Properties of distal objects are directly detected by the perceptual system from the stimulus array, and are perceived without further mediation by mental processes. The physical input coming to the perceiving organism from the world is credited with providing adequate information for the perception of distal objects and events. Because the stimulus array is lawfully shaped by the behavior of objects producing the physical stimulation, the stimulus contains systematic patterns of energy distribution that are assumed to be sufficient for the invariant specification of the underlying objects in the environment and their physical behavior, i.e., events. The ecological approach places a major emphasis on temporal variations in the energy distribution of the stimulus array because the dynamic behavior of objects is reflected in the temporal patterning of the array.

The information processing view, in contrast, considers the sensory input to the perceptual system to be impoverished with respect to the distal objects in the

environment, and, in and of itself, inadequate for accurate perception of distal objects and their physical behavior. Even though it is agreed that the input signals can be lawfully shaped by the behavior of distal objects, the sensory input is viewed as requiring additional processing by the perceptual system in order to obtain an accurate perception of distal objects. Therefore, perception is proposed to be a mediated multi-stage process in which an accurate percept of an object and event is gradually derived from the initially impoverished input through some kind of matching process. During matching, certain characteristics of the input are evaluated against internal representations of distal objects such as templates or prototypes (McAdams, 1993; Martin, 1999). Thus, in principle, the information processing approach emphasizes the mental processes that operate on the input and transform it into percepts, rather than finding information in the input itself that specifies the distal objects and events in the outside world.

Despite some major conceptual differences between the two views of perception, both approaches are but different ways of answering two central questions: What is the information in the physical signal that specifies the identity and behavior of objects and events in the environment?, and How is this information perceived by the organism? The information processing approach focuses mainly on the mental processes that transform an impoverished input into an accurate percept of a distal object or event, while the ecological approach emphasizes the information in the stimulus that defines the distal objects and events to the perceiver. Presently, however, the two approaches to perception are best viewed as two alternative guides to posing research questions and examining experimental data, rather than definitive explanations of how perception works. For one, invariant patterns of change in the stimulus that are said to specify objects and events

have not been specified for the majority of objects and events that organisms actually perceive. Similarly, perceptual processes responsible for matching input stimuli with mental representations are usually described only in terms of what they are supposed to do, rather than how they do it. Currently, as Handel (1989) observes, the choice of a theoretical platform seems more a question of faith than science. Nevertheless, both approaches may provide useful conceptual tools in environmental sound research.

2.1.1. *Ecological approach*

Overall, the ecological approach, being the more recent of the two, has been developed to a lesser extent. The main thrust of this ecologically oriented research has been to determine the nature of information about the environment people are able to obtain from environmental sounds (Repp, 1987; Freed, 1989; Li, Logan & Pastore, 1991), and challenging more traditional approaches to auditory perception (Fowler, 1990; 1991; Carello et al., 2003). Overall, these works have undeniably succeeded in demonstrating that individual environmental sounds can convey a wealth of information about various attributes of their physical sources.

In addition, ecologically oriented studies have highlighted several flaws in the purely analytical approach to understanding the perception of source attributes. In one example, provided by Carello et al. (2003), Gordon and Webb (1996) have argued that it is impossible to judge the shape of a struck drum based on their mathematical analysis of the physical event. However, in an experiment that involved human listeners attending to such sounds, Kunkler-Peck and Turvey (2000) demonstrated that listeners could identify the shape significantly above chance. Thus, the necessary idealization of the physical

representations of real world events involved in their mathematical modeling may provide a rather shaky ground for establishing the physical basis of real world perception. On the other hand, the empirically derived knowledge of human perceptual abilities provides new challenges for physicists who attempt to explain the physical basis of perceptual phenomena, and may lead to the development of new analytical models that are more ecologically valid.

Some pioneering work in this direction has been done by Gaver (1993b), who has developed a number of synthesis algorithms for various environmental sounds. These algorithms are particularly interesting because they are controlled by the physical parameters of sound-producing objects and events, which serve as inputs, rather than by statistically derived direct specification of the intended acoustic structure. Gaver's algorithms incorporate both the perceptual characteristics of the sound-producing objects and events obtained from listeners' reports and the physical analysis of the sound-producing behavior of the source objects. Some of the algorithms are built around physical models of sound-producing objects' behavior, e.g., impacts and scraping, while others are based mainly on their ability to elicit desired percepts, e.g., dripping, machine sounds. Although his algorithms have never been formally tested, my own experience with them as well as others anecdotal reports (Gaver, 1993b; Gygi, 2001) suggest that they can often convincingly mimic sounds produced by specific natural sources, and convey interpretable information about the source. The question, however, remains whether the sounds obtained through Gaver's or similarly developed synthesis methods convey perceptual information in the same way as natural sounds do. That is, while the synthesized sound may convey sufficient information about particular properties of the

sound source, it is an empirical question whether this information is conveyed using the same acoustic structure. Given that the acoustic structure of naturally produced sounds is by default much richer than that of synthesized ones (Jenkins, 1985; Carello et al., 2003), it would not be surprising if the acoustic basis for the perception of particular source attributes were somewhat different in each case. For instance, it is a well-established finding in speech perception research that information about stop consonant voicing can be conveyed by several different acoustic cues (Borden, Harris & Raphael, 1994). By analogy, it may be possible that information about some material properties of an object may be also specified by several different acoustic parameters.

Overall, a perceptually-oriented mathematical modeling of the sound-producing behavior of objects in the environment provides a very promising method for studying the perception of the sources of environmental sounds. Hypotheses about the physical basis of the perception of the environment through sound can be tested through the kind of synthesis algorithms developed by Gaver. This approach is useful in determining how physical source attributes are specified in the acoustic signal, and what parameters of the acoustic signal support the perception of these source attributes.

2.1.2. *Information processing approach*

The information processing approach is comprised of a number of different theoretical models that can be applied to environmental sound perception. In general, these models emphasize the processing performed on the sensory input by the perceptual system. To date, one of the most detailed and comprehensive frameworks dealing with perception of distal objects and events through sound is the Auditory Scene Analysis

(ASA) framework (Bregman, 1990). In ASA, it is proposed that all sounds entering the listeners' ears are subjected to two different kinds of perceptual processes: primitive scene analysis and schema guided processes. The primitive processes are defined as involving neither voluntary control, nor prior learning on the part of the listener. They are the auditory reincarnation of the Gestalt principles of perceptual organization of sensory information, e.g., proximity, similarity, common fate, set, continuity, and closure. Primitive processes perform initial parsing of all acoustic input to the auditory system. Parsing by primitive processes results in auditory streams that represent ways in which different elements of sensory input are grouped according to their physical and perceptual characteristics. An auditory stream is further viewed "as a computational stage on the way to the full description of an auditory event."(Bregman, 1990: p.10). The term, auditory event, while never explicitly defined by Bregman, seems to signify the end-product description of a distal environmental event in the mind of the listener. In addition to primitive processes, perceptual parsing of the acoustic input can also be performed by schemas. Unlike primitive processes, schemas are acquired, and their operations can be controlled by the individual to some extent. Schemas are employed for the perception of familiar sounds in the environment such as speech, music, and environmental sounds. The formation of schemas depends on the work of primitive processes. The relationships between different clusters of perceptual qualities provided by the primitive processes may constitute an emergent property for a given auditory event. Schemas then are higher-order representations of auditory events, which comprise not only individual perceptual qualities but also a particular set of relationships among these qualities. However, while primitive processes are instrumental in the formation of

schemas, the operations of the primitive and schema-driven processes are not seen as mutually exclusive or happening in a set order. Rather, they are viewed as different perceptual analyses of the acoustic input, where both schemas and primitive processes can determine the final outcome.

Attempts have been made to formalize and apply the ideas of the ASA framework in a related area of research known as Computational Auditory Scene Analysis (CASA). A number of computational models have been developed to test and better explicate perceptual processes proposed in ASA (Ellis, 1996; Martin, 1999; Bates, 2000). However, over a ten-year period since CASA's origin (Ellis, 1996), these attempts have met with very limited success (Bates, 2000). This may have partly resulted from assumptions about what the different components of computational systems are supposed to do at different stages of processing, without fully understanding how this can be done. This leaves the designer of such a system with little more to build on than his/her intuition, and places solutions to perceptual problems into the hands of insightful engineers rather than experimental psychologists. This state of affairs has led some CASA specialists to suspect that the fundamental questions of CASA have not been adequately formulated, and to question anew the basic acoustic information that the perceptual system exploits (Bates, 2000).

Another exemplar of the information processing approach is the model of natural computation that has been applied to environmental sound perception by Richards (1988, Wilde & Richards, 1988; Li et al., 1991). It is assumed that the perceptual system conducts sequential inference tests on the sensory input to determine the identity and behavior of distal objects. This view shares an important assumption with the direct

realist view in that the perceptual dimensions used in object recognition are based on the lawfully structured signal that emanates from the object. Because object behavior is thus lawfully represented in the signal it produces, it can be perceived by a listener with a perceptual system that has an internal model describing the systematic relationships between the properties of physical objects and the acoustic signals they produce. Thus, despite conceptual differences between Richard's model and the ecological approach, there is no clear procedural difference that can be used to contrast these two views empirically (Gygi, 2001).

2.2 Experimental paradigms and tasks

A number of experimental paradigms have been applied to studying the perception of environmental sounds and their sources. Generally, the experimental paradigms are not unique to this area of investigation, but come from research in related fields such as language processing and lexical recognition, musical instrument identification, voice quality assessment, speaker recognition, and speech perception research. Often researchers attempt to combine several paradigms in a single study, in order to achieve the most comprehensive understanding of perceptual phenomena (Ballas, 1993; Marcell, 2000; Gygi, 2001).

2.2.1. *Direct identification*

Perhaps the most widely used method in investigating the perception of environmental sounds involves direct identification. Direct identification encompasses a variety of experimental tasks and procedures. These include open response identification

(listeners are simply asked to describe what they hear), or constrained identification (listeners have to select a label from a closed set of sound names, or classify the sound based on a given set of category names). In addition, researchers can vary instructions to listeners to focus their attention on one or more specific attributes of a sound or its source such as, for instance, the length of a dropped object (Carello et al., 1998). Alternatively, listeners can be instructed to categorize the sounds they hear in a free categorization task without any specific description of the characteristics of stimuli sounds.

Several studies have attempted to establish what information is conveyed by environmental sounds by asking listeners to describe what they hear. In one early study, Vanderveer (1979) asked her subject-listeners to write a short phrase describing each of the 30 recorded environmental sounds used in the study. It was found that listeners tended to describe the stimuli in terms of the sound-producing behavior of the source objects, and referred to the auditory qualities of the sounds only when their sources could not be identified. Overall, listeners' responses showed fairly high accuracy in identifying the nature of the sources responsible for the acoustic stimuli with mean identification accuracy of 82.1% correct. The error analysis of the responses also showed that listeners tended to confuse some sounds based on similarities in their temporal patterning. For example, 'dropping a book' could be confused with 'clapping', but not with 'shuffling paper'. However, Vanderveer (1979) did not formally define temporal patterning, and did not conduct any quantitative acoustic analysis on her stimuli. Consequently, as Gygi (2001) notes, her conclusions about the acoustic structure are based primarily on individual observations.

A similar study was conducted by Gaver (1988). He asked listeners to describe in as much detail as possible each of the 17 prerecorded environmental sound stimuli. The results generally agreed with Vanderveer's (1979) findings of high identification accuracy of the source objects and events, and confirmed that environmental sounds tended to be identified primarily in terms of their sources. An examination of listeners' errors also highlighted possible reasons for perceptual confusions, and indicated ways in which the perceptual system might process environmental sounds. For example, the sound of opening and closing a file drawer was often confused with the sound of a bowling alley. This confusion adds credence to Vanderveer's idea that temporal patterning of the sound plays an important role in the identification of source events, although explicit formal measures of "temporal patterning" still have not been formulated.

These and other (Marcell et al., 2000; also see Gygi, 2001) studies that employed an open set identification task demonstrated generally high accuracy in listeners' identification of the sources of environmental sounds. The data are consistent with the claim that what listeners report about environmental sounds are primarily source objects and events, rather than auditory qualities. However, while listeners' introspections and intuitions about environmental sounds are interesting and revealing, they are idiosyncratic and difficult to quantify. They do not provide a finite set of generative perceptual dimensions and attributes, but rather a list of possible ways to interpret a given sound (Gaver, 1993a).

An attempt to address this shortcoming has been made by Marcell et al. (2000). In the first of a large series of studies, the authors obtained listeners' spontaneous descriptions of 120 environmental sounds including some snippets of music. Listeners'

descriptions of the test sounds were then tabulated in order to derive a normative set of naming and scoring guidelines for distinguishing incomplete or erroneous responses from idiosyncratically phrased "correct" responses with the same meaning. The scoring guidelines were based on the frequency of a particular verbal description given to each sound across 25 listeners. Consequently, some stimuli were scored as correctly identified when only the name of the object making the sound was given, e.g., 'cat' for cat meowing, while correct identification of other sounds required both the name of the object and its behavior during sound production to be described, e.g., 'baby crying'. This approach represents a more formal and constrained way of scoring identification accuracy. Rather than using the researcher's own judgments, which are obviously biased by the knowledge of each test sound and repeated listening to the sound during the preparation of an experiment, listeners' naming responses are taken as a baseline for determining the accuracy of identification. On the other hand, such scoring guidelines can fail to indicate whether a particular piece of information was perceived by a listener and not reported in the response, or whether it was not perceived and not reported.

In another study conducted in the same project, Marcell and colleagues asked 38 listeners to categorize each sound in a free categorization task. All category labels were then collapsed across subjects. Twenty three category names that appeared at least 33% of time in the general pool of provided category labels were selected for the next experiment. Some of the selected category names referenced the source of the sound, e.g., 'water/liquid', while others referenced the context, e.g., 'household'. The selected categories were then offered to another group of listeners whose task was to sort the same 120 experimental sounds using the 23 category names provided. Both categorization

tasks revealed that different category labels could be used to classify the majority of the sounds in the sample, although some category names clearly outweighed others in the number of times listeners selected them for a given sound. Thus, while in some cases listeners seemed to prefer one category label over others, the categories themselves were not mutually exclusive. They provided alternative and in some cases complimentary ways to describe each sound.

One limitation of direct identification studies in which listeners are presented with unmodified recordings of environmental sounds is in their reliance on the listener's introspections about test sounds. Thus, only consciously accessible aspects of the sounds can be reported. While this method may provide a variety of useful information about what listeners can report about the sounds and their sources, it is quite possible that some aspects of environmental sound perception and cognitive organization go unnoticed because they are not accessible to consciousness. Information conveyed by environmental sounds about the nature and behavior of their sources may, in principle, render certain acoustic aspects of these sounds impossible to examine consciously, as is often the case with speech perception (Repp, 1984). To compensate for this limitation, direct identification is often combined with other experimental paradigms to further investigate the effects of acoustic structure on perception.

One such paradigm consists of three main components: (1) an account of the physical properties of the sound-producing objects being examined, (2) an account of the acoustic consequences of the sound-producing behavior of these objects, (3) an account of what information about the sound-producing objects and their behavior can be perceived by the listener from the ensuing acoustic signal (Li et al., 1991; Gaver, 1993b).

The information from each of these accounts is then combined to determine the parameters of the acoustic structure responsible for the perception of particular properties of the objects of interest and their sound-producing behavior. An hypothesis about the effect of a given acoustic parameter, or a combination of such parameters, on the perception of the source objects and events is tested by systematically varying the parameters of the acoustic stimulus signal, and monitoring the effect of this variation on perception. For example, Warren & Verbrugge (1984) tested the hypothesis that information that distinguishes an object's bouncing vs. breaking after a fall is conveyed by the time-varying pattern of spectral peaks. This hypothesis was tested by making the spectral peaks in different frequency bands in the stimulus either occur at the same instances in time, or to occur asynchronously, i.e., with peaks at each frequency band having independent time courses. It was found that listeners heard the object as bouncing when the peaks occurred at the same times, while they heard it as breaking when the peaks in each frequency band had asynchronous time courses in relation to the peaks in other bands. Similarly, Li et al. (1991) successfully altered perception of the gender of a walker by manipulating the spectral slope and the spectral mode of the footsteps' signal.

This method (by and large very similar to the analysis-by-synthesis method in speech research) provides an efficient way of establishing the correspondence between the acoustic parameters of the signal and the perceived properties of the source. However, a major drawback of applying this paradigm to the perception of environmental sounds lies in the limited ability to generalize the results. The number of sound sources in the environment is very large and individual sound sources often have a unique set of perceptual characteristics. Therefore, detailed knowledge of the correspondence

between various parameters of acoustic signals and their perceptual effects cannot be easily extrapolated across different environmental sounds. For instance, the acoustic parameters that correlate with perceived hand configurations during clapping are not the same as those for objects' bouncing vs. breaking. By the same token, the kind of information listeners can obtain from claps is different from that obtained from falling objects, even in spite of some gross similarities in the sounds' waveforms (e.g., temporal patterns of energy envelopes, rapid onset). Eventually, some generalizations can be made if the number of individual environmental sounds and sources explored in this somewhat sporadic manner is sufficiently large and diverse. Nevertheless, it is not clear at present whether there can be a finite set of acoustic parameters that can be used to derive any source characteristic from an environmental sound, or whether such parameters are always specific to individual environmental sound-source pairs.

Another approach to examining the perceptually relevant acoustic structure of environmental sounds through their direct identification involves modifying sound stimuli by filtering. The researcher selectively filters out the energy in some frequency regions of the acoustic signal, while leaving some frequency regions unaffected, and monitors the effect of this manipulation on perception. In one such study (Gygi, 2001), four listeners were presented with 70 environmental sounds which were high or low pass filtered at frequency cutoff points between 300 Hz and 8000Hz in octave spaced intervals (i.e., 300Hz, 600Hz, 1200Hz, 2400Hz, 4800Hz, 8000Hz). Listeners, previously trained with unmodified sound recordings, were asked to identify each sound by selecting the most suitable name for it from the list of response options provided. All response options in the list were names of possible sound sources. Overall, identification performance

tended to be better for highpass than for lowpass filtered sounds, and generally reached near perfect scores after several days of testing with the filtered stimuli. The greater contribution of higher frequency regions to identification accuracy suggests that more information about sound source identity is conveyed in higher frequencies. The crossover point, that is the frequency point which divides the spectrum into two equally intelligible halves, was determined to be at or slightly above 1200Hz. This value lies at the lower end of reported crossover points of equal intelligibility found in studies with filtered speech (French & Steinberg, 1947). It was also found that the general advantage of higher frequency regions for sound source identification, nevertheless, varied across individual sounds. Such sounds as 'Thunder' and 'Waves' were described as being "very difficult to identify even at the more moderate highpass filters, such as F(cutoff) = 600Hz and 1200Hz, even though most of the other sounds were easily recognized (Gygi, 2001: p. 95)." On the other hand, identification of other sounds (e.g., 'Glass breaking') was severely degraded even by moderate low pass filtering. Thus, despite the overall greater contribution of high frequency information to source identification, there is considerable variability among individual sounds in terms of the frequency regions bearing most of the perceptually relevant information. Because it is unknown if the experimental stimulus set can be considered representative of environmental sounds in general, it is possible that the high frequency advantage might be lost if a different set of environmental sounds were similarly tested.

In order to further examine the contributions of specific frequency regions, Gygi (2001) conducted another filtering study. The same 70 environmental sounds were bandpass filtered into 6 filter bands logarithmically spaced between 150Hz and 9600 Hz.

Eight new listeners, highly familiar with the unmodified sounds, were tested with the filtered versions of the stimuli. Five presentations of the filtered stimuli presented in random order were conducted over a period of nine days. Overall, all subjects showed the same pattern of identification performance across the six filter band conditions. The stimuli filtered with the two lower bandpass settings which spanned the frequency region between 150Hz and 600Hz, were identified the worst, with mean percent correct of 31% and 51%, respectively. In contrast, the mean identification performance for the stimuli with the four highest filter settings, spanning the frequency region between 600Hz and 9600Hz, was 70%-80%. These results are in general agreement with those from the highpass and lowpass filtering study in terms of the advantage of higher frequency regions for the sound source identification. However, as in the previous study, some sounds were best perceived only under a particular filtering condition and did not follow the general pattern in which better identification performance was found in higher frequency bands.

The bandpass filtering approach to studying the perception of environmental sounds preserves the dynamic frequency information within the passband. However, the role of dynamic frequency variation in the perception of the sources of environmental sounds may also be examined with the spectral smearing of the signal. Spectral smearing can be performed in ways that are very similar to some of the current vocoder-based cochlear implant processors. This approach was taken by Gygi (2001) in an experiment where (1) all frequency information was removed from the acoustic signal but the overall energy envelope was preserved, and (2) the energy envelopes of six frequency bands

derived from the original stimuli were used to modulate an equal number of noisebands which were then combined and presented to listeners.

The methodology of the first signal processing scheme consisted of modulating white noise (that by definition has a uniform spectrum) by the energy envelope of each environmental sound. The signal modulated noise stimuli were presented to 8 experienced listeners who previously took part in the bandpass filtering study. The original stimuli were the same 70 stimuli as in the filtering study. Overall, these experienced listeners demonstrated 46% identification accuracy. While significantly lower than performance accuracy in most of the bandpass filtered conditions, this result shows that some perceptually relevant information about sound sources still remained in the modulated-noise signals devoid of any spectral detail. This result, however, was somewhat confounded because of the listeners' high familiarity with the sounds from the previous experiments. Thus, at least some of their identification judgments could have been based on such arbitrary cues as overall duration of the sound, or the dynamic energy pattern specific to that particular token of the sound, rather than another acoustic token of the same sound-producing event. This may partially account for the finding that another group of 8 naïve listeners who did not participate in the previous studies performed very poorly in an identification task with the same stimuli, achieving only 13% identification accuracy. However, on the second day, after they were presented the original unmodified stimuli, their identification performance improved to 23%. While 23% is still a very low identification score, the exposure to the signal-modulated noise stimuli on the first day of testing and the original unmodified sounds of the second day apparently resulted in a small, but statistically significant improvement in identification accuracy. It remains

unclear whether performance would have improved had listeners not heard the original unprocessed sounds, and listened only to the signal-modulated noise stimuli on the first and the second day of testing.

The signal processing method adopted for the six channel signal-modulated noise condition is reminiscent of the signal processing performed by the classic vocoder (Duddley, 1940), and is currently utilized in a number of cochlear implant processing strategies. After the original signal is filtered into frequency bands, dynamic energy envelopes are obtained for each band. These envelopes are then used to modulate white noise, which is then filtered according to the same filter specifications as were used to filter the original signal. The modulated noise bands thus obtained are combined into a single waveform and presented to listeners. This method of processing preserves the overall dynamic energy pattern specific to each frequency band; however, it destroys fine spectral variation present within individual frequency bands of the original signal. Thus, the spectral information within each band is smeared and caricatured in the dynamic energy pattern.

In this experimental condition, a new group of eight listeners with no previous experience with the unprocessed sounds, achieved 36% identification accuracy on the first day of testing. This score was significantly higher than that obtained by the listeners with the signal modulated noise stimuli even on the second day of testing, suggesting that the six channel stimuli contained substantially more perceptually relevant acoustic information about the sources of the test sounds. However, this score was still lower than that obtained by the experienced listeners even without any spectral information at all in the single channel condition. On the second day of testing with the six channel signal-

modulated noise stimuli, subjects' performance improved dramatically to 66 % correct, surpassing even that of the experienced listeners who had been tested with the single channel stimuli of the previous condition. These results indicate that while increasing the degree of spectral degradation adversely affected identification of the sources of environmental sounds, with some experience, listeners were able to improve their identification performance, possibly by making use of some previously unexploited acoustic information. An interesting follow up to these results would involve investigating whether the improvement in identification performance would generalize to similarly processed but not previously tested environmental sounds.

Another interesting finding of this project (Gygi, 2001) was that some sounds with strong temporal patterning were identified fairly well even when no spectral information was present in the single channel condition. Sounds like 'helicopter', 'rain', and 'gallop' apparently needed no spectral information to be identified with above 50% accuracy even by naïve listeners on the first day. This suggests that time-varying energy information is sufficient for accurate identification of the sources of some environmental sounds but not others. However, Gygi's efforts to account for listeners' overall identification performance using a variety of temporal- and spectral- based measures of the signals met with a limited success. Taken together, various spectral and temporal properties of the signals accounted for 60% of the variance for the experienced listeners, and 39% of the variance for the naïve listeners in the single channel condition. Spectral and temporal properties also accounted for 50% of the variance for the naïve listeners in the six-channel condition.

2.2.2. *Qualitative ratings of perceptual and cognitive attributes*

Qualitative ratings have been used to determine the underlying perceptual dimensions of environmental sounds and their sources. Typically, listeners are asked to rate each sound they hear on a variety of scales representing various physical properties or perceptual qualities of the sound or its source. These rating scales, in most cases, are pre-selected by the researcher. When choosing anchor labels for the rating scales, the researcher attempts to include a broad range of potentially relevant perceptual dimensions. For example, labels may be selected from semantic (e.g., sharp - dull, dark - light), emotional (e.g., pleasant - unpleasant), or physical (e.g., small - large, heavy - light) dimensions. The choice of labels for rating scales is, of course, arbitrary to a large extent, which poses a danger of missing the actual perceptual dimensions used by listeners. On the other hand, a broad range of labels increases the chances of capturing the relevant dimensions.

Listeners' judgments on rating scales typically undergo further statistical analysis to find the perceptual dimensions underlying their judgments of the stimuli on different scales. These analyses may include multidimensional scaling (MDS), multiple regression, factor analysis, and principal component analysis. The analyses effectively reduce the initial number of perceptual dimensions represented by each rating scale to fewer implicit dimensions that account for most of the variance in listeners' ratings. At that point, the researcher's task becomes explaining the nature of the dimensions obtained by relating them to cognitive, perceptual, or physical properties of the stimuli.

Early auditory perception research that utilized qualitative ratings avoided the use of environmental sounds. Instead, researchers concentrated on electronically generated

sounds which either did not have an identifiable source, e.g., pure tone complexes, (Von Bismark, 1974a; 1974b), or related to the source in arbitrary ways, e.g., sonar sounds, (Solomon, 1958). These studies explored semantic involvement in auditory processing, and, in some cases, demonstrated a relationship between acoustic structure and semantic attributes. For instance, von Bismark (1974a; 1974b) found that ratings of electronically generated stimuli on the "sharp - dull" dimension correlated significantly with the location of the spectral centroid, while ratings on the "compact - scattered" dimension could indicate whether the stimulus was harmonic or inharmonic. However, only a small proportion of variance could be accounted for by the semantic or acoustic factors investigated, and the implications of the results for naturally occurring sounds remained unclear.

Later studies that used qualitative ratings focused on environmental sounds. Bjork (1985) investigated perceived auditory and emotional attributes of 15 environmental sounds. In an effort to decrease the information unrelated to the sounds' auditory and emotional qualities, the stimuli were presented to listeners backwards. However, according to the author, even after reversal, the sources could still be identified. Factor analysis revealed that five factors could account for 91% of the variance, with both emotional and auditory components associated with the individual factors. Factor 1 seemed to encompass various emotional qualities such as 'pressing', 'tense', 'unpleasant', etc. Factor 2 was correlated with 'sharpness' and 'pitch' as well as 'boring – interesting' ratings. Factor 3 was correlated with loudness. Factor 4 and 5 were correlated with 'simplicity' and 'fast-slow' ratings respectively. However, the

author warned that the "results may be restricted to the persons involved and the nature of the stimuli and scales used... (page187)"

Ballas (1993) took a more naturalistic approach to exploring the perception of environmental sounds, and did not manipulate sound recordings. He utilized rating scales in Experiment 3 of his comprehensive study. A sample of 41 sounds selected from sound- effects records were rated on 22 scales. The scales referred to auditory (e.g., loudness, timbre), semantic (e.g., interesting – boring), and other cognitive and perceptual factors that had been found valuable in the previous studies (e.g., ease of naming the sound, or imagining preceding and following sounds in a hypothetical sequence). Importantly, some of the scales assessed the perception of the sources of sounds as well without asking the listeners to identify the sources. For instance, listeners were asked "How easily does the mental picture of the person or object which caused this sound come to mind?", or "How many events can you think of that could have caused this sound?" A principal component analysis of the ratings revealed that three factors accounted for 87% of the variance in the rating data. These factors were separately composed of the ratings encompassing sound identifiability, timbral properties, and the sound's uniqueness for a given category. In order to further investigate whether test sounds could be grouped based on the rating judgments, the author also conducted a higherarchical cluster analysis using the factor scores. The analysis revealed four major clusters of sounds: (1) mostly water -related sounds (e.g., 'Water drip', 'Toilet flush', 'Oar rowing'), (2) signaling sounds (e.g., 'Doorbell', 'Car horn'), (3) modulated noise sounds (e.g., 'Lawn mower', 'Sawing'), (4) transient sounds (e.g., 'Light switch', 'Cork pop'). However, the general descriptive labels provided by the author to the four clusters

(e.g., mostly water sounds, signaling sounds, etc.) applied only to some of the sounds in the clusters. Other sounds did not fit well under their cluster names. For instance, 'Cigarette lighter' was clustered with water sounds, while 'Power saw' belonged to the signaling sounds.

More recently, Kidd & Watson (1999, described in detail in Gygi, 2001) obtained listeners' ratings of 145 environmental sounds which came from CD sound effects libraries. The sounds were rated on 20 scales representing auditory and emotional factors as in previous studies. Some scales also represented source property characteristics (e.g., large – small). A principal component analysis performed on the rating data revealed that four factors accounted for 88% of the variance. These factors were interpreted as reflecting harshness, size, quality, and complexity of the sounds. On the other hand, the 13 acoustic characteristics of the sounds, some of which were based on waveform statistics (e.g., centroid, skewness) and some on listeners judgments (e.g., pitch), when combined could account for only 60% of the variance. This result was interpreted as indicating that either more elaborate and complex acoustic factors might account for much of the variance, or that other non-acoustic cognitive factors play a major role in listeners judgments.

2.2.3. *Similarity ratings*

Similarity ratings have also been used to find perceptually relevant characteristics of environmental sounds. Unlike qualitative ratings, however, listeners are not given any indication of what the underlying perceptual dimensions might be, and, instead, have to base their judgments on internally derived dimensions they find useful in distinguishing

two sounds. This leads to clustering of the stimuli based on their similarity - dissimilarity with each other. The researcher's task is again to explain why stimuli tend to cluster the way they do, or, in other words, why a given stimulus is more similar to some stimuli than to others. The reasons for clustering are similarly sought in perceptual, cognitive, and physical aspects of the stimuli and their sources. The similarity rating paradigm appears to have more ecological validity than the qualitative ratings method because the perceptual dimensions are not pre-selected by the experimenter, but are, rather, listener based. However, that does not mean that such internal dimensions are uniform across listeners or generalizable to sounds not included in the stimulus set.

One of the early environmental sound similarity studies was conducted by Cermak & Cornillon (1976). The authors were interested in what makes recorded traffic noise segments more or less acceptable and more or less similar for listeners. Twenty-two female listeners attended to pairs of 1-minute traffic noise segments. Their judgments were analyzed using a multidimensional scaling procedure. The obtained similarity-dissimilarity results indicated that the differentiation of traffic sounds was made on the basis of two main criteria. The first was the perceived intensity of the sounds, which was highly correlated with the mean energy equivalent sound level of the stimuli (dBA scale). The second criterion included information about the sound source (e.g., the perceived proximity of the source, the proportion of the time a truck or bus could be identified). However, despite some success in being able to explain the perceptual basis of the similarity judgments, the authors warned that the results were not necessarily applicable to other traffic sounds which may include motorcycles, hot-rod cars, and defective vehicles. Thus, only limited knowledge was gained about how

perceptual organization of even a single subclass of environmental sounds, i.e. traffic noise, is structured in the mind of the listener.

Recently a similarity rating study was carried out by Gygi and colleagues (Gygi, 2001). One hundred stimuli were randomly arranged in pairs and presented to listeners. Half of the stimulus set contained alternative sound tokens of the same source event as the other half. The tokens were selected to be "as different acoustically as possible (based on the experimenter's subjective judgments) while still representing the same event (p. 86)." In addition to judging the acoustic stimuli, listeners also judged the sound similarity based solely on their memory or knowledge of the sounds, and source similarity based on memory of the source. In the two later conditions, listeners were not presented stimulus sounds. Instead, they were presented with pairs of stimulus names. For the condition where listeners heard acoustic stimuli, it was found that two tokens of the same event were consistently judged more similar to each other than to any other stimulus that referenced a different source event. The 2-dimensional MDS solutions for each of the three conditions were also similar in many ways. Across all three conditions, stimuli tended to cluster (loosely) into (1) animal sounds (e.g., 'rooster', 'cat') , (2) quasi-periodic temporally patterned sounds (e.g., 'footsteps', 'ping-pong'), and (3) water-based sounds (e.g., 'toilet', 'water pouring'). However, as in Ballas' (1993) cluster analysis based on qualitative rating data, the clusters were not clearly defined or homogeneous in any of the three solutions. A multiple regression solution for over twenty acoustic measures accounted for 66% of the variance in the ordering on the first dimension of the MDS solution. The mean saliency (which is a measure of confidence in the pitch of the signal), standard deviation of the spectrum, and the standard deviation of centroid

velocity gave the largest correlations with the first dimension of MDS. Some acoustic measures also correlated with the second dimension, although the correlations were much weaker than those for the first. When combined, they accounted for 48% of the variance. These results were interpreted as suggesting that, while there was some acoustic basis for similarity judgments on a diverse sample of environmental sounds, other, possibly, semantic factors also played a role in their perceptual organization.

2.3 Stimuli

The selection of environmental sound stimuli for a given study is largely a function of the goals of the study, the practical constraints associated with obtaining necessary sounds, and, to a great extent, the skill and creative talents of the researcher who designs the study. In general, the stimuli employed in empirical research of environmental sound perception fall into four large categories.

2.3.1. *Single type of environmental sound*

Some researchers (Warren & Verbrugge, 1984; Repp, 1987; Li et al.,1991) have investigated the perception of a single type of sound-producing object or event such as bottle's bouncing vs. breaking, handclaps, or footsteps. Environmental sound tokens are presented to listeners in isolation, and listeners' perception is investigated only with respect to some specific attribute of the sound-producing object or event. As previously mentioned, these studies demonstrated that individual environmental sounds can convey a variety of information about the characteristics of corresponding objects and events (Repp, 1987; Carello et al., 1998; Freed, 1989). It has also been shown that information

about the object obtained from the acoustic signal it generates is not limited to the object's basic physical properties. Listeners can also identify more complex object characteristics such as a walker's gender from the sound of his/her footsteps (Li, Logan, Pastore, 1991). However, the knowledge obtained in such studies may be specific to the type of sound source being investigated. While these studies are invaluable in highlighting specific aspects of perceptual processing involved in source identification, their results cannot be confidently extrapolated to the perception of other types of sound sources.

2.3.2. *Environmental sound inventories*

Others (Vanderveer, 1979; Ballas, 1993; Gaver, 1988) have investigated the perception of sets of environmental sounds. The test sounds are selected to correspond to a wide variety of sound-producing objects and events, but they are presented to listeners separately. Listeners' responses are collected for a single environmental sound at a time. Research with many distinct isolated environmental sounds has led to the following general conclusions. (1) Listeners tended to identify and categorize environmental sounds in terms of events and objects, rather than sound auditory characteristics (Gaver, 1988, Vanderveer, 1979). (2) For the environmental sounds tested, listeners were generally accurate in assigning a given sound to its source object and event (Vanderveer, 1979; Gaver, 1988; Ballas, 1993; Marcell et al., 2001). (3) Mistakes were not random, but corresponded to similarities in the physical properties of the confused objects and events, and/or similarities in the temporal patterning in the acoustic signals (Vanderveer, 1979; Gaver, 1988). (4) Listeners' ease and accuracy in object and event identification was

related to both the acoustic properties of the sound and the frequency of its occurrence in the environment (Ballas, 1993). (5) Even when listeners were asked about sound characteristics other than those of source objects and events (e.g., cold - warm, pleasant - unpleasant), their responses tended to cluster around both the physical properties of the sources and the acoustic characteristics of the signals (Ballas, 1993; Gygi, 2001).

2.3.3. *Temporally ordered sequences of environmental sounds*

A different approach consists of presenting listeners with sequences of environmental sounds (Ballas & Howard, 1987; Fowler, 1990). Several environmental sounds are presented to listeners in a temporal sequence. After listening to such a sequence, listeners are asked to interpret the sounds in the sequence in terms of source objects and events (Ballas & Howard, 1987), or judge a specific property of such an object or event (Fowler, 1990). Research with sequences of sounds revealed that (1) the perceptual interpretation of a sound-producing object or its property represented by an individual environmental sound can change as a function of the preceding and following sounds in the sequence (Ballas & Howard, 1987; Fowler, 1990), and (2) the structured patterns of environmental sounds based upon "logical relationships" among individual sounds (e.g., valve open, drip, short noise, water flush) are easier to learn than unstructured ones (e.g., clang, short noise, drip, valve open) (Ballas & Howard, 1987).

2.3.4. *Mixtures of environmental sounds*

An approach to stimulus selection and perceptual testing which more closely resembles the way environmental sounds occur in the world outside the laboratory

examines listeners' perception of environmental sounds in acoustic mixtures. The objects and events causally related to individual environmental sounds in such a mixture may exist and behave independently from each other, and their acoustic signals may overlap in time and frequency. For instance, as I am writing these lines I can hear a car driving by my window, footsteps of a passerby on the street, the buzzing of an annoying fly, and my fingers striking the keys – all happening at the same time. Curiously, to date the only investigation of human ability to recognize sound-producing objects and events in mixtures, that I am aware of, composed exclusively of environmental sounds has been conducted in order to evaluate the performance of a computational system designed to handle this task (Ellis, 1996). Not surprisingly, human listeners outperformed the computational system in identification of objects and events from the acoustic mixture of environmental sounds.

2.3.5. *Remarks about stimulus selection*

Several concerns need to be addressed regarding the selection of environmental sounds for perceptual testing. One arises from the prevalence of environmental sound studies that use isolated sounds presented to listeners "out of context." In natural environments, sounds emanating from different sound-producing physical sources rarely, if ever, occur in isolation. Most often they are embedded in mixtures with other sounds. Thus, if the ultimate goal is to understand how humans actually perceive environmental sounds in their natural environment, not in an acoustically tightly controlled one, one future research goal is to explore the perception of environmental sounds in mixtures. The preference given to isolated environmental sounds in perceptual research is partially

the result of the reductionist view that aims to understand the whole as the sum of its parts. In many ways, this approach is warranted. Investigating the perception of environmental sounds in isolation effectively reduces the number of variables to account for in a study, and makes research questions somewhat easier to tackle. The point, however, is that the researcher needs to be aware that understanding the perception of isolated environmental sounds is not the same as understanding the perception of these sounds in more natural acoustic settings.

Another important aspect of stimulus selection concerns the quality and ecological validity of the sounds presented to listeners. Leaving aside the synthesized or purposely edited sounds, there still remains the question of whether sound recordings that are commonly substituted for actual original sounds may impose some constraints on how listeners perceive them. Even provided that the physical characteristics of the sounds are accurately and faithfully preserved in recordings, it is still no more than an assumption that removing such sounds from their original environmental context does not somehow influence their perception (Truax, 2001). For instance, hearing a mooing cow in the rural area may not be quite the same as hearing the same sound through the earphones in a sound-proof booth, as far as what it tells the listener about the state of affairs in the environment. That is, when the mooing cow is heard over earphones or loudspeakers in the booth, no typical adult listener would expect the cow to be around. Given that the function of environmental sounds is to inform the listener about the state of his/her environment, it is conceivable that perceptual processing of the same sound may be different in the natural and unnatural contexts. While there are very convincing reasons why cows, diesel trucks, and airplanes are not brought to the perceptual

laboratory, the issue of the perception of natural vs. recorded sounds seems to warrant consideration at least during experimental design.

2.4. Listeners

By default, the majority of listeners in environmental sound research are young normal-hearing college-age adults. Only limited use has been made so far of the listening skills of other potentially important listening populations such as (1) hearing impaired listeners and users of various electronic aids to hearing such as hearing aids and cochlear implants, (2) children and elderly adults, (3) expert listeners, (4) animals, and (5) machine listeners.

CHAPTER 3

CURRENT STUDY: ENVIRONMENTAL SOUND SOURCE IDENTIFICATION WITH A VARYING NUMBER OF SPECTRAL CHANNELS

A convenient and useful method for achieving the practical and theoretical aims of the current project comes from previous research on speech perception with simulated cochlear implants. In order to avoid various confounding factors that result from surgical implementation of cochlear implants, factors which cannot be adequately controlled across implanted listeners, e.g., electrode placement, depth of insertion, and electrode cross talk, a simulated model of a cochlear implant can be used (Loizou, 1998). In the simulation, acoustic signals are processed in the manner similar to that used by some cochlear implants. However, the processed output is acoustic, rather than electric. The acoustic output can be presented to normally hearing adults in order to determine how source identification changes as a function of number-of-channels used. In addition to not being invasive, and thus posing no health risks to listeners, this approach makes it possible to evaluate the perceptibility of the output of a cochlear implant processing strategy under "ideal" circumstances without any influence of confounding factors which may have negative effects on perception (Loizou, 1998).

The major goal of the experiment described below was to determine how systematic changes in the degree of spectral resolution, represented by a varying number of channels, affects the identification of the sources of a large sample of familiar environmental sounds. Another goal of the experiment was to determine similarities and differences among individual environmental sounds in the number of channels required for source identification.

3.1. Stimuli and Signal Processing

Original sounds used in this study were 60 familiar environmental sounds with easily identifiable sources. They were selected from an extensive royalty-free CD library of environmental sounds commonly used in television and radio postproduction (SoundIdeas: General 6000). The original sounds were stereo recordings stored in the form of digital .wav files sampled at 44,100 Hz and quantized at 16 bits. For the purposes of this study only left channel signals from each selected original file were used. The sounds were also shortened in duration when the length of an original sound was greater than 10 seconds. Care was taken to preserve any onset and offset information. Deletions were made only on redundantly repeating portions of each edited sound (e.g., car engine idling). Stimuli were further processed to include a linear 15ms amplitude ramp at the beginning and the end of each file.

The selected sounds represented a variety of sound sources and included (a) human and animal vocalizations and bodily sounds, (b) mechanical sounds of interacting inanimate solids, (c) water related sounds, (d) aerodynamic sounds, and (e) electric and acoustic signaling sounds (see Appendix A for a complete listing of sounds arranged by category). Many of these sounds were also similar to the types of sounds used in previous environmental sound research with normal-hearing (Ballas, 1993; Marcell et al., 2000; Gygi, 2001) and hearing-impaired listeners (Owens, Kessler, Raggio & Schubert, 1985). Response labels denoting sound sources were selected from open set responses given to each sound by ten native English speakers during pilot testing.

All 60 selected sounds were digitally processed to obtain six different number-of-channel conditions. These conditions were chosen to provide a large range of spectral

resolution (Figure 1). Thus, the pre-selected files were processed using 2, 4, 8, 16, 24, and 32 channels. The signal processing algorithm and parameters were adapted from previous research that investigated minimal spectral resolution required to understand speech (Dorman et al., 1997). Each original sound was filtered into a given number of frequency bands equal to the desired number of channels using 6^{th} order Butterworth filters. The frequency ranges of individual bands were logarithmically spaced between 300 Hz and 5500 Hz (Appendix B). An energy envelope of each band was obtained by, first, half-wave rectifying the signal, and, then, low-pass filtering it with a 2^{nd} order Butterworth filter with a 160 Hz frequency cutoff. Next, the envelope of each band was excited with white noise, and, filtered using the same band-filter settings as had been used to obtain the same frequency band from the original signal. The energy level in each noise band thus obtained was equated with the energy level in the corresponding band of the original signal. Finally, all the modulated noise bands were added together and low-pass filtered at 5000 Hz with a 6^{th} order elliptic filter. This represented a processed output signal.

In addition to 60 processed stimuli in each channel condition, another set of 60 stimuli was obtained by bandpass filtering the original sounds between 300 – 5500 Hz using a 6^{th} order Butterworth filter, and then additionally filtered with a 6^{th} order low-pass elliptic filter with a 5000Hz cutoff. This set of stimuli with minimal spectral degradation provided a baseline condition for comparing listeners' performance with spectrally smeared stimuli. The matching of the bandwidth of the unprocessed and processed stimuli was done to insure that any differences in identification performance were due to spectral smearing. The removal of spectral energy outside 300 to 5000 Hz region insured

that the differences in identification performance between stimuli with and without spectral smearing were not due to the information in the spectrum which was outside the frequency range of the channels.

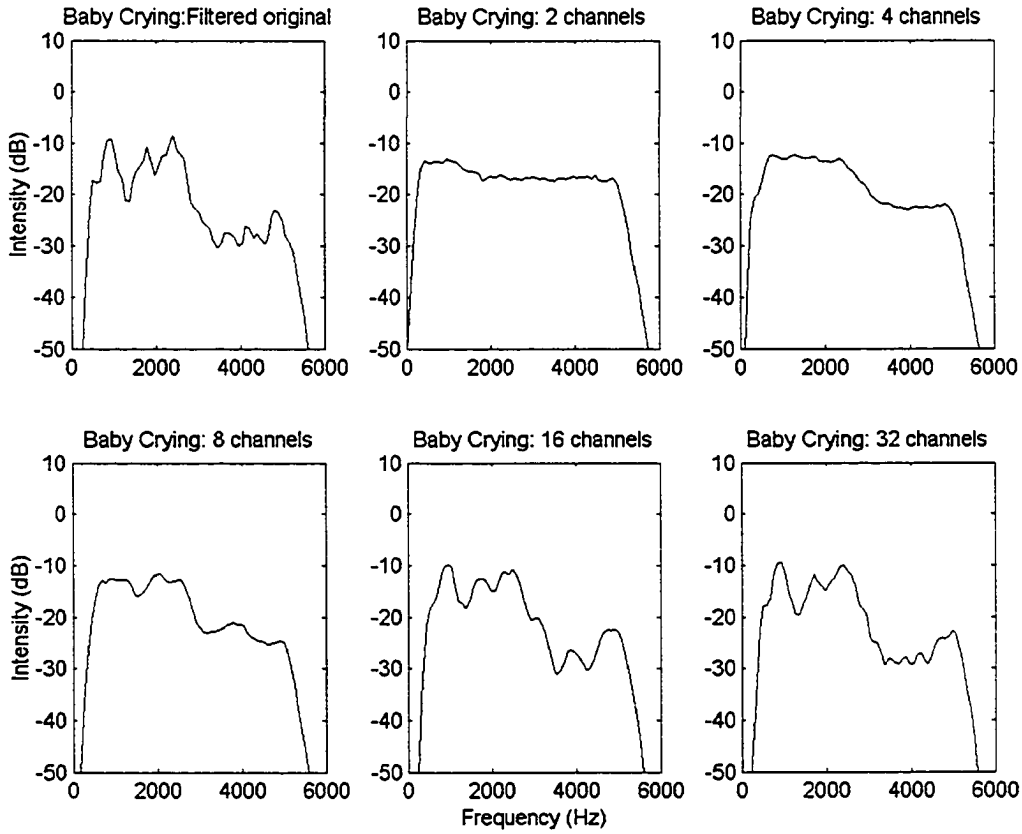

Figure 1: An example of variation in spectral resolution for different number-of-channels conditions. Each panel represents a long-term average spectrum of the sound of baby crying for the filtered original and different channel conditions.

3.2. Design and Procedure

A Latin square design was used to insure that learning effects would not influence listeners' performance in different channel conditions. Listeners were divided into six

groups of ten. Each group of listeners heard ten different stimuli from every channel condition, followed by all 60 broadband filtered original sounds. Thus, each listener heard a total of 120 stimuli. Sounds were randomly assigned to each set of ten stimuli, but the types of sounds in each set of ten were never repeated in any other channel condition within the same group of ten listeners (Table I). None of the listeners heard a processed token of the same original type of environmental sound more than once.

Table I: An illustration of the Latin square design used in the study.
(Group = A listener group of ten participants, Condtn. = condition, chan. = channels, S = stimulus)

	Listener group					
Condtn.	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
2 chan.	S1 – S10	S11 – S20	S21 – S30	S31 – S40	S41 – S50	S51 – S60
4 chan.	S11 – S20	S21 – S30	S31 – S40	S41 – S50	S51 – S60	S1 – S10
8 chan.	S21 – S30	S31 – S40	S41 – S50	S51 – S60	S1 – S10	S11 – S20
16 chan.	S31 – S40	S41 – S50	S51 – S60	S1 – S10	S11 – S20	S21 – S30
24 chan.	S41 – S50	S51 – S60	S1 – S10	S11 – S20	S21 – S30	S31 – S40
32 chan.	S51 – S60	S1 – S10	S11 – S20	S21 – S30	S31 – S40	S41 – S50

Processed stimuli were randomized for each new listener within each set of ten stimuli. The presentation order of channel conditions also randomly varied for each listener to avoid possible order effects. All 60 unprocessed stimuli were randomized each time they were presented to a new listener.

Listeners were tested in a closed-set response format. Initially, they were provided with a description of the categories of environmental sounds they would hear, and given verbal examples of possible electronic distortions to environmental sounds such as found in telephone and radio. Next they were presented with nine tokens of processed and unprocessed environmental sounds not included in the test stimulus set. The listener's task was to select the best name for the source of each sound he/she heard from a list of 60 labels displayed on the screen in an alphabetical order.

Prior to experimental testing, listeners were trained to enter their responses using the experimental interface, and familiarized with each response label and its position on the screen. During the training session, the name of each of the 60 stimuli was displayed on the screen, and each listener had to respond to it as if the stimulus were presented aurally.

Each listener was tested individually in a quiet room. Stimulus presentation was conducted using a custom-built computer program. Original sound files were pre-processed according to the specifications of each channel condition and presented to participants via Sennheiser HD 414 headphones at a comfortable listening level which ranged from 88 dB rms for the most intense sound to the 67 dB rms for the least intense one[1]. After the presentation of a new stimulus, every listener had an option to hear the

[1] To obtain the range of presentation levels, two 1KHz sinewaves were synthesized. Their rms levels were matched to those of the test sounds with the highest and the lowest rms levels corrected for silent intervals. The sinewaves were then delivered through the playback system that was used in the study and recorded through a DB – 100 occluded ear simulator with a DB – 050 ear canal extension mounted on the KEMAR manikin (Knowles Electronics) with small pinnae using Etymotic ER11 microphones and preamplifier. The differences in voltage between a calibration signal with a known sound pressure level and each of the sinewaves were used to determine the sound pressure levels of the two sounds with the minimum and maximum rms levels.

stimulus once again before indicating his/her response. A typical experimental session took approximately one hour.

3.3. Participants

Sixty-five listeners took part in the experiment. Data from five listeners were excluded from further analysis because these listeners' accuracy on the unprocessed original sounds was below 90% correct. Of the remaining sixty participants, twenty were male and forty were female. Thirty-seven of the sixty participants were native speakers of English, and twenty-three were fluent but non-native speakers of English. All sixty participants passed a hearing screening at 25 dB HL on at least one ear. The age of the participants ranged from 18 to 47 years, with the mean participant age of 29 years.

CHAPTER 4

RESULTS

Data were analyzed to determine (a) the overall source identification accuracy for the spectrally smeared and band-limited original sounds; (b) similarities and differences in identification performance for individual sounds across channel conditions; (c) grouping of individual sounds by the number of channels required for source identification; (d) clustering of the stimuli and responses in a psychological space based on the listeners' identification errors.

4.1. Identification accuracy on the band-limited original sounds

First, source identification scores were analyzed for the band-limited original sounds to determine whether listeners were familiar with the sounds used in the experiment, and could identify their sources. The average identification accuracy of the 60 band-limited sounds across all 60 listeners was 97% correct, with a standard deviation of 2.6%. Apparently, limiting the sounds' frequency range to a 300 – 5500 Hz bandwidth had little effect on identification performance. The accuracy analysis of individual sound tokens revealed that 57 of the 60 sound stimuli were identified at 90% correct or above. The lowest mean accuracy score for any undistorted stimulus was 85% ('water bubbling'). The high accuracy scores on original sounds suggest that all listeners were familiar with the vast majority of the test sounds when presented without major distortions. Moreover, there were no differences in mean identification performance of males and females on the unprocessed sounds. Both scored at 97% correct, with a standard deviation of 2.7% in each case. There was, however, a small difference in mean

identification accuracy between native and non-native English speakers, with natives scoring 98% and non-native participants scoring at 96% on unprocessed sounds, with standard deviations of 2.3% and 2.7%, respectively.

The differences in responses of native and non-native English participants were further examined in a Median test to find out whether the two groups of listeners differed significantly from each other in identification accuracy. The results of the Median test revealed a significant difference ($\chi^2 = 7.22$, $p < 0.01$) between native and non-native English participants. However, follow-up Median tests performed in each channel condition failed to show any significant differences between the two groups. Therefore, data from both groups were treated together in all subsequent analyses because (a) the main questions of this study addressed the perception of environmental sounds under the conditions of decreased spectral resolution (i.e., channel conditions), in which responses of native and non-native participants did not significantly differ, and (b) performance of individual listeners in either group was higher than 90% on the original sounds, indicating that both native and non-native listeners were able to identify most of the sound sources from the band-limited stimuli.

4.2. Identification accuracy across channel conditions

In order to determine how varying degrees of spectral resolution affected listeners' identification performance, identification accuracy was assessed across listeners and across stimuli for each channel condition. Cross-listener analysis was based on the mean percent correct values of individual listeners in each channel condition (Table II), while cross-item analysis was based on the mean percent correct values of individual

sounds in each channel condition (Table III). Overall, the mean identification accuracy tended to improve as the number of channels increased (Figure 2). However, accuracy scores in individual channel conditions were distributed differently when examined across listeners or across individual sounds. Even though listeners did not respond to the same stimuli in each channel condition due to the Latin square design used in the study, listener accuracy scores seemed to be nearly normally distributed around the mean in each channel condition as indicated by the small differences between condition means and medians (Table II). This result demonstrates that the population of listeners responding to the stimuli was indeed homogeneous, despite the fact that some listeners responded to different stimuli in the same channel condition. Therefore, parametric statistical analysis could be performed on accuracy rates across listeners.

Table II: Identification accuracy across listeners for each channel condition (all 60 sounds).

	Channel condition						
Measure	2 chan. % corr.	4 chan. % corr.	8 chan. % corr.	16 chan. % corr.	24 chan. % corr.	32 chan. % corr.	Originals % corr.
Mean	32	42	59	66	67	65	97
Median	30	40	60	70	65	60	98
Standard deviation	13	13	16	18	19	14	3

Table III: Identification accuracy across stimuli for each channel condition (all 60 sounds).

	Channel condition						
Measure	2 chan. % corr.	4 chan. % corr.	8 chan. % corr.	16 chan. % corr.	24 chan. % corr.	32 chan. % corr.	Originals % corr.
Mean	32	42	59	66	67	65	97
Median	20	40	70	70	80	70	98
Standard deviation	33	34	31	32	30	33	4

A univariate between – subject analysis of variance (ANOVA) revealed a highly significant main effect of channel condition, $F(5, 20) = 27.4$, $p < 0.001$. This confirmed the observation that increasing spectral resolution improved listener identification performance. However, follow-up independent pairwise t -tests demonstrated that differences in performance between adjacent channel conditions were significant ($p < 0.05$) only up to sixteen channels; listeners' performance did not significantly change with further increases in the number of channels.

The ANOVA also showed a significant main effect of stimulus set $F(5, 20) = 7.7$, $p < 0.001$. This indicates that there was a variation in the response accuracy patterns elicited by different stimulus sets even for the same channel condition. This was most likely the result of (a) random assignment of stimuli to conditions in the Latin Square

design, and (b) the relatively small number of stimuli per condition for every Latin square stimulus set. As will be shown later in this chapter, different environmental sounds seem to require different numbers of channels to be identified. Ten stimuli may not be sufficient to provide a balanced random set of different kinds of environmental sounds. The main effect of stimulus set may thus be due to sampling error in the stimuli. That is, some stimulus sets might have had more sounds requiring only few channels for identification, while others might have included more sounds requiring a large number of channels.

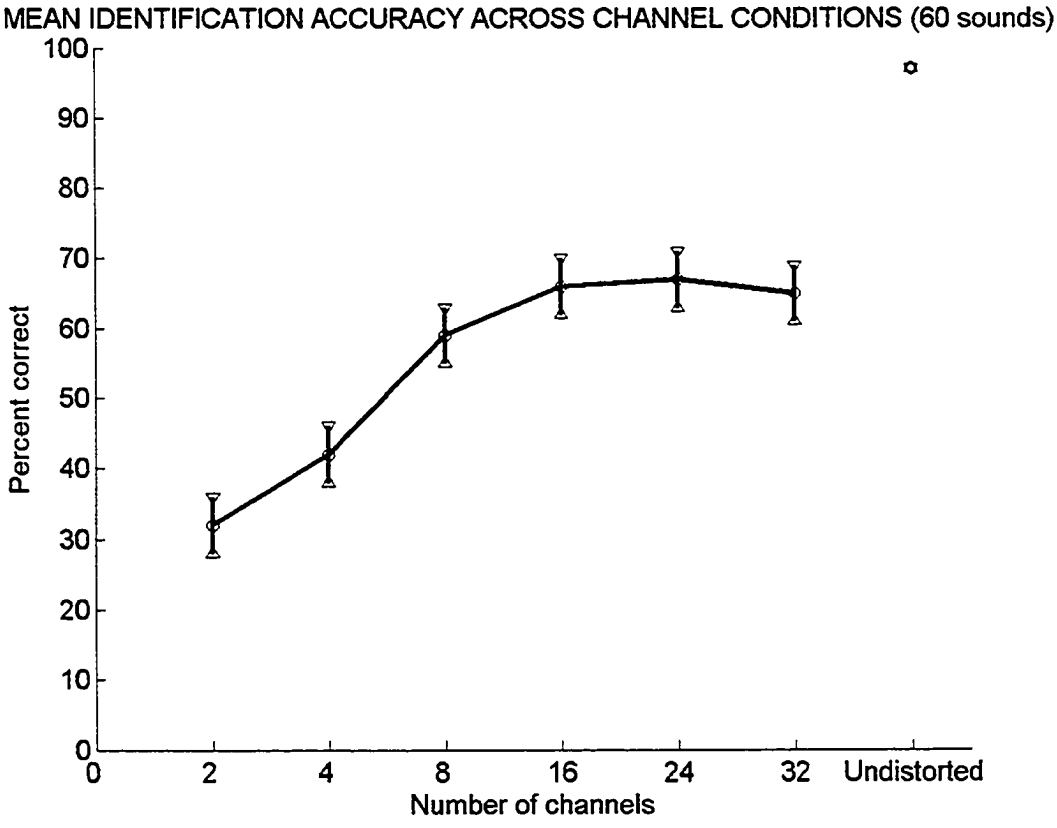

Figure 2: Stimulus mean identification accuracy as a function of the number of channels. Error bars represent the standard error of the mean.

In contrast to accuracy scores across listeners, the distributions of accuracy scores across sounds were less symmetrical and more heterogeneous (Table III). The distribution of the accuracy scores was positively skewed in the 2-channel condition and negatively skewed in 8- and 24-channel conditions. In addition, the variance in response accuracy across stimuli was markedly higher than that across listeners. These findings indicate a higher degree of variability in accuracy scores across degraded sound stimuli than across listeners. They also suggest that, within different channel conditions, some sound sources were identified much better than others, and that the identification performance varied widely among individual sounds. Because the distributions of accuracy scores were skewed in several channel conditions and the assumption of homogeneity of variance could not be maintained, nonparametric analysis methods were used to assess the significance of the effect of channel condition on identification using sounds as the sampling variable.

An analysis of accuracy across sounds in each channel condition further confirmed the previous result obtained across listeners that identification performance tended to improve as the number of channels increased. A Friedman test of rank ordered accuracy values across six channel conditions was significant ($\chi^2 = 76.2$, $p < 0.0001$), suggesting that an overall improvement in performance with increasing number of channels was not due to chance. However, an improvement in identification performance was also evident only up to sixteen spectral channels. Wilcoxon Signed Rank tests on adjacent condition pairs revealed a significant change in performance accuracy between 2 and 4 channels ($Z = -3.5$, $p < 0.01$), four and eight ($Z = -2.5$, $p < 0.02$), and 8 and 16 channels ($Z = 2.4$, $p < 0.02$). Performance did not change significantly between 16 and

24 channels, 24 and 32 channels, or 16 and 32 channels. On the other hand, even though the change in performance accuracy was significant between 8 and 16 channels, the magnitude of the mean performance change was small (i.e., 7%), while the median accuracy scores in these conditions were both equal to 70%. This suggests that, functionally, the increase in spectral resolution from 8 to 16 channels provided very little benefit.

Table IV: Eleven sounds that were not perceived at 70% correct or more in any condition listed in ascending order of highest accuracy reached at any channel condition.

	Channel condition					
Sound source names	2 chan. % corr.	4 chan. % corr.	8 chan. % corr.	16 chan. % corr.	24 chan. % corr.	32 chan. % corr.
Train whistle	0	0	0	0	0	0
Blowing nose	30	30	0	10	0	0
Car horn	0	0	0	0	0	30
Zipper	0	10	40	20	10	10
Clearing throat	0	10	10	50	10	10
Man drinking	0	0	50	20	20	0
Water draining	0	0	40	20	40	50
Burp	0	10	40	20	60	40
Cork pop	40	60	50	60	60	20
Wind blowing	30	30	50	60	50	50
Woman sighing	0	0	10	50	50	60

Another important finding was that the overall identification performance for any of the channel conditions never exceeded 67% for these stimuli. In fact, 11 of the 60 sounds (Table IV) were never identified correctly by more than 60% of the listeners in any experimental condition, while all the undistorted original tokens of these sounds were identified correctly in the 90% - 100% correct range. For five of these 11 sounds, identification accuracy tended to improve as the number of channels increased (i.e., 'burp', 'car horn', 'water draining', 'wind blowing', 'woman sighing'), and for one sound (i.e., 'train whistle') it remained at 0% percent correct throughout all channel condition. This result suggests that the identification of these six sounds may require an even higher spectral resolution than that achieved with thirty-two spectral channels. On the other hand, for the remaining five sounds, identification accuracy tended to decline with increases in spectral resolution, suggesting that there might have been some unfavorable distortions introduced during the processing of the stimuli.

4.3. Decline in identification accuracy with increasing spectral resolution

An unexpected finding of the study was that the identification accuracy of 19 sounds, including five which were never identified correctly above 60%, decreased as the number of spectral channels increased (Table V). The criterion for classifying a sound as having a declining identification accuracy pattern was a drop in performance of 30% or more relative to the maximum accuracy reached at a lower number of channels. Even though the 30% error criterion was by necessity an arbitrary one, it was chosen to insure that an accuracy decrease was not likely to be accidental. Specifically, the 30% or more decrease meant that 3 fewer listeners could identify the source of a sound correctly at this

Table V: Nineteen sounds with declining mean accuracy (%) at higher number of channels. The sounds are listed in order of descending magnitude of the decline in identification. Percent magnitude of the overall decline for each sound is shown in bold.

Sound source names	2 chan.	4 chan.	8 chan.	16 chan.	24 chan.	32 chan.	Decline
Clapping	100	100	90	60	40	0	**100**
Horse trotting	40	100	90	100	50	20	**80**
Biting, chewing apple	20	70	40	70	60	10	**60**
Page turning	20	60	90	40	30	30	**60**
Footsteps	70	70	60	70	30	20	**50**
Shoveling dirt	60	60	70	30	20	20	**50**
Typing on keyboard	100	100	100	100	80	50	**50**
Clearing throat	0	10	10	50	10	10	**40**
Cork pop	40	60	50	60	60	20	**40**
Snoring	10	70	50	20	30	30	**40**
Thunder	70	90	70	70	50	50	**40**
Blowing nose	30	30	0	10	0	0	**30**
Door closing	60	80	80	90	90	60	**30**
Ice cubes into glass	30	20	80	40	70	50	**30**
Jackhammer	20	40	70	50	50	40	**30**
Man drinking	0	0	50	20	20	0	**30**
Stone splashing water	40	50	100	80	90	70	**30**
Water bubbling	0	0	80	60	100	70	**30**
Zipper	0	10	40	20	10	10	**30**

condition as compared to the highest accuracy obtained for this sound in any lower number-of-channels condition.

A decline in identification accuracy for the 19 sounds was especially evident in conditions with high numbers of channels. For 16 of these 19 sounds, the highest accuracy was obtained with 2, 4, or 8 channels. Only 3 sounds reached the highest accuracy values at 16 or 24 channels. None of the 19 sounds had a better identification accuracy in the 32-channel condition compared to any other condition with a lower number of channels. This suggests that, for these 19 sounds, increases in spectral resolution beyond 8 channels led to little or no improvement in source identification. Those sounds that reached the highest identification accuracy with only 2 channels (i.e., 'clapping', 'typing on keyboard', 'footsteps', and 'blowing nose') showed a more or less gradual decline as the number of channels increased. The cross-channel accuracy pattern of the remaining sounds was "Λ" shaped, with identification accuracy rising to a peak value, and than falling as the number of channels increased. Consequently, the overall magnitude of a drop in identification accuracy was most pronounced for those sounds that reached the highest identification accuracy with the spectral resolution of 8 channels or fewer.

The finding of declining accuracy with improved spectral resolution was surprising because the band-limited tokens of these sounds, with one exception of 'water bubbling', were identified correctly at 90% or more. In principle, a decline in identification accuracy should not have been the case because increasing the number of channels should lead to improved spectral resolution of the sound, while preserving its temporal structure. However, a comparison of waveforms for these sounds across the

number of channel conditions indicated that their temporal structure was, in fact, altered during processing. Figure 3 presents an example of temporal smearing observed in the waveforms (left – hand panel) and in the spectrograms (right – hand panel) of the affected sounds at higher number of channels. The top part of the figure shows a 3 second segment of the waveform and spectrogram of 'footsteps'. The middle part shows the waveform and spectrogram of the same segment after processing with 2 channels, and the bottom part shows its waveform and spectrogram after processing it with 32 channels. As can be seen in the waveforms, the amplitude peaks are distinct and are preserved between the original and the 2-channel version of the sound. On the other hand, the peaks are smeared in the 32-channel version of the sound compared with the other two. This observation is further confirmed through examining the spectrograms of these sounds across channel conditions. As can be seen in the spectrograms, the onset of energy in lower frequency regions of the 32-channel version of the sound is delayed relative to the 2-channel version and the original.

The reason for the temporal distortions are thought to lie in the group delays introduced by the filters during the signal processing performed on the stimuli. Because filters had different frequency ranges and a relatively high order of coefficients, the group delays across filters varied significantly across channel frequencies (see Appendix C). Low frequency filters had particularly small frequency ranges and slow release times when the number of channels was high. Thus, when the output signals from all spectral channels were combined to produce a stimulus, the frequency dependent components among channels became asynchronous relative to their relationships in the original signal. This led to the temporal smearing observed in the stimuli at high numbers of channels.

On the other hand, informal listening suggests, when group delays are compensated for during the addition of individual channel signals, or when lower order filters with smaller group delays are used, the sound sources of the nineteen sounds with decaying accuracy become much more identifiable.

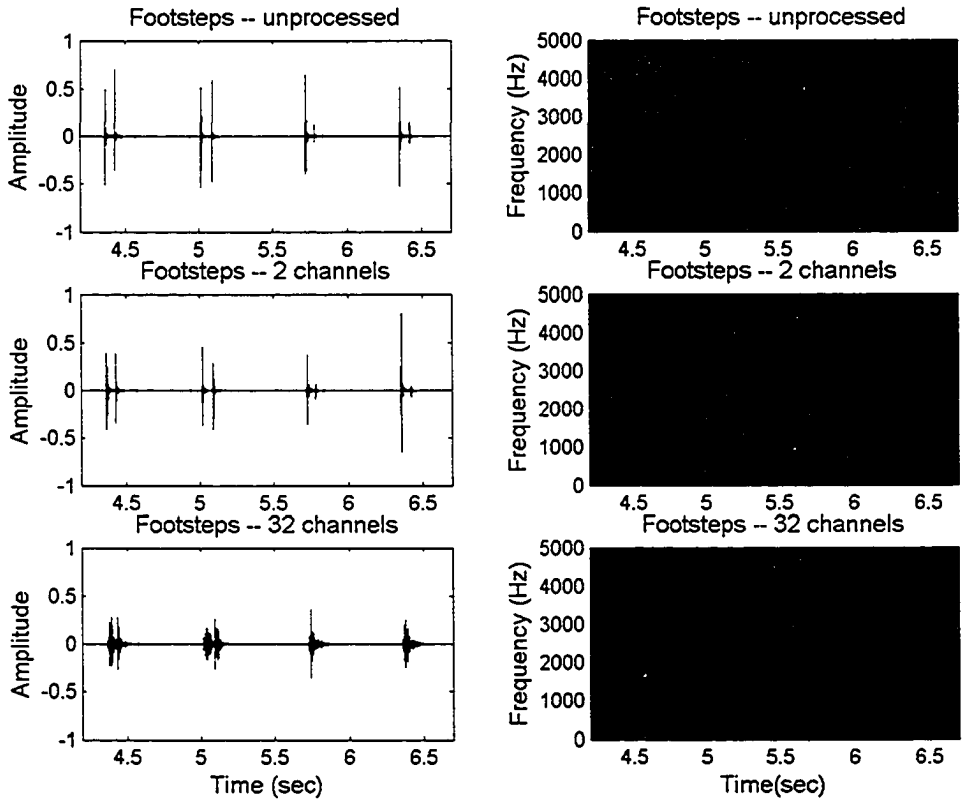

Figure 3: An example of temporal smearing as seen in the waveforms (left – hand panel), and spectrograms (right – hand panel) of 'footsteps'.

The finding of declining identification of almost a third of all sounds at higher number of channels raised doubts as to whether the lack of improvement in identification performance beyond 16 channels was not due to the decline in identification accuracy of these sounds. To address this concern, mean percent correct values were once again obtained for the 41 sounds that did not show a decline in accuracy. The accuracy scores of 30 of these 41 sounds monotonically improved as the number of channels increased,

while the performance on 11 of these sounds remained asymptotic across all channel conditions, using a previously specified 30% error margin. As can be seen in Figure 4 and Table VI, when only sounds with non-declining identification patterns were included, listeners' mean identification performance continued to improve by the average of 10% between 16 channels and 32 channels. Follow-up Wilcoxon Signed Rank tests on adjacent channel conditions, using sounds as the sampling variable, showed that there was a significant improvement in identification of sounds between 16 and 32 channels ($Z = -2.06$, $p < 0.04$). However, there were no significant improvements between either 16 and 24 channels, or 24 and 32 channels. This result confirms that the slope of identification accuracy function tends to flatten beyond 16 channels even when the sounds that show a decline in accuracy at higher number of channels are excluded from the analysis.

Table VI: Identification accuracy across stimuli for each channel condition (41 sounds with a nondecreasing accuracy pattern).

	Channel conditions						
Measure	2 chan. % corr.	4 chan. % corr.	8 chan. % corr.	16 chan. % corr.	24 chan. % corr.	32 chan. % corr.	Originals % corr.
Mean	30	37	57	71	76	81	98
Median	20	30	60	80	90	90	98
Standard deviation	33	33	33	33	26	23	3

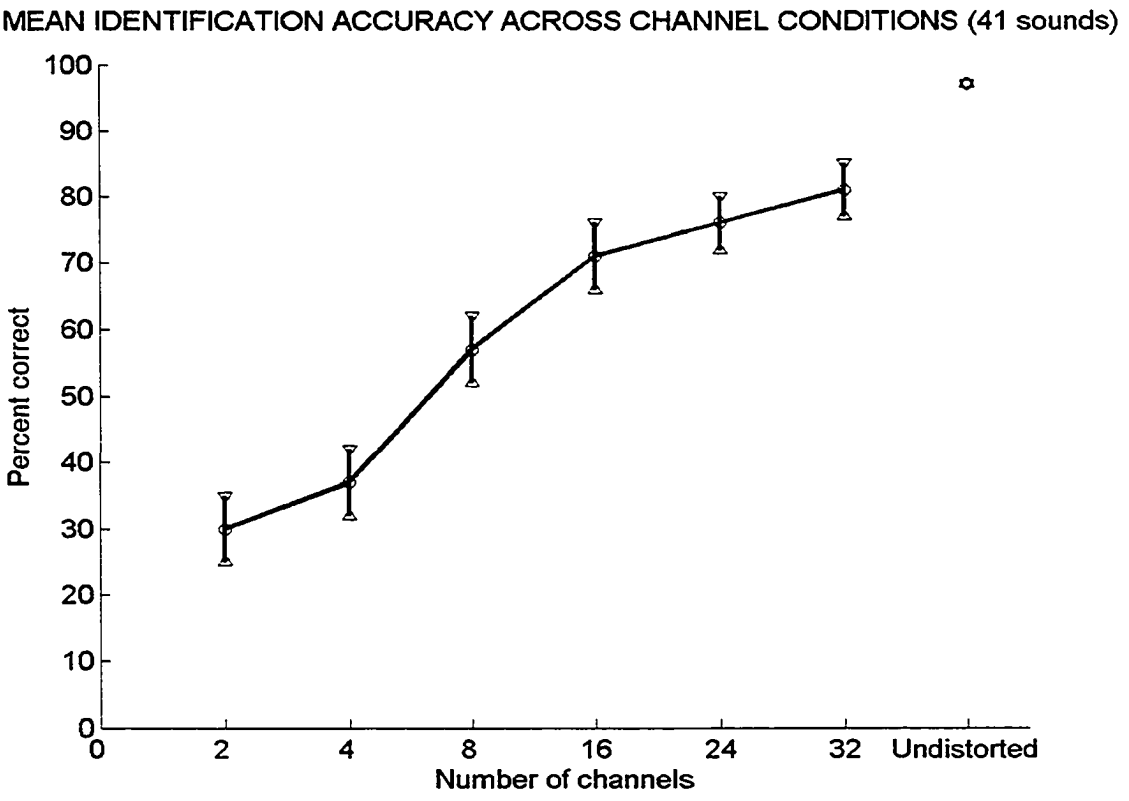

Figure 4: Stimulus mean identification accuracy as a function of the number of channels (41 sounds). Error bars represent the standard error of the mean.

4.4 Grouping of sounds by the number of channels required for source identification

A further insight into the nature of information available in each channel condition was gained by grouping the test sounds by the number of channels required for sufficiently accurate source identification. Sufficiently accurate source identification was functionally defined as identification accuracy of at least 70% correct. The 70% correct identification criterion was chosen because it was sufficiently high to insure that the majority of listeners identified the source of a given sound correctly (i.e., seven out of ten listeners), and because it was sufficiently low to permit the inclusion of the majority of the test sounds. As was previously noted, listeners' mean accuracy across sounds did not

exceed 67% correct in any channel condition. Therefore, a higher accuracy threshold would require many sounds to be discarded from this analysis.

Table VII: Sounds identified at 70% correct or more grouped by channel condition.

2 channels	4 channels	8 channels	16 channels	24 channels	32 channels
clapping	biting apple	birds chirp	baby crying	airplane	church bell
footsteps	bowling	brush teeth	doorbell	child cough	man yawn
heartbeat	car starting	camera	gargling	cow moo	
helicopter	clock ticking	barking	rooster	glass break	
machine gun	dog panting	horse neigh	toilet flush		
ping-pong	door closing	ice cubes			
pool balls	horse trotting	jackhammer			
busy signal	pouring soda	man panting			
thunder	snoring	page turning			
train motion		shovel dirt			
typing		siren			
whip		splash water			
		water bubble			
		water drip			
		water running			
		woman laugh			
		sneezing			

Individual sounds were grouped by the lowest number of channels in which their identification accuracy reached at least 70% correct (Table VII). As can be seen, stimuli perceived with 70% or higher accuracy with 2 and 4 channels are primarily sounds whose unprocessed tokens could be characterized by temporally patterned brief energy bursts (e.g., 'footsteps', 'clapping', 'horse trotting'), single impact sounds (e.g., 'pool balls colliding', 'bowling', 'door closing'), or broadband relatively slow energy fluctuations ('thunder', 'car starting', 'snoring'). Sounds whose sources were identified correctly with 8 channels represent a mixture of more temporally patterned sounds with small energy variation across spectrum (e.g., 'jackhammer', 'shoveling dirt'), and sounds whose unprocessed tokens contain several time-varying narrow band components in the spectrum (e.g., 'birds chirping', 'horse neighing', 'siren', 'water bubbling', 'woman laughing'). Many of the sounds perceived correctly at higher number of channels than 8 may also be described as having a dynamic resonant structure (e.g., 'baby crying', 'cow mooing', 'man yawning'). Although some of these sounds also seem to have distinct patterns of energy change over time (e.g., 'doorbell', 'child coughing', 'glass breaking'), these patterns alone did not provide a sufficient cue to the identity of the source. Instead listeners appeared to rely more heavily on the spectral characteristics of the sounds, which became available only at higher numbers of channels. Thus, to a first approximation, it seems plausible to conclude that sources of temporally patterned sounds with broadband energy spectrum (e.g., 'helicopter', 'machine gun') required fewer spectral channels to be identified than sources of sounds containing dynamic narrow band components (e.g., 'rooster crowing' or 'cow mooing', or 'church bells'). A more precise differentiation of sounds requiring low vs. high spectral resolution may become possible

after conducting a comprehensive acoustic analysis of the stimuli in each channel condition and the original undistorted sounds.

Table VIII: Cross channel correlations (Spearman R) in identification performance.

	2 chan.	4 chan.	8 chan.	16 chan.	24 chan.	32 chan.	Originals
2 chan.	1	0.86*	0.48*	0.36*	0.18	0.03	0.01
4 chan.		1	0.54*	0.40*	0.15	-0.01	-0.07
8 chan.			1	0.66*	0.56*	0.34*	0.11
16 chan.				1	0.78*	0.67*	0.21
24 chan.					1	0.78*	0.25
32 chan.						1	0.21
Originals							1

* correlation is significant at the 0.01 level (two-tailed)

These findings suggest that the nature of information about sound sources available at low spectral resolution (i.e., 2 and 4 channels) may be different from that available to listeners at higher spectral resolution. To infer the extent to which the information in each channel condition may be similar or different to that in other channel conditions, a correlational analysis was performed on the identification accuracy of individual sounds for each channel condition (Table VIII). Because accuracy scores were not normally distributed, Spearman correlations were obtained on ranked accuracy scores across channel conditions. As can be seen in Table VIII, accuracy scores in 2 and 4 channel conditions were strongly and significantly correlated with each other.

Correlations of 2 or 4 channel conditions with 8 or 16 channel conditions were also significant, but had a lower magnitude. On the other hand, accuracy scores in the 8 and 16 channel conditions correlated significantly with scores in the 24 and 32 channels, which in turn significantly correlated with each other, but did not correlate with 2 or 4 channel stimuli.

The distribution of Spearman correlations on identification accuracy suggests that stimuli presented in 2 and 4 channel conditions contained highly redundant information about corresponding sound sources. This conclusion also applies to stimuli presented with 24 and 32 channels. Stimuli in 8 and 16 channel conditions seemed to share some similarities with stimuli with lower number of channels and some with stimuli with higher number of channels. On the other hand, correlations between low channel conditions and high channel conditions were not significant and had very low magnitude. This suggests that there were considerable differences in the accuracy of responses to the stimuli between the low and high channel conditions. Taken together with the grouping of sounds by channel conditions, the results of the correlational analysis support the conclusion that stimuli identified with relatively high accuracy with a low number of channels share many perceptual characteristics and, probably, acoustic parameters, while they differ from sounds identified with a high number of channels.

This interpretation is further confirmed by examining correlations among response frequencies across channel conditions (Table IX). For this analysis, the frequency of each of the 60 response alternatives was determined across all six channel conditions. Because the distribution of response frequencies closely approximated the normal distribution, Pearson product-moment correlations were obtained. These

correlations indicate that the distribution of listener response preferences was similar within the low channel conditions and within the high channel conditions. However, listener response preferences were different between low and high channel conditions as indicated by the low-magnitude, nonsignificant correlations between them. The 8 channel condition seems to be in the middle as far as response preferences are concerned, being moderately correlated with both low and high channel conditions.

Table IX: Cross channel correlations (Pearson product - moment) in response frequencies.

	2 chan.	4 chan.	8 chan.	16 chan.	24 chan.	32 chan.	Originals
2 chan.	1	0.87*	0.50*	0.10	-0.05	-.10	-.02
4 chan.		1	0.57*	0.20	0.06	-0.05	-0.05
8 chan.			1	0.63*	0.57*	0.45*	0.01
16 chan.				1	0.80*	0.74*	0.14
24 chan.					1	0.88*	0.30
32 chan.						1	0.25
Originals							1

* correlation is significant at the 0.01 level (two-tailed)

The response frequency correlations are consistent with the differences in modal responses derived from response frequencies. For instance, while the modal response in the 2 and 4 channel conditions was 'thunder', the modal responses in higher channel conditions were always water based sounds such as 'stone splashing water' at 8 channels,

'toilet flushing' at 16 channels, and 'water draining' at 24 and 32 channels. The differences in modal responses across channel conditions demonstrate general response preferences that participants had when listening to stimuli in a given condition. That is, based on the examination of modal responses, it appears that stimuli in the 2 and 4 channel conditions sounded more thunder-like to listeners, while the stimuli in higher channel conditions sounded more water-like. These general response preferences may reflect the differences in spectral structure of sounds in low and high channel conditions due to the changes in spectral resolution.

4.5 Error analysis

Source identification errors were analyzed in order to determine perceptual characteristics of spectrally degraded stimuli and psychological dimensions underlying the perception of sound sources. First, a 60 x 60 stimulus-response frequency matrix was obtained for each channel condition. Every row of such a matrix contained information on what responses were given to a single stimulus, and how often each response was given (regardless of how many of these responses were correct or incorrect). Similarly, every column of this matrix contained information on what stimuli elicited a given response, and how often. Unfortunately, due to the sparsity and the asymmetry of the matrices across conditions, it was difficult to obtain any meaningful error patterns or trends directly from the frequency data. A suitable statistical procedure for the direct analysis of stimulus-response matrices directly could not be found.

Second, the data in every matrix were transformed to obtain similarity measures among responses and, separately, among the stimuli. A procedure to measure similarity

among responses and among stimuli was developed. A response similarity index (RSI) was derived for every possible response pair in the following way.

- Let *A* and *B* be a pair of response labels.
- Let *a* be the frequency of response *A*, and *b* be the frequency of response *B* when responses *A* and *B* are both given to a single stimulus.
- Let *T* be the total number of responses given to all stimuli.
- Then, $RSI_{(A \& B)} = (a / b) * (\{a + b\} / T)$ for a single stimulus.
- The total $RSI_{(A \& B)} = \sum [(a / b) * (\{a + b\} / T)]$ across all stimuli.

A stimulus similarity index (SSI) was derived using the same procedure, but with *A* and *B* representing a pair of stimuli, and *a* and *b* representing the frequencies with which the two stimuli elicited the same response across the complete response set. It can be seen that the first term of the equation, i.e., (a / b), evaluates the inter-relationship of A and B (the largest of the two frequencies was always used as the denominator). If *a* is small and *b* is large, *A* and *B* are not likely to be highly related. If, however, *a* and *b* are the same, then *A* and *B* are highly related because they are equally likely response labels for a given stimulus, or stimuli that elicited the same response, in the case of SSI. The second term of the equation assesses how much of the total response space *A* and *B* are covering together. If both *a* and *b* are small, they cover a small amount of the total response space. If the sum of *a* and *b* is large, *A* and *B* can account for a larger portion of the total response space, and their total RSI or SSI value becomes larger.

Overall, RSI measures were obtained from the information contained in the rows of the original stimulus-response matrix, while SSI measures are obtained from the information contained in the columns of the original matrix. Because different responses

could be given to the same stimulus, and different stimuli could elicit the same response, the RSI and SSI measures are conceptually independent from each other and provide two different kinds of information. RSI measures indicate how similar two response labels are in the mind of the listener in a channel condition. On the other hand, the SSI measures how similar two stimuli sound to listeners in a given condition. Thus, the RSI can be viewed as a measure of proximity in abstract mental representations of sounds, while the SSI can be viewed as a measure of proximity in perceptual qualities of actual sounds.

Both RSI and SSI measures derived from each original stimulus-response frequency matrix resulted in two new matrices for each original one. One, the RSI matrix, represented symmetrically arranged similarity measures among all response pairs, while the SSI matrix represented symmetrically arranged similarity measures among all stimulus pairs in a given condition. This manipulation resulted in six RSI and six SSI matrices across all channel conditions. Unfortunately, the matrices still could not be meaningfully compared with each other because, as the number of channels increased, listeners made fewer errors. Consequently, matrices in different channel conditions were affected by differences in the number of errors that listeners made (i.e., the number of nonempty cells), and had unequal densities. This resulted in unbalanced RSI and SSI matrices in terms of the total number of data points they contained. Therefore, in order to be able to interpret the information contained in the response and stimulus errors, each group of six matrices was averaged to obtain a single RSI and a single SSI matrix.

Data in each of the two matrices were subjected to an Alscal multidimensional scaling (MDS) procedure in order to examine how responses and stimuli were distributed

in a two-dimensional perceptual space (Figure 5 & 6). Distances for MDS solutions were derived directly from RSI and SSI by dividing one by each similarity value. Zeros in the data were replaced with very small values of 0.0000001 that were several orders of magnitude less than

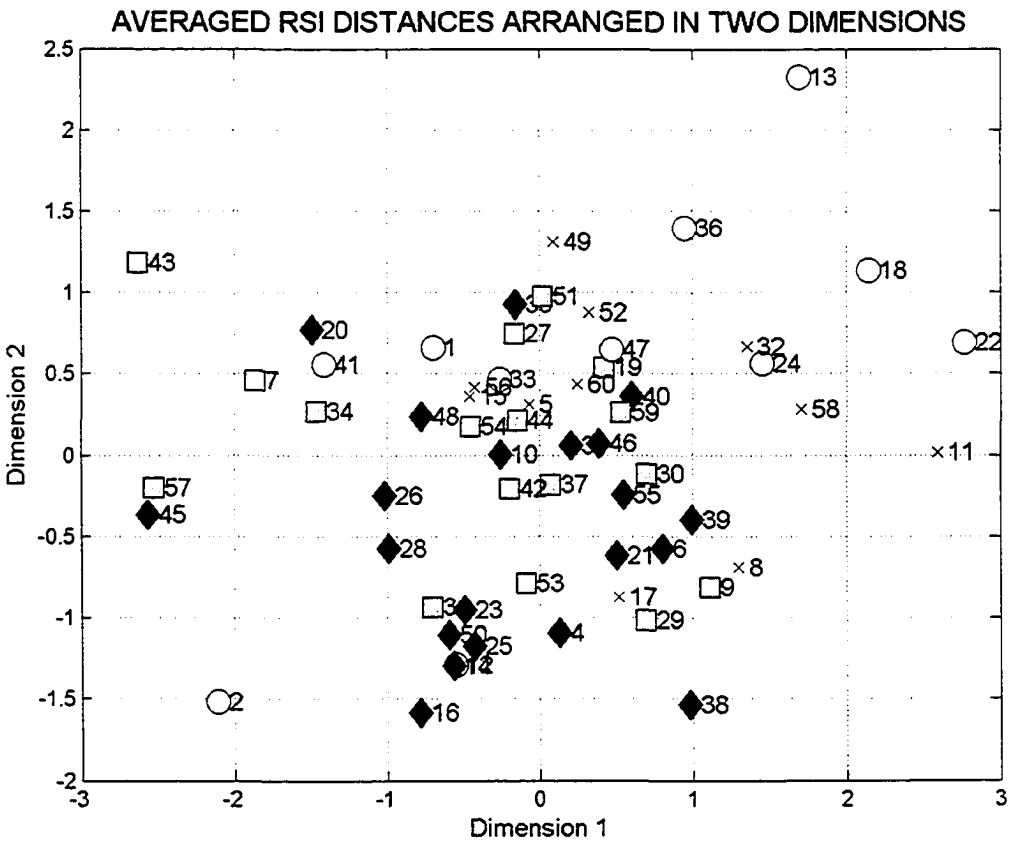

Figure 5: Averaged RSI distances arranged in two dimensions (RSQ = 0.81). Sound names corresponding to each number on the figure are listed in Appendix D. Legend: Black diamonds represent sounds identified at or above 70% correct with 2 or 4 channels; Gray squares represent sounds identified at or above 70% correct with 8 channels; White circles represent sounds identified at 70% correct or above with 16, 24, or 32 channels; x-marks represent sounds that were not identified at or above 70% correct in any of the channel conditions.

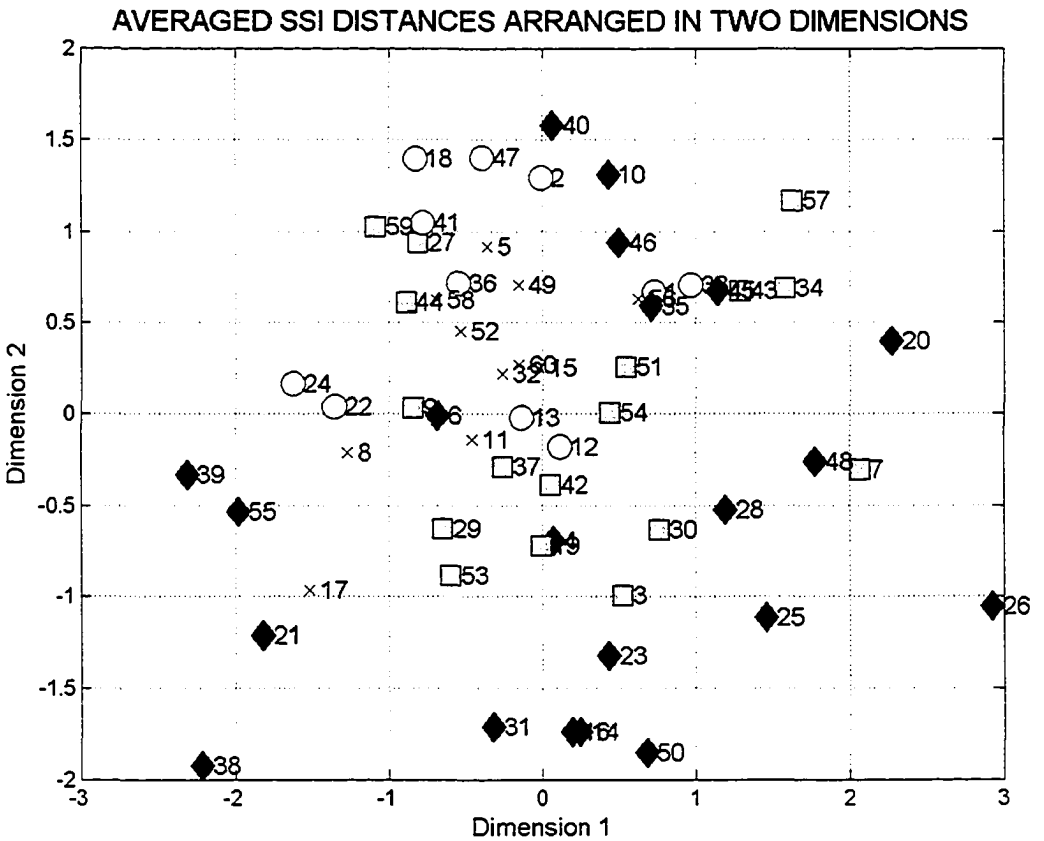

Figure 6: Averaged SSI distances arranged in two dimensions (RSQ = 0.73). Sound names corresponding to each number on the figure are listed in Appendix D. Legend: Black diamonds represent sounds identified at or above 70% correct with 2 or 4 channels; Gray squares represent sounds identified at or above 70% correct with 8 channels; White circles represent sounds identified at 70% correct or above with 16, 24, or 32 channels; x-marks represent sounds that were not identified at or above 70% correct in any of the channel conditions.

the smallest RSI or SSI value. These manipulations were based on the assumption that the highest similarity index reflects the smallest perceptual distance and vice versa. Despite the unavoidable distortions introduced by these somewhat crude transformations

of the original error frequency data, the two MDS solutions indicate certain patterns in the grouping of stimuli and responses. As can be seen from the MDS solutions obtained from both stimuli (Figure 6) and the response data (Figure 5), reoccurring broadband transients such as 'typing'[50], 'footsteps'[23], and 'heartbeat'[25], tend to cluster together towards the bottom of Dimension 2 on both RSI and SSI solutions. On the other hand, sounds with dynamic resonant spectral structure such as human and animal vocalizations (e.g., 'man yawning' [36], 'cow mooing' [18]), water sounds (e.g., 'water bubbling' [51], 'water draining' [52]), and signaling sounds (e.g., 'siren' [43], 'doorbell' [22], 'church bells' [13]) are mostly located closer to the top of Dimension 2. Dimension 1 is harder to interpret, and it may represent a composite of several underlying dimensions.

Furthermore, Figures 5 and 6 also show that sounds identified with accuracy of 70% correct or above with low number of channels (i.e., 2 or 4) tend to occupy different areas of the RSI and SSI spaces than sounds identified with high number of channels (i.e., 16, 24, or 32). Sounds requiring a medium number of channels (i.e., 8) tend to be positioned between the high and low number of channel sounds. Although these clusters too are not well defined spatially, they seem to reflect some perceptual segregation of sounds requiring different degrees of spectral resolution. Thus, they support a previous interpretation (see section 4.4) that sounds requiring a low number of channels are perceptually similar to each other, while they differ from sounds requiring a high number of channels for accurate identification.

A quantitative assessment of the relationships of the perceptual spaces for stimuli and responses revealed that the ranked coordinates of data points on Dimension 1 of the

RSI solution correlated significantly (Spearman R = - 0.74, $p < 0.01$) with ranked data coordinates on Dimension 1 of the SSI solution. Similarly, data coordinates on Dimension 2 of the RSI solutions correlated significantly (Spearman R = 0.60, $p < 0.01$) with data coordinates on Dimension 2 of the SSI solution. This analysis suggests that the perceptual spaces for stimuli (SSI) and responses (RSI) were related to each other. Finally, to explore whether the differences observed among the sounds in the multidimensional space are related to the variability among sounds in the number of channels needed for sufficiently accurate source identification, stimulus and response coordinates on Dimension 1 and Dimension 2 were rank ordered and correlated with ranked number of channels values which produced the best identification performance for each sound. The only significant correlation on Dimension 1 was for the RSI solution and was of very low magnitude (Spearman R = -0.26, $p < 0.05$). On the other hand, correlations on Dimension 2 of both RSI and SSI solutions had a moderate magnitude and were significant (Spearman Rs = - 0.42 and – 0.39, respectively, $ps < 0.01$). Although neither correlation was very strong, they suggest that some portion of the variance on Dimension 2 could be related to the spectral resolution of the sounds required for source identification.

CHAPTER 5

DISCUSSION

The results of the study can be used to address several practical and theoretical issues in environmental sound perception. Each of these questions will be considered separately below.

5.1. Optimal number of channels

The results demonstrate that source identification performance continuously improves with increased spectral resolution when sounds with declining accuracy are excluded from analysis (Figure 4). However, the improvement in performance is most pronounced only up to sixteen channels. The rate of improvement decreases with further increases in spectral resolution. Moreover, the biggest gain in identification accuracy was achieved between four and eight channels for both the entire 60-item set and the 41-item set of sounds which showed a continuous improvement with increasing number of channels (Table III, Table VI). Thus, while 32 channels do not seem to provide spectral resolution necessary to achieve identification accuracy equivalent to that of undistorted sounds, it appears that relatively high accuracy in the range of 57% to 71% correct can be achieved with 8 to 16 spectral channels.

Compared with the number of channels required for high speech intelligibility, identification of environmental sounds may require a higher degree of spectral resolution. Only 4 to 5 channels are necessary to understand spoken sentences in quiet with an accuracy rate of more than 90% correct (Dorman et al., 1997; Shannon et al, 1995). More difficult speech materials require a higher number of channels for the same level of

intelligibility. Eight channels are necessary to achieve asymptotic performance of about 90% correct with isolated multi-talker or synthetic vowels (Dorman et al., 1997). On the other hand, identification of environmental sounds at 8 channels was still below 60% correct, and continued to improve with further increases in the number of channels.

There may be several explanations for these empirical discrepancies in studies of the spectral resolution needed to perceive environmental sounds and speech. Some of the differences may be explained by variation in methodologies of the studies, while others may reflect a more global difference in speech and environmental sound perception. First, listeners in the speech studies were familiarized with spectrally nondegraded stimuli prior to testing. The amount of exposure varied between several presentations of all stimuli with corresponding response labels (Dorman et al., 1997) to 8 to 10 hours of practice (Shannon et al., 1995). In contrast, listeners in the present study received no prior exposure to the test materials and were only familiarized with response labels, but not the original sounds. Thus, it is possible that the difference in performance might have been smaller if the listeners received a similar amount of exposure to undistorted test materials prior to testing.

Second, there exists no established baseline for comparing environmental sound identification with speech perception even without distortions. Because scores on speech perception tests involve a count of perceptual units (i.e., phonemes or words) which may be inherently different from individual environmental sounds, the discrepancies should not be considered a reliable indicator of the differences in speech and environmental sound perception. Rather, they indicate a potential for a difference in perception between these types of sounds.

Moreover, depending on the task, perception of speech sounds may proceed along a somewhat different perceptual path than perception of non-speech environmental sounds. Speech sounds originate from one type of sound source – the vocal tract[2]. Although different speech sounds may involve different articulators within the vocal tract, the possible types of sound-producing objects and the kinds of events that produce speech sounds are, nevertheless, highly constrained compared with a great variety of sound sources that produce non-speech environmental sounds. Thus, faced with a task of identifying a speech sound, the listener already knows the type of sound source and is able to concentrate on the cues to its physical behavior. In contrast, when asked to identify the source of an environmental sound, the listener has a more difficult task, i.e., to identify the sound source and its sound-producing behavior. A more ecologically valid comparison between speech and environmental sound perception could come from a test where non-speech environmental sounds would be presented along with speech sounds, with listeners having no prior knowledge of which sounds are speech and which are not. The results of such a test may provide a more balanced estimate of the difference in spectral resolution required for speech and environmental sound perception, and more closely reflect listening tasks of everyday life.

Furthermore, speech sounds are, by far, the most important sounds listeners hear and produce in the environment. The amount of exposure to and practice with the sounds

[2] It can be argued that speech sounds may also be synthesized "from scratch" by electronic means. In this case, they do not originate in the vocal tract. However, electronically synthesized speech sounds merely mimic the regularities found in the acoustic signal of naturally produced speech, or follow articulatory models. Therefore, as far as their perception is concerned, synthesized sounds can be treated as electronically reproduced vocal tract sounds. At least, it is highly unlikely that the human perceptual system has evolved a different mechanism for perceiving natural vs. synthetic speech.

of speech undoubtedly exceeds that of any non-speech environmental sound. Some researchers also believe that speech perception is accomplished with a specialized perceptual module which comprises perceptual processes unique to speech and which functions independently of general auditory perception (Liberman & Mattingly, 1985). Although the results of the current study cannot be used to decide whether or not this is the case, they tentatively suggest that perception of speech sounds may be more resistant to spectral smearing than perception of non-speech environmental sounds. However, this questions must be explored further before any definitive conclusions can be reached.

Finally, the number of channels required for accurate source identification performance may be considerably affected by the kinds of environmental sounds used as test materials. The definition of a representative set of environmental sounds always involves a number of arbitrary assumptions due to the sheer variety of environmental sounds and the lack of a clear taxonomy. Thus, different degrees of spectral resolution may be required to achieve high identification performance with different types of environmental sounds. In fact, more than one third of the 60 environmental sounds tested in the study were identified with accuracy of 70% or above with only 4 spectral channels or fewer, while almost two thirds were identified at 70% or above accuracy with 8 channels or less (Table VII). On the other hand, the remaining 22 test sounds required 16 channels or more to be identified at 70% correct or above, with some 11 of these sounds never reaching 70% accuracy with as many as 32 channels (Table IV). It is possible that some of these sounds could have been identified more accurately in the absence of the asynchronous temporal distortions introduced to their spectra by filter group delays. However, the spectral resolution required to identify their sources would still likely be

relatively high, compared to the sounds identified with only 2 or 4 channels. Overall, results demonstrate that different degrees of spectral resolution may be optimal for different subsets of environmental sounds.

5.2. Sound variability and classification based on spectral resolution

Variability in spectral resolution required to achieve the same identification performance for different environmental sounds provides an empirically based parameter useful in classifying a large number of environmental sounds. Results indicate that environmental sounds systematically vary in the amount of spectral detail needed for source identification. As shown earlier, the perceptual characteristics and, probably, acoustic parameters of sounds identified correctly by the majority of the listeners with 2 and 4 channels seem to differ from those of sounds identified with higher numbers of channels (Table VII).

The distribution of sounds identified correctly with lower and higher spectral resolution is similar to that obtained in previous research (Gygi, 2001). Gygi's research showed that sounds identified most accurately by naïve listeners with only 1 spectral channel were broadband temporally patterned energy bursts (e.g., 'helicopter', 'rain', 'gallop', 'beer can opening', 'gun shot', 'ping-pong ball bouncing', 'clock ticking'). As spectral resolution was increased to 6 channels, an independent group of naïve listeners was able to identify correctly sounds with dynamic narrow band resonant components in the spectrum (e.g., 'bird calling', 'baby crying', 'dog barking', and 'bubbles'). The consistency among the types of sounds identified with low and high spectral resolution in both Gygi's (2001) and the present study demonstrate that sound variability in required

spectral resolution is a robust perceptual effect that can be demonstrated in spite of the differences in particular sound tokens used, or differences in the signal processing methods and the experimental tasks.

The classification of environmental sounds based on the spectral resolution required for source identification provides a framework for exploring acoustic parameters and cognitive factors that may distinguish sounds requiring low vs. high number of channels. If acoustic parameters can reliably distinguish these sounds in accordance with required spectral resolution, they may be incorporated into a model for predicting the required spectral resolution for novel environmental sounds. However, it is likely that, in addition to acoustic parameters, there are cognitive factors that can influence the identification of environmental sounds. In addition to such previously identified factors as sound familiarity and the speed of source name retrieval (Marcell et al., 2000; Ballas, 1993), identification accuracy of a given sound may also be affected by the number of other sounds that share similar acoustic properties. A sound that has one or more unique properties may be easier to identify than a sound that has several properties similar to those of other sounds. For instance, sounds of a cracking whip or a bouncing ping-pong ball seem to have a very distinct pattern of energy change over time, while the dynamic envelopes of gargling or a flying airplane may be similar to those of several other sounds. Future research must examine how different acoustic cues to source identity are weighted across different environmental sounds based on their spectral and temporal properties.

5.3. Perceptual clustering of environmental sounds

Classifying sounds by required degree of spectral resolution highlights the salient acoustic parameters and perceptual characteristics involved in source identification. Another approach to determining perceptual similarities and differences among sounds is through examining the positioning of sounds in a psychological space. The moderate to strong correlations between the data coordinates on the corresponding response (i.e., RSI) and stimulus (i.e., SSI) dimensions indicate that the perceptual characteristics of sounds obtained by SSI measures are related to the knowledge about sound sources that listeners have as measured by RSI. This result supports the previous finding by Gygi (2001) who demonstrated very similar distributions of sounds across three MDS spaces obtained from (a) acoustic tokens, (b) imagined sounds, (c) imagined sound sources. This result is particularly noteworthy because there were considerable methodological differences between Gygi's and the present study in the kind of perceptual data from which the MDS solutions were derived. Gygi's data comprised direct similarity ratings for actual sounds or imagined sound pairs, while the present data were indirectly obtained from perceptual confusions produced by the listeners. Taken together, these findings indicate considerable similarity between the perceptual characteristics of the acoustic signals of environmental sounds and the representations of the perceptual characteristics of sounds and sources in the mind of the listener.

Further, the distribution of sounds in the two MDS spaces (Figure 5 & 6) also seems to contain several identifiable, if overlapping, clusters. Although the coordinates of individual cluster members seem to vary considerably between the RSI and SSI space, at least three different clusters can be distinguished in each case. These are (1) brief

sounds with one or two transient components (e.g., 'cork popping'[17], 'door closing'[21], 'pool balls colliding'[39], 'whip'[55]), (2) rhythmic sounds of temporally patterned transients (e.g., 'typing'[50], 'clock ticking'[16], 'footsteps'[23], 'clapping'[14], 'heartbeat'[25]), and (3) water based sounds (e.g., 'water bubbling'[51], 'water draining'[52], 'water running'[54], 'water splashing'[44], 'man gargling'[33]). In addition, there seem to be two small distinct clusters of signaling sounds. One consists of 'siren'[43] and 'telephone'[45], which can be found in the vicinity of each other. The other contains 'car horn'[11], 'church bells'[13], and 'door bell'[22].

At least two of these clusters are reminiscent of those described by Gygi (2001) who found that water-based sounds and temporally patterned rhythmic sounds tended to form clusters. Unlike Gygi's clusters, however, there is an additional cluster of brief transient sounds, and no identifiable cluster of animal sounds. Instead, animal sounds are distributed throughout the two MDS spaces, although in the SSI space, three animal vocalizations are found in close proximity (i.e., 'rooster crowing'[41], 'horse neighing'[27], and 'cow mooing'[18]) and are surrounded by human vocalizations (i.e., 'baby crying'[2], 'man yawning'[36], 'woman sneezing'[59]). The similarities in sound clustering found in Gygi's and the present results further support the interpretation that sounds are, indeed, distinguishable in a psychological space based on the properties of their sources.

The present cluster of brief transient sounds may also correspond to the cluster of transients described by Ballas (1993), which included sounds of door closing and cork popping among others that were not tested in this project. In addition, Ballas' (1993) cluster of water-related sounds also seems to partially overlap with the present water

based sound cluster, although his cluster also included sounds produced not by water but in a water related context (i.e., 'foghorn', 'boat whistle'). The partial correspondence between Ballas' (1993) and the present clusters is of particular interest because Ballas' clusters were based on ratings of sound characteristics without asking listeners to explicitly identify sound sources. The finding that clusters of similar sounds can be formed from data that do not reference sound sources and the data that specifically reference sound sources implies that perceptual properties of sounds are closely integrated with those of their sources, and, at least in some instances (e.g., water-based sounds and transients), the perception of sounds always implies the perception of its source.

The clustering patterns observed in this study also resemble the source-based environmental sound categories proposed by Gaver (1993). Although none of the present or previously reported (i.e., Gygi (2001), Ballas (1993)) clusters is as clearly defined as might be expected based on Gaver's taxonomy, the overall clustering patterns are able to provide tentative evidence that clustering of environmental sounds in a psychological space is related to the physical properties of sound producing objects and events. A further elucidation of this relationship may become possible through an analysis of acoustic parameters, source properties, and perceptual characteristics shared by the member of different sound clusters.

5.4. Effects of spectral asynchrony on sound source identification

The processing of the stimuli inadvertently caused a systematic spectral asynchrony due to unmatched filter group delays across frequency bands. The spectral

energy in the lower frequency channels was thus delayed relative to the higher frequency channels for stimuli with a large number of channels. The amount of the delays gradually increased as the number of channels increased (see Appendix C). The delays, however, did not seem to have a noticeable effect on source identification up to 16 channels. With sixteen channels, the time difference between the lowest frequency channel and the highest frequency channel reached18 ms, on average. Further increases in the number of channels also resulted in increases in group delays. With greater than 16 channels, source identification declined by 30% or more for 19 environmental sounds.

Spectral asynchrony affected individual sounds differently. The identification accuracy of the majority of the test sounds continued to rise or remained asymptotic with an increase in the number of channels. On the other hand, the identification accuracy of other sounds declined. Most of the sounds noticeably affected by spectral asynchrony were temporally patterned sequences of brief energy bursts ('clapping', 'horse trotting', 'footsteps', 'jackhammer', 'typing on keyboard', 'biting and chewing apple', 'ice cubes into glass', 'shoveling dirt', 'snoring'), and sounds with only one transient component (e.g., 'stone splashing water', 'door closing', 'page turning', 'cork popping from bottle'). Thus, it seems that the effects of spectral asynchrony are most detrimental for sounds consisting of brief components that, presumably, contain information about the source.

Interestingly, the error responses given to the affected sounds suggest that spectral asynchrony mostly affects the identification of the sound-producing object, while main event characteristics are preserved. For instance, the modal confusion of 'clapping' in the 32-channel condition was 'clock ticking', while the modal confusion of 'horse trotting' was 'helicopter flying.' In both cases, the object is different, but the event still

contains the repetitive rhythm of the original sound source, although its description is changed to accommodate the new type of object (i.e., helicopters do not trot, and clocks do not clap).

Taken at face value, this result seems contradictory to the finding of Warren & Verbrugge (1984) who argued that changing the position of spectral peaks across 4 frequency bands mainly affects the identification of the sound producing event. However, this discrepancy may result from the fact that, in the present case, the spectral peaks were consistently delayed relative to each other by a relatively small amount. On the other hand, in Warren & Verbrugge's study the spectral peaks were either in synchrony across the bands, or were randomly positioned in time relative to each other. It is thus possible that even in their study, the application of consistently arranged, rather than random, asynchronies would result in the perception of the same event, but a different object. Furthermore, the identity of the event may also be related to the overall dynamic energy pattern of the sound. The small systematic time delays across bands did not affect the overall smoothed dynamic envelope of the sound. In contrast, it is likely that random vs. synchronous positioning of spectral peaks across frequency bands in Warren & Verbrugge's study produced a bigger change in how the energy varied over time.

The effects of spectral asynchrony across channels on the source identification of environmental sounds also suggest that frequency-specific time delays may be more detrimental for the identification of environmental sounds than speech. Fu & Galvin (2001) found that the intelligibility of high – context sentences processed through a 16-channel noise-band processor similar to the one used in the present study was not

significantly affected by spectral asynchronies of less than 160ms. On the other hand, the identification accuracy of four environmental sounds declined by more than 30% in the 16-channel condition with the maximum cross channel time delay of only 18 ms. Longer channel delays led to a decline in the identification accuracy of even more sounds, and had a noticeable effect on the overall identification performance, despite the fact that these sounds had greater spectral resolution. Although methodological discrepancies prevent straightforward comparisons of the effects of spectral asynchrony on speech and environmental sound perception, they warrant a closer investigation of the effects of spectral asynchrony on environmental sound source identification.

5.5. Implications for cochlear implants

The signal processing method used in this study simulates the processing of acoustic signals by multiple-channel cochlear implant processors. Thus, the results of the study have several implications for the perception of environmental sounds by cochlear implant users. First, the results suggest that accurate perception of a large number of familiar environmental sounds, without training, requires a higher number of channels than the perception of speech. On the other hand, the majority of environmental sounds can be perceived with 70% accuracy with 8 channels or fewer, which corresponds to the optimal number of channels needed to reach asymptotic performance with speech sounds (Dorman et al., 1997). The rate of improvement in identification performance declines considerably with the spectral resolution higher than 16 channels. Therefore, in the case of cochlear implants, the practical difficulties associated with implementing a greater

number of channels may outweigh the benefits of a slightly improved identification accuracy.

The present results were based on the identification performance of naïve listeners who did not have any prior exposure to spectrally smeared environmental sounds. It is possible that identification performance may improve if listeners have more practice in identifying environmental sounds with limited spectral resolution. Previous research provides some support for this expectation. The identification performance of Gygi's (2001) naïve listeners, whose identification accuracy was 13 % on single channel stimuli and 36 % on 6 channel stimuli, improved to 23 % and 66 % correct, respectively, when tested a second time, after a single presentation of the original undistorted sounds.

Other evidence that cochlear implant users may be able to adapt their perceptual skills to the reduced spectral resolution of environmental sounds is suggested directly by the identification performance of cochlear implant users. Tyler, Moore & Kuk (1989) found that the mean identification accuracy of cochlear implant patients using a single channel 3M/Vienna device on a set of 18 environmental sounds was 41% correct. This accuracy level corresponds to the accuracy achieved with 4 channels by naïve listeners tested in this study, notwithstanding the differences in the number of environmental sounds tested in each case. Interestingly enough, the cochlear implant listeners were highly accurate not only with such temporally patterned sounds as 'footsteps', but also with more harmonic ones such as 'piano', 'birds chirping' and 'whistling.' The finding that harmonic sounds can be perceived without any spectral detail based solely on the dynamic envelope suggests that (a) these sounds are potentially distinguishable by the use

of envelope cues alone, and (b) that listeners are able to learn to rely solely on these temporal cues in source identification.

In the same study, the users of a 4-channel Symbion device identified the same set of environmental sounds with the mean accuracy of 83 % correct, surpassing the identification performance of present listeners even in the 32 channel condition. Curiously, the Symbion users significantly outperformed users of cochlear implant devices that utilized a higher number of channels (i.e., Chorimac, Duren/Cologn, and Nucleus). This suggests that spectral resolution is only one among many factors that affect the identification performance of cochlear implant users.

Another factor that may affect the identification of environmental sounds by cochlear implant users consists of frequency-dependent temporal distortions of the original sound's spectrum. Such distortions may be introduced naturally by listening to sounds in reverberant environments, or artificially by having unequal filter group delays across channels. Which ever may be their source, the present results indicate that frequency-dependent temporal distortions have a considerable effect on identification of the sources of many environmental sounds. Previous research (Fu & Galvin, 2001) indicates that reduction in the number of frequency channels increases the negative effects of cross-channel spectral asynchrony on speech perception. The present results suggest that the identification of environmental sounds may be even more susceptible to the negative effects of spectral asynchrony than speech. Therefore, it is possible that the reduction in fine spectral detail of environmental sounds coupled with the variations in temporal alignment of spectral energy across channels makes the task of identifying their sources even more difficult for the listener.

CHAPTER 6

SUMMARY, CONCLUSIONS, AND FUTURE RESEARCH

6.1. Summary and Conclusions

The results of the present study showed that both spectral and temporal parameters play an important role in identification of the sources of environmental sounds. Without significant temporal distortions, identification accuracy continuously improved with increases in spectral resolution. However in this study, the rate of improvement in performance was maximal between 4 and 8 spectral channels. Increases in spectral resolution beyond 16 channels produced only minor improvement in identification accuracy.

In general, perception of environmental sounds may be less resistant to the effects of spectral and temporal distortions than perception of speech sounds. Cross-channel asynchronies on the order of 20 ms. were sufficient to produce a negative effect on source identification of several spectrally smeared environmental sounds, while much higher asynchronies are required to produce a significant effect on speech intelligibility (Fu & Galvin, 2001). In addition, given comparable degrees of spectral smearing, speech sounds are identified more accurately than environmental sounds. These findings suggest that a greater number of spectral and temporal cues may be available to the listener for perceiving speech sounds than for perceiving environmental sounds. The behavior of the source may be represented more redundantly in the acoustic signal of speech sounds than nonspeech environmental sounds. On the other hand, it is also possible that the human perceptual system is better adapted to processing speech sounds than environmental sounds, and that it can more readily "pick-up" information about the behavior of sound-

producing objects and events from speech sounds than from environmental sounds. However, these findings and their potential significance have to be considered with caution in the absence of any direct comparisons of speech and environmental sound perception under the same experimental conditions.

There was a large degree of variability among individual environmental sounds in the amount of spectral detail required for source identification. Sources of many environmental sounds could be identified quite well with very little spectral resolution, while others required spectral resolution of 32 channels or higher to be identified correctly. Further, there seemed to be systematic differences in the perceptual characteristics and, probably, acoustic parameters of sounds requiring low vs. high spectral resolution for accurate source identification.

Finally, analysis of listener errors demonstrated a close link between perceptual characteristics of environmental sounds and those of their source objects and events. Present results confirmed previous findings (Vanderveer, 1979; Gaver, 1993a; Gygi, 2001) which demonstrated regularities in listener confusions of the sound sources along the spectral and temporal dimensions. Thus, although every type of sound produced in the environment may have a unique set of physical parameters and perceptual characteristics, there are notable regularities in how environmental sounds and their sources are organized in the mind of the listener. This study showed that some of these regularities are based on the information contained in the spectrum of environmental sounds, while others are likely to be based on temporal parameters and perceptual characteristics which further link environmental sounds with their sources.

6.2. Future Research

By design and by accident, this study has highlighted several previously unexplored and potentially important venues of investigation in environmental sound perception. Future research will tackle questions raised by this study but left unanswered due the limitations in its scope and design. Several topics for follow-up research are discussed below.

Perhaps the most direct and specific question that begs to be addressed based on the present results relates to the observed detriment in source identification performance under varying spectral resolution produced by cross-channel spectral asynchrony. In the present design, the variation in performance produced by spectral asynchrony is confounded with that produced by variations in spectral resolution. The higher the number of channels, the greater was the relative difference in group delays across channels, and the bigger was the cross-channel asynchrony. Thus, a modified signal processing method is needed to separate the effects of spectral asynchrony from those of spectral resolution. One way to examine this question is to use high order FIR filters with a similar frequency response as the current Butterworth filters but with equal group delays across channels. If perceptual testing is done using the same design and stimuli as in the present study, this signal processing method will allow for a direct comparison of the effects of relative spectral asynchrony on source identification. Depending on the results, further research can then be carried out to assess the relative perceptual effects of spectral asynchrony on perception of speech, environmental sounds and musical instruments.

Another immediately apparent topic for further investigation is determining acoustic parameters which could account for the grouping of environmental sounds by the number of channels required for source identification (Table VII). As previously discussed, the perceptual qualities of sound sources identified correctly with a low number of channels appear distinct from those sounds whose sources were identified correctly only with a high number of channels. A set of acoustic parameters that could distinguish low spectral resolution sounds from high spectral resolution ones would provide theoretically and practically useful information. Theoretically, it may reveal acoustic parameters that may be indicative of the perceptual organization of environmental sounds, i.e. the acoustic correlates of perceptual qualities. Practically, it would provide a basis for a testable model for predicting the required number of channels for novel environmental sounds. Such a model, if sufficiently robust, may aid in the design of novel cochlear implant processing strategies designed to enhance environmental sound perception.

In a similar vein, the extent to which the present results may be indicative of actual cochlear implant users' performance must be investigated by testing actual cochlear implant users. If relatively inexperienced users of cochlear implants with a similar signal processing strategy demonstrate comparable performance patterns across channels, it would validate the acoustic simulation approach to evaluating implant performance, at least for this set of environmental sounds. If obtained, such findings would open doors to further simulation studies which could examine other signal processing strategies and parameters noninvasively as well as venture into an unexplored area of perceptual learning.

Indeed, available cochlear implant literature (Tyler et al., 1989) suggests that experienced users of implants utilizing even a small number of frequency channels can be remarkably accurate in environmental sound perception. On the other hand, anecdotal reports indicate that environmental sound perception is very difficult shortly after the implant is turned on. Therefore it is likely that listeners' ability to identify environmental sound sources with limited spectral resolution can improve with practice. It is thus important to know the time course of perceptual learning, the acoustic information that listeners learn to use to identify sources with limited spectra, and the factors that influence their learning process. This may, in turn, lead to the development of comprehensive environmental sound testing and training materials which could be used for research and therapeutic ends.

Appendix A: Types of sounds selected for testing in the study arranged by source category.

Animal sounds

1. Birds chirping
2. Cow mooing
3. Dog barking
4. Dog panting
5. Horse neighing
6. Horse trotting
7. Rooster crowing

Human sounds

8. Baby crying
9. Biting, chewing apple
10. Blowing nose
11. Brushing teeth
12. Burp
13. Child coughing
14. Clapping
15. Clearing throat
16. Footsteps
17. Heartbeat
18. Man drinking
19. Man gargling
20. Man panting
21. Man snoring
22. Man yawning
23. Woman laughing
24. Woman sighing
25. Woman sneezing

Mechanical interactions of inanimate solids

26. Bowling strike
27. Camera taking picture
28. Car starting
29. Clock ticking
30. Door closing
31. Glass breaking
32. Helicopter flying
33. Ice cubes into glass
34. Jackhammer
35. Page turning
36. Ping-pong ball
37. Pool balls colliding
38. Shoveling dirt
39. Train in motion
40. Typing on keyboard
41. Whip
42. Zipper

Water sounds

43. Pouring soda into cup
44. Stone splashing water
45. Toilet flushing
46. Water bubbling
47. Water draining
48. Water dripping
49. Water running

Aerodynamic sounds

50. Airplane flying
51. Cork popping from bottle
52. Machine gun
53. Thunder
54. Wind blowing

Signaling sounds

55. Car horn
56. Church bells
57. Doorbell
58. Siren
59. Telephone busy signal
60. Train whistle

Appendix B: Lower frequency cutoff of every channel in each channel condition. (The higher frequency cutoff was always the lower frequency cutoff of the consecutive channel. The higher frequency cutoff of the highest channel was always 5500 Hz).

Condition / Channel #	2 channels	4 channels	8 channels	16 channels	24 channels	32 channels
1	300	300	300	300	300	300
2	1285	621	432	360	339	329
3		1285	621	432	382	360
4		2658	893	518	432	394
5			1285	621	487	432
6			1848	745	550	473
7			2658	893	621	518
8			3823	1071	701	567
9				1285	791	621
10				1541	893	680
11				1848	1008	745
12				2216	1138	815
13				2658	1285	893
14				3188	1450	978
15				3823	1637	1071
16				4586	1848	1173
17					2086	1285
18					2355	1407
19					2658	1541
20					3000	1687
21					3387	1848
22					3823	2024
23					4316	2216
24					4872	2427
25						2658
26						2911
27						3188
28						3491
29						3823
30						4187
31						4586
32						5022

Appendix C: Averaged group delays (in ms.) for each channel relative to the highest frequency band.

Condition / Channel #	2 channels	4 channels	8 channels	16 channels	24 channels	32 channels
1	1	3	8	18	28	38
2	0	2	7	18	30	42
3		1	5	16	27	40
4		0	3	13	25	36
5			2	11	22	33
6			1	9	19	31
7			1	7	17	28
8			0	6	15	25
9				4	13	23
10				4	11	21
11				3	10	19
12				2	8	17
13				1	7	15
14				1	6	14
15				1	5	12
16				0	4	11
17					4	10
18					3	9
19					2	8
20					2	7
21					2	6
22					1	5
23					1	5
24					0	4
25						3
26						3
27						2
28						2
29						2
30						1
31						1
32						0

Appendix D: Correspondences between RSI and SSI point numbers on Fig. 5 and Fig. 6 and sound source labels.

1	Airplane flying
2	Baby crying
3	Birds chirping
4	Biting, chewing apple
5	Blowing nose
6	Bowling
7	Brushing teeth
8	Burp
9	Camera
10	Car starting
11	Car horn
12	Child coughing
13	Church bells
14	Clapping
15	Clearing throat
16	Clock ticking
17	Cork popping
18	Cow mooing
19	Dog barking
20	Dog panting
21	Door closing
22	Doorbell
23	Footsteps
24	Glass breaking
25	Heartbeat
26	Helicopter flying
27	Horse neighing
28	Horse trotting
29	Ice cubes into glass
30	Jackhammer
31	Machine gun
32	Man drinking
33	Man gargling
34	Man panting
35	Man snoring
36	Man yawning
37	Page turning
38	Ping-pong ball
39	Pool balls colliding
40	Pouring soda into cup
41	Rooster crowing
42	Shoveling dirt
43	Siren
44	Stone splashing water
45	Telephone busy signal
46	Thunder
47	Toilet flushing
48	Train motion
49	Train whistle
50	Typing on keyboard
51	Water bubbling
52	Water draining
53	Water dripping
54	Water running
55	Whip
56	Wind blowing
57	Woman laughing
58	Woman sighing
59	Woman sneezing
60	Zipper

REFERENCES

Ballas, J. A. & Howard, J.H. (1987). Interpreting the language of environmental sounds. *Environment and Behavior*, 19(1), 91-114.

Ballas, J.A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19 (2), 250-267.

Bates, J.K. (2000). How to hear everything and listen to anything. Retrieved 10/15/2001 from http://home.computer.net/~jkbates/thela1.htm

Bismark, G. von (1974a). Timbre of steady sounds: a factorial investigation of its verbal attributes. *Acoustica*, 30,146-157.

Bismark, G. von (1974b). Sharpness of steady sounds. *Acoustica*, 30, 160-171.

Bjork, E. A. (1985). The perceived quality of natural sounds. *Acoustica*, 57, 185-188.

Borden, G.J., Harris, K.S. & Raphael, L.J. (1994). *Speech science primer*. Williams & Wilkins.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, Massachusetts: MIT Press.

Cabe, P. A. & Pittenger, J.B. (2000). Human sensitivity to acoustic information from vessel filling. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 313-324.

Carello, C., Wagman, J.B. & Turvey, M.T. (2003). Acoustic specifications of object properties. In J. Anderson & B. Anderson (Eds). *Moving Image Theory: Ecological Considerations.* Carbondale, IL: University of Southern Illinois Press.

Carello, C., Anderson, K. L., Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science*, 9(3), 211-214.

Cermak, G. W. & Cornillon, P.C. (1976). Multidimensional analyses of judgments about traffic noise. *Journal of the Acoustical Society of America*, 59(6), 1412-1420.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, 76, 1636-1647.

Delhrone, L.A. & Reed, C.M. (1998). Identification of environmental sounds through cochlear implants, *International Sensory Aid Conference*, Portland, ME.

Dorman, M.F. (1993). Speech perception by adults. In Tyler, R.S. (Ed.)., *Cochlear Implants: Audiological Foundations*, Singular Publishing Group, San Diego, CA.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America*, 102(4), 2403-2411.

Dowell, R.C., Brown, A.M., Seligman, P.M. & Clark, G.M. (1985). Patient results for a multiple-channel cochlear prosthesis. In R.A. Schindler & M.M. Merzenich (Eds.), *Cochlear Implants*, Raven, New York.

Doyle, J., Doyle, J., & Turnbull, F. (1964). Electrical stimulation of the eighth cranial nerve. *Archives of Otolaryngology*, 80, 388-391.

Duddley, H. (1940). Remaking speech. *Journal of the Acoustical Society of America*, 11, 169-177.

Edgerton, B.J., Prietto, A. & Danhouer, J.L. (1983). Cochlear implant patient performance on the MAC battery. *Otolaryngologic Clinics of North America*, 16(1), 267-280.

Eisenberg, L.S., Berliner, K.I., House, W.F. & Edgerton, B.J. (1983). Status of the adults' and children's cochlear implant programs at the House Ear Institute. *Annals of the New York Academy of Science*, 405, 323-331.

Ellis, D. P. (1996). *Prediction-driven computational auditory scene analysis.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Fowler, C. A. & Rosenblum, L. D. (1990). Duplex perception: a comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4), 742-754.

Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, 88(3), 1236 – 1249.

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.

French, N.R. & Steinberg, J.C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 19, 90 – 119.

Fu, Q. & Galvin, J. (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users. *Journal of the Acoustical Society of America*, 109 (3), 1166 –1172.

Gantz, B., Tyler, R., Knutson, J., Woodworth, G., Abbas, P., McCabe, B., Hinrichs, J.,

Tye-Murray, N., Lansing, C., Kuk, F. & Brown, C. (1988). Evaluation of five different cochlear implant designs: Audiological assessment and predictors of performance. *Laryngoscope*, 98, 1100-1106.

Gaver, W. (1988). *Everyday listening and auditory icons*. Dissertation Abstracts International. 50(4-B), 1669.

Gaver, W. W. (1993a). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1), 1-29.

Gaver, W. W. (1993b). How do we hear in the world?: Explorations in ecological acoustics. *Ecological Psychology*, 5(4), 285-313.

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

Gordon, C. & Webb, D. (1996). You can't hear the shape of a drum. *American Scientist*, 84, 46-55.

Gygi, B. (2001). *Factors in the Identification of Environmental Sounds*, Unpublished doctoral dissertation, Indiana University, Bloomington, Indiana. Retrieved 02/20/02 from http://www.indiana.edu/~k300bg/dissall.pdf

Handel, S. (1989). *Listening*. The MIT Press, Cambridge, Massachusetts.

International Community of the Auditory Display (ICAD) (1997). Sonification Report: Status of the Field and Research Agenda. Retrieved 4/20/2002 from http://www.icad.org/websiteV2.0/References/nsf.html

Jenkins, J. J. (1985). Acoustic information for objects, places, and events. In W. H. Warren & R. E. Shaw (Eds.), *Persistence and Change*, pp. 115-138. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kidd, G. R. & Watson, C.S. (1999). Sound quality judgments of everyday sounds, *Journal of the Acoustical Society of America*, 106(4), pt. 2, 2267.

Kunkler-Peck, A. J. & Turvey, M.T. (2000). Hearing shape. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 279-294.

Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America*, 90(6), 3036-3049.

Liberman, A. M. & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1-36.

Loizou, P. (1998). Mimicking the human ear. *IEEE Signal Processing Magazine*, 15(5),

101 –130. Retrieved 10/22/01 from
http://www.utdallas.edu/~loizou/cimplants/tutorial/tutorial.htm

Loizou, P. C., Dorman, M. & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, 106(4), Pt. 1, 2097-2103.

Marcell, M.M., Borella, D., Greene, M., Kerr, E. & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, 22(6), 830-864.

Martin, K. D. (1999). *Sound-Source Recognition: A Theory and Computational Model.* Unpublished doctoral dissertation, MIT, Cambridge, Massachusetts.

Mecklenburg, D. J. & Brimacombe, J. A. (1985). The Nucleus 22-channel cochlear implant system. *Hearing Instruments*, 36(6), 35-38.

McAdams, S. (1993). Recognition of auditory sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in Sound: The Cognitive Psychology of Human Audition.* Oxford University Press.

Owens, E., Kessler, D. K., Raggio, M.W & Schubert, E. D. (1985). Analysis and revision of the Minimal Auditory Capabilities (MAC) battery. *Ear & Hearing*, 6, 280 – 290.

Repp, B. H. (1984). Categorical Perception: Issues, Methods, Findings. *Speech and Language: Advances in Basic Research and Practice*, 10, 243 – 335.

Repp, B. H. (1987). The sound of two hands clapping. *Journal of the Acoustical Society of America*, 81(4), 1100-1109.

Repp, B. H. (1988). Integration and segregation in speech perception. *Language and Speech*, 31(3), 239-271.

Richards, W. (1988). *Sound Interpretation.* In W. Richards (Ed.), Natural Computation, MIT, Cambridge, MA.

Schindler, R. A. & Kessler, D. K. (1987). The UCSF/Sorz cochlear implant: patient performance. *American Journal of Otolaryngology*, 8, 247 – 255.

Shannon, R. V., Zeng, F., Kamath, V., Wigonski, J & Ekelid, M. (1995). Speech recognition with primarily temporal clues. *Science*, 270, 303-304.

Shannon, R. V., Zeng, F. & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America*, 104(4), 2467-2476.

Simmons, B. (1966). Electrical stimulation of the auditory nerve in man. *Archives of Otolaryngology*, 84, 2-54.

Solomon, L. N. (1958). Semantic approach to the perception of complex sounds. *Journal of the Acoustical Society of America*, 30(5), 421-425.

Truax, B. (2001). Acoustic Communication. Ablex Publishing, Westport, Connecticut.

Tye-Murray, N., Tyler, R., Woodworth, G., & Gantz, B. (1992). Performance over time with a Nucleus and Ineraid cochlear implant. *Ear and Hearing*, 13(3), 200 – 209.

Tyler, R.S., Gantz, B.J., McCabe, B.F., Lowder, M.W., Otto, S.R. & Preece, J.P. (1985). Audiological results with two single-channel cochlear implants. *Annals of Otology, Rhinology and Laryngology*, 94(2), 133-139.

Tyler, R.S., Lowder, M.W., Otto, S.R., Preece, J.P., Gantz, B.J. & McCabe, B.F. (1984). Initial Iowa results with the multichannel cochlear implant from Melbourne. *Journal of Speech and Hearing Research*, 27, 596-604.

Tyler, R.S., Moore, B., & Kuk, F. (1989). Performance of some of the better cochlear-implant patients. *Journal of Speech and Hearing Research*, 32, 887-911.

Tyler, R. S. (1993). *Cochlear Implants: Audiological Foundations*. Singular Publishing Group, San Diego, CA.

Vanderveer, N.J. (1979). Ecological acoustics: Human perception of environmental sounds. *Dissertation Abstracts International*, 40, 4543B.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*,**167**, 392-393.

Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events. *Journal of Experimental Psychology: Human Perception and Performance*, 10(5), 704-712.

Weaver, W. & Shannon, C. E. (1949). *The Mathematical Theory of Communication*, Urbana, Illinois: Universtity of Illinois Press.

Wilde, R. & Richards, W. (1988). Recovering material properties of sound. In W. Richards (Ed.), *Natural Computation*, MIT Press, Cambridge, MA.

Working Group on Communication Aids for the Hearing-Impaired. (1991). Speech-perception aids for hearing-impaired people: current status and needed research. *Journal of the Acoustical Society of America*, 90(2), Pt. 1, 637-685.