6

Recognition of sound sources and events

Stephen McAdams

6.0 INTRODUCTION

Imagine playing for listeners an acoustic demonstration of a pile of ceramic dinner plates sliding off a counter, tumbling through the air knocking against one another, and finally crashing on to a relatively hard surface upon which all but one of the plates break—the unbroken one is heard turning on the floor and then finally coming to rest.[1] Then we ask the listeners the following questions. What kinds of objects were heard? How many were there? What was happening to them? And finally, did they all stay in one piece throughout the whole event? Even in the absence of visual cues or a situational context that might predict this event, any listener from a culture that makes use of such objects in such surroundings would easily describe what has been heard, recognizing the nature of the global event, the individual objects that played a role in it, and the transformation of most of them from a whole to a shattered state.

Humans have a remarkable ability to understand rapidly and efficiently aspects of the current state of the world around them based on the behaviour of sound-producing objects, or sound sources, even when these sources are not within their field of vision. For instance we recognize things knocking together outside the window, footsteps approaching from behind, and objects that have been dropped unexpectedly. Though, of course, we use all of our sensory systems together, the above examples suggest that listening can contribute significantly to the perception of the environment in order to act appropriately with respect to it. This ability is due in part to the existence of processes of perceptual organization as discussed by Bregman (1990, Ch. 2 this volume). Once the acoustic information stimulating the ears has been analysed into mental descriptions of sound sources and their behaviour through time, their recognition may be considered to be one of the primary, subsequent tasks to be accomplished based on information supplied by the auditory system (Schubert 1975).

Thinking in sound: the cognitive psychology of human audition, ed. S. McAdams and E. Bigand. Oxford University Press, 1993, pp. 146–98.

Recognition means that what is currently being heard corresponds in some way to something that has already been heard in the past, as when a voice on the telephone, or the footsteps of someone walking down the hall, or a piece of music on the radio, are each recognized. Recognition may be accompanied by a more or less strong sense of familiarity, by realizing the identity of the source (e.g. a car horn), and often by an understanding of what the source being heard signifies to the listener in his or her current situation, thereby leading to some appropriate action. For instance, you might be on the verge of walking carelessly across the street when a horn beeps. You immediately understand it as a very specific danger signal and quickly jump back off the road.

According to the *information processing* approach to psychology (cf. Anderson 1985; Lindsay and Norman 1977), the link between the perceptual qualities of the sound source, its abstract representation in memory, its identity, and the various meanings or associations it has with other objects in the listener's environment are hypothesized to result from a multi-stage process. This process progressively analyses and transforms the sensory information initially encoded in the auditory nerve. Recognition is accomplished by matching this processed sensory information with some representation stored in a lexicon of sound forms in long-term memory. The degree of match may determine whether the sound is recognized or not and may also determine the degree of familiarity that is experienced. The form(s) the sound event may assume and various semantic information associated with it through experience are also linked in memory (cf. Frauenfelder 1991). The activation of these associations follows from the activation of the particular entry in the lexicon. Which associations are activated and which subsequent actions are taken depend on the local context. For example, a listener's experience of the significance of a car horn would be different depending on whether he or she were crossing a street absent-mindedly, sitting in a cinema, or waiting at a stop light. So the recognized identity and significance of the sound event are the result of analysis, matching, and association processes.

Another approach is that of *ecological psychology* (see Gibson 1966, 1979 for the original writings on this approach, and Michaels and Carello 1981 for an excellent, accessible introduction). Ecological theory hypothesizes that the physical nature of the sounding object, the means by which it has been set into vibration, and the function it serves for the listener (as well as its name, presumably) are perceived directly, without any intermediate processing. That is, perception does not pass through an analysis of the elements composing the sound event and their reconstitution into a mental image that is compared with a representation in

[1] The demonstration cited may be found on track 5 of the compact disc *Sound Effects 7* (Dureco 1150582).

memory. The perceptual system itself is hypothesized to be tuned to those aspects of the environment that are of biological significance to the listener or that have acquired behavioural significance through experience. In a sense the claim that the recognition of the function of an object in the environment is perceived directly without processing would seem to evacuate the whole question of *how* (i.e. by what neurophysiological or mental process) organisms in possession of auditory systems that are stimulated by sound vibrations come to be aware of the identity of a sound source or how such sources acquire identity and significance for these listeners. The most appealing aspect of this approach, however, concerns the endeavour to develop descriptions of the structure of the physical world that make evident the properties that are perceived as being invariant, even though other properties may be changing. For example, you may still be able to recognize your grandmother's voice though she may have a cold, or be speaking over a noisy telephone line, or be speaking rapidly and in a higher pitch register because she is excited about something. Once such invariants have been isolated, the subsequent (psychological) task would then be to determine how listeners detect these properties. As such, ecological acoustics places more emphasis on the structure of acoustic events that are relevant tò a perceiving (and exploring) organism than has been the case in the information processing tradition (at least in non-verbal hearing research). Being interested in the mechanisms by which recognition occurs, and given the relative paucity of research in ecological acoustics (aside from a growing body of research on speech perception), this chapter will primarily adopt the information processing approach in the discussion that follows, making reference to possible contributions by the ecological stance where these are appropriate.

The terms 'recognition' and 'identification' have been operationally distinguished in experimental psychology. Recognition is often measured as the proportion of times a subject correctly judges whether a stimulus item has been heard before, usually within the time frame allotted for one (or a series of) experimental session(s). A simple 'old item/ new item' type of judgement may be accompanied by a rating of the degree of confidence in the judgement or by a rating of the degree of familiarity with the item. Identification experiments, on the other hand, require listeners to name or label the item, e.g. 'that's an oboe', or 'that's intensity number 5'. In a certain sense, identification can be considered a more narrowly focused kind of recognition that involves access to a (perhaps hierarchically organized) lexicon of names. For example, imagine that you hear a familiar bebop trumpet player on the radio, and you know that you have heard this person playing often while you were in college, but the name of the person escapes you. We would certainly say that recognition has occurred. We could even say that certain degrees of recognition have taken place, since you have recognized the sound source as a trumpet, the music as jazz, probably as being of a particular period of bebop style, and even more specifically as being played by a person you listened to in college. It is only the last level of the classification hierarchy that has not given rise to lexical activation, i.e. the name of the musician.

The purpose of the present chapter is to examine aspects of auditory representations and the nature of the processes operating on them that result in the recognition of sound sources and events. For the purpose of illustration, the discussion will be primarily confined to musical instruments and a few simple, 'natural' acoustic events. Relatively little work has been done on systematically investigating the various stages of the recognition process in audition. Therefore, I am obliged to examine a number of experiments that were directed at other experimental issues in an attempt to glean, indirectly, information that will help us piece together a picture of non-verbal auditory recognition. While the purpose of this chapter is to examine auditory recognition in general, most of the experiments that have dealt directly with the recognition of musical instruments and natural sound events have been confined to identification experiments.

The structure of the rest of this chapter is as follows. First I will consider in abstract terms the stages of processing that may be imagined to contribute to recognition as conceived within the information processing approach and extract from this consideration a number of important issues that should be analysed. Next I will examine experimental data from the literature on perception and identification of musical instrument tones and natural acoustic events. These experiments variously involved asking listeners to discriminate between sounds, to judge the degree of similarity among them, to classify them, or to identify them by name or by an arbitrarily chosen label. The data obtained from these experiments will be used to complete a more concrete, if tentative, picture of the stages of auditory processing previously discussed in abstract terms. Finally, I will discuss the properties of a number of models of recognition drawn from the domains of nonverbal audition, speech, and visual form recognition and analyse them in terms of what they might contribute to further clarification of the process of auditory recognition.

6.1 STAGES OF PROCESSING IN AUDITORY RECOGNITION

The recognition process may be conceived as involving several hypothetical stages of auditory information processing. This conception is shown schematically in Fig. 6.1. Different models of recognition hypothesize different stages of processing (some of the ones in Fig. 6.1 being

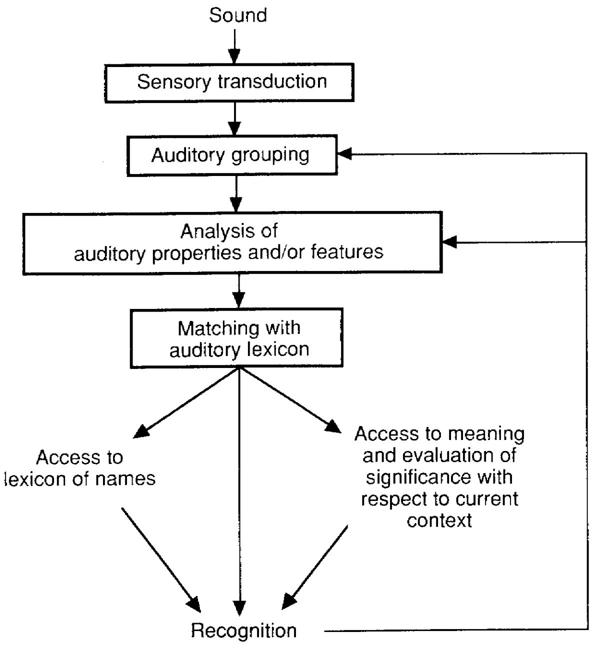
Sound
Sensory transduction
Auditory grouping
Analysis of auditory properties and/or features
Matching with auditory lexicon
Access to lexicon of names
Access to meaning and evaluation of significance with respect to current context
Recognition

Fig. 6.1 Schematic diagram of the stages of processing involved in recognition and identification. The representation in flow-diagram form does not necessarily imply that the processing of each type is organized sequentially. Some models view things otherwise, as is discussed in the text. Further, some models regroup two or more 'types' of processing under one mechanism or bypass them altogether.

completely bypassed, for example). They also differ in the degrees to which the results of later stages feed back to influence the operation of earlier ones.

6.1.1 Peripheral auditory representation of the acoustic signal

The first stage is the 'raw' representation of the acoustic signal in the peripheral auditory nervous system (labelled *sensory transduction*) (cf. Aran *et al.* 1988; Pickles 1982; Roman 1992). This process involves the transmission of vibrational information to the cochlea in which the signal sets into motion different parts of the basilar membrane depending on its frequency content. Higher frequencies stimulate one part of the membrane and lower ones another part. The movement of the basilar membrane at each point is transduced into neural impulses that are sent through nerve fibres composing the auditory nerve to the brain. They then undergo a series of operations in the central auditory processing centres. Each nerve fibre encodes information about a relatively limited range of frequencies, so the acoustic frequency range is essentially mapped on to the basilar membrane and encoded in the array of auditory nerve fibres. This tonotopic mapping is to some extent preserved through all of the auditory processing centres up to cortex. At low intensity levels the band of frequencies encoded within a given fibre is very narrow. With increasing intensity, this band increases in extent. At very high intensities, the frequency selectivity of an individual fibre is rather broad. In spite of this change of frequency selectivity with intensity, the spectral (or frequency) aspects of a sound are thus represented in part by the degree of activity present in each nerve fibre. However, each fibre also encodes information about temporal characteristics of the incoming waveform in its own frequency region by way of the detailed timing of neural impulses. As the part of the basilar membrane that is sensitive to a given frequency region moves, the pattern of activity in the population of nerve fibres connected to that part also fluctuates such that the nerve firings are to some degree time-locked to the stimulating waveform. Taken together, these two general forms of encoding (average rate of activity and temporal fine-structure of the nerve firing pattern) ensure that the auditory system represents a large variety of acoustic properties. If we consider that the array of frequency-specific fibers represents a kind of spectral dimension and their average level of activity across different populations of nerve fibres represents a kind of intensive dimension, the ensemble of activity as it evolves through time can be thought of as a mean-rate neural spectrogram. The global pattern of spectral evolution encoded in the auditory nerve would be in part captured in this representation. Another possible visualization of the auditory representation would plot the evolution through time of the degree of synchrony among the neural impulse patterns of neighbouring fibres as a function of the frequency to which each group is most sensitive, resulting in a kind of neural synchrony spectrogram. This representation would capture aspects of the temporal fine structure of the sound stimulus that are reliably encoded in the auditory nerve. Limitations in the spectral and temporal resolving power of the auditory transduction process in turn impose limits on the aspects of an acoustic waveform that are encoded in the auditory nerve.

6.1.2 Auditory grouping processes

Bregman (Ch. 2, this volume) discusses principles according to which this array of time-varying activity is then believed to be processed in order to constitute separate auditory representations of the various sound sources present in the environment (labelled *auditory grouping* in Fig. 6.1). One principle that arises from work on auditory grouping is that the properties of a sound event cannot be analysed as such until its constituent components have been integrated as a group and segregated from those of other sound events. This principle represents the primitive (or 'bottom-up') processes that are primarily driven by an analysis of the incoming sensory information. One should not rule out, however, a possible contribution of schema-driven (or 'top-down') processes by way of which more familiar or well-known sound events (such as one's name) would be more easily separated from a noisy sound environment than less well-known events (represented by the feedback arrows in Fig. 6.1). Most experiments on perception of auditory qualities and on auditory recognition present isolated sounds to listeners. In such cases, this particular level of processing has little influence on the final result, though of course the situation is quite different in everyday listening.

6.1.3 Analysis of auditory properties and features

Once the sensory information has been grouped into representations of sound sources, a series of processes are engaged that, according to some theories of recognition, progressively analyse the perceptual features or properties that are relevant to listening activity at a given moment (labelled *analysis of auditory properties and/or features* in Fig. 6.1). This analysis process may extract information over both local (milliseconds to centiseconds) and global (centiseconds to a few seconds) time spans (subsequently referred to as micro- and macrotemporal properties).

The analysis of *microtemporal properties* would be concerned with simple sound events such as one plate sliding over, or striking, another plate, or like a single note of a musical instrument. These properties are determined by the nature of the resonant structure being set into vibration (resulting from the geometry and materials of the sound source), as well as by the means with which it is excited (sliding, striking, bowing, blowing, etc.). The resonance properties (such as those of a vibrating crystal glass or a violin string and body) generally give us information about the physical structure of the sound source itself and will later serve to allow us to identify it. This class of properties would be involved in the detection of what ecological psychologists call the *structural invariants* of a sound source, i.e. those aspects of the acoustic structure that are 'shared by objects that, at some level of description, can be considered the same' (Michaels and Carello 1981, p. 25). For example, a violin will often be identifiable as such whether it plays in the low or high register, whether it is plucked or bowed, whether it was made by Stradivarius or is a factory-made training instrument. It should be noted, however, that the exact cues concerning the properties that are present in the acoustic waveform are far from clearly delineated at this point, although some progress is being made as will be summarized in Section 6.3. Alternatively, a plucked string event (i.e. the way the string is excited) can be recognized whether it is a guitar, violin, harpsichord, or wash-tub bass. In the terminology of ecological psychology, this class of properties would involve a listener's detection of *transformational invariants*, i.e. those acoustic properties that specify what is happening to the sounding object. Each of these sets of properties remains constant in spite of variation in other properties. The role of psychological experimentation is to isolate the classes of invariants to which the listener is sensitive and to try to understand how the auditory system extracts and makes use of these invariants, at least in the case of the information processing framework.

The analysis of *macrotemporal properties* is concerned with the rhythmic and textural aspects of a whole environmental event that allow us to identify, for instance, a set of dinner plates sliding, tumbling, and crashing on the floor. The identification of a breaking event is interesting in that it involves a change in the resonant properties of the object in question, since its size and geometry are changed. It also involves a change in the number of sound sources. Here more than a single excitation of a plate is involved, since there is a whole series of frictional and impact events that have an identifiable macro-organization. These kinds of events are also identified on the basis of transformational invariants.

So structural invariants are those properties that specify the nature of the object or group of objects participating in an event, while transformational invariants are the patterns of change in the stimulus that specify the nature of the change occurring in or to the object. In Section 6.3 we will examine the nature of the micro- and macrotemporal properties that play a role in auditory recognition.

6.1.4 Matching auditory properties to memory representations

According to the diagram in Fig. 6.1, the initial auditory representation has, by this stage, been recoded as a group of abstract properties that characterize the invariants in the sound source or event (such as spectrotemporal structure of the onset of a musical tone, or the accelerating rhythmic pattern of a bouncing ball slowly coming to rest). This group of properties (which constitutes the input representation to the recognition process) is then matched to classes of similar sound sources and events in memory (labelled *matching with auditory lexicon* in Fig. 6.1). The stimulus is recognized as the class that gives the closest match to the auditory representation. Two kinds of matching process have been conceived and will be discussed in more detail in Section 6.4. Briefly, some researchers conceive of this matching as a *process of comparison* whereby the auditory feature representation is compared with stored memory representations and the most closely matching one is selected. Such a comparison process would be mediated by a kind of executive agency (homunculus) or mental algorithm that performs the comparison and then returns the result. In other conceptions, access to memory structures involves the *direct activation* of memory representations of sources and events that are sensitive to configurations of features in the perceptual representation. This latter mechanism has a certain elegance and computational economy in being able to avoid postulating an outside agent that must receive the analysed sensory data, search available memory while comparing the two, and then decide upon which bit of memory best corresponds to the sensory data. It also avoids having to postulate a memory 'storage' that is independent of the auditory processing elements (see Crowder, Ch. 5 this volume, for more on this proceduralist approach to memory). Recognition of the sound event is then determined by the memory representation that receives the highest degree of activation by the auditory representation and occurs when the memory representation achieves some threshold of activation. Memory representations of this sort are probably of a relatively abstract nature, i.e. you may never have heard this particular sound of breaking plates before but you recognized it as such anyway. For both the mediated and direct approaches, if no category is matched, or if too many categories are matched with about the same degree of fit, no recognition can occur.

6.1.5 Activation of the verbal lexicon and associated semantic structures

The selection or activation of an appropriate representation in memory may then be followed by the activation of items in the listener's lexicon of names, concepts, and meanings that are associated with that class of sound events. It would be at this point that identification would take place since the listener would then have access to the name of the sound source or event (e.g. a struck plate, a trotting horse). In the sound example described above, it would be the identification of ceramic objects and their behaviour: there were several of them, they were dinner plates, they slid and fell to the ground, knocking against one another on the way down, and broke on the hard surface, except for one that was heard turning at the end. Lexical activation gives access to associated knowledge about properties of the class as it relates to the perceiver and the local situation. These associations would allow the listener to plan appropriate action without having to verbalize what was heard. Children and animals can recognize sources and events (one might even say 'identify', in the sense of recognizing the identity of them), and act appropriately with respect to them without having mastered language skills. Once language is available, however, the recognition process also gives access to a verbal lexicon that allows the listener to name or describe the event verbally. At and beyond this stage, the processing is no longer purely auditory in nature.

6.1.6 Interactions between stages of processing

Information processing does not proceed uniquely from the bottom up, that is, from sensory transduction to recognition. For example, it seems likely that under certain conditions, knowledge of the form of a sound event can help with separating it from background noise. Everyone has had the experience of being in a noisy environment where it is difficult to understand what other people are saying, and yet to suddenly hear one's name emerge out of the din. Another anecdotal example experienced by many bilingual people is that the noise levels at which speech remains intelligible can be higher for one's mother tongue than for a second language; ingrained knowledge of one's mother tongue assists in recognizing words on the basis of partial information. These phenomena are examples of what Bregman (1990) calls schema-based processing in auditory organization, and they reflect influence from later stages of processing both on auditory grouping and on the analysis of auditory properties and features. One of the most clear-cut examples of top-down influence on phoneme recognition is the phonemic restoration illusion (Warren 1970). If a phoneme is cut out of a word and replaced with silence, interrupted speech is heard. If it is replaced with a burst of noise, listeners have the impression of having heard the missing phoneme 'behind' the noise. If one postulates a partial match of the available phonetic material plus the perceptual assumption that the missing material is truly there but has been masked, the restoration demonstrates a reconstructed analysis of perceptual elements that are completed on the basis of recognition. Other similar effects of this nature are well known in speech recognition (cf. Frauenfelder and Tyler 1987; Segui 1989).

6.2 IMPORTANT ISSUES IN AUDITORY RECOGNITION RESEARCH

Based on this speculative summary description of the recognition process, let us now isolate a few important issues that need to be considered. These concern the auditory input representations to the matching process, long-term memory representations of classes of sound sources and events, and the process by which the analysed auditory input is matched to these memory representations.

6.2.1 Auditory input representations

One of the most important problems concerns determining the nature of the auditory representation of sound stimuli at different stages of processing. At any given stage, the representation of sounds may be conceived as specific values or distributions along *continuous* dimensions in a multidimensional 'space' of auditory properties or they may be analysed into a configuration of *discrete* features or elements. Further, the representation may be transformed from continuous to discrete at some particular stage of processing. Some models of recognition postulate a transformation of continuous information into discrete features at a very early (preperceptual) stage. This featural representation would then be used to compare the current stimulus with the contents of long-term memory (Massaro 1987; McClelland and Elman 1986). In other models, the segmentation of the perceived source or event into parts or components occurs quite late in processing, i.e. the stimulus is categorized as a whole only at the point of recognition (Ashby and Perrin 1988; Braida 1991; Klatt 1989; Nosofsky 1986). It is equally important to determine the nature of the represented properties or features, i.e. the information in an acoustic signal that is necessary and sufficient for perception and recognition. And finally, we need to evaluate how these various descriptions capture the appropriate invariants in the acoustic structure so that it is correctly categorized.

6.2.2 Long-term memory representations

What is the functional form of representation of previous auditory experience in long-term memory? Some recognition models postulate abstract representations of categories in terms of rules (propositions), descriptions, or patterns (Massaro 1975, 1987; Klatt 1989; Marr 1982; McClelland and Elman 1986). These kinds of model often imply a separation between generic (abstract) and episodic (specific) memories. For other models, a more diffuse representation is hypothesized where categories are conceived as bounded regions in a continuous auditory-parameter space (Durlach and Braida 1969; Ashby and Perrin 1988; Miller 1982), or else categories are composed of a set of memory traces of individual episodes that are each specific instances of a given category (Hintzman 1986; Nosofsky 1986). These latter two may not be so different in form since the set of all points in a continuous space that are closer to a given exemplar than to other exemplars could also be considered to constitute a bounded region (cf. Braida 1991). The issue of the form of representation in memory is crucial since it could be argued to impose strong constraints on the sensory information that is useful in recognizing familiar sound events as well as new instances of a familiar class of

sound events that have never been specifically encountered (e.g. a different oboe or words spoken by a person speaking with a foreign accent). Associated with this problem is the fact that sound sources remain recognizable despite enormous variability in the acoustic signals they produce, so it might be presumed that the forms of the memory representations used in recognition are isomorphic at some level of description with the stimulus invariants available to the auditory system.

6.2.3 Matching auditory to memory representations

What is the nature of the process by which newly formed auditory representations are matched to memory representations of previous experience? Is the matching of incoming sensory information with the contents of long-term memory performed by some specialized outside (executive) agent or by a process of direct activation by the contents of the information? What is the relation between the kinds of errors people make in recognition and identification tasks and the nature of the matching process? And how does the nature of this process give rise to the experience of differing degrees of familiarity and recognizability of sound events?

With these issues in mind let us now examine the available evidence concerning the perception and recognition of sound sources and events.

6.3 EXPERIMENTS ON PERCEPTION AND RECOGNITION OF SOUND SOURCES AND EVENTS

In this section, a number of experiments will be examined whose aim was to study perceptually relevant aspects of the auditory representation and identification of acoustic sources and events. The usefulness of the various potential cues for auditory recognition in general will then be considered.

Several kinds of sound sources and events have been studied in terms of the dimensions and features that contribute to their comparison and recognition. The ones that will be considered here include musical instruments (as an example of a class of sound *sources*) and complex acoustic *events* other than speech and musical sounds. Recognition of speech and speakers' voices will not be discussed since the emphasis of this book as a whole is on non-verbal audition, though some models of speech recognition will be included in Section 6.4. An introduction to the representation and identification of simple speech sounds and speakers' voices may be found in Handel (1989, Chs 8 and 9). Other areas of audition that have received much attention in both psychoacoustics and music psychology include pitch and rhythm perception and recognition.

It may be argued that both are as much meaningful 'events' (at least in music) as a musical instrument is a meaningful 'source'. These areas acquire additional interest by the fact that what listeners represent and compare mentally when listening to them are patterns of relations within a sequence (cf. Dowling and Harwood 1986; Sloboda 1985). Melodies and rhythms are thus extended auditory events of considerable complexity. However, aspects of their recognition are dealt with in other chapters in this volume (Warren, Ch. 3, Crowder, Ch. 5, Peretz, Ch. 7, and Trehub and Trainor, Ch. 9), so they will not be discussed here.

The process of non-verbal recognition in audition has not been systematically studied aside from a few studies on source or event identification. In order to evaluate the framework proposed in Section 6.2, therefore, we are obliged to consider research that has posed other experimental questions. As such, I will examine a number of different experimental tasks that are cogent to the present discussion and will then describe the results obtained with these tasks in listening to musical instrument tones and natural acoustic events.

6.3.1 Experimental tasks used to study source and event perception

Discrimination

Discrimination performance is measured for sounds that have been modified in some way to determine which modifications create significant perceptual effects. If no one was capable of hearing the difference between an original sound and a somewhat simplified version of that sound, we might conclude that the information that was removed from the sound was not represented in the auditory system, so there would be no sense in taking that level of detail into consideration in trying to explain auditory recognition. In one version of a discrimination task, listeners are presented with a pair of sounds that are identical or different in some way and are asked to decide whether they are the same or different. The experimenter varies the degree of difference between the two sounds in an attempt to understand the listener's sensitivity to the amount of change along a given stimulus dimension. Or the experimenter might make some kind of structural modification to a complex sound event in order to find out if the listener is sensitive to the modification when comparing it with the original sound. What this task is not good for is indicating a listener's perception of invariance in a class of stimuli, since the members of a class can each be discriminably different from one another, but would be treated as perceptually equivalent in some other kind of perceptual situation (such as listening to music or trying to understand what someone is saying or trying to detect a certain class of nuclear submarine on the basis of sonar signals). Some models do relate discrimination performance to identification performance in an explicit way, however, in the sense that sensitivity to change along a stimulus dimension can influence the way in which a listener categorizes the perceptual continuum (cf. Durlach and Braida 1969; Macmillan 1987; Rosen and Howell 1987).

Psychophysical rating scales

Unidimensional rating scales have been used in psychophysics since its inception by Fechner in the latter part of the 19th century (Fechner 1966). They have been particularly popular in attempting to describe the relation between physical quantities and perceptual values, such as the relation between sound intensity and loudness (cf. Thurstone 1927; Stevens 1956, 1957). In essence they may be considered to determine some psychological parameter that characterizes a given sensory continuum (Shepard 1981). Typically listeners are presented with a set of stimuli, one at a time, that vary along some analytic acoustic parameter (such as intensity) or along some more complex physical continuum (such as hardness of a percussion mallet). They are asked to rate each sound with respect to the continuum on some kind of numerical scale. The scale can be fixed (by comparison with a standard stimulus of predetermined value) or freely chosen by the subject. The experimenter then tries to establish the relation between the ratings (as a measure of perceived value) and some physical measure of the sound stimulus. If lawful relations of this kind can be established, the experimenter is in a position to make hypotheses about the auditory representation of the physical continuum being varied.

Similarity ratings

Similarity (or dissimilarity) ratings are used to discover the salient dimensions that underly the perceptual experience of a small set of sounds. A typical experiment involves presenting all possible pairs from a set of sound stimuli to a listener who is asked to rate how dissimilar they are on a given scale (say 1 to 8, where 1 means very similar and 8 means very dissimilar). In one variant of the analysis technique (called 'multidimensional scaling'; cf. Kruskal and Wish 1978; Schiffman *et al.* 1981), the ratings are then treated as psychological distances between the judged items and a computer program tries to map the distances on to a spatial configuration in a given number of dimensions. This mapping yields a geometrical structure that is interpreted as reflecting the perceptual qualities listeners used to compare the sounds, or, alternatively, as reflecting the structure of mental representations that allows them to make orderly comparisons. The interpretation of these structures is often focused on giving a psychoacoustic meaning to the spatial representation by relating the dimensions of the space to acoustical properties of the tones. What the structures may be considered to represent are the common *salient* dimensions to which listeners pay attention in the context of such an experiment. It is quite likely that the dimensions on which listeners do focus are determined by the set of sounds used in the experiment, i.e. their representations may be coded with respect to the stimulus context provided within an experimental session. In addition to giving us an idea of the auditory representations that listeners use in comparing sounds, these kinds of results can also contribute to an analysis of the processes underlying recognition when compared with identification tasks. A strong correlation has been demonstrated between similarity structures and the kinds of confusions people make among stimulus items in identification tasks (Shepard 1972; Grey 1977; Nosofsky 1986; Ashby and Perrin 1988), i.e. perceptual similarity and confusability are constrained by similar (or perhaps the same) limitations in the representation and comparison of auditory events. More recent techniques are beginning to refine this theoretical relation between similarity and identification within a signal detection framework (Braida 1991).

Matching

A matching task can be used to investigate recognition without requiring the listener to attach a verbal label to the sound source or event (Kendall 1986). A test stimulus is presented and then several comparison stimuli are presented, one of which is of the same class or category as the test stimulus. The listener is asked to say which of the comparison stimuli matches the test stimulus. The comparison stimuli may vary along perceptual dimensions that are irrelevant to the identity of the source or event being tested. This task would appear to be quite useful for answering questions about perceptual invariance within classes of stimuli as well as for investigating which aspects of the stimulus structure are used by listeners to effect an appropriate match. For example, the experimenter might perform various simplifying modifications on the test stimuli with respect to the (unmodified) comparison stimuli and estimate the degree to which such changes affect the listener's ability to make correct matches. As such, this task could give insight into the issues of both auditory and long-term memory representations and their comparison in the act of recognition.

Classification

A classification task consists of a listener being presented with a set of sound stimuli and being asked to sort them into classes on the basis of which ones go best together. A free classification task places no constraints on the number of classes, i.e. it is up to each listener to decide the appropriate number of classes and their contents. Some classification tasks may involve predefined (and named) classes within which the listener is to sort the sound events, this latter task being somewhat closer to an identification task. Based on the sorting results, the experimenter must then try to relate the class structure to properties of the stimuli, usually looking for structural similarities among stimuli that are classed together and differences among those that have been placed in separate classes. For the purposes of studying auditory recognition, these relations would indicate something about the properties that listeners use in organizing classes of sound sources and events in memory. While not a very refined technique, this experimental paradigm is easily performed by listeners and can help sketch out in rough detail the kinds of sound properties that are worth investigating more systematically.

Identification

In an identification experiment a set of sound stimuli is presented to listeners and they are asked to assign names or labels to them, one at a time. In free identification, listeners are required to draw upon whatever lexicon of names or descriptions they have acquired in their lives up to that point. More restricted identification tasks provide listeners with a list of labels for the items from which they are to choose (such as names of musical instruments, or arbitrary labels such as the numbers one to ten). Analysis of identification performance often proceeds by creating a confusion matrix in which the number of times a given item was called by a given name is recorded. This allows a detailed analysis of the confusions that listeners make among sound sources or events. Such confusions are informative about stimulus features that different items may share in their internal representations. In some studies, sound stimuli are modified in various ways and the degradation in their identification is measured as a function of the amount of modification. The main idea is that if an element in a sound is removed or is degraded too much, and that element is essential to the successful comparison of the sound with a representation in long-term memory, identification performance will deteriorate. A comparison between discrimination and identification experiments using similar stimuli is often fruitful for qualifying the utility of discrimination results for understanding recognition. The extent to which an event can be simplified without affecting identification performance is the extent to which the information is not used by the listener in the identification process, even if it is discriminable.

6.3.2 Musical instruments

A number of experiments have investigated the representation of auditory properties that distinguish the timbres of musical instrument sounds and allow listeners to identify them. In the discussion that follows the focus on timbre is not intended as a study of the perceptual quality for its own sake, but rather on its role as a perceptual vehicle for the identity of a particular class of sound sources. Musical instruments are good candidates for this kind of research since they have been thoroughly studied from the standpoints of both physical acoustics and experimental psychology (cf. Benade 1976; Leipp 1980; Barrière 1991). The search for the dimensions and features used in perception and recognition can be justified as a search for an economy of description of the stimulus and, ultimately, as an attempt to relate experimental data to models of perceptual processing. The focus in this section will be on the auditory cues listeners use to compare and identify acoustic and synthesized instrument tones. A discussion of the issues of memory representation and the matching process is deferred until Section 6.3.4.

Discrimination and identification studies

Experiments that have helped reveal the cues that are encoded in auditory representations of musical tones have studied the discrimination and identification of instrument tones that have been simplified in one way or another. One methodological issue in this kind of experiment concerns choosing the appropriate (i.e. auditory perceptual) conception of an 'element' or 'part'. While some of the studies described below have rather simplistic views of what constitutes an appropriate part of a sound event, to examine them will nevertheless be instructive.

The most simplistic of these approaches arbitrarily considers a musical sound event as composed of three parts: an attack (or onset) portion, a middle sustain (or relatively stable) portion, and a final decay (or offset) portion. Based on this conception of the sound event, manipulations involve removing or modifying various portions and determining whether any appreciable perceptual effects result. The modifications have been achieved either by cutting and splicing magnetic tape on to which sounds have been recorded (Saldanha and Corso 1964) or by editing them on a computer in digital form (Luce 1963; Kendall 1986), where the cuts and splices can be performed with much greater precision and run less risk of introducing discontinuities in the signal which would perturb listeners perceptions of them.

Saldanha and Corso (1964) investigated the identification of conventional musical instruments of the Western orchestra playing isolated tones both with and without vibrato. They sought to evaluate the relative importance, as cues in musical instrument identification, of onset and offset transients, spectral envelope of the sustain portion, and vibrato. Identification performance was surprisingly poor for some instruments even without modification, perhaps indicating that some of the information listeners normally use to identify instruments is accumulated across several tones. On the whole, however, Saldanha and Corso found that the attack and sustain portions were important for identification, i.e. performance decreased when these portions were removed, whereas cutting the decay portion had very little effect. The largest drop in performance occurred when the attack was cut out, leaving only the sustain portion. For tones with vibrato, on the other hand, the reduction was not as great with a cut attack as for tones with no vibrato. This result indicates that important information for instrument tone identification exists in the very first part of the sound event, but that in the absence of this information, additional information still exists in the sustain portion and is augmented slightly when a pattern of change that specifies the resonance structure of the source is present, as occurs with vibrato. McAdams and Rodet (1988) have shown that vibrato helps listeners extract information about the spectral envelope of synthesized vowel-like sounds. One might conclude that information in the tone that indicates how the instrument is set into vibration (i.e. the pattern of transients in the attack) contains the most important cues, followed by information concerning the global spectral structure that can be extracted during the sustain portion of a tone.

Another kind of simplification involves performing a fine-grained acoustic analysis of instrument tones and then resynthesizing them with modifications. Instruments from the string, woodwind, and brass families were employed. Grey and Moorer (1977) and Charbonneau (1981) presented listeners with different versions of each tone: the original recorded tones (Fig. 6.2(a)) and resynthesized versions of each one with various kinds of modifications. For each instrument, musician listeners were asked to discriminate among the versions and to rate how different they were. Charbonneau also used the rating scale to assess recognizability of the modified tones. These experiments showed that simplifying the pattern of variation of the amplitudes and frequencies of individual components in a complex sound (Fig. 6.2(b)) had an effect on discrimination for some instruments but not for others. Tones in which the attack transients were removed (Fig. 6.2(c)) were easily discriminated from the originals, confirming with greater precision the results from the cut-and-splice studies. Applying the same amplitude variation to all of the components (thus replacing the individual variations normally present; not shown in Fig. 6.2) grossly distorted the time-varying spectral envelope of the tone and was easily discriminated though it did not always adversely affect recognizability. Complete removal of frequency change during the tone (Fig. 6.2(d)) was also easily discriminated, although

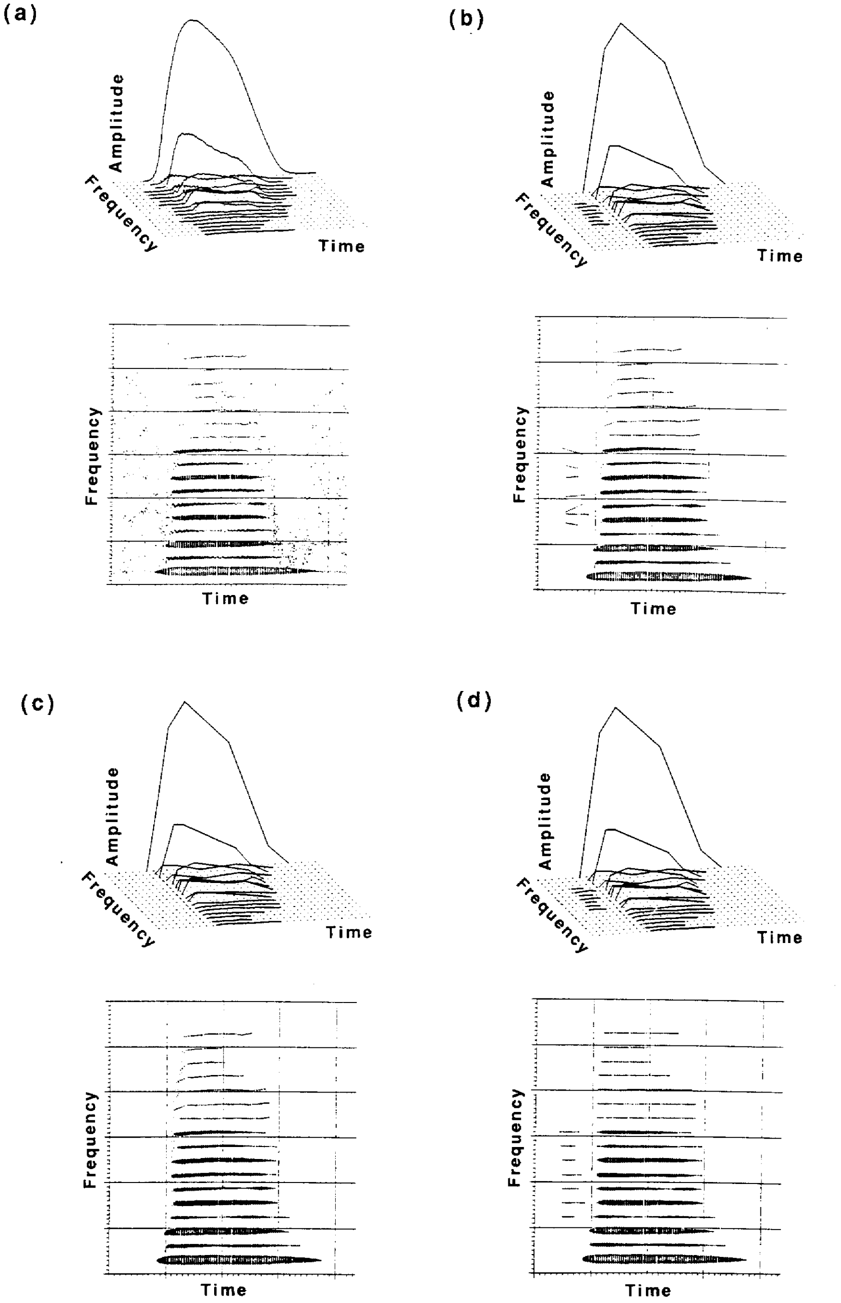

Fig. 6.2 Time–frequency perspective plots and spectrograms illustrating the modifications to a bass clarinet tone. The curves in the time–frequency plots represent the amplitude envelope for each frequency component. (a) The original tone. (b) Line-segment approximations of the amplitude and frequency variations. (c) Line-segment approximations with deletion of initial transients. (d) Line-segment approximations with flattening of the frequency variations. (From Figs 2, 3, 4, 5, Grey and Moorer (1977) pp. 457–8. © Acoustical Society of America, 1977. Reprinted with permission.)

applying a common frequency variation to all components (not shown) had only a weak effect on discriminability and none on recognizability. This latter result suggests that small modifications to the coherence of change among the frequencies of the components, while noticeable, do not much affect the identification of the instrument, which is more based on spectral structure and attack characteristics.

One difficulty in generalizing these results to everyday situations should be mentioned. All of the experiments described above employed isolated tones played at the same pitch, loudness, and duration. The temporal and spectral features of musical instrument tones change quite noticeably across pitch and loudness levels and so the features that contribute to identification of individual tones at one pitch and loudness may not be the same as those that contribute at another. In addition, the way an instrument's tones vary with pitch and loudness may be a strong cue in itself for recognition when listening to the instrument in a musical context. Two studies have investigated discrimination and recognition of modified instrument tones within musical contexts.

Grey (1978) used the kind of simplified tones shown in Fig. 6.2(b) for three instruments (bassoon, trumpet, and clarinet). He created notes at other pitches by transposing the instrument spectrum to higher or lower frequencies. He then asked listeners to discriminate the simplifications of a given instrument for either isolated tones or the same tones placed in musical patterns that differed in the number of simultaneous melodic lines, rhythmic variety, and temporal density. An effect of musical context would thus be measured as a difference in discrimination performance among the conditions. An increasingly complex musical context did not affect discrimination between original and modified versions of the bassoon, but hindered such discrimination for the clarinet and trumpet. An acoustic analysis of the bassoon tone showed that the simplification involved a change in the spectral envelope, which was not the case for the other instruments. Changes for the bassoon were described by listeners as differences in 'brightness' or 'high frequencies' which seem to be related to the spectral envelope, while those for the clarinet and trumpet were in 'attack' or 'articulation'. It seems that small spectral differences were slightly enhanced in single-voice contexts compared with isolated tones and multi-voiced contexts, though discrimination remained high. Articulation differences, on the other hand, were increasingly disregarded as the complexity and density of the context increased. These results suggest that in cases where demands on perceptual organization and the storing and processing of sequential patterns are increased, fine-grained temporal differences are not preserved as well as spectral differences. Since such differences are not used for discrimination in musical contexts, it seems unlikely that they would be available to the recognition process.

One possible confounding factor in Grey's (1978) study is the fact that the different pitches were created by transposing a single tone's spectrum and then concatenating and superimposing these tones to create the musical patterns. This removes any normal variation of spectral envelope with pitch as well as any articulation features that would be involved with passing from one note to another in a melody. Kendall (1986) controlled for these problems in an experiment in which the tone modifications were of the digital cut-and-splice variety. In his experiment, listeners heard two different melodies played in legato (connected) fashion. The first one was an edited version of the melody played by one of three instruments (clarinet, trumpet, or violin). The second melody was then played in unedited form by each of the three instruments in random order. Listeners had to decide which of the instruments playing the second melody matched the one playing the first melody. Several kinds of modifications were presented: normal tones, sustain portion only (cut attacks and decays), transients only (with either a silent gap in the sustain portion, or an artificially stabilized sustain portion). The results suggest that transients in isolated notes provide information for instrument recognition when alone or coupled with a natural sustain portion, but are of little value when coupled to a static sustain part. They are also of less value in continuous musical phrases where the information present in the sustain portion (most probably related to the spectral envelope) is more important. This conclusion confirms that of Grey (1978) with a recognition task and stimuli that include more realistic variations.

From these studies of the effects of musical context on discrimination and recognition we can conclude that the primacy of attack and legato transients found in all of the studies on isolated tones is greatly reduced in whole phrases (particularly slurred ones). The spectral envelope information present in the longer segments of the sustain portion is thus of greater importance in contexts where temporal demands on processing are increased.

A caveat on the interpretation of studies in which attack and decay transients are excised should be mentioned. The cutting out of an attack is quite simplistic in its conception of an instrument tone's morphology. It does not really remove an attack altogether; it replaces the original attack with another one. The new attack may be considered strong perceptual evidence *against* the instrument in question and so a reduction in identifiability is not surprising. This phenomenon may be related, to some extent, to that of phonemic restoration (Warren 1970). When a phoneme is removed and replaced by silence the word is less well identified than when the silence is replaced by a noise burst of appropriate spectral composition, i.e. the noise could have masked the phoneme were it actually present. Silence in this case is strong evidence against the phoneme, since abrupt stops and starts in the signal are encountered. The masking noise burst is only weak evidence against the phoneme since the phoneme might, in fact, still be present. It is possible that different results might be obtained in the musical tone studies mentioned above, if similar procedures were used. Having a wrong attack (i.e. the cut attack) plus the right sustain may give lower identification performance than having a (potentially) masked attack plus the right sustain. In the latter case, the auditory recognition process, not being confronted with contradictory evidence, would be able to make better use of the sustain portion of the sound in making a partial match to the memory representation for the tone.

Multidimensional scaling studies

Several studies have performed multidimensional scaling analyses on dissimilarity ratings for musical instrument tones or synthesized tones with characteristics that resemble those of musical instruments (Plomp 1970, 1976; Wedin and Goude 1972; Wessel 1973; Miller and Carterette 1975; Grey 1977; Krumhansl 1989). In all of these studies, the perceptual axes have been related either qualitatively or quantitatively to acoustic properties of the tones as will be described below.

Grey (1977) recorded, digitized, and then analysed tones from 16 instruments played with equal pitch, loudness, and subjective duration. Listeners then rated the dissimilarities for all pairs of tones. Grey settled on a three-dimensional structure as capturing the greatest amount of the variation in the data structure while not having so many dimensions as to make the structure difficult to interpret. The final perceptual structure is shown in Fig. 6.3.

Grey qualitatively related the axes to acoustic properties in the following way. The first dimension represents the spectral energy distribution in the sound (or its spectral envelope) and is primarily a spectral dimension. This can be thought of as representing the degree of 'brightness' of the sound quality (Wessel 1979). Instruments with low brightness are the French horn and the cello played *sul tasto* (a technique of bowing over the fingerboard that gives a soft, velvety kind of sound). Instruments with high brightness include the oboe as well as the trombone played with a mute (which gives it a strident quality). The difference in brightness is illustrated by comparing the instrument spectrograms or frequency–time perspective plots of FH (French horn) and TM (trombone, muted) in Figs 6.4 and 6.5. Note primarily the difference in number of harmonics. The second dimension is related to a combination of the degree of fluctuation in the spectral envelope over the duration of the tone and the synchrony of onset of the different harmonics. This is a spectro-temporal dimension that has been called 'spectral flux' by Krumhansl (1989). Instruments with high synchronicity and low fluctuation include

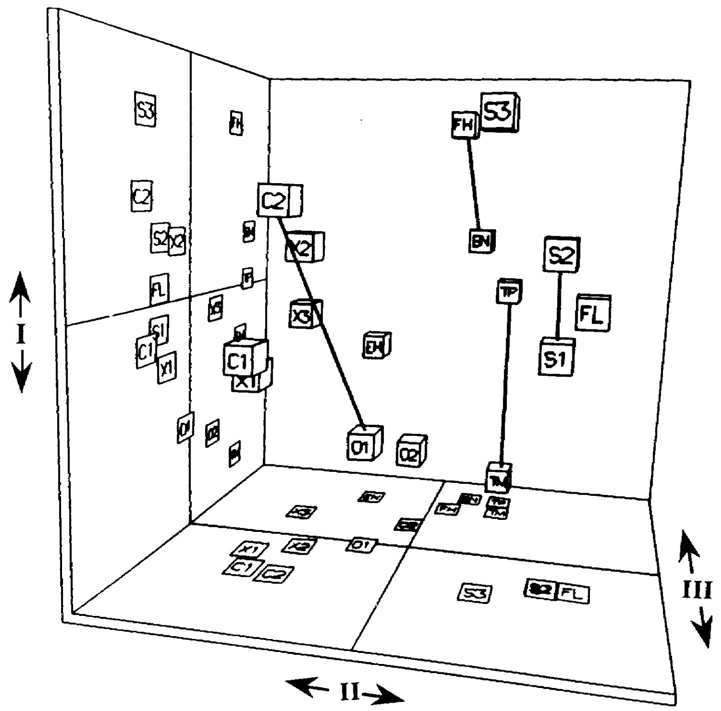

Fig. 6.3 Three-dimensional scaling solution for 16 musical instrument tones equated for pitch, loudness, and perceived duration. Two-dimensional projections of the configuration appear on the wall and floor. Abbreviations for the instruments: 01 and 02, two different oboes; C1 and C2, E♭ and bass clarinets; X1, X2, and X3, saxophones playing softly and moderately loud, and soprano saxophone, respectively; EH, English horn; FH, French horn; S1, S2, and S3, cello playing with three different bowing styles: *sul tasto, normale, sul ponticello*; TP, trumpet; TM, muted trombone; FL, flute; BN, bassoon. Dimension I (top–bottom) represents spectral envelope or brightness (brighter sounds at the bottom). Dimension II (left–right) represents spectral flux (greater flux to the right). Dimension III (front–back) represents degree of presence of attack transients (more transients at the front). Lines connect pairs of timbres that were modified by Grey and Gordon (1978) (see text and Fig. 6.6). (From Fig. 3, Grey and Gordon (1978) p. 1496. © Acoustical Society of America, 1978. Adapted with permission.)

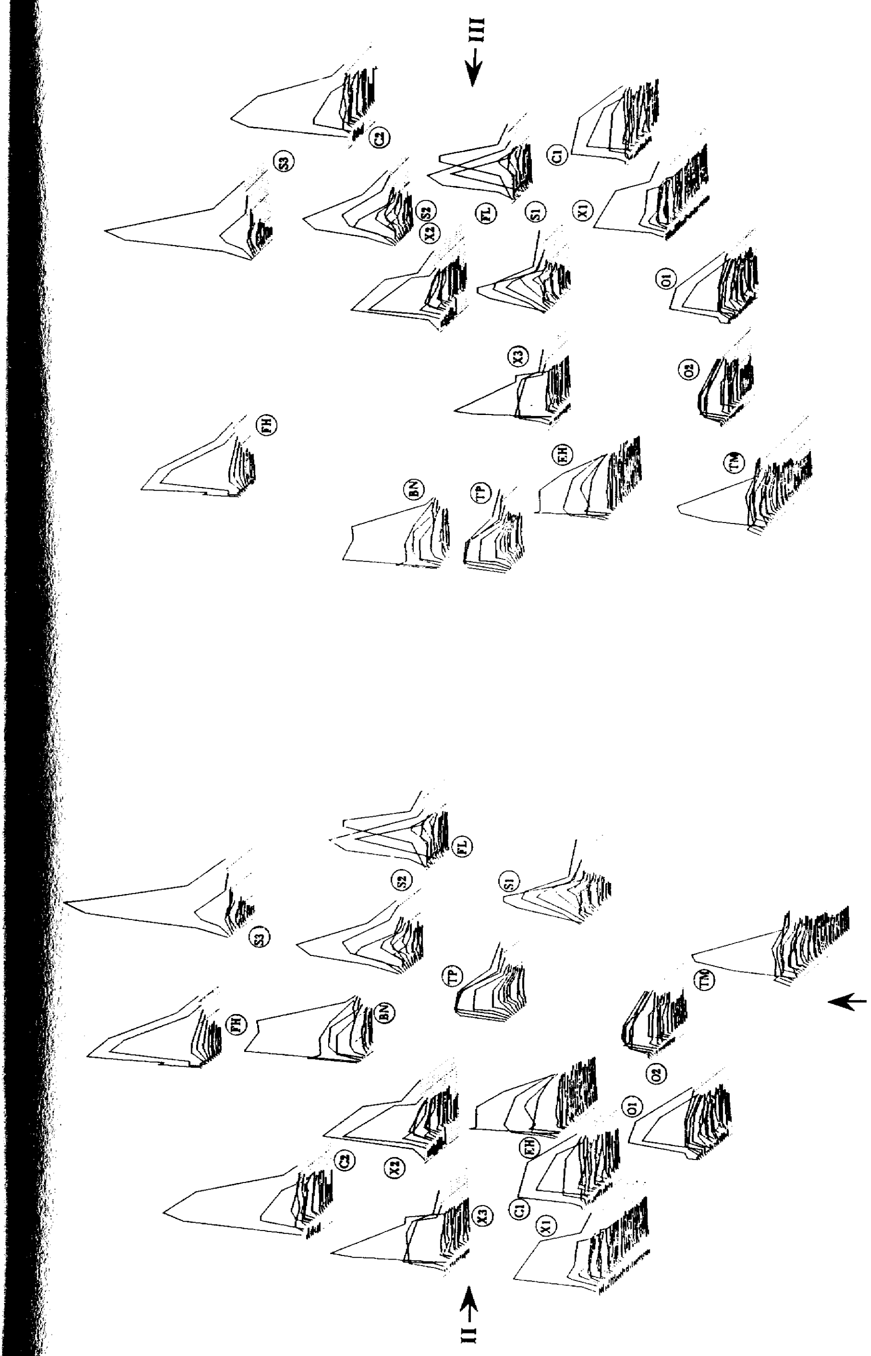

Fig. 6.4 Two-dimensional projections of the three-dimensional solution in Fig. 6.3 on to the dimension-I–dimension-II plane (left) and the dimension-I–dimension-III plane (right). The circle with the abbreviation for the instrument name (see Fig. 6.3 caption) indicates the position in the plane. Next to each label is the time-frequency perspective plot for that instrument tone. (From Fig. 2, Grey (1977) p. 1273. © Acoustical Society of America, 1977. Adapted with permission.)

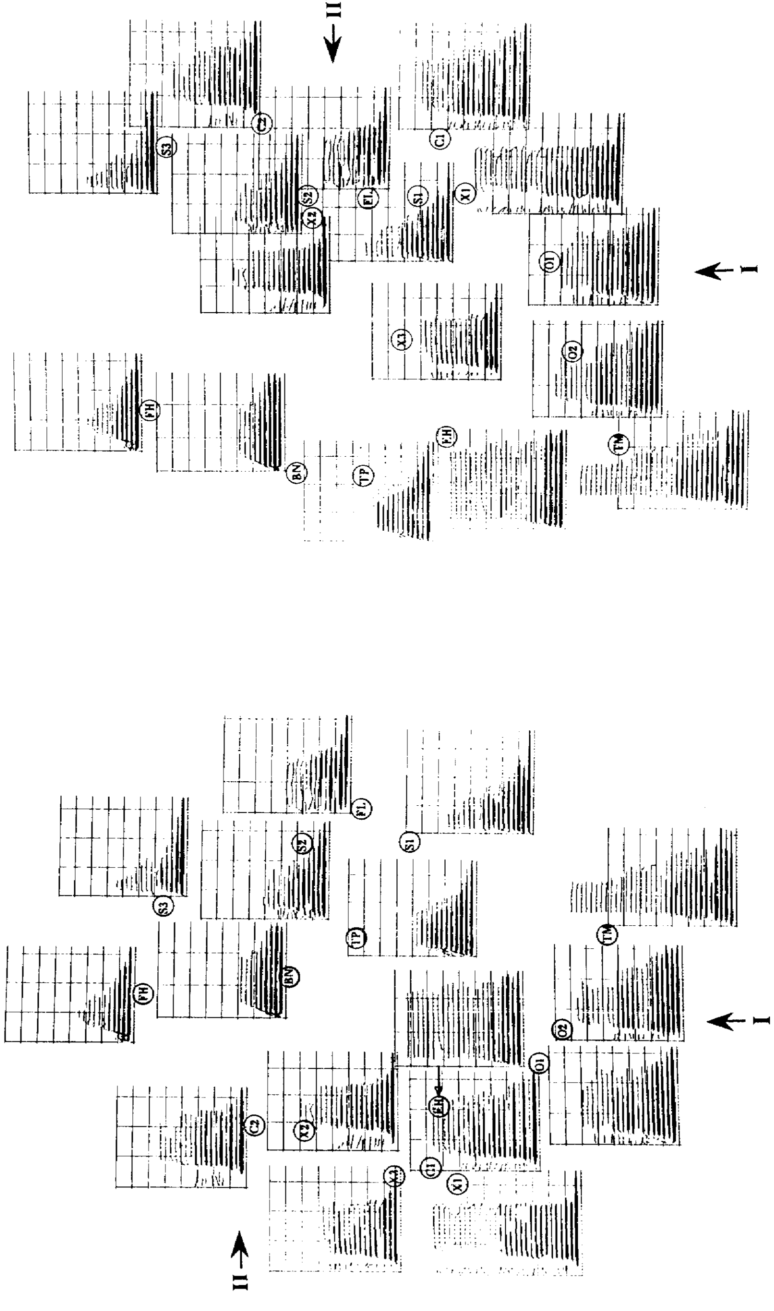

Fig. 6.5 Two-dimensional projections as in Fig. 6.4, except that the spectrogram for each instrument tone is shown here. (From Fig. 3, Grey (1977) p. 1274. © Acoustical Society of America, 1977. Adapted with permission.)

clarinet and saxophone whose harmonics start, stop, and fluctuate together in amplitude. Instruments with low synchronicity and high fluctuation include the flute and cello. The brass (trumpet, trombone, French horn) and double reed (oboe, English horn, bassoon) instruments are somewhere in the middle. The difference in flux is illustrated in Fig. 6.4 by comparing X1 (saxophone) and FL (Flute). Note the relative homogeneity of the envelopes for X1, whereas those for FL are quite varied. The third dimension represents the relative presence of inharmonic transients in the high frequencies just before the onset of the main harmonic portion of the tone. This dimension might be called the 'attack quality' dimension. It is primarily temporal in that the temporal position is crucial, as is the relative lack of periodicity in the transient waveform. Instruments that rate high on this dimension include the strings, flute and single reeds (clarinet, saxophone), whereas the brass, bassoon and English horn have lower ratings. The difference in attack quality is illustrated by comparing EH (English horn) and C1 (E♭ clarinet) in Fig. 6.5 (right). Note the absence of preliminary transients for EH and their abundance in C1. Grey found that there was a strong degree of correspondence among listeners' ratings in this study as well as across two sets of judgements on the same set of stimuli, indicating that the mental representation of the set of timbres is relatively stable and more or less shared by the listeners. He also found a very high correlation between the matrix of similarity judgements and the matrix of confusion errors listeners made when asked to identify the sounds, thus indicating the utility of examining similarity rating studies for developing an understanding of the auditory representations that contribute to recognition processes.

Another study, conducted by Grey and Gordon (1978), hypothesized that since the perceptual dimensions seem closely correlated with acoustic properties, modifying the acoustic properties for a single perceptual dimension in systematic ways ought to cause changes of the position of a given tone along that dimension. To test this hypothesis, they selected four pairs of tones from among the original 16 and exchanged their spectral envelopes, trying to leave the other properties as intact as possible. They then reinserted the modified tones into a multidimensional scaling study with the other eight unmodified tones. The results demonstrate that in all cases the tones exchanged places along the brightness dimension, though in some cases displacements along other dimensions also occurred (compare the connected pairs of instruments in Figs 6.3 and 6.6). These displacements still respected the nature of the perceptual dimensions—envelope changes that also modified the way in which the spectral envelope varied with time for a given tone resulted in appropriate changes along the dimension of spectral flux.

One may question the validity of the assumption that extremely complex sounds like those corresponding to musical instrument tones really differ

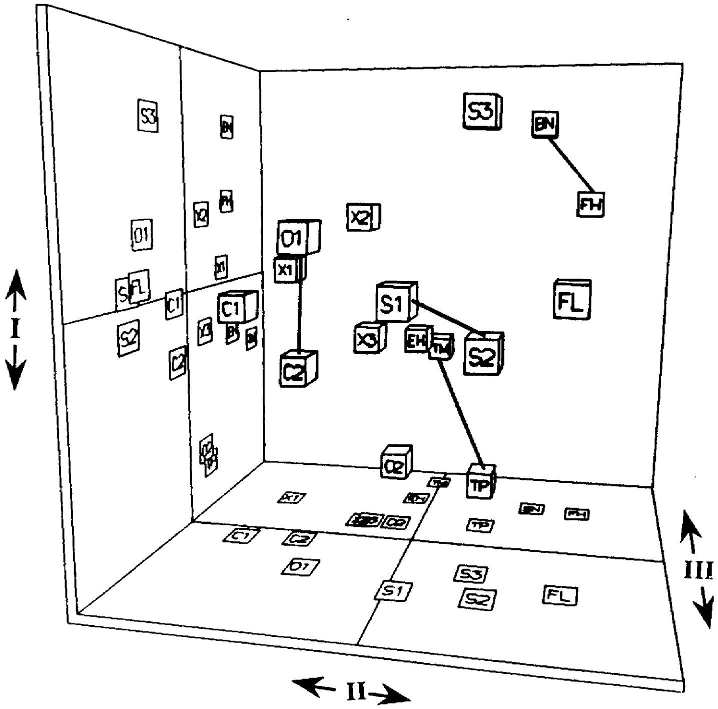

Fig. 6.6 Three-dimensional spatial solution for 16 instrument tones, four pairs of which swapped spectral envelopes. These pairs are connected by lines. Dimensions and abbreviations are described in Fig. 6.3. (From Fig. 2, Grey and Gordon (1978) p. 1496. © Acoustical Society of America, 1978. Adapted with permission.)

in terms of only a few underlying (or common) perceptual dimensions. Each timbre may also have unique characteristics that are not easily coded along continuous dimensions, such as the returning bump of the hopper on a harpsichord, the odd-harmonic structure of the clarinet spectrum, or the 'blatt' of a trombone attack, and so on. This possibility is evidenced by a certain degree of variability in the similarity data from the early scaling studies on timbre which is not accounted for by their scaling solutions. Krumhansl *et al.* (1988) used a set of 'instruments' created by digital sound synthesis (Wessel *et al.* 1987) in which many of the sounds imitated those of musical instruments, while others were intended to simulate hybrids of instruments, e.g. the 'vibrone' is a cross between vibraphone and trombone. A multidimensional scaling analysis technique developed by Winsberg and Carroll (1989*a,b*) was used which, in addition to uncovering the common perceptual dimensions shared by the tones, would allow for specific dimensions or features that applied only to individual timbres. The three-dimensional solution was remarkably similar to those found in previous studies. Within these common dimensions the hybrid instruments were almost always situated somewhere between their two imagined progenitors. Further, the analysis of specificities showed that a significant amount of variability in the similarity judgements, not attributable to the common dimensions, could be accounted for by postulating unique features for some of the instruments, such as harp, harpsichord, clarinet, and vibraphone. Further acoustic and perceptual analyses will be needed to relate these specific features to the acoustic and perceptual properties of the individual instrument tones, but this technique seems promising for tagging sounds that have special perceptual features that may in turn contribute significantly to identification performance. That such is the case is suggested by the studies of Strong and Clark (1967*a,b*) who found that the unique spectral envelope of some instruments (such as clarinet and trumpet) contributed more to identification than the amplitude envelope. However, when the amplitude envelope was unique (as in trombone and flute), it had a greater importance. Therefore, listeners appeared to use whatever characteristic was likely to specify the instrument with the least ambiguity and were not constrained to listening for a single cue across all possible sources.

Summary of studies on musical instruments

A number of characteristics of instrument tones presented in isolation seem to be used for their discrimination (Grey and Moorer 1977; Charbonneau 1981). These include, in decreasing order of importance, information present in the attack portion, information concerning the spectral envelope and its evolution through time extracted during the sustain portion, and the presence of small, random variations in the component frequencies. Other characteristics were only weakly discriminable at best, including the degree of coherence in frequency variation on the components, the degree to which the amplitude and frequency variations were simplified (though this was more discriminable for some instruments than for others), and small variations in the temporal pattern of onsets of the frequency components. When similar kinds of tones were placed in musical contexts of various degrees of complexity, temporally based cues seemed to lose importance for discrimination in favour of more spectrally based cues (Grey 1978; Kendall 1986).

Multidimensional scaling studies revealed the importance of spectral envelope distribution, attack character, and spectral evolution in comparisons of the degree of similarity among tones (Grey 1977; Krumhansl 1989), though the perceptual dimensions found to be important are probably quite sensitive to differences in the set of sound events presented to listeners (cf. Miller and Carterette 1975; Plomp 1976). Further analyses have revealed that some tones may possess unique features or dimensions, though the acoustic and perceptual nature of these specificities remain to be determined (Krumhansl 1989; Winsberg and Carroll 1989*a*,*b*).

Identification studies have shown a strong inverse correlation with similarity ratings (tones judged as being similar are more often confused in labelling tasks (Grey 1977)). Such studies have also confirmed the importance of the attack portion of the sound event as well as the patterns of change in the sustain portion that signal the nature of the spectral envelope. They showed that the decay portion of the tone contributes relatively little to an instrument's identity on the basis of an isolated tone (Saldanha and Corso 1964). Other studies confirm the importance of spectral envelope and temporal patterns of change for identification of modified instrument tones (Strong and Clark 1967*a*,*b*).

6.3.3 Natural acoustic events other than speech and musical sounds

A large class of sounds that is only beginning to be studied, primarily because the ability to synthesize and control them has been severely limited by the available techniques, is comprised of the complex acoustic events of our everyday environment. What often distinguishes many of these sounds is the complexity of their spectral and temporal structure on the scale of the microproperties seen in connection with musical instrument tones, as well as the complexity of their temporal structure on a larger time scale. The breaking plates example discussed in the introduction demonstrates how the textural and rhythmic evolution over the entire event specifies both the nature of the sources involved as well as the interactions among them that give rise to the global event. Below I will review four studies on the perception and recognition of such complex events, two concerning brief events and two concerning more complicated event structures.

Freed (1990) studied listeners' abilities to rate on a unidimensional scale the relative hardness of a mallet striking metal pans of four different sizes. Each pan was struck with six mallets differing in hardness (metal, wood, rubber, cloth-covered wood, felt, felt-covered rubber). He performed acoustic analyses and extracted four abstract 'timbral predictor' parameters that were derived from a simulated peripheral auditory representation of the acoustic signal. These predictors included measures of overall energy, spectral distribution, rate of spectral evolution, and rate of decay of the event. Listeners' ratings increased with mallet hardness and were completely independent of the kind of pan being struck. Since the sounds contain information both about the nature of the pan (the resonator) and the nature of the mallet (the exciter), it appears that listeners are able to abstract the nature of the mallet alone from the combined

acoustic information. The timbral predictors selected by Freed were reasonably well correlated with the mallet hardness ratings. However, while the subjective ratings were independent of the pan type, the timbral predictors varied as a function of both the pan and the mallet type, indicating that they did not succeed in extracting psychoacoustic invariants that specified *only* the mallet. A great deal of research remains to be done on the invariant cues that allow listeners to separately characterize and recognize resonators and their sources of excitation.

Repp (1987) studied the sound of two hands clapping. He was primarily interested in what the spectral information in a hand clap contributed to the identification of both the clapper and the configuration of the clapping hands. He recorded 20 people clapping individually and analysed acoustically the spectral structure of each person's average clap. He performed a data reduction analysis on the individual clap spectra and attempted to recover the principal spectral features that describe the ensemble of analysed claps. The main idea was that if these features exhaustively described all of the claps, each clap should be constructable by mixing together the main features with differing weights, e.g. 50 per cent of feature 1 + 20 per cent of feature 2 + 5 per cent of feature 3 + 25 per cent of feature 4. Another person with another hand configuration would have different weights. It is then the task of the analyst to determine the physical origins of each feature. Unfortunately, as is often the case with this kind of analysis, it is not completely clear what physical properties underly each component. Analysis of various hand configurations produced by the author (Fig. 6.7) suggested that about half the variation in spectral features can be specifically associated with hand configuration, e.g. a low-frequency peak seems to be associated with a palm-to-palm resonance and a mid-frequency peak appears to be associated with a palm-to-finger resonance. Others factors such as hand curvature, fleshiness of palms and fingers, and striking force may also contribute to the variation in spectral features across clappers, but were not specifically analysed in this study.

Listeners had all participated as clappers and knew one another. They were asked to identify the clapper from a list of participants, including themselves. Overall identification of specific individuals was quite poor though people recognized their own claps better than those of their colleagues. From these identifications Repp looked to see if there was any consistency among listeners for identifying a clapper as male or female. They were quite consistent at assigning a male or female person to a given clap, but the judgements showed no relation whatever to the actual sex of the clapper. There seem to be certain features that are used by people to evaluate sex. For example, series of claps that were faster, softer, and had higher resonance frequencies were more often judged as being produced by females. Rather than actually representing perception

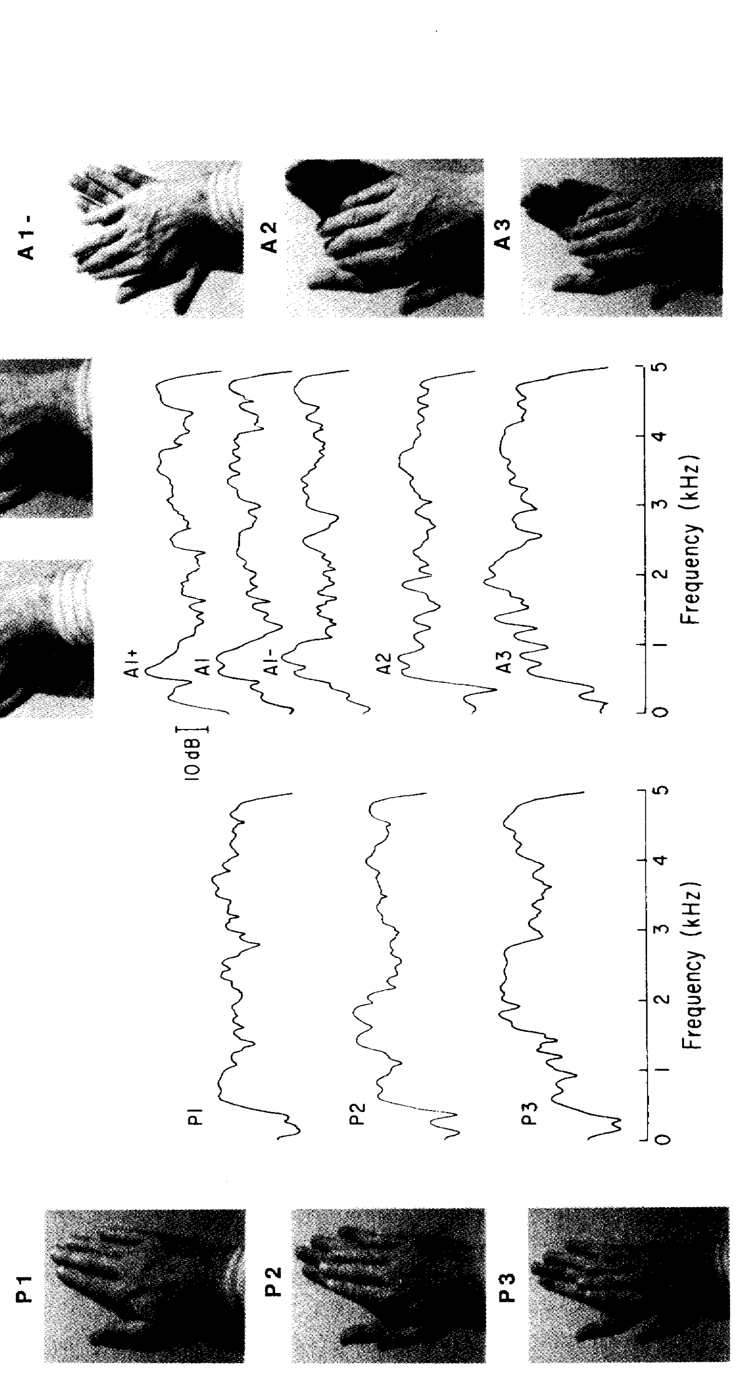

Fig 6.7 Photographs of hand-clapping configurations and their averaged, normalized spectra. (From Figs 3, 4, Repp (1987) pp. 1103–4. © Acoustical Society of America, 1987. Adapted with permission.)

of maleness and femaleness of claps, these results may more reflect certain auditory cultural stereotypes people have about how males and females clap. This result nevertheless suggests that certain classes of acoustic characteristics are associated, in the listener's memory, with gender characteristics. When asked to identify claps in terms of hand configuration, listeners were quite good at the task when all of the claps were performed by the same person. Repp concluded that sound emanating from a source conveys perceptible spectral information about the configuration of that source which the listener can then use to recognize or identify the configuration. However, when the 20 original clappers were intermixed, the configuration was difficult to recover, which weakens this conclusion and indicates that acoustic information specifying hand configuration is not invariant across clappers.

Warren and Verbrugge (1984) conducted an experiment in which listeners were asked to classify sound events as representing either 'bouncing' or 'breaking' glass objects or as events that did not correspond to either class. They were particularly interested in the importance of time-varying properties of sound patterns in specifying mechanical events in the world. In the first experiment, stimuli consisted of recordings of glass jars falling from various heights on to a hard surface. Several features distinguished the bouncing and breaking events: bouncing events are specified by simple sets of resonances which have the same accelerating rhythm pattern, and breaking events are specified by several sets of different resonances with different rhythmic patterns. In the bouncing event, each impact sets the entire bottle or jar into vibration and the different vibration modes are heard as resonances that decay rapidly until they are restimulated by the next impact. This spectral structure stays relatively constant across the several impacts. The rhythm of the impacts is a kind of irregular but accelerating pattern (called damping) where the impacts follow closer and closer together and decrease in amplitude as the object is pulled to rest by gravity. In the breaking event, a noise burst is present at the beginning which corresponds to the initial rupturing of the object. Then a series of overlapping, independent accelerating rhythms are heard which correspond to each piece of the broken object bouncing and coming to rest. The broken pieces do not have as much resilience and so come to rest more quickly which gives a much shorter total duration to the breaking than to the bouncing event. The spectrum for a breaking event also covers a wider range of frequencies. Results for these recorded natural sounds showed a 99 per cent correct classification of the original bouncing and breaking events.

In another experiment, synthetically reconstructed sound events were derived from the recordings. In the artificial events, the rhythm pattern over the course of the event was modified. These artificial sounds were designed to control for effects of duration and width of the frequency spectrum present in the natural sounds. In other words, Warren and Verbrugge (1984) wanted to test for the importance of the global rhythmic pattern in identification, by keeping the spectral and temporal microproperties constant across conditions. Four broken pieces were individually dropped and their bouncing behaviour was recorded. They were then digitally combined such that their successive impacts were either synchronous, following the same rhythmic pattern to simulate bouncing, or completely independent, to simulate breaking. Results for these constructed sequences showed an 89 per cent correct identification of the predicted categories based on the rhythmic behaviour. A single damped sequence with all resonances on each impact was heard as bouncing since there was a spectral invariance across the impacts as well as an appropriate damping behaviour. Overlapping damped sequences at different rates with different resonances on each sequence were heard as a breaking event due to the rhythmic diversity in the overall pattern. In some additional demonstrations, Warren and Verbrugge showed that a single resonance in a natural bounce sequence (i.e. with accelerating rhythm) but without the normal decreasing amplitude pattern was still heard as bouncing since the appropriate rhythmic pattern was combined with an invariant spectral structure. Also, a periodic sequence with all resonances heard on each event, but which did not accelerate, was not heard as bouncing in spite of a decreasing amplitude pattern, since this sequence did not reproduce the appropriate rhythm. These results demonstrate that combined spectral and temporal patterning provides the information necessary to distinguish between the two event categories.

Van Derveer (1979) studied the free identification and classification of a series of complex acoustic events such as shuffling cards, jingling keys, knocking, and sawing. In the free identification task, listeners most often named mechanical events and only reported abstract perceptual qualities when they could not recognize the source. In the classification task, the events were clustered into categories such as repetitive, percussive sounds (hammering, knocking), crackling sounds (crumpling or tearing paper, shuffling cards), or clinking sounds (glasses struck together, a spoon striking porcelain). When these classification judgments were combined with confusion errors in the identification task, they showed grouping by temporal patterns and spectral content. Confusions were made primarily within classes and only rarely occurred across classes, e.g. hammering might be confused with knocking (lower frequency sounds with a regular rhythm), but rarely with jingling (higher frequencies with an irregular rhythm). According to Handel (1989, Ch. 8), it would seem most natural to refer to sounds by the object and mechanical event that produce them, which suggests the possibility that listeners judged the similarity of the actions that produced the sounds (presumably on the basis of transformational invariants). This hypothesis would seem, however, to ignore the fact that the groups could also be partially interpreted with respect to the acoustic properties of the materials involved (i.e. structural invariants), which would certainly entail differences in the constraints on spectral and temporal behaviour, e.g. pounding or grating on wood, shuffling, tearing and crumpling paper, jingling metal, striking glass or ceramic. The question of whether these phenomena truly demonstrate a direct auditory perception of the physical cause of the event, or a more likely post-auditory semantic reconstruction based on recognizable acoustic characteristics of the source materials and knowledge of the ways they are excited, cannot be answered within the experimental framework of this study.

In summary, it seems clear that the way complex sound events are produced gives rise to spectral and spectro-temporal properties (Repp 1987; Freed 1990) as well as more global patterns of change (Van Derveer 1979; Warren and Verbrugge 1984) that are used by the auditory system in the process of acoustic event recognition. What remains to be elucidated are the precise cues used for various sound sources and the way they are used in various sound contexts. It is also important to conduct more systematic research on the way the auditory system is capable of independently extracting cues that specify the resonators or exciters that are involved in mechanical sound-producing events (cf. Huggins 1952). Recent progress in the development of physical models for digital sound synthesis should provide the necessary tools and impetus to carry this important work forward.[2]

6.3.4 Discussion of the experimental evidence

Auditory representations

From the set of studies that we have looked at, we can summarize some of the acoustic properties that appear to be useful for the auditory comparison and recognition of sound sources and events. It should be kept in mind that, according to the information-processing approach adopted here, the discussion of *useful* acoustic properties implies that they are somehow successfully represented in the auditory system, at least to the level of input to the matching process. Two large classes of properties can be distinguished: microproperties and macroproperties, though the boundary between the two remains a bit fuzzy. The two classes are primarily distinguished by the rates at which things occur.

Microproperties would be extracted over relatively short time periods

[2] For more information on digital sound synthesis using physical models see Adrien (1991), Florens and Cadoz (1991), Cadoz (1992) as well as a forthcoming special issue (in preparation) of the *Computer Music Journal* on 'Physical models of instruments'.

(tens to hundreds of milliseconds) corresponding to a single percussive or continuous excitation of a resonant body. *Spectral microproperties* define the shape of the spectrum at given points in time (or the average shape over a short time period). They include the form of the spectral envelope (related to general resonance properties of the sound source) as well as the frequency content of the spectrum (harmonic, inharmonic, noise: related both to the nature of the exciter and to the relations between the modes of vibration of the object). All of these microproperties have been shown to contribute to sound source comparison and identification in the studies described above. *Temporal microproperties* concern the variation over time of the amplitude or frequency of the sound or of the spectrum as a whole. They include ongoing fine-grained temporal characteristics (related to rapid amplitude and frequency modulations that give rise to perceptions of roughness and jitter or vibrato), the form of the amplitude envelope (related to articulation style and onset rate), and the presence of low-amplitude, inharmonic transients at the beginning of an event which are due to non-linear behaviour of the vibrating object as it is set into motion and before it settles into a stabilized oscillatory behaviour. *Spectro-temporal microproperties* describe the change in shape of the spectrum over time. They include the pattern of onsets of frequency components in a sound event (synchrony of onsets), and fluctuations in the spectral envelope during the course of a sound event. The relative importance of the various cues has been shown to depend on the stimulus context. Much more work is needed to refine our knowledge of the degree to which these different cues are necessary and sufficient for recognition and of *how* the local stimulus context might influence their necessity and sufficiency. For example, the change in relevance of temporal and spectral cues when musical instruments must be discriminated and recognized on the basis of single tones or full melodies suggests that reliable sensory information accumulated over time is based more on spectral than on temporal cues, although this depends on the instrument being studied . It may also be, as suggested by the ecological position, that reliable detection of stimulus invariants requires variation in irrelevant cues so that the listener can detect what does not change.

Macroproperties would be extracted over longer periods of time (hundreds of milliseconds to a few seconds) corresponding to multiple stimulations of the sound sources participating in the event as in bouncing, knocking, jingling, and so forth. *Temporal patterning macroproperties* represent the rhythmic and textural structure of an extended event and are related either to the gesture by which an object (or a group of objects) is set into vibration, or to changes in the state of integrity of an object (such as a breaking plate or glass jar). *Spectral variation macroproperties* have been found to correspond to the nature of the material being stimulated as well as to the transformations of its geometry. They would include the presence of temporally co-ordinated resonances or, conversely, the diverse unco-ordinated resonances that indicate the presence of multiple sources (e.g. a single bouncing jar versus several jar pieces bouncing independently).

The experimental data on identification confusions among musical instrument tones suggest that a continuous dimensional representation is in general more appropriate than a discrete featural representation. Salient common dimensions that have been shown to be used in comparing instrument timbres and which are highly correlated with data on confusions among timbres include brightness, attack quality, and spectral flux. Furthermore, results from experiments that treated tones as consisting of parts (attack, sustain, decay) can most likely be explained in terms of these dimensions. On the other hand, the new techniques of specificity analysis for similarity data (Winsberg and Carroll, 1989*a*,*b*) may require us to refine this position since they may allow us to isolate unique properties in order to conduct further experimentation to determine whether they are continuous or featural. For example, the spectral envelope is a continuously varying (multiple) 'dimension', whereas properties like the 'bump' of the hopper on the harpsichord may be featural within the context of musical instrument recognition, since the 'bump' is either present or absent (cf. Garner 1978). Once their featural or dimensional status has been established, the extent to which such features contribute to recognition remains to be determined. For the moment, though, models that are based on continuous representation would seem to be sufficient to explain the majority of the available experimental evidence.

The application of the notion of continuous dimensions to more complex sound events such as bouncing bottles and jingling keys may also be appropriate. These events can be characterized in part by their macrotemporal patterning, i.e. rhythmic and textural patterns. It seems clear that the application of discrete features to such patterns would be quite difficult. What remain to be determined in greater detail are the nature of the dimensions of representation underlying these patterns and the characteristics which allow listeners to distinctly classify them.

Long-term memory representations

The importance of continuous dimensions in auditory representation suggests their prominence in the representation of sound events in long-term memory as well. The strong correlation between perceived similarity and identification errors found by Grey (1977) supports this notion. It seems intuitively obvious that the more similar two sounds are, the more likely it should be for them to be confused with one another. These results would argue in support of models of recognition that explicate the relation between stimulus similarity and identification errors for stimuli represented along continuous dimensions.

Studies of stimulus modifications such as cutting out bits of sound (Saldanha and Corso 1964; Kendall 1986), simplifying amplitude and frequency behavior on harmonic components (Grey and Moorer 1977; Grey 1978), filtering (Pollack *et al.* 1954), or changing aspects of the resonance structure of a sound source (Kuwabara and Ohgushi 1987) have shown that identification performance degrades progressively over a certain range of variation in these parameters. That performance degradation is progressive rather than abrupt could either reflect the fact that categories have large, fuzzy boundaries or indicate that a large number of cues contribute to identity. It also suggests that if an individual characteristic does not match in the comparison between an auditory and a memory representation, a certain degree of identifiability is maintained by other properties, which would support Handel's (1989, Ch. 8) suggestion that recognition is supported by the accumulation and comparison of multiple cues.

The positive effect of frequency modulation (vibrato) on identification performance in the absence of attack transients (Saldanha and Corso 1964; McAdams and Rodet 1988) as well as the predominance of spectral cues in musical contexts provides evidence that dynamic musical instrument tones may give rise to a representation of the resonance structure that is accumulated over time and then represented as an abstract form in memory. Categories derived from spectral envelopes are ubiquitous in speech (see work on vowels by Macmillan *et al.* 1988) and have also been shown to contribute to the identification of hand configuration in clapping (Repp 1987).

Residual acoustic information following removal or simplification of fine-grained variations in different acoustic parameters (Grey and Moorer 1977; Charbonneau 1981) has been shown to have a lesser or greater degree of impact on identification performance depending on the property that was modified. This further supports the idea that event representation in audition involves abstraction since not all of the detail is preserved. Conversely, since many of these variations can be discriminated, indicating that the detail remains present at least to a level of processing that allows comparison (and, for example, judgement of quality of playing style, Grey and Moorer 1977), it may be that abstraction takes place at the moment of retrieval. For complex sound events, such as those produced by bouncing and breaking objects, there seem to be prototypic spectral and temporal properties (or transformational invariants) that characterize a class of sound events, such as the unitary accelerating rhythm and invariant spectral structure that specifies bouncing and the multiple accelerating rhythms, each with different spectral structures, that specify breaking. These spectro-temporal forms would need to be generalized over a large variety of bouncing rates, degrees of irregularity in the rhythm, and spectral characteristics of the fallen object or the newly created pieces if the object breaks. More systematic research is needed to clarify the nature of the transformational invariants that listeners are able to detect and the form in which they are represented in long-term memory.

The matching process

None of the experiments reported above give a clear indication of the nature of the matching process itself. Work on speech and visual form recognition has approached this problem and will be discussed briefly in Section 6.4. A couple of issues specific to non-verbal audition are raised, however.

Experiments on the multidimensional scaling of judged similarities among instrument tones have usually settled on two- or three-dimensional solutions as adequately describing the data. Intuitively this number of dimensions would seem to be quite small compared with the number of ways the tones can differ. One possible reason for this observed limit is that in making comparisons between successively presented tones, listeners are capable of using (or paying attention to) only a limited number of dimensions at a time. In experimental situations where a large number of stimuli are to be compared, certain dimensions or features may acquire a higher psychological weight that subsequently influences the pattern of similarity judgments. This conclusion is supported by the work of Braida (1988) on identification of multidimensional vibrotactile stimuli in which it was shown that when two dimensions must both be identified, sensitivity for a given stimulus dimension is generally reduced. This reduction in sensitivity is partially responsible for the fact that multidimensional stimuli do not tend to transmit as much information as performance for the individual dimensions would suggest. Another possibility is that the limitation in the number of dimensions that can be taken into account by a listener is not in the degree of detail with which the stimuli are encoded but in the extraction of a longer-term representation of distinguishing properties of the entire set of stimuli. One wonders to what extent this latter limitation may also reflect constraints on the process by which auditory representations are matched with memory representations. If the limitation is due to the matching process itself rather than to the auditory analysis process, it would follow that under some conditions not all the features or dimensions encoded in a category's long-term representation can be activated at any given time, i.e. that there are limits in the parallel matching of the multiple auditory cues that specify the sound source or event. However, the fact that listeners' similarity judgements can be better explained by positing specific dimensions or features attached to individual stimulus items in addition to the common dimensions (Krumhansl 1989; Winsberg and Carroll, 1989*a,b*), suggests that these limitations are perhaps at least partially methodological and that the encoding and matching processes are perhaps not the bottleneck. Unfortunately, no recognition or identification data have been collected on stimuli for which the specificities have been analysed in order to determine the role they might play in such tasks.

With this summary of experimental evidence pertaining to auditory recognition in mind, let us now examine a few models of the recognition process.

6.4 SURVEY OF RELEVANT RECOGNITION MODELS

A number of models have been developed by researchers in cognitive psychology which simulate the processes of acoustic dimension or feature analysis, similarity evaluation, categorization, recognition, and identification. They derive from work in intensity perception, speech, visual form perception, and vibrotactile perception. These models embody to varying degrees the stages of processing outlined in Section 6.1 and implement them in different ways. I will briefly describe the main characteristics of these models in terms of the main issues in non-verbal auditory recognition and will attempt to evaluate their relevance for the experimental data presented above.

6.4.1 Perceptual representations

The primary factors that distinguish the way different models simulate the internal perceptual representations prior to matching with memory are

(1) the degree to which the peripheral sensory representation is further analysed into elementary dimensions, properties, or features, and

(2) the degree to which the internal representation of the stimulus is continuous or discrete.

The design issues underlying these factors involve the search for an economical encoding of stimulus information in a form that captures the appropriate invariants.

In Klatt's (1979, 1989) speech recognition model, the only level of auditory representation is the neural spectrogram encoded in the auditory nerve (see Section 6.1.1). This model postulates a representation based on short-term spectra that describe the relative energy present at a given moment distributed across different frequency bands (see Fig. 6.8). Sequences of these templates are considered to adequately represent important phonetic transitions in the fine structure of the stimulus without

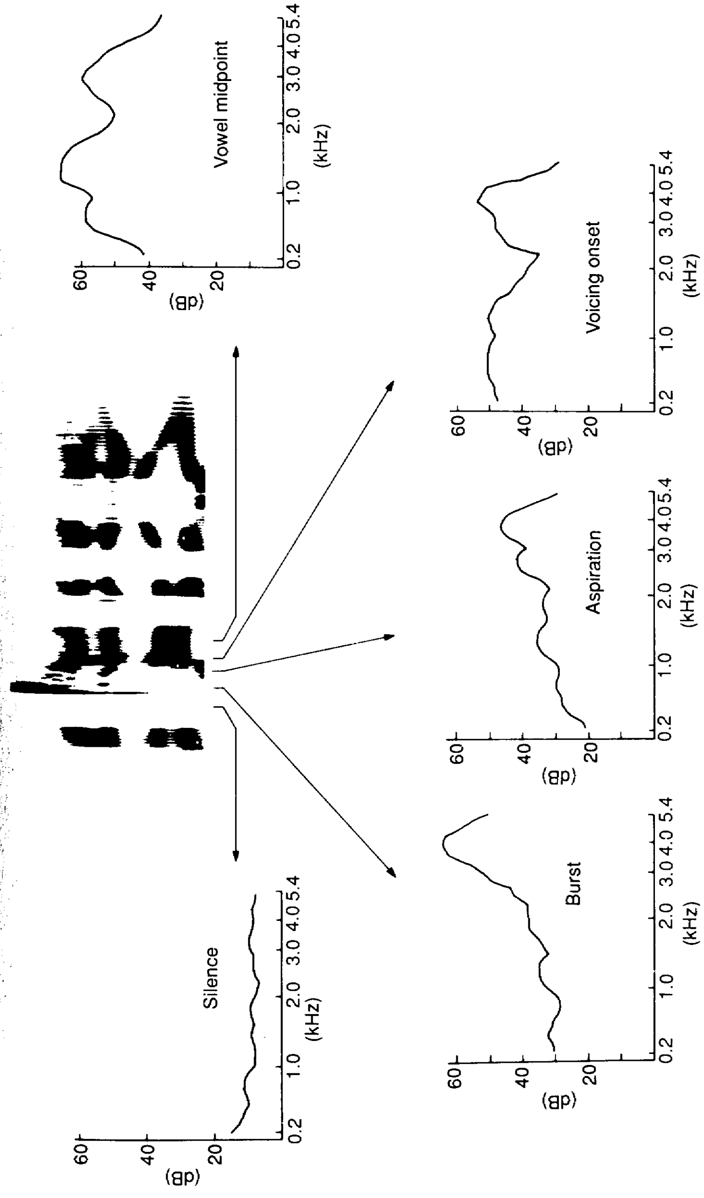

Fig. 6.8 A spectrogram of a speech sample is shown in the upper centre of the figure. A sequence of five static spectra, as represented in the auditory system, approximate the phonetic transition from the middle of the consonant/t/ to the middle of the vowel /a/. At a higher level of the auditory system these spectra would be matched to a series of spectral templates that would recognize the word containing /ta/. (From Fig. 7, Klatt (1989) p. 193. © MIT Press, 1989. Reprinted with permission.)

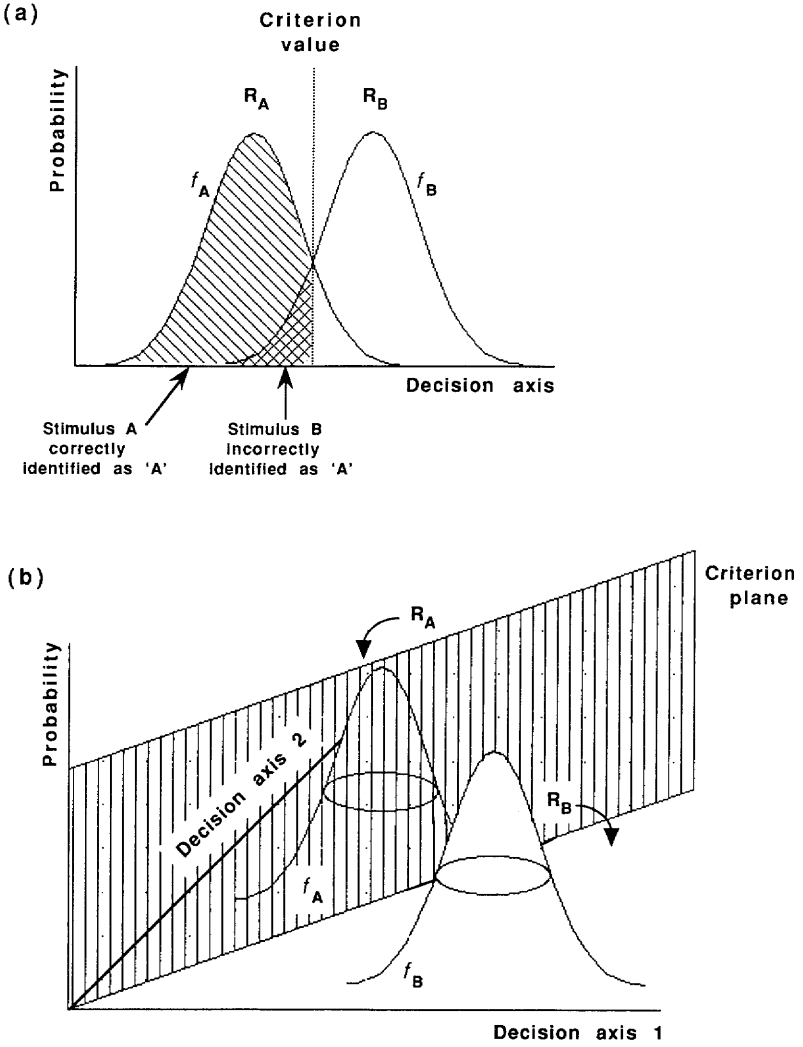

Fig. 6.9 Schematic illustrations of the representations of two categories in a statistical decision model. According to this class of models, stimuli A and B each give rise to a perceptual representation which is a random decision variable. The probability that a stimulus be represented by a given internal value on the decision axis (or axes) is represented by a probability density function, f_A or f_B. The form of this distribution depends on a number of sensory and cognitive factors, depending on the model. The single peak in the function indicates that the perceptual effects of a given stimulus tend to group around a single value. In an identification task, an optimal criterion point is chosen (by the subject) to maximize the likelihood of correctly choosing the appropriate category. This criterion point divides the decision axis into response regions, R_A and R_B. The degree of perceived similarity and confusability is conditioned by the degree of overlap of the distributions. (a) Example of two categories for stimuli that vary along a single dimension. The area under the curve f_A to the left of the dotted criterion line represents the probability that stimulus A will result in the response 'A'. The area under curve f_B to the left of the criterion line represents the probability that stimulus B will result in the incorrect response 'A'. (b) Example of two categories for stimuli that vary along two dimensions. The optimal criterion in this case is defined by a planar surface passing between the two peaks. (From Figs 1, 5, Ashby and Perrin (1988) pp. 128, 134. © American Psychological Association, Inc., 1988. Adapted with permission.)

imposing a segmentation into discrete features at any stage prior to accessing the lexicon of available words in the listener's memory. The advantage claimed for this representation is that information contained in the fine structure of the stimulus known to be useful in word recognition is not discarded too early in the processing chain. It is unclear, however, what auditory process would have to be postulated to adequately 'sample' the incoming sensory information, since more templates would need to be extracted during consonants (where the spectrum changes rapidly) than during sustained vowels (where the spectrum changes more slowly). If such a model were to be applied to non-verbal stimuli, similar questions would arise. How is the auditory system to be provoked into extracting more templates during attack, transition, and dynamically varying sustain periods than during relatively stable periods?

Some models postulate a continuously valued representation of the stimulus that, if multidimensional, is analysed into separable cues corresponding to the relevant dimensions of stimulus variation (such as angle and size of geometric forms or the frequencies of the first two formants for vowels, etc.). This representation along continuous dimensions can either be deterministic with respect to the stimulus (Nosofsky 1986; Nosofsky *et al.* 1989) or probabilistic (Durlach and Braida 1969; Ashby and Perrin 1988; Braida and Durlach 1988). The probabilistic representation is due in part to the introduction of noise in the sensory transduction process. Instead of a given physical stimulus value being encoded as a single value in the sensory system on all occurrences, it is encoded with a certain degree of variability that is modelled as a random process, e.g. as a Gaussian (bell-shaped) probability density function which describes the probability that a given physical value gets encoded as a given sensory value (see, for example, the function f_A in Fig. 6.9; the representation is unidimensional in (a) and two dimensional in (b)). All perceptual decisions are then performed with respect to this noisy representation. The original model of Durlach and Braida was unidimensional, modelling identification of arbitrarily defined intensity categories that are learned by listeners through feedback concerning correct responses. Subsequent versions have been extended to model the identification of more naturally categorized stimuli with multiple auditory cues, such as vowels and consonants (Macmillan *et al.* 1988), as well as the identification of consonants from multimodal (auditory, visual lip-reading, vibrotactile) stimulation (Braida 1991). The visual form recognition models of Nosofsky *et al.* (1989) and Ashby and Perrin (1988) are also multidimensional. While the models treat the dimensions as essentially independent, i.e. they are represented as being orthogonal in a multidimensional space, it is not clear to what extent each dimension is initially analysed separately and then integrated, or whether the stimulus is simply represented as a multidimensional unit from the start. This question has been explicitly addressed by Braida (1988) in work on the identification of three-dimensional vibrotactile stimuli, which has shown that multidimensional identification performance can be less than the sum of identification performance as measured for each dimension individually. This implies limitations of the perceptual representation, memory representation, or the matching process. His work on multimodal consonant identification demonstrated that a model that integrates sensory information prior to assigning a category label better predicts identification performance than a model that derives a label for each sensory modality and then tries to integrate the decisions (Braida 1991). It remains to be seen whether such a model can be successfully applied to the recognition of acoustic sources and events, though its apparent generality is promising. For example, the multidimensional version could easily be applied to similarity and recognition data for musical instrument tones.

Another class of models proposes that the sensory input is analysed into a number of discrete features that are then integrated prior to matching with memorized categories. In the spoken-word recognition model of McClelland and Elman (1986), sub-phonetic features, phonemes, and words are represented as nodes in a pseudo-neural (connectionist) network. Auditory preprocessing results in appropriate nodes being activated according to the presence of given features. The outputs of these feature nodes converge on phoneme nodes at a higher level of the network, and the phoneme nodes subsequently converge on word nodes. The activation spreads from lower to higher levels according to the degree of converging activity at each level. So if all the features that signal the phoneme /b/ are present, they would all be active and this activation would converge on the node for /b/. An additional feature of this network provides for mutual inhibition of nodes at the same level of representation, e.g. if a /b/ is stimulated, it inhibits the /d/ and /g/ nodes, phonemes that share similar features. A different kind of feature-processing architecture is proposed in the (multimodal) speech recognition model of Massaro (1987). In this model, features are assigned a value according to their relative strength of presence (or according to the degree to which the sensory information specifies their presence). The resulting 'fuzzy' truth values vary between 0 and 1 in the model rather than simply being labelled as present (1) or absent (0) (hence the label 'fuzzy logical model of perception'). These values are used to calculate the probability that a configuration of features specifies a given syllable in the lexicon (the unit of analysis being the syllable rather than a whole word in this model, for reasons that are supported by experimental research, cf. Massaro 1975; Segui 1989; Segui *et al.* 1990). So for both of these models, primitive features are extracted in some way at relatively early stages and are the unit of auditory representation (either as nodes in a neural network or as strength values in a logical propositional system). In both cases the strength of each feature is also represented (as degree of activation for McClelland and Elman or as a fuzzy truth value for Massaro).

Given the current state of knowledge of non-verbal auditory recognition, the models that have the greatest intuitive appeal are those that maintain some kind of multidimensional continuous representation of the stimulus information until relatively late stages of processing. This is suggested, for example, by the continuous nature of the perceptual dimensions revealed for musical instrument tones as well as the reasonable correspondence between this continuous representation of interstimulus relations and identification errors. However, a mixed model combining continuous common dimensions and discrete unique features for certain stimulus items may turn out to be the most valid simulation of a psychological representation of musical instrument tones, for example (Winsberg and Carroll 1989*a*,*b*). More systematic research is needed to critically analyse the relative contributions of different characteristics of each of these models to the recognition process as a whole.

6.4.2 Memory representations

The models examined above tend to fall into three basic classes with respect to the way stimulus categories are represented in long-term memory. For the sake of simplicity, these will be referred to as categorized continua, connectionist networks, and propositional descriptions, though any given model may have features that correspond to aspects of more than one of these classes.

Categorized continuum models posit the existence of a continuous auditory representation of the stimulus along one or more dimensions (Nosofsky 1986; Ashby and Perrin 1988; Braida and Durlach 1988). Category boundaries are in some way established along these continuous dimensions dividing the representational space into regions. A stimulus is categorized according to the region within which its auditory representation falls. Categories are defined by the subject according to how the

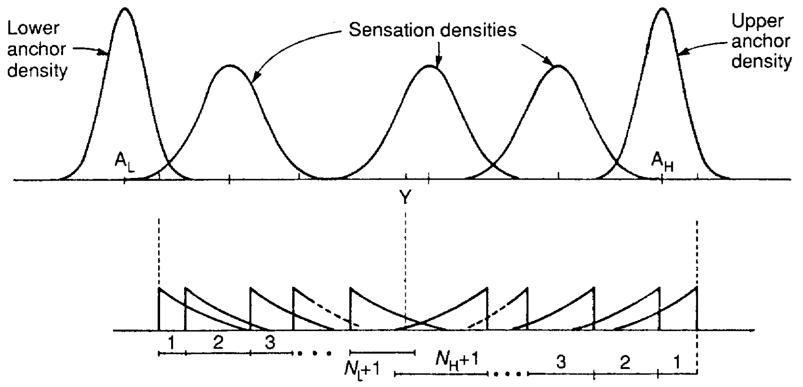

Fig. 6.10 Schematic diagram of the context-coding mechanism of Braida and Durlach (1988). In the upper part of the diagram, the noisy memory representation for the sensations produced by stimuli as well as the perceptual anchors used to categorize them are shown. The sensation Y is encoded by measuring its position relative to the noisy anchors A_L and A_H using a 'ruler' whose step sizes are also noisy. The ruler is illustrated in the lower part of the diagram. The steps are represented as exponentially distributed random variables. The distances of Y from the anchors, A_L and A_H, are measured, and the likely position of Y is estimated by the number of (noisy) steps from each anchor (N_L and N_H). (From Fig. 4, Braida and Durlach (1988) p. 565. © Neurosciences Research Foundation, Inc., 1988. Reprinted with permission.)

stimuli cluster in the space, i.e. subjects try to optimize the placement of category boundaries such that clusters of stimulus representations fall inside the boundaries. Nosofsky (1986) hypothesizes that episodic memory traces of individual experiences of a category can be represented as points in multidimensional space. Multiple exemplars of the same category tend to cluster together in the representational space. The relative weights of the individual dimensions can be varied by selective attending to one or more dimensions in order to maximize correct categorization of the stimuli. In the case of the 'optimal processing' models (Ashby and Perrin 1988; Braida and Durlach 1988), the stimulus representation is probabilistic and the subject must therefore try to optimize category boundaries so as to decrease the likelihood of identification errors given that the representations of stimuli with neighbouring values along one or more dimensions may overlap (see Fig. 6.9). The Braida and Durlach model proposes a context-coding memory mode in which listeners establish perceptual anchors along a given dimension and use these to help define and remember the category boundaries (see Fig. 6.10). In the case of identification of arbitrary categories along the intensity dimension, two anchors are placed just outside the range of intensities presented. For natural categories such as phonemes, different kinds of memory strategies

seem to be used, i.e. the anchors are placed at category *boundaries* for vowels and at the *centre* (or prototypical value) of the category for consonants (Macmillan *et al.* 1988). This latter result indicates that different strategies for encoding category boundaries may be used for different classes of stimuli. In the model, a number of factors have been shown to influence the representation of the category boundaries and, consequently, identification performance. Increases in stimulus range result in decreases in identification performance due to increased noise in the encoding of category boundaries that are further away from the perceptual anchors. Variations in the frequency of occurrence of the stimulus categories during an experiment result in shifts in category boundaries to take advantage of higher probabilities of occurrence of members of some categories. The episodic clustering and probability density function representations may not be that different, functionally, since in a formal sense both define bounded regions in a perceptual space.

The connectionist model of McClelland and Elman (1986) represents categories in the network as nodes upon which inputs from the configurations of features that compose them converge. These nodes correspond to phonemes and words. In the model of Klatt (1989), nodes consist of spectral templates organized into sequential networks. A given node is connected backward to nodes for spectra that are expected to precede it and forward to nodes for spectra that are expected to follow it. Specific sequences of such nodes represent words in all their possible acoustic–phonetic variations. As a listener encounters new pronunciations of a word, new elements can be added locally in the sequence decoding network to account for the new variations. As such the network explicitly enumerates all the possible spectral sequences for all possible word combinations of a given language.

Propositional description models represent stimuli as logical propositions concerning the conjunction of features that compose any given category. In the 'fuzzy logical model of perception' (Massaro 1987), the logical terms are the probability of presence (or strength) of a given subphonemic feature (such as voicing—which distinguishes /d/ from /t/—or tongue placement—which distinguishes among /b/, /d/, and /g/, etc.). These are then integrated and the configuration is compared with the proposition for a given syllable, e.g. syllable /ba/ is voiced, has labial rather than dental or velar placement at the beginning of the event, and so on. So syllables are represented as configurations of fuzzy truth values for the individual features that compose them.

Compared with the experimental evidence for auditory recognition of non-verbal sources and events, the models that seem most directly applicable are those proposing categorized continua, primarily because no experimental work has addressed the existence of discrete perceptual features in non-verbal sound events. Even the claim of purely discrete features as the basis for categorical perception of speech phonemes is the object of serious criticism (Macmillan 1987; Massaro 1987). However, the paucity of experimental data from appropriately conceived experiments does not allow us to rule out an eventual role of other types of model at this point.

6.4.3 The matching process

The kinds of matching processes specified or implied by the models discussed above depend strongly on the sensory and memory representations postulated. They may be classed according to whether matching takes place by activation of memory traces or nodes, by the evaluation of the 'goodness-of-fit' between sensory and memory representations, or by a statistical decision process that evaluates the probability of a given sensory representation falling within appropriate category boundaries.

McClelland and Elman's (1986) model is a trace activation model. Klatt's (1989) model can also be formalized in these terms since spectral template nodes activated in an appropriate sequence represent particular words in the lexicon. In a trace activation model, the sensory representation nodes send a wave of activation to configurations of nodes that represent categories in memory. The most highly activated node or sequence of nodes results in recognition of that category (a word in both models). As with lower levels in McClelland and Elman's model, an activated category inhibits other categories according to its degree of activation, i.e. an activated word inhibits its potential competitors for recognition. As such, if sensory information is degraded and several candidates in the lexicon are moderately or weakly activated, ambiguity of recognition results and errors may occur. In the same manner, if several category items are similar in structure, sharing many lower level features, their levels of activation will be similar and errors may also occur.

Massaro (1987) uses a goodness-of-fit measure to compare the auditory representation with a stored prototype description. The fuzzy truth values of all the features composing a syllable are multiplied and normalized to give its goodness-of-fit score. The syllable with the highest score is the one recognized. As in the trace activation models, syllables that share similar feature configurations are more likely to be confused since their goodness-of-fit scores would be comparable. The Klatt (1989) model could also be formalized in terms of goodness-of-fit. According to his conception of the matching process, the input auditory spectra are compared with all available spectral templates using a spectral distance metric. This metric measures how similar the input is to a given template. The spectral template that has the smallest distance score best matches the input spectrum. High scoring templates can subsequently be reduced in number according to whether they satisfy the sequencing constraints that represent allowable phonetic transitions and, ultimately, the best-fitting sequence in the lexicon will give rise to the recognized word.

The 'optimal processing' models (Ashby and Perrin 1988; Braida and Durlach 1988) hypothesize a statistical decision making process that estimates the category within which the sensory information falls. As mentioned in the previous section, an optimal placement of the boundaries would minimize the identification-error rate. For Ashby and Perrin, since the stimulus representation is a multidimensional probability density function, the degree of overlap of the distributions, for stimuli that are adjacent in the space, gives a measure of how similar they are and how probable confusion errors between the two are. So a given analysed sensory representation is compared with the available category boundaries and categorized accordingly. Identification errors result from imprecision in the sensory representation or from biases in the estimation of optimal boundary placement. For Braida and Durlach, the stimulus representation is similar, but the matching process is a bit different. For unidimensional stimuli, category boundaries are remembered with respect to perceptual anchors that are just beyond the end-points of the stimulus continuum. Presuming a stimulus continuum comprised of equally spaced values (such as intensity values every 5 dB across the range from 40 to 90 dB, for example), matching consists of 'measuring' the distance of the sensory representation from the nearest anchor point with a somewhat noisy 'ruler', the units of which are the category boundaries. The further the to-be-identified value is from an anchor point, the noisier its distance measurement is, and thus the greater the probability that it will be incorrectly identified as some neighbouring category (see Fig. 6.10). What this aspect of the model captures is the part of the identification error that is due to the matching process, which is combined in a complete analysis with the error due to stimulus encoding as well as with the error that accumulates in the auditory representation in working memory as the memory trace of the stimulus decays.

6.5 CONCLUSIONS

In reviewing the few studies that have been conducted on the processes of non-verbal auditory recognition and identification, it becomes clear that much remains to be done in the realms of both experimentation and modelling, even just to bring auditory research to the same level of development as research on visual shape recognition. This is particularly true for work concerning memory representations and the matching process involved in non-verbal auditory recognition. The field is certainly wide open and, given the new possibilities of digital signal analysis and synthesis, should see fruitful growth in the years to come.

We have seen that the acoustic micro- and macroproperties to which listeners are sensitive seem to be primarily continuous in nature with the possible exception of a number of unique features that may be represented in discrete fashion. More research on the exact nature of the acoustic cues that are necessary and sufficient for auditory recognition is needed. This work must address the fact that the cues uniquely specifying a given sound source may be quite different from one source to another, as was demonstrated in the work on musical instruments. Further, the utility of these cues can vary a great deal depending on the stimulus context within which recognition takes place. An important experimental question concerns determining how the context modulates the auditory representation that serves as input to the recognition process.

The prominence of continuous auditory representations and the close relation of similarity and identification studies suggest at least some form of fine-grained information encoded in the long-term representation of auditory categories. Furthermore, results that indicate the accumulation of information used for recognition over time and over stimulus variation (dynamic sustained portions of sound events that help define the resonance structure of sound sources, multi-note musical contexts, etc.) suggest the importance of determining the nature of the physical invariants that may also be accumulated in long-term memory over repeated exposure to such sound sources.

Conclusions concerning the matching process can only be speculative at this point given the lack of experimentation directed at this question. Among the more pressing issues might be cited the problem of determining how the matching process constrains the types of recognition error that are made and which cannot be completely accounted for by imprecision in the auditory representation of stimuli or in the long-term memory representation. The analytic approach of Braida and colleagues, who succeed in decomposing the various contributions to identification error (cf. Braida and Durlach 1988), seems the most fully developed at present and presents an interesting framework within which to begin testing hypotheses that are addressed more specifically to the process of auditory recognition of sound sources and events.

ACKNOWLEDGEMENTS

This chapter benefited from the insightful remarks of four *ad hoc* referees, including S. Handel, B. H. Repp, and E. D. Schubert, as well as the students of the fourth (anonymous) reviewer. Comments by E. Bigand, M.-C. Botte, L. D. Braida, A. de Cheveigné, E. F. Clarke, W. M. Hartmann, A. Gorea, J. Grose, J. Segui, and M. Stroppa are also greatly appreciated.

REFERENCES

Adrien, J.-M. (1991). The missing link: modal synthesis. In *Representations of musical signals* (ed. G. De Poli, A. Piccialli, and C. Roads), pp. 269–98. MIT Press, Cambridge, MA.

Anderson, J. R. (1985). *Cognitive psychology and its implications*. Freeman, New York.

Aran, J.-M., Dancer, A., Dolmazon, J.-M., Pujol, R., and Tran Ba Huy, P. (1988). *Physiologie de la cochlée*. INSERM / EMI, Paris.

Ashby, F. G. and Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, **95**, 124–50.

Barrière, J.-B. (ed.) (1991). *Le timbre : métaphores pour la composition*. Christian Bourgois, Paris.

Benade, A. H. (1976). *Fundamentals of musical acoustics*. Oxford University Press.

Braida, L. D. (1988). Development of a model for multidimensional identification experiments. *Journal of the Acoustical Society of America*, **84**, S142(A).

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*, **43A**, 647–77.

Braida, L. D. and Durlach, N. I. (1988). Peripheral and central factors in intensity perception. In *Auditory function: neurobiological bases of hearing* (ed. G. M. Edelman, W. E. Gall, and W. M. Cohen), pp. 559–83. Wiley, New York.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. MIT, Cambridge, MA.

Cadoz, C. (ed.) (1992). *Modèles physiques : création musicale et ordinateur*. Collection Recherche, Musique et Danse, Vol. 7, 8, 9. Maison des Sciences de l'Homme, Paris.

Charbonneau, G. R. (1981). Timbre and the perceptual effects of three types of data reduction. *Computer Music Journal*, **5**, 10–19.

Dowling, W. J. and Harwood, D. L. (1986). *Music cognition*. Academic, Orlando, FL.

Durlach, N. I. and Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, **46**, 372–83.

Fechner, G. T. (1966). *Elements of psychophysics* (transl. from German, *Elemente der Psychophysik*, Vol. 1, Breitkopf und Härtel, Leipzig, 1860). Holt, Rinehart and Winston, New York.

Florens, J.-L. and Cadoz, C. (1991). The physical model: modeling and simulating the instrumental universe. In *Representations of musical signals* (ed. G. De Poli, A. Piccialli, and C. Roads), pp. 227–68. MIT Press, Cambridge, MA.

Frauenfelder, U. H. (1991). Une introduction aux modèles de reconnaissance des mots parlés. In *La reconnaissance des mots dans les différentes modalités sensorielles: Etudes de psycholinguistique cognitive* (ed. R. Kolinsky, J. Morais, and J. Segui), pp. 7–36. Presses Universitaires de France, Paris.

Frauenfelder, U. H. and Tyler, L. K. (1987). The process of spoken word recognition: an introduction. *Cognition*, **25**, 1–20.

Freed, D. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *Journal of the Acoustical Society of America*, **87**, 311–22.

Garner, W. R. (1978). Aspects of a stimulus: features, dimensions, and configurations. In *Cognition and categorization* (ed. E. Rosch and B. B. Lloyd), pp. 99–133. Erlbaum, Hillsdale, NJ.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton-Mifflin, Boston.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton-Mifflin, Boston. (Republished by Erlbaum, Hillsdale, NJ, 1986).

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

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

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

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

Handel, S. (1989). *Listening: an introduction to the perception of auditory events*. MIT Press, Cambridge, MA.

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, **93**, 411–28.

Huggins, W. H. (1952). A phase principal for complex-frequency analysis and its implications in auditory theory. *Journal of the Acoustical Society of America*, **24**, 582–9.

Kendall, R. A. (1986). The role of acoustic signal partitions in listener categorization of musical phrases. *Music Perception*, **4**, 185–214.

Klatt, D. H. (1979). Speech perception: a model of acoustic–phonetic analysis and lexical access. *Journal of Phonetics*, **7**, 279–312.

Klatt, D. H. (1989). Review of selected models of speech perception. In *Lexical representation and process* (ed. W. D. Marslen-Wilson), pp. 169–226. MIT Press, Cambridge, MA.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In *Structure and perception of electroacoustic sound and music* (ed. S. Nielzen and O. Olsson), (Excerpta Medica 846), pp. 43–53. Elsevier, Amsterdam.

Krumhansl, C. L., Wessel, D. L., and Winsberg, S. (1988). Multidimensional scaling with specificity analysis for 21 synthesized tones from a Yamaha TX802 FM Tone Generator. Unpublished data. IRCAM, Paris [reported in Krumhansl (1989)].

Kruskal, J. B. and Wish, M. (1978). *Multidimensional scaling*. Sage, Beverly Hills, CA.

Kuwabara, H. and Ohgushi, K. (1987). Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech. *Acustica*, **63**, 120–8.

Leipp, E. (1980). *Acoustique et musique*, 3rd edn. Masson, Paris.

Lindsay, P. H. and Norman, D. A. (1977). *Human information processing: an introduction to psychology*, 2nd edn. Academic, New York.

Luce, D. (1963). *Physical correlates of nonpercussive musical instrument tones*. Ph.D. thesis, Massachussetts Institute of Technology. Cambridge, MA.

McAdams, S. and Rodet, X. (1988). The role of FM-induced AM in dynamic spectral profle analysis. In *Basic issues in hearing* (ed. H. Duifhuis, J. W. Horst, and H. P. Wit), pp. 359–69. Academic, London.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1–86.

Macmillan, N. A. (1987). Beyond the categorical/continuous distinction: a psychophysical approach to processing modes. In *Categorical perception* (ed. S. Harnad), pp. 53–87. Cambridge University Press, New York.

Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (1988). Resolution for speech sounds: basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, **84**, 1262–80.

Marr, D. (1982). *Vision*. Freeman, San Francisco.

Massaro, D. W. (1975). Language and information processing. In *Understanding language* (ed. D. W. Massaro), pp. 3–28. Academic, New York.

Massaro, D. W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry*. Erlbaum, Hillsdale, NJ.

Michaels, C. F. and Carello, C. (1981). *Direct perception*. Prentice-Hall, Englewood Cliffs, NJ.

Miller, J. D. (1982). Auditory-perceptual approaches to phonetic perception. *Journal of the Acoustical Society of America*, **71**, S112(A).

Miller, J. R. and Carterette, E. C. (1975). Perceptual space for musical structures. *Journal of the Acoustical Society of America*, **58**, 711–20.

Nosofsky, R. M. (1986). Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39–57.

Nosofsky, R. M., Clark, S. E., and Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **15**, 282–304.

Pickles, J. O. (1982). *Introduction to the physiology of hearing*. Academic, London.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In *Frequency analysis and periodicity detection in hearing* (ed. R. Plomp and G. F. Smoorenburg), pp. 397–414. Sijthoff, Leiden.

Plomp, R. (1976). *Aspects of tone sensation*. Academic, London.

Pollack, I., Pickett, J. M., and Sumby, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, **26**, 403–6.

Repp, B. H. (1987). The sound of two hands clapping. *Journal of the Acoustical Society of America*, **81**, 1100–9.

Roman, R. (ed.) (1992). *Le système auditif central: anatomie et physiologie*. INSERM/EMI, Paris.

Rosen, S. and Howell, P. (1987). Auditory, articulatory, and learning explanations of categorical perception in speech. In *The psychophysics of speech perception* (ed. M. E. H. Schouten), pp. 113–60. Nijhoff, The Hague.

Saldanha, E. L. and Corso, J. F. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, **36**, 2021–6.

Schiffman, S. S., Reynolds, M. L., and Young, F. W. (1981). *Introduction to multidimensional scaling: theory, methods, and applications*. Academic, Orlando, FL.

Schubert, E. D. (1975). The role of auditory perception in language processing. In *Reading, perception, and language*, pp. 97–130. York, Baltimore.

Segui, J. (1989). La perception du langage parlé. In *Traité de psychologie cognitive*, Vol. 1 (ed. C. Bonnet, R. Ghiglione, and J.-F. Richard), pp. 199–234. Dunod, Paris.

Segui, J., Dupoux, E., and Mehler, J. (1990). The role of the syllable in speech segmentation, phoneme identification, and lexical access. In *Cognitive models of speech processing: psycholinguistic and computational perspectives* (ed. G. T. M. Altmann), pp. 263–80. MIT Press, Cambridge, MA.

Shepard, R. N. (1972). Psychological representation of speech sounds. In *Human communication* (ed. E. E. David and T. B. Denes). McGraw-Hill, New York.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: on the status of 'direct' psychophysical measurement. *Journal of Mathematical Psychology*, **24**, 21–57.

Sloboda, J. A. (1985). *The musical mind: the cognitive psychology of music*. Oxford University Press.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes—loudness. *American Journal of Psychology*, **69**, 1–25.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, **64**, 153–81.

Strong, W. and Clark, M. (1967*a*). Synthesis of wind-instrument tones. *Journal of the Acoustical Society of America*, **41**, 39–52.

Strong, W. and Clark, M. (1967*b*). Perturbations of synthetic orchestral wind-instrument tones. *Journal of the Acoustical Society of America*, **41**, 277–85.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273–86.

Van Derveer, N. J. (1979). Acoustic information for event perception. Unpublished paper presented at the celebration in honour of Eleanor J. Gibson, Cornell University, Ithaca, New York [cited in Warren and Verbrugge (1984) and in Handel (1989)].

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, **167**, 392–3.

Warren, W. H. and Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 704–12.

Wedin, L. and Goude, G. (1972). Dimension analysis of the perception of instrument timbres. *Scandinavian Journal of Psychology*, **13**, 228–40.

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

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, **3**, 45–52.

Wessel, D. L., Bristow, D., and Settel, Z. (1987). Control of phrasing and articulation in synthesis. *Proceedings of the 1987 International Computer Music Conference*, pp. 108–16. Computer Music Association, San Francisco.

Winsberg, S. and Carroll, J. D. (1989*a*). A quasi-nonmetric method for multidimensional scaling of multiway data via a restricted case of an extended INDSCAL model. In *Multiway data analysis* (ed. R. Coppi and S. Belasco), pp. 405–14. North-Holland / Elsevier, Amsterdam.

Winsberg, S. and Carroll, J. D. (1989*b*). A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. *Psychometrika*, **54**, 217–29.