Parncutt, R. (1989). *Harmony: A psychoacoustical approach*. Berlin: Springer-Verlag.

# 2. Psychoacoustics

Psychoacoustics investigates relationships between the physical properties of sounds (waveform, spectrum, level, frequency, ...) and the way sounds are experienced (loudness, pitch, timbre, salience). The first stage of auditory perception involves spectral analysis in the cochlea, with specific time and frequency characteristics. Thereafter, analytical information is extracted by categorical perception, and holistic information (which can ambiguous, depending on context) is extracted by pattern recognition. In a psychoacoustical approach, the perception of complex tones (and hence of ordinary environmental sound sources) involves the spontaneous recognition of harmonic patterns among the pitches of audible pure tone components. Consequently, the pitch of complex tones (and even of pure tones) can be ambiguous. Pitch may be measured and perceived on continuous scales (in psychoacoustics) and categorical scales (in music); the latter case includes the recognition of both intervals (relative pitch) and notes (perfect pitch) by musicians.

## 2.1 Philosophy of Perception

### 2.1.1 Hardware and Software

Within limits, it is useful to draw an analogy between the brain and the hardware of a computer. The way we perceive, by this analogy, is like a computer program – a software package for the brain [cf. Lilly 1974].

There is no sharp boundary between hardware and software in computing. A lot of what is called hardware is in some sense programmed to perform specific transformations on input signals. The same may be said for perception and behaviour.

The software of perception develops quite differently from contemporary computer software. It is acquired ("learned") as the organism actively explores and interacts with its environment. In this respect, the brain may be said to be *self-programming*. The program by which it programs itself is "innate" or "instinctive". The self-programming process involves interaction of the whole organism with its various environments; it begins before birth (Sect. 3.1.2), and continues throughout life.



Hardware and software can be remarkably independent of one another; the same computer can run completely different kinds of program (i.e. perform completely different algorithms), and the same program can be performed on completely different kinds of computer (e.g. serial versus parallel processors). Similarly, the nature of perception may be largely independent of the particular ways in which the human brain stores and processes information.

In particular, music perception does not necessarily depend on brain physiology; Roederer's suspicion [1987, p. 82] that "... 'universal' characteristics of music are ... the result of built-in physiological or neuropsychological functions of the auditory system" probably applies only to the physiology of the ear (e.g., its frequency analyzing property). Instead, the nature of music would appear to depend primarily on the way the auditory system interacts with sound, considered as a part of the interaction of the organism with its environment [Gibson 1979]. Most aspects of the perception of music may be satisfactorily explained in terms of familiarity with environmental and musical sounds (Sect. 3.1).

### 2.1.2 Matter, Experience and Information

A useful philosophical basis for the study of music perception is the *three worlds* concept of Karl Popper [Popper and Eccles 1977; Terhardt, personal communication]. World 1 is the world of *matter* (and energy): it comprises physical objects, states and processes, and includes musical instruments, tones, the ear and the brain. World 2 is the world of *experience,* or states of consciousness. It includes all aspects of musical experience – sensations of tone, harmony, rhythm, consonance and tonality, as well as the emotions evoked by a piece of music. The contents of world 3 may be variously described as symbols, descriptions, language, "objective knowledge", or simply *information.* World 3 includes thoughts and ideas, literature, computer programs, musical scores, and music theory.

The degree to which correspondences exist between the three worlds is limited; each world is, to some extent, autonomous. The limited correspondence between worlds 1 and 3 (matter and information) is reflected by Heisenberg's uncertainty principle in quantum mechanics – a special case of the general rule that you can't measure something without in some way changing what you are measuring. The limited correspondence between worlds 2 and 3 (experience and information) is reflected by the existence of "feelings which cannot be put into words". In the case of worlds 1 and 2 (matter and experience), brain states and associated experiences are measured and expressed in fundamentally different ways, involving physical measurements (expressed in physical units) on the one hand and observers' introspective reports (expressed in natural language) on the other.

There is no clear a priori justification for the belief that all aspects of experience may someday be predictable on the basis of physiological measure-

ments, no matter how sophisticated such measurements might become in the future. In the words of Gibson [1979, p. 306],

"Perception cannot studied by the so-called psychophysical experiment if that refers to physical stimuli and corresponding mental sensations. The theory of psychophysical parallelism that assumes that the dimensions of consciousness are in correspondence with the dimensions of physics and that the equations of such correspondence can be established is an expression of Cartesian dualism. Perceivers are not aware of the dimensions of physics. They are aware of the dimensions of information in the flowing array of stimulation that are relevant to their lives."

Moore [1982] in his book aimed to specify the relationships between sounds and sensations "in terms of the underlying mechanisms", seeking to "understand how the auditory system works, at well as to look as what it does" (p.1). The phrase "underlying mechanism" betrays Moore's belief in concrete relationships between stimulus and sensation at the level of brain function. In the light of Gibson's comments (above), Moore may well be asking unanswerable questions.

It is widely believed that only the physical world *really* exists, and that physical states and processes underlie both experience and information. This raises some thorny questions. If experiences don't really exist, for example, what is the point of funding the arts? And if information does not really exist, exactly what *was* it that Mozart bequeathed to humanity? A contrasting (and equally valid) view is that experience is the foundation and final arbiter of knowledge [Clifton 1983]. According to this view, the existence of the physical world is just a hypothesis based on everyday experience of, and theories about, the environment. If this is the case, however, why is it that the physical world can be described and measured more precisely than the worlds of experience and information? In Popper's approach, philosophical problems such as these are avoided by regarding matter, experience and information as equally real.

Gödel's theory in mathematics may be interpreted to imply that no theory or philosophy can explain itself: all abstract systems incorporate inconsistencies [Hofstadter 1980]. Popper's three worlds concept is no exception. For example, a *thought* may be regarded as either a piece of information or an experience. On the other hand, all scientific research relies on some kind of paradigm [Kuhn 1962]. The three worlds concept is chosen as a paradigm on which to base a theory of music perception, not because it is perfect (it isn't), but because it clarifies the multidisciplinary mosaic of music perception research.

An example of the application of the concept in the case of music performance is Nakamura's [1987] study of the relationships between "the dynamics of a piece of music that professional performers intend to convey to listeners [world 3]; . . . the intensity of tones produced by the performers [1]; and . . . the listeners' perception of the dynamics of performances [2]" (abstract). Further examples are described below.



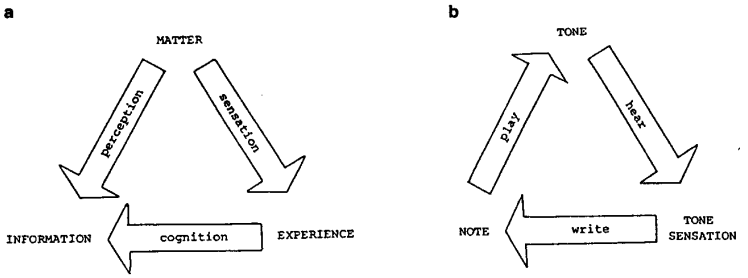


**Fig. 2.1 a, b.** Popper's three worlds and music perception. **a** Perception, sensation and cognition. **b** A cycle of musical creation

### 2.1.3 Perception, Sensation and Cognition

Perception is an active process by which organisms extract information from and interact with their environments [Gibson 1966]. Sensation, by contrast, is passive. It involves experiencing or being aware of sensory input, without necessarily focussing on environmental objects.

In traditional psychological paradigms, perception is regarded as a two-stage process involving the subprocesses sensation (as studied in psychophysics) and cognition (Fig. 2.1 a). In the first stage, physical stimuli are "converted" into sensations. In the second stage, hypotheses about the environment are made on the basis of available sensations; the results may be called *percepts* of environmental objects. In the traditional approach, then, sensation is regarded as an essential prerequisite for perception.

Gibson [1966] observed that most environmental interaction is almost entirely "automatic", occurring with little or no awareness of the analytic complexity of associated sensory patterns, cognitive processes and motor responses. This suggests that the perceptual extraction of information from the environment occurs much more directly than in the traditional two-stage model. Gibson consequently demoted sensations from their traditional status as prerequisites for perception to the more realistic status of mere *byproducts* of perception.

Perceptual theories may be divided into three kinds: those based on psychophysics (the interaction between worlds 1 and 2), cognition (2 and 3) and direct or ecological perception (1 and 3). In the present study, psychophysical and (to a lesser extent) direct or ecological explanations of music perception are generally preferred to cognitive ones. Psychophysical and direct perceptual explanations have the advantage that they involve the physical world directly, and the physical world is more experimentally measurable and precisely specifiable than the worlds of experience and knowledge. Because cognitive theories relate the "subjective" worlds of experience and knowledge

to each other, they lack the stability of being anchored to "objective" physical measurements.

The music listener experiences both sensations and percepts. Musical percepts correspond to physically real objects like singers, instruments and loudspeakers. Musical sensations include tone sensations, and their pitch, loudness, and timbre. Rhythm, melody and harmony also fall under the heading of sensations, as they have no specific correlates in the everyday, non-musical environment.

### 2.1.4 Tone, Tone Sensation and Note

The word "tone" in this study refers to a *physical* entity: a periodic acoustical disturbance which can evoke pitch. A "note" is an *instruction* to play a tone. In addition, the term "tone sensation" is used to refer to the *experience* accompanying the perception of a tone.

In experimental acoustics, the basic measurement of a tone is its *pressure waveform*: a function of oscillatory pressure against time, recorded at some point in space by means of a microphone. The amplitude and phase spectra of a tone, obtained by Fourier (spectral) analysis of its waveform, may be used to recreate the original waveform by adding component waveforms.

*Tone sensations* (or "sensory tones" [Terhardt 1979b]) are defined to be experiences associated with the perception of tone sources, such as people speaking, musical instruments being played, and so on. Tone sensations have the attributes salience (perceptual importance), pitch, timbre, (apparent) onset time and (apparent) duration.

*Notes* belong to the world of information. The attributes of a note correspond not to the physical attributes of the tone to be played but to its *perceptual* attributes, expressed by means of labelled *categories* (C#, quarter note, etc.). Pitch categories in Western music (Sects. 2.5.3–2.5.5) are normally specified relative to each other and to tonally important pitch categories by means of interval categories expressed in semitones. Time categories, used to express both the onset time and duration of a note, are specified as rational numbers (fractions of whole numbers) relative to a prevailing pulse or metre. In performance, the actual sizes of notated pitch and time intervals depend on performer and context [Sundberg 1982; Gabrielsson et al. 1983; Sundberg et al. 1983]. Different performances of the same music notation nevertheless remain within certain limits (category boundaries) implied by the notation.

In Western music, loudness and timbre are notated separately from pitch, onset time and duration. Loudness is indicated categorically by *dynamics* corresponding to ordinary words such as "loud", "very soft", etc. Timbre is indicated categorically in music notation by *orchestration*: the names of musical instruments, and instructions for the use of mutes, special techniques, and effects such as pizzicato and flutter tonguing. The actual (experienced) loudness and timbre and corresponding physical characteristics of a tone played



according to a particular musical score and on a particular instrument depend on pitch, context, player and so on.

The relationship between tones, tone sensations and notes may be regarded as a cycle of musical creation which links the performers, audiences and composers of Western music (Fig. 2.1 b). Composers write instructions to performers in the form of scores. In sight-reading, the performer sees a musical note, "thinks" it, and then plays it; the result is called a tone. The performance of musical tones is controlled by a kind of feedback mechanism by which the performer hears what has been played, checks whether its sensory attributes correspond to those required by the notation, the performer's concept of the music, and expectations of a real or imagined audience, and then makes appropriate adjustments to the performance.

In the case of improvisation (e.g. in the Baroque period, and in jazz), the word "write" in the figure may be replaced by "decide". Improvisers decide which notes to play on the basis of the kinds of sounds they have just created, and the direction they wish to take in the music. Similarly, the word "note" in the figure may be interpreted as a kind of self-instruction on the part of an improviser, referring (in a rather analytical way) to decisions made and executed during improvisation.

Decisions made during musical improvisation need not be conscious, and experienced note-readers are not necessarily conscious of the individual notes of a score as it is being sight-read. The idea of unconscious decisions, regulated partly by tone sensations experienced during a musical performance, may be used to explain the kind of feedback mechanism by which both music readers and improvisers control their performances. An account of psychological aspects of music reading and improvisation is given by Sloboda [1985].

From the physicist's point of view, tones are more important than the tone sensations which they evoke and the notes by which they are played. From the psychologist's point of view, tone sensations are basic, for without them we would never have developed the concepts of "tone" or "note". From the musician's point of view, notes are basic, because without notes no musical tones would be played and no musical tone sensations heard; even improvisers may be thought to imagine notes before playing them. These three views are internally coherent, but nonetheless limiting. It is preferable to assign tones, tone sensations and notes equal importance in an objective analysis of music perception.

## 2.2 Auditory Sensation

### 2.2.1 Loudness and Timbre

The sensory attributes of tonal sounds (i.e. simultaneities) most commonly investigated in psychoacoustics are loudness, pitch, and timbre. In the case of individual tones perceived within tonal simultaneities, one speaks in psycho-

acoustics of *salience* (sensory importance) rather than loudness. Like all psychoacoustical parameters, the sensory attributes of tonal sounds depend on listener and context.

The loudness, pitch and timbre of an isolated tone *all* depend on *all* the corresponding physical parameters: intensity, frequency and spectral composition (as a function of time) of the tone [Fletcher 1934]. So any physical change in a sound is likely to produce a change in all its sensory attributes. For example, changing frequencies of the pure tone components of a sound changes its loudness and its timbre.

Of all the sensory attributes of tonal sounds, pitch is the most important for harmony, and is dealt with in detail in later sections. Loudness and timbre are not so important for harmony, and are discussed briefly below.

The (subjective) loudness of a pure tone depends on its frequency as well as its sound pressure level [Fletcher and Munson 1933]. The loudness of a complex tone or sound also depends on its spectral distribution [Zwicker 1960] (Sect. 4.3.1). Loudness is measured in psychoacoustics by comparing the loudness of a test sound with that of a standard reference tone [American Standards Association 1960; Zwicker 1982]. The *loudness level* in *phon* of a sound is defined as the sound pressure level (SPL) in decibels (dB) of a (standard) 1 kHz pure tone when the sound and the standard tone are judged to be equally loud. For example, a sound which is just as loud as a pure tone of frequency 1 kHz and SPL 60 dB has a loudness level of 60 phon. Loudness level is an accurate, but not a proportional, measure of loudness. Doubling the (apparent) loudness of a sound doesn't double its loudness level, but increases it by about 10 phon. The corresponding proportional scale is called simply *loudness*, and is measured in *sone*, such that a (test) sound of loudness $n$ sone is judged to be $n$ times louder than a pure tone of frequency 1 kHz and SPL 40 dB (the standard). A loudness of 1 sone corresponds to a loudness level of 40 phon, 2 sone corresponds (approximately) to 50 phon, 4 sone to 60 phon, 8 sone to 70 phon, and so on.

Timbre (tone quality) is associated with the identification of environmental sound sources [Bregman and Pinker 1978], including musical instruments [Saldanha and Corso 1964]. Like vowel quality, timbre depends on the absolute frequencies and amplitudes of pure tone components. In addition, the physical characteristics of the onset of a musical tone are crucial for timbre and instrument identification [Berger 1964]. A powerful technique for the understanding of timbre is analysis-by-synthesis [Risset 1978]. Like pitch (tonality) and loudness (dynamics), timbre can be used to delineate musical forms in contemporary styles [McAdams and Saariaho 1985].

Timbre is multidimensional [Wedin and Goude 1972]. It may be quantified on various sensory scales such as "brightness" and "richness", and studied by multidimensional scaling of similarity ratings [Grey 1977]. Sensory dimensions of timbre which are important for the theory of harmony are *roughness*, associated with beating between pure tone components, and *tonalness*, the degree to which a sound has the sensory properties of a single complex tone such as



a speech vowel [Terhardt 1983] (Sect. 3.2.2). Bad musical intonation (tuning) causes roughness to increase and tonalness to decrease; this explains the finding of Madsen and Geringer [1981] that deliberate mistuning in flute/oboe duets is often misinterpreted by listeners as bad tone production on the part of the performers.

### 2.2.2 Spectral Analysis

According to Fourier's theorem in mathematics, any waveform of finite duration (not necessarily periodic) may be expressed as a sum of component waveforms which are sinusoidal over the same duration. In acoustical terms, this means that any sound may be expressed as a sum of pure tone components. Note that these components are not directly measurable, so – strictly – they do not exist as physical entities. Instead, they are found by subjecting the waveform of a sound to a *mathematical* procedure: spectral analysis.

The relationship between sound input to the ear and the information conveyed to the brain is essentially the same as the relationship between a sound and its pure tone components. In this sense, the ear subjects incoming sounds to spectral analysis [Ohm 1843; Helmholtz 1863; Terhardt 1972, 1985]. This may be regarded as an early stage in the extraction of information from sound, in order to enable and facilitate interaction with the environment.

The cochlea is a bony, snail-like hollow in the petrous bone. The basilar membrane, which it houses, may be regarded as the receptor surface of the peripheral auditory nervous system. The basilar membrane is tapered: broad at one end and narrow at the other. When a pure tone is detected, waves travel along the membrane, reaching maximum amplitude at a point depending on the frequency of the tone [Bekesy 1947]. This spectral information is maintained in the peripheral nervous system [Evans 1975].

The importance of place on the basilar membrane in determining the pitch of pure tone sensations is supported by work with the partially deaf. Damage to part of the basilar membrane can cause deafness in a corresponding frequency range [Crowe et al. 1934], and electrodes implanted at different places in the auditory nerve of a deaf person produce tone sensations of different pitch [Simmons et al. 1965]. However, some experimental pitch data cannot be accounted for by place alone: it appears that both place information (e.g. which parts of the basilar membrane experience maximum displacement) and temporal information (e.g. the rate at which a particular part of the membrane oscillates) contribute to the pitch of pure tone components [Moore, 1982]. For example, below about 50 Hz, the position of maximum amplitude is independent of frequency [Bekesy 1947]. In this region, the pitch of a pure tone may depend on the rate of neuron firing in the auditory nerve [cf. Wever and Bray 1937]. In any case, Ohm's acoustical law, as described above, holds regardless of how the complex motion of the basilar membrane is translated into the pitch of pure tone components.

Like any spectral analysis system, the ear has limited frequency resolution. Simultaneous pure tone components must differ in frequency by a certain minimum amount before they can be resolved (or discriminated). Such minimum frequency differences are determined by the effective time constants (i.e. effective durations of the analysis interval) of the ear, which vary as a function of frequency [Terhardt 1985]. Simultaneous pure tones must be at least 1.0–1.5 semitones apart (considerably more than this at low frequencies) to be resolved, i.e. to produce distinct tone sensations [Plomp 1964; Terhardt 1968a].

The ear is not a perfect spectrum analyzer. Under certain spectral conditions, single pure tones of high sound level can produce harmonic distortion [Egan and Klumpp 1951], and simultaneous pure tones can produce combination tones (Sect. 1.2.5).

The output of the ear's spectral analysis is influenced mainly by that part of the incoming waveform immediately preceding the time of observation; earlier and earlier parts of the waveform influence perception less and less. An appropriate mathematical procedure for modelling this kind of spectral analysis is the *Fourier-t-transform* (or FTT), in which sound waveforms are multiplied by an exponential decay function, and spectral analysis is subsequently performed on a window extending from negative infinity to the present. In psychoacoustical applications, the variable amplitude and frequency dependencies of the FTT may be adjusted to fit those of the auditory system [Terhardt 1985]. When this is done, only audible pure tone components are output by the procedure, i.e. masking is automatically accounted for.

The *masked threshold* (or audiogram) of a pure tone is a graph of the sound pressure level (SPL) of a second, simultaneous, barely audible pure tone, as a function of its frequency [Wegel and Lane 1924]. It is roughly triangular in shape, peaking at the frequency and amplitude of the first tone: the closer the second tone lies to the first in frequency, the more it is masked, so the higher its SPL needs to be before it can be heard. For pure tones above about 500 Hz ($C_5$), the gradient of the lower-frequency side of the masked threshold is constant at roughly 9 dB per semitone.

As a rule, a change can be heard in a sound if part of its masked threshold undergoes a vertical shift of 0.5–1.0 dB [Riesz 1928; Zwicker 1970]. This implies that a change can be heard in a pure tone if it is shifted in frequency by 0.06–0.12 semitone. Difference thresholds of frequency as low as 0.02 semitone for the best listeners under ideal conditions [König 1957; Fastl and Hesse 1984] may be due to the added role of temporal information in pitch perception. Alternatively, they may be explicable in terms of musical experience: small pitch changes are more important in music than small loudness changes, and discrimination improves with practice [E. J. Gibson 1953].



### 2.2.3 Sensory Memory

Sensory memory is spontaneous memory, i.e. memory in the absence of attention, noticing, categorization, abstraction, semantic processing, etc. In a sense, this is not memory at all – it is a kind of spontaneous decay characteristic of the sensory system for which "memory" is the conventional psychological metaphor. To measure the duration of sensory memory it is necessary to ensure that a stimulus remains unnoticed for a specified time after its real-time occurrence.

The duration of visual sensory memory is about 0.1–0.2 s [Averbach and Coriell 1961]. Decay times in this range are also characteristic of forward masking effects (masking between sequential sounds) in psychoacoustics [Moore 1982; Zwicker 1982]. Auditory sensory memory, otherwise known as echoic memory [Neisser 1967] or precategorical acoustic storage [Crowder and Morton 1969] lasts much longer than both visual sensory memory and acoustical masking effects. Eriksen and Johnson [1964] estimated its duration at 10 s. Later researchers reported lower values such as 5 s [Glucksberg and Cowen 1970] and 2 s [Crowder 1970], suggesting that Eriksen and Johnson's experiment was influenced by ordinary, non-sensory memory.

Sensory memory linkage may be regarded as an essential prerequisite for the spontaneous perception of pitch relationships between sequential sounds (pitch commonality and proximity, Sect. 3.2.3). This is no problem in music, as the chords that make up chord progressions are normally much less than 2 s apart. In experiments to investigate pitch relationships (Chap. 5), the pairs of sounds presented in each trial followed each other at time intervals much shorter than 2 s. On the other hand, the time intervals between different trials in the experiments generally exceeded 2 s, so that sensory interference between trials was unlikely to affect results.

The duration of auditory memory increases considerably if sounds are noticed as they occur in real time. Sensory material persists longer in memory the more it is "processed through semantic levels" [Craik and Lockhart 1972], i.e. the higher it is abstracted in a perceptual hierarchy. Wickelgren [1969] found evidence of both sensory or "short-term" and categorical or "intermediate-term" memory for pitch, with durations of about 2 and 20 s respectively.

Memory for a particular sound is disrupted by intervening sounds [Wickelgren 1966; Massaro 1970; Deutsch 1972a; Dewar et al. 1977; Olsen and Hanson 1977]. Duration of memory for tones in an unfamiliar musical context tends to fall as the apparent rate of sensory information in that context increases. These effects are neglected in the present study, which is mainly concerned with sensory auditory memory in the absence of interference.

## 2.3 Extraction of Information

### 2.3.1 Noticing and Salience

To notice something is to become aware or conscious of it. This often involves assigning a verbal label to it. There is a large grey area between "noticed" and "unnoticed", in which objects and stimuli influence experience and environmental interaction, but are not necessarily assigned verbal labels.

In this study, the salience of an environmental object or stimulus is defined quantitatively as the probability that it will be noticed. In other words, the salience of the corresponding percept or sensation is its probability of occurring. If a sensation or percept already exists, then its salience may be regarded as a measure of its apparent importance or strength. For example, a chord may evoke several tone sensations, but some may sound more important than others.

The pure tone components of a complex tone are seldom directly noticed, yet each contributes to the perception of the tone as a whole (Sect. 2.4.3). The degree to which each contributes depends on its salience (Sect. 4.4.2). Similarly, the degree to which (unnoticed) tone components contribute to the strength of sequential pitch relationships depends on their salience (Sects. 4.6.1, 4.6.2).

Relatively salient tone sensations in a musical chord normally correspond to actual tones, and are recognized as such by musicians. Tone sensations with low salience do not normally correspond to actual tones, but to implied or harmonically related pitches such as the root of a chord in inversion (Sect. 6.1.5).

### 2.3.2 Categorical Perception

Categorical perception refers to the division of a perceptual continuum into labelled categories, specified by their centres and widths, or by the positions of their boundaries. Categorical perception may be regarded as the most elementary or analytical way of extracting information from a perceptual continuum.

The concept of categorical perception was originally developed to explain phoneme boundaries in speech sounds [Liberman et al. 1957, 1961]. Perceptual discrimination is normally easier across category boundaries. In other words, stimuli are more likely to be judged as "different" if they fall into different perceptual categories.

A familiar example of categorical perception is the perception of colour. Electromagnetic radiation in particular frequency bands evokes particular colours. The band of frequencies corresponding to a particular colour (red, orange, yellow, etc.) corresponds to a perceptual category.

The position of the boundary between two neighbouring perceptual categories is always somewhat vague or flexible. In a rainbow, for example, one cannot see exactly where "red" stops and "orange" begins. The position of the



boundary between two categories also depends on the observer and on the context in which a stimulus is presented. For example, the colour aqua will sometimes be called blue, sometimes green, depending on observer and background colour.

The positions of category boundaries may be either innate or learned. Boundaries between colours appear to be primarily innate (due to the physiology of the eye). Boundaries between speech phonemes appear to be primarily learned by exposure to speech: adults' discrimination at phoneme boundaries is sharper than infants' [Eimas et al. 1971]. Similarly, the musical interval discrimination functions of musicians are sharper than those of untrained listeners, implying that boundaries between musical scale degrees are also learned [Burns and Ward 1978]. Innate forms of categorical perception are universal. For example, primary colour labels have similar or identical meanings in different languages. Learned forms, such as the categorical perception of speech vowels and musical intervals, are culture-specific.

The width of a perceptual category generally exceeds one difference threshold (or just noticeable difference, or difference limen). For example, optical frequencies which can be distinguished in only 50% of experimental trials may be regarded as one difference threshold apart; in ordinary perception, such frequencies normally fall in the same category, i.e. they have the same colour.

### 2.3.3 Holistic Perception and Pattern Recognition

Holistic (synthetic, global) perception is the perception of whole objects or scenes. It involves the direct extraction of high-level information from the environment. By contrast, analytic perception occurs only when a specific object or stimulus, or part thereof, is attended to. How holistically or analytically an even will be perceived depends on the observer [see e.g. Zenatti 1985] and on the context of the event.

Both percepts and sensations may be either holistic or analytic. An analytic sensation is defined to be the experience accompanying the "sensing", with an analytic attitude, of a stimulus. Holistic sensations are generally more meaningful than analytic sensations. They are also more likely to be linked to environmental objects, in which case they become "holistic percepts".

Holistic perception normally occurs quite spontaneously, with little or no apparent effort on the part of the observer. This is readily explained in the direct perceptual approach of Gibson [1966], according to which holistic sensations are merely experiences accompanying the direct perception of whole objects. Analytic perception requires an "analytic attitude", and can be quite difficult, even though the information being sought is more closely related to the information output by the sense organs than that sought in holistic perception. For example, it is quite difficult to hear out the harmonics of a complex tone.

Traditional psychophysics tends to regard analytic sensations as more fundamental than holistic sensations. This is because psychophysics is concerned

with the relationship between sensations and the stimuli (such as light and sound patterns) which evoke them. This relationship is held to be mediated first by the physical-physiological transducing properties of the sense organs, and secondly by *perceptual grouping* processes, by which analytic sensations corresponding to physiological output of the sense organs are grouped by stages into holistic sensations.

General principles of perceptual grouping were described by Wertheimer [1923] and Koffka [1935]. The principles cover the grouping of both simultaneous and sequential events in music, i.e. both chords and melodies. Applications in hearing and music have been described in detail by Deutsch [1982b] and Moore [1982].

If the same sensation occurs at different times, the two events may be perceived to be related (and therefore to be likely candidates for perceptual grouping) due to their *identity*. Different stimuli are perceived as identical if their difference is not perceived, i.e. if they are close enough to be assigned to the same perceptual *category* (Sect. 2.3.2).

Sensations in different categories may be grouped by *proximity* if they are close on some psychophysical scale. Visual sensations are grouped if corresponding regions of excitation are nearby on the retina. For example, a dotted or broken line is perceived as such because the dots or line segments making it up are close to each other. Stars which in three dimensions are relatively far from each other are nevertheless perceived as constellations because corresponding points on the retina are close to each other. Spontaneous grouping of auditory sensations by proximity is called *streaming* (Sect. 2.4.6).

Grouping of sensations by *familiarity* is called *pattern recognition*. Familiar patterns of sensations correspond to regularities or invariances in the environment [Gibson 1966; Bregman 1981]. Pattern recognition normally occurs quite spontaneously, with no conscious effort by the observer. The recognition of familiar patterns is an essential ingredient in the interaction of an organism with its various environments.

Instinctive behaviour in animals and humans is evidence that some aspects of pattern recognition are innate. However, most perceptual patterns become familiar by spontaneous learning and exploration in early life, implying that most aspects of pattern recognition are acquired. Later in life, pattern recognition processes become increasingly resistant to change: new perceptual patterns become increasingly difficult to learn and recognize.

Patterns may be recognized if they are incomplete, or if extra components are included. For example, a written word may still be recognized if some letters are added or taken away (i.e. if it is misspelled); the more letters are added or deleted, the less likely it is that the original word will be recognized. Melodies may be recognized if appropriate pitches are heard at appropriate times, in spite of missing or added notes: a melody whose notes are interleaved with distractor notes can still be recognized [Dowling et al. 1987].

The recognition of incomplete or superposed patterns may be modelled by *template matching* [cf. Uhr 1963]. A template (or prototype) is an idealized



representation of the perceptually relevant features of a familiar pattern of sensations. Pattern recognition may be regarded as a process whereby matches are sought between the components of a template and configurations of sensations occurring in real time. The more components of the real-time configuration match those of the template, the more likely it is that the corresponding pattern will be recognized. Note that pattern recognition templates exist only as parts of perceptual models; they have no actual physiological correlates in the peripheral or central nervous system.

The classification of perceptual grouping criteria into identity, proximity and familiarity is not always clear cut. Familiar patterns are identical to or close to previously experienced patterns, and there is no sharp dividing line between identical and proximate sensations, due to the flexibility of perceptual categories.

### 2.3.4 Ambiguity, Multiplicity and Context

A stimulus is *ambiguous* if it may be interpreted in two or more different ways. Consider again the example of a misspelled word. The more letters are added or taken away from the original word, the more ambiguous the interpretation of the word becomes – unless, of course, a new word is formed with a new, unambiguous meaning. In the template approach to pattern recognition, a stimulus pattern is ambiguous if it may be matched by a number of different templates, or by the same template in a number of different ways.

Perceptual ambiguity is normally associated with holistic perception, in which a perceptual event can have only one meaning at a time. In analytical perception, an event is analyzed into a number of simultaneous percepts: the event exhibits perceptual *multiplicity*. In the case of a written word, for example, the reader's attention can switch from holistic to analytical perception, resulting in awareness of individual letters.

The same stimulus may be ambiguous or multiple or both, depending on its *context*. A single word on a blank page (e.g. "can") is ambiguous – it has several possible meanings. It is also multiple in the sense that one's attention is focused on the individual letters of the word. In context (e.g. "I drank a can of beer after work"), both ambiguity and multiplicity are reduced. Similarly, the pitch of a single complex tone in isolation may correspond to the pitch of its first or second harmonic (or both, or a number of other possibilities: see Sect. 6.1.5), but in the context of a melody the pitch rarely differs from that of the fundamental.

By reducing ambiguity, context facilitates comprehension. Letters are easier to read in words than they are in isolation, and words are easier to read in grammatical than in non-grammatical phrases [Cattell 1886]. Similarly, musical notes are easier to read in more "grammatical" tonal contexts [Sloboda 1976].

Ambiguity is relatively unusual in perception and language. In ordinary settings, perceptual patterns are overspecified: much of the information in the

patterns is redundant [Garner 1970]. In natural language, ambiguity is normally avoided, for obvious reasons. In music, however, ambiguity plays an important role, maintaining interest and generating multiple expectations [Meyer 1973; Thomson 1983]. From this point of view, it is inappropriate to describe music as a language. If music *is* a language, it is more similar to poetry than to prose.

## 2.4 Tone Sensation

### 2.4.1 Terminology

Pipping [1895] distinguished between two kinds of pitch: the pitch of an individual pure tone component, such as a harmonic of a complex tone (which he called "tone pitch") and the overall pitch of a complex tone, corresponding to the fundamental frequency ("clang pitch"). Pipping thought that clang pitch was due to nonlinear distortion in the form of difference tones. Schouten [1940] observed that a complex tone appears to have two sensory components at the pitch of the fundamental, "one of which, having a pure tone-quality is identical with the fundamental tone, whereas the other, having a sharp tone quality and great loudness, is of different origin" (p. 358). Schouten called this additional subjective component the *residue*, hypothesizing that its pitch corresponded to the periodicity of upper, unresolved components of the complex tone.

Terhardt [1972, 1974a] made the same distinction as Pipping and Schouten, but used different terms and a different explanation for the two kinds of pitch. He proposed that *virtual* (clang/residue) pitch was formed by the spontaneous recognition of the familiar pattern of *spectral* (tone) pitches of a complex tone. The term virtual pitch added to a whole array of names for clang/residue pitch which had come into use in the meantime, among them fundamental pitch, periodicity pitch, and low pitch.

According to the American Standards Association [1960], there is only one kind of pitch: "Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high". This definition states that pitch is an *attribute* of auditory sensation – not a sensation in itself. The definition implies that there may be different kinds of *sensation* which *have* pitch, but there is only one kind of pitch. It is therefore appropriate in the above discussion to refer to two kinds of tone sensation rather than two kinds of pitch.

For this purpose, I have coined the terms *pure tone sensation* and *complex tone sensation*, as they refer directly to the types of tone which normally produce the two kinds of tone sensation. Spectral pitch may be defined using this terminology as the pitch of a pure tone sensation; virtual pitch, as the pitch of a complex tone sensation.



Pure and complex tone sensations, like all tone sensations, also have the attributes timbre and salience. This fact is hard to express using the terms spectral and virtual pitch. To refer to the timbre or salience of a spectral or virtual pitch is to refer to an attribute of an attribute. It is more logical to speak instead of the timbre or the salience of a (pure or complex) tone sensation.

A complex tone, in the proposed terminology, may evoke several different pure tone sensations (corresponding to its audible harmonics) and several different complex tone sensations (corresponding to implied fundamentals of different groups of harmonics). However, a (pure or complex) tone sensation is in all cases a single entity in the experience of the listener, with just one pitch, one timbre and one salience.

### 2.4.2 Pure Tone Sensations

Pure tone sensations are single sensations normally evoked by pure tones or pure tone components. They may also be produced by noise, because noise can evoke pitch [Fastl 1971]. Narrower bands of noise are more tone-like or "tonal" than wider bands [Aures 1984].

Complex tones are overwhelmingly heard as single wholes. The hearing out of pure tone components requires an unusually analytical listening attitude. Consequently, most people are unaware that this is possible. As hearing out of pure tone components is rarely necessary in musical performance, even musicians do not always develop the skill. For example, Rameau developed his theory of the *basse fondamentale* by experimenting with a Pythagorean monochord, and only afterwards learned that the harmonics of a tone could be individually heard [Christensen 1987].

Interestingly, Rameau believed that octave multiples of the fundamental frequency (the second, fourth, eighth, ... harmonics) were inaudible in ordinary complex tones. Scientists such as Helmholtz and Stumpf agreed that the second and fourth harmonics were harder to hear out than the third and fifth [Plomp 1964]. This effect has not (to my knowledge) been backed up by experimental data. Perhaps it is due to musical conditioning, via octave equivalence. In any case, the effect is neither expected nor explained on the basis of Terhardt's [1972, 1974a] pitch theory.

The configuration of pure tone sensations in a sound may be represented by a graph of salience (perceptual importance) against time, called the *spectral pitch pattern*. This may be regarded as the ultimate basis for the sensory attributes (pitch, timbre, salience, etc.) of complex tone sensations [Stoll 1982]. The spectral pitch pattern may be modelled as a continuous function of time by Fourier time transform [Terhardt 1985; Heinbach 1986]. The recognition of patterns (and hence sound sources) among the contours of the spectral pitches in the pattern is remarkably analogous to the recognition of visual objects from the contours of their edges and boundaries [Terhardt 1986; cf. Gibson 1979].

The salience of a pure tone component (and hence of a pure tone sensation) may be defined as its probability of being noticed, or the degree to which it contributes to the perception of complex tones (Sect. 4.4.2). It depends on audibility (level above masked threshold) and, to a lesser extent, on frequency (Sect. 6.1.2). It also depends on context. For example, pure tone components are easier to hear, and therefore more salient, if they move relative to each other [Brink 1982]: this indicates that they do not come from the same source (e.g. they are unlikely to be harmonics of the same fundamental [McAdams 1984]).

The exact pitches of pure tone components within a complex sound depend not only on frequency but also on level and masking [Terhardt 1972, 1979a; Hesse 1987] (Sect. 3.3.3). Variations of pitch with masking and changes of level are called *pitch shifts*. The pitch of a low-frequency pure tone falls slightly as its level is increased [Stevens 1935]. For example, the pitch of the electric bass of a rock band can sound sharp relative to the rest of the music when the music is damped to barely audible level by walls and/or distance. Two simultaneous pure tones which partially mask each other have the effect of pushing each other apart in pitch by a small (but perceptible) amount [Walliser 1969c]. The effect of masking on pitch is seldom noticeable, as the pure tones concerned are normally perceived as components of complex tones, and the pitch of a complex tone is affected relatively little by masking [Stoll 1985].

### 2.4.3 Complex Tone Sensations

Complex tone sensations are generally associated with percepts (or "auditory images" [McAdams 1984]) of complex tones, such as people talking, and musical instruments being played. Most tone sensations in music and in everyday sounds are of the complex kind.

With respect to pure tone sensations, complex tone sensations are holistic: they are associated with the grouping or "fusion" ("Verschmelzung" [Stumpf 1898]) of pure tone sensations. Complex tone sensations may themselves combine to form other sensations such as chord and melody sensations in music. With respect to such higher order sensations, complex tone sensations are analytical.

A complex tone may be perceived as a whole even if its fundamental is missing, i.e. if it is a *residue tone*. Schouten [1940] theorized that the pitch of a residue tone depends on the *periodicity* of unresolvable higher frequency components – the "residue". This inspired decades of psychoacoustical research into the detection of periodicity among spectral components of complex tones [Moore 1982 and references therein]. Periodicity was supposed to be detected in time intervals between peaks in the fine structure of the waveform of a sound, and coded as synchronies (*phase-locking*) in neural firing patterns.

The periodicity model explains the spectral pitch of low-frequency pure tones, and the residue pitch produced by (apparently) unresolvable high harmonics. However, the model has some serious drawbacks. The underlying assumption that a direct correspondence exists between experience and brain states or processes is unscientific, or at best premature (Sect. 2.1.2). No physiological or anatomical evidence has been found for an appropriate time measuring mechanism. And as yet no one has been able to establish a model based on periodicity which makes sensible predictions concerning the pitch properties of complex sounds in the general way that the model of Terhardt et al. [1982b] does (although plans for such a model are described by Moore [1982]).



According to Terhardt [1972, 1974a] and others [Goldstein 1973; Wightman 1973], the (virtual) pitch of a complex tone results from the recognition of a harmonic pattern among the (spectral) pitches of its resolvable (i.e. audible) pure tone components. Terhardt's model differs from the others in that it is based on *familiarity* with the pitch pattern produced by ordinary complex tones [cf. Whitfield 1967]. In Terhardt's version of pitch pattern recognition, the physiology of the perception of pure tones – in particular, whether their pitch is determined by place or time information on the basilar membrane – is not relevant (Sect. 2.2.2). The basic data of the model are in no sense the "temporal patterns of firing in different groups of auditory neurones", as suggested by Moore [1982, p. 127]. The pattern recognition part of the model is concerned with the functional relationship between two sets of *experiential*, not physical parameters: the (spectral) pitches and audibilities of the pure tone components of a sound, and the (virtual) pitches and saliences of the complex tone sensations it evokes.

Why does the pitch of a complex tone correspond to the lowest component of the pattern (the fundamental) rather than some other component? A possible reason is that the pitch of the fundamental corresponds to the period of the complex tone's waveform [Rasch and Plomp 1982]. According to the pattern recognition model, however, the auditory system is not sensitive to the period of the waveform as a whole; temporal patterns are reflected by the *roughness* of a tone, not its pitch [Terhardt 1969]. Another possible reason is that the fundamental is normally the most audible (or salient) of the harmonics of a typical complex tone: it is only masked from one side, and from a considerable pitch distance (an octave), whereas the other harmonics are masked from both sides, and at smaller intervals [Terhardt 1979a]. (Note that the fundamental does not necessarily have the highest sound pressure level, SPL. Often, a higher harmonic has the highest SPL, e.g. if it falls in the centre of a speech vowel formant.) However, the audibility of the fundamental differs from that of the other components only by degree. Perhaps the unique property which distinguishes the lowest component from the others is simply that it is the lowest [Terhardt, personal communication]. The harmonic number of the highest audible component of a typical complex tone varies over a wide range – say, from about 5 to 15 (see Sect. 6.1.1) – but the harmonic number of the lowest audible component is almost always one.

The recognition of harmonic patterns among spectral pitches may be modelled by means of a *harmonic template* incorporating the salient features

of the spectral pitch pattern of a typical complex tone (Sects. 2.3.3, 4.4.1) [cf. Cohen 1984]. The pitch distances between the components of the template are slightly stretched relative to a harmonic series of frequencies, due to pitch shifts [Terhardt 1979a]. The dependence of virtual pitches on spectral pitches [Houtsma and Rossing 1987] may be modelled by shifting the template across the pitch range and looking for matches between template components and real-time spectral pitches. The pitch of modelled complex tone sensations corresponds to that of the lowest template component; their salience, to how many spectral pitches match template components and how closely they match. Salience depends also on the context of other (pure and complex) tone sensations. Optimal fit is more important in the *spectral pitch dominance region* between about 300 and 2000 Hz [Terhardt et al. 1982b] (Sect. 6.1.2) than in higher or lower regions. So the virtual pitch of a complex tone does not necessarily correspond exactly to the spectral pitch of its fundamental, especially if the spectrum of the tone is slightly inharmonic. The template approach may be used to explain why and how, and to estimate to what extent, complex tones exhibit pitch shifts [Stoll 1984; Terhardt and Grubert 1987].

The recognition of harmonic pitch patterns in ordinary complex tones is universal. A remarkably analogous *cultural* aspect of tone perception is the assignment of tones in a musical context to particular steps of a diatonic scale, i.e. the recognition of *diatonic* pitch patterns. Jordan and Shepard [1987] studied this by presenting listeners with major scales whose intervals had been uniformly stretched (so that the octave was noticeably larger than normal) or equalized (to produce 7-tone equal temperament). The resultant shifts in the pitches of other scale steps (notably the tonic) could be explained by postulating a rigid *diatonic template* (or tonal schema) consisting of scale steps separated by the familiar intervals of the major scale.

Harmonic and diatonic pitch pattern recognition are similar in the following ways. Features of the pattern-recognition template are acquired by experience of regularly recurring pitch patterns; pitch intervals between template elements remain the same in spite of irregularities in input stimuli; modelling involves finding the *best fit* between the template and some configuration of pitches heard in real time [Moore et al. 1985]; pitch ambiguity effects (of both complex tone sensations and tonics) may be explained in terms of alternative template fits; and pitch shift effects (again, of both complex tone sensations and tonics) may be accounted for in terms of the lining up of template components. In both cases, it should be emphasized that the template is no more than part of a model and has no physiological reality.

The musical pitch of tone sensations becomes difficult to judge above a frequency of 4–5 kHz [Bachem 1948]. Proponents of the periodicity approach to pitch perception believe this is due to uncertainty in the time at which nerve impulses begin, which prevents phase-locking above 4–5 kHz [Rose et al. 1967]. Proponents of the pattern-recognition approach point out that speech harmonics rarely have audible harmonics above 4–5 kHz (e.g. the eighth harmonic of 500 Hz) and so the auditory system is not familiar with harmonic



pitch patterns in this region [Terhardt 1979a]. Whatever the reason, pure tone sensations above 4–5 kHz (i.e. above the top end of the modern piano) practically never play a harmonic role in music. The tones of the top two octaves of the piano are normally heard not as complex but as pure tone sensations, corresponding to their fundamental pure tone components; the second and higher harmonics of these tones contribute to timbre, but not to pitch (Sect. 6.1.2).

A clear complex tone sensation may be evoked by three successive harmonics not including the fundamental [Fletcher 1924]. The complex tone sensation evoked by two such harmonics is weaker, but can still be heard under suitable conditions [Sutton and Williams 1969; Smoorenburg 1970; Houtsma 1979]. There is even evidence for the existence of subharmonic complex tone sensations of single pure tone sensations [Houtgast 1976] (Sect. 2.4.5).

### 2.4.4 Pitch Ambiguity of Complex Tones

The frequencies of the lower harmonics of a complex tone do not automatically specify the tone's fundamental frequency. For example, the second and fourth harmonics could be the first and second harmonics of a complex tone an octave higher, or the first and second harmonics could be the second and fourth harmonics of a (partially masked) tone an octave lower. This explains why the virtual pitch of a complex tone in isolation is *ambiguous* (Sect. 2.3.4): it may lie an octave or (occasionally) some other consonant interval (e.g. a fifth) above or below the main pitch [Terhardt 1972, 1974a; Terhardt et al. 1986].

The degree of ambiguity of the pitch of a complex tone – the likelihood of hearing a virtual pitch which doesn't correspond to the fundamental – depends on the spectrum of the tone, the presence of maskers, and the context in which the tone is heard. The probability that the second harmonic will be heard to be the fundamental increases as the relative audibilities of the upper harmonics increase. The influence of spectral envelope (timbre) on the octave position of the pitch of a complex tone has been demonstrated in musical interval recognition experiments [Hesse 1982]. Complex tones in the presence of maskers evoke fewer pure tone sensations than unmasked tones, and therefore specify the position of the fundamental less clearly, especially if the fundamental is absent [Terhardt et al. 1986]. Tones in context (e.g. in a melody) are less ambiguous with respect to pitch than isolated tones: context normally suggests which pitch register a tone is most likely to be heard in.

The pitch ambiguity of complex tones has several important consequences for music. Inexperienced performers sometimes try to tune their instruments to a frequency an octave away from that of a reference tone such as a piano tone; once they realize their mistake, they find it easier to attend to the "correct" (main) pitch of their instrument and of the reference tone. Children and people with little musical training or ability sometimes sing a fourth, fifth of octave away from the correct pitch of a melody. The first and fifth scale

degrees of a melody are sung more faithfully than other scale degrees by young children [Francès 1972]. Sequential tones at intervals such as octaves, fifths and fourths are perceived to be related (Sects. 5.4–6), and notes an octave apart are given the same (or similar) names in written music (Sect. 3.3.1).

### 2.4.5 Subharmonic Pitches of Pure Tones

Pure tones seldom occur in isolation in the everyday environment, but sometimes only one of the pure tone components of a complex sound is audible, due to masking by other sounds. For example, loud background noises may make speech almost inaudible, so that occasionally only one of the harmonics of a speech vowel can be heard. Such a vowel may still evoke a complex tone sensation, whose (virtual) pitch lies in the range suggested by the context in which the vowel is heard, i.e. the range of pitch of previous speech vowels, or the range of pitch characteristic of a particular speaker. The *exact* pitch of the vowel may be determined by lining up the (spectral) pitch of the audible harmonic with one of the upper components of the harmonic template described in Sect. 2.4.3: it is a "sensory subharmonic" of the (spectral) pitch of the audible harmonic. A pure tone can thus be perceived either in the ordinary way, as a pure tone sensation, or – in certain contexts – as a complex tone sensation at a subharmonic pitch.

Experiments on the perception of melodies of pure tones in which selected tones are displaced by octaves [Deutsch 1973; Kallman and Massaro 1979] suggest that pure tones an octave apart can have similar melodic functions. Since pure tones appear in musical contexts in these experiments, it is likely that musical experience strongly influences the results, so it is not clear how much the results might be influenced by subharmonic pitches.

Similarity ratings of sequential pure tones by nonmusicians [Allen 1967; Thurlow and Erchul 1977] (Sects. 5.5, 6) suggest that pure tones an octave apart sound more similar than pure tones separated by slightly larger or smaller intervals. This effect could be due either to the lining up of the main pitch of one tone with the suboctave pitch of another (pitch commonality), or (again) to musical experience. By contrast, Kallman [1982] found no effect at all at the octave when pure tones were compared for similarity.

Experiments on the pitch of pure tones [Houtgast 1976] have demonstrated the existence of subharmonic pitches in the presence of background noise. Noise may allow non-existent harmonics to be spontaneously imagined [Brink 1982], and thence to contribute to the formation of complex tone sensations. In the absence of noise, Houtgast found no evidence for the existence of subharmonic pitches.

Demany and Armand [1984] familiarized infants aged three months with short melodic fragments of pure tones, and then presented new fragments in which selected tones had been shifted away from their original frequencies. The infants displayed fewer novelty reactions when tones were shifted through octave intervals than they did for other intervals. The authors concluded that



pure tones an octave apart sounded similar to these infants. The results of Demany and Armand may be explained by the hypothesis that the subharmonic tone sensations of pure tones are more salient for infants than they are for adults, due to infants' lack of familiarity with isolated pure tones. Infants have extensive experience of complex tones (e.g. speech vowels), and so are sensitive to simultaneous patterns of pure tone sensations, but they normally have relatively limited experience of sound sources which are capable of producing pure tones, and hence evoking *single* pure tone sensations (e.g. loudspeakers in psychoacoustical laboratories).

In experiments on *octave stretch* [Ward 1954; Terhardt 1970], listeners are required to tune the octave interval between alternating high and low pure tones by adjusting the frequency of one of the tones. The frequency ratio between the tones is generally found to exceed 2:1 by a small but significant margin. Terhardt [1972, 1974a] explained this result as follows: each pure tone evokes a single, unambiguous pitch, and the listener compares the interval between these pitches with the interval normally heard between the lower two pure tone sensations evoked by a complex tone, an interval which is slightly stretched due to pitch shifts. An alternative (and, for practical purposes, equivalent) explanation is that the two tones are perceived to be similar due to their pitch commonality (Sect. 3.2.3): the listener lines up the suboctave pitch of the higher tone with the main pitch of the lower tone.

### 2.4.6 Melodic Streaming

A complex tone sensation may be regarded as a grouping of pure tone sensations resulting from the spontaneous recognition of a familiar, harmonic pattern. A melodic stream is another kind of perceptual grouping, either of pure or of complex tone sensations, due to *proximity* in one or more tonal attributes (loudness, pitch, timbre, duration) or in time.

Streaming of pure tone sensations due to proximity in pitch and time occurs both for adults [Miller and Heise 1950; Noorden 1975] and for infants [Demany 1982]. A directly analogous effect occurs in vision: a pair of lights switched on and off in alternation in a dark room look like a single, moving light, provided they are close enough and they alternate fast enough [Kubovy 1981].

Complex tones, when perceived as wholes, may stream if they are similar in *timbre* [Bregman and Pinker 1978; Wessel 1979]. In orchestration, woodwind parts blend better if they are "dovetailed" (e.g. if one oboe plays higher than the first clarinet and one lower) as this inhibits streaming by timbre. In ambiguous cases, a tradeoff occurs between streaming of pure tone sensations by pitch and of complex tone sensations by timbre, depending on the relative saliences of the tone sensations [Singh 1987].

Two or three auditory streams may be heard simultaneously, but it is difficult to attend to more than one [Bregman and Campbell 1971]. It is difficult to separate streams that cross over in pitch: interleaved melodies, in which the

tones of two different melodies alternate, merge into a single, unrecognizable sequence if the melodies overlap in pitch [Dowling 1973]. However, the interleaved melodies are easier to recognize if they are already familiar.

Streaming is affected by timing and source direction. Simultaneous sounds are easier to discriminate (i.e. to segregate into different streams) if their onset times are not quite the same [Rasch 1978], as is usually the case in musical performance [Rasch 1979]. Sounds from similar directions stream (e.g. the "cocktail party effect", stereo reproduction of orchestral music). Perception of the direction of a sound source is assisted by head movements and vision [Lippman 1963b; Gibson 1966].

Sound sources (e.g. musical instruments against an orchestral texture) may often be identified by coherent variation of physical characteristics such as amplitude and frequency of harmonics [McAdams 1984; Bregman et al. 1985] otherwise known as *vibrato*. Vibrato makes it easier to follow a particular voice against to contrapuntal background. In Romantic opera, for example, vibrato enables solo voices to penetrate loud (or thick) orchestral textures. However, vibrato also inhibits blending of voices. This may explain why less vibrato was used in Baroque opera, where harmonizing was more important, and the music less passionate [Galliver 1969]. The blending of vibrato voices is improved if the vibrato is synchronized (e.g. in string quartets).

Like complex tone perception, melodic streaming may be regarded as consequence of *familiarity* with the auditory environment. In general, sounds stream if they appear to come from the same source [Bregman 1981]. Such sounds are often close in tonal attributes (loudness, pitch, timbre) and in time and direction, but not always. For example, the timbre of the clarinet differs markedly between its registers, but this does not necessarily inhibit the streaming of clarinet tones in music.

Sounds from different sources can stream, if it appears that they could have originated from the same source. In musical *hocket* the notes of a melody are played alternately on different instruments or sung by different voices. Because the timbre of the instruments or voices varies relatively little, the result sounds like a single melody. Examples are to be found in some African and Indonesian musics and, in the West, in medieval music (including Gregorian chant) and among the compositions of Webern [Dalglish 1978; Erickson 1982].

When two tones of different pitch and loudness alternate in *legato* (i.e. with no silent gap between them), the quiet tone may be perceived to remain sounding through the loud tone, even though it is physically absent [Thurlow and Elfner 1959]. This is an example of the effect called *closure* by the Gestalt psychologists. In this example, closure occurs only if the louder tone would have completely masked the quieter tone had the quieter tone actually been present. Intelligibility of speech in a noisy environment is enhanced by the effect of closure [Miller and Licklider 1950]. Like other streaming effects, this effect arises from familiarization with the audible world. It differs from other streaming effects in that it is determined not solely by regularities of the auditory environment but to a large extent by physiological limitations of the



ear (i.e. masking). In other words, it is determined by the nature of the interaction between the organism and its environment.

After considering the available experimental evidence on streaming, Sloboda [1985, p. 162] concluded that "pitch streaming is a real 'pre-musical' phenomenon, although musical knowledge may interact with and modify its effects." Streaming may thus be regarded as a sensory basis for melodic perception, and so for the theory of counterpoint [Wright 1986]. Just as melodic streaming occurs when sounds are somehow close in their sensory attributes, melodic continuity and unity are enhanced in musical performance by maintaining a relatively constant dynamic (loudness); and in composition, melodic continuity and unity are maintained by the use of small pitch and time intervals, and by maintaining a particular orchestration. In mainstream music theory and practice, wide leaps are avoided in melodies and in the voices making up a harmonic progression; when wide leaps do occur, their disruptive effect is reduced by resolving the second note by stepwise movement in the direction of the first note.

## 2.5 Pitch Perception

### 2.5.1 Dimensionality

The generally accepted definition of pitch (Sect. 2.4.1) implies that it is a one-dimensional sensory continuum. The psychological reality of such a continuum is apparent from such elementary perceptual skills as the ability to identify the higher of two pure tones (an ability which is shared by infants [Trehub 1987]) and the ability to estimate the magnitude of the pitch distance between two pure tones [Stevens et al. 1937].

Musical pitch may be described as *multidimensional*, its two main dimensions being pitch height (as in the one-dimensional model) and tone chroma [Shepard 1964, 1982; Idson and Massaro 1978]. This may be concluded from multidimensional scaling solutions of experimental results on the similarity or relative height of complex tones (Sect. 1.3.2). However, such experimental results depend on the spectra of the tones presented to listeners [Ueda and Ohgushi 1987]. It would seem more straightforward to describe the pitch of complex tones (including octave-spaced tones) as *ambiguous* relative to a *one*-dimensional pitch continuum [Terhardt 1972, 1974a, 1979b]. This clarifies the distinction between sensory and cultural effects in musical pitch perception, especially the perception of octave equivalence (cf. Sect. 3.3.1; Chap. 5), and makes it unnecessary to postulate the existence of "cognitive structures" in order to account for experimental results.

### 2.5.2 Continuous Pitch Scales

Fletcher [1934] proposed measuring the pitch of a sound in terms of the frequency of a pure reference tone of constant loudness, whose pitch is judged to be the same as that of the sound. This measure of pitch, which may be called *equivalent frequency*, was also used by Terhardt [1972, 1974a]. Terhardt's method was identical to that of Fletcher, except that he held the sound pressure level of the pure tone constant instead of its loudness. This procedure is analogous to that for measuring *loudness level* (Sect. 2.2.1), in which the frequency of a pure reference tone is held constant (at 1 kHz) and its SPL is varied; loudness level could be called "equivalent SPL".

Equivalent frequency and equivalent SPL are not proportional scales: doubling equivalent frequency does not necessarily make a sound seem twice as high, nor does doubling equivalent SPL make a sound seem twice as loud. Stevens et al. [1937] developed a proportional pitch scale (analogous to *loudness* in sone) called the *mel scale*, in which equal scalar intervals (measured in *mel*) corresponded to equal apparent interval sizes [see also Stevens and Volkmann 1940; Beck and Shaw 1963]. The mel scale is roughly proportional to the logarithm of frequency above about 1 kHz, and approaches a linear relationship with frequency at low frequencies [Zwicker and Feldtkeller 1967]. Like loudness in sone (Sect. 2.2.1), pitch in mel is quite imprecise; it is unsuitable for measuring small pitch effects such as pitch shifts.

The mel scale is an appropriate measure of pitch only when *pure* tones are heard in a *non-musical* context by musically untrained listeners. For example, Attneave and Olson [1971] found the scale to be inappropriate for the musical task of melodic transposition. Moreover, the mel scale, as originally defined by Stevens et al. [1937], only applies to the apparent size of intervals between pure tones of *equal loudness*, not equal SPL as in the experiments of Elmasian and Birnbaum [1984].

Pitch in mel may be scaled by the rule that equal sensory intervals contain roughly equal numbers of difference thresholds [Terhardt 1968c]. The difference threshold of frequency depends considerably on the listener, both for pure tones [Fastl and Hesse 1984] and complex tones [J. Meyer 1979]. On average, the difference threshold of frequency for sequential pure tones is around 0.05 semitones in the region above 500 Hz (cf. 1.0–1.5 semitones for simultaneous pure tones: Sect. 2.2.2). At lower frequencies, it is about 1 Hz [Fastl and Hesse 1984].

The difference threshold of frequency for complex tones is about the same as that for pure tones, with the exception that it retains its lowest (high-frequency) value (about 0.05 semitones) right down to about 100 Hz, or $G_2$, the bottom line of the bass clef [Walliser 1969b]. This may be understood in terms of spectral pitch dominance. The pitch of a complex tone of fundamental frequency 100–400 Hz normally depends only on the pitches of the 3rd, 4th and 5th harmonics [Ritsma 1967]. Complex tones with fundamental frequencies lower than about 100 Hz no longer have dominant harmonics above 500 Hz.



So the pitch difference threshold for a complex tone, measured in semitones, increases as its frequency falls below 100 Hz, i.e. as the frequencies of its dominant harmonics fall below 500 Hz.

Above 100 Hz, the apparent size of melodic intervals is proportional to their size in semitones. The log frequency or *frequency level* scale is therefore appropriate for the pitch of complex tones across almost all of the musical range. Only when melodies are transposed into the deep bass (below the bass clef) do melodic intervals sound smaller than normal.

### 2.5.3 Categorical Pitch Perception

The diatonic scales (major and minor) are familiar to members of Western culture, musicians and non-musicians alike. Also familiar are the non-scale notes which occur in diatonic music. In other words, the entire chromatic scale is familiar, provided a certain subset of that scale (the diatonic scale) is emphasized. Consequently, a pitch interval of random size is perceived by musicians to belong to a particular semitone category (m2, M2, m3, etc.) [Siegel and Siegel 1977a, b; Burns and Ward 1978]. Similarly, a complex tone of random frequency, presented in a tonal musical context, is perceived as belonging to a particular scale step.

The perceptual categorization of musical pitches and intervals may be regarded as a prerequisite for the understanding of pitch relationships and structures. Categorization reduces the amount of information carried by the pitches of a passage of music to a manageable level (Sect 2.3.2), removing information about the precise tuning of a pitch or interval, and retaining only its semitone category. Even under ideal listening conditions, mistunings of 0.1–0.3 semitones (depending on the interval) are acceptable [Moran and Pratt 1926; Vos 1982; Hall and Hess 1984]; even larger variations are acceptible in musical performances [Burns and Ward 1978, and references therein]. Perceptible out-of-tuneness does not necessarily affect musical meaning and function. For example, an out of tune subdominant chord still has a subdominant function within its key context, provided, of course, that it is not so out of tune that it is perceived as another chord. Mistuning is more disturbing for more salient pitches, e.g. those of a melody as opposed to its accompaniment [Rasch 1985].

The categorical perception of musical pitch begins when the auditory system "decides" whether a particular audible harmonic belongs to a complex tone [Moore et al. 1985]. Terhardt et al. [1982b] accounted for this decision-making process by assigning a "harmonicity" value to the interval between an audible component of a complex tone and its fundamental. In their model, calculated harmonicity falls gradually to zero when an interval is mistuned by 8% in frequency, or a little over a semitone. The harmonicity of the interval between a tone component and an assumed fundamental may be regarded as a measure of the probability that the component will be perceived as belonging to a complex tone with that fundamental.

In well-tuned Western harmonic progressions, the frequencies of audible pure tone components are close enough to equal temperament that all spectral pitches may be unambiguously assigned to degrees of the chromatic scale. The model in Chap. 4 takes advantage of this by defining all pitches and intervals (including intervals between harmonics and fundamentals) relative to the pitch categories of the chromatic scale. This simplifies the above decision-making procedure: the probability that a particular tone component is perceived as belonging to a particular complex tone in the model is effectively either 100% or zero. Categorization of pitch is also appropriate for modelling pitch commonality and pitch distance (Sects. 4.6.1, 2). Pitch commonality is concerned with sequential tone sensations in the same pitch category; pitch distance, with sequential tone sensations in different pitch categories.

### 2.5.4 Musical Training

As Western musicians belong to the class of Western audiences, all aspects of pitch perception discussed above apply for Western musicians. In addition, musicians' pitch perception is conditioned by their knowledge of the relationship between music as it sounds and as it is written and played, i.e. by their knowledge of music theory.

Only musicians have the opportunity to learn to recognize by name the notes, intervals, chords and keys of Western music from the sound alone. A prerequisite for the recognition of such musical elements is their categorical perception (Sect. 2.3.2).

Intervals are recognized primarily on the basis of their absolute sizes, not on the basis of their consonance. Evidence for this is that confusions between intervals in recognition experiments normally occur between intervals of similar size rather than similar consonance: for example, fifths are confused with sixths much more often than with octaves [Plomp et al. 1973; Terhardt et al. 1986]. Consonant intervals are nevertheless easier to recognize than dissonant ones [Terhardt et al. 1986]. This effect may be regarded as sensory (tones spanning consonant intervals have pitches in common, Sect. 3.2.3), cultural (consonant intervals occur more often in melodies [Jeffries 1974]), or indirectly sensory (consonant intervals occur more often in melodies because of their pitch commonality).

Differences between the perception of music by musicians and non-musicians are important for experimental reasons (Chap. 5). Musically trained participants in psychoacoustical experiments can respond in quite different ways from untrained listeners. They may be "better" participants in that their responses are more consistent and more sensitive to subtleties. If sensory properties of sounds are of paramount interest, as in this study, musicians may be "worse" participants in that their responses are more strongly influenced by musical conditioning.



### 2.5.5 Perfect Pitch

Everyone has *absolute pitch* in that they can discriminate male and female adult voices by their pitch alone (e.g. on the telephone). This kind of absolute pitch has an uncertainty or category-width of, say, three to six semitones. Like other aspects of absolute pitch, it is based on experience. Experiments with infants [Clarkson and Clifton 1985] suggest that the minimum uncertainty of "universal" absolute pitch is probably about three semitones. The fact that this is about the same as the width of a critical band in the most important range of pitch (Sect. 4.3.1) is probably coincidental: critical bandwidth is only important for simultaneous tones, whereas absolute pitch applies to isolated tones.

In Western music, the pitch continuum is divided up into absolute pitch categories much smaller than those of speech, corresponding to steps of the chromatic scale (Sect. 3.3.2). Normally, only the performer of a piece of music is aware of the names of these categories ("F", "A*b*", etc.). Therefore, only musicians are in a position to develop that kind of absolute pitch, called *perfect pitch*, in which pitch is identified absolutely in semitone categories [Harris and Siegel 1975]. Perfect pitch is normally acquired in childhood [Sergeant 1969]. With sufficient practice, it can also be acquired later in life [Cuddy 1968; Brady 1970].

There are many theoretical approaches to the origins and nature of perfect pitch [e.g. Ward 1963; Ward and Burns 1982; Costall 1985; Heyde 1987]. They mostly concentrate on perfect pitch in Western music. However, Western music is more highly developed regarding *relative* pitch (i.e. harmony) than absolute pitch. It may therefore be fruitful to look at perfect pitch (i.e. absolute pitch identification with an accuracy of a semitone or less) in musical cultures where harmony is less important. For example, unaccompanied melodies in Australian aboriginal music are sung in different places and at different times at the same frequencies, with an uncertainty of less than a semitone [Ellis 1967].

A surprising thing about perfect pitch is that so few musicians develop the ability, considering that absolute identification of stimulus properties is normal in the other senses [Watt 1917]. The reason why so few Western musicians have perfect pitch may be due in part to a conflict between the *spontaneity* of absolute pitch judgments and the *analytical* attitude to pitch required of the Western musician. The verbalization of musical note names requires quite an analytical attitude, perhaps because there are so many notes to distinguish between [Miller 1956]. In spite of this, perfect pitch often occurs quite spontaneously, in the following ways. Folk and ethnic musics in which a kind of perfect pitch is in evidence tends to be performed in a more spontaneous manner than Western art music. Perfect pitch is aided by other, relatively spontaneous experiences such as strong emotive associations [Ellis 1985, p. 65], chromesthesia or "colour hearing" [Peacock 1984; Rogers 1987], and the realization that a familiar passage of music is being played in the right or the wrong key [Terhardt and Ward 1982; Terhardt and Seewann 1983].

Musicians who identify familiar musical tones (e.g. piano tones) with great reliability are not so good at identifying the pitches of pure tones [Lockhead and Byrd 1981], suggesting that *timbre* plays an important role in perfect pitch. Absolute timbre perception and absolute pitch perception are hard to separate; this may be regarded as an example of the general fuzziness of the distinction between pitch and timbre (Sect. 2.2.1).

In a "direct perception" approach [Gibson 1966], absolute pitch involves the identification of sound *sources* according to their pitch. This explains the spontaneity of absolute pitch judgments. Further, since sound sources which differ in pitch (e.g. different piano strings) also differ in timbre, it also explains why timbre sometimes interferes with absolute pitch judgments.

Terhardt's [1972, 1974a] approach to pitch perception yields new insights into perfect pitch. According to Terhardt, musical tones exhibit octave ambiguity (the octave position of the pitch of an isolated complex tone is somewhat uncertain) and pitch shifts (the pitch of a complex tone is slightly different from that of pure tone of the same frequency). As perfect pitch possessors "memorize" the sensory properties of musical tones, they inevitably "memorize" the tones' octave ambiguity and pitch shifts, and these properties inevitably affect absolute pitch judgments. This readily explains the octave and semitone errors found by Balzano [1984] in experiments on the absolute pitch of pure tones.