4
The Role of Memory in Auditory Perception

Laurent Demany and Catherine Semal

1. Introduction

Sound sources produce physical entities that, by definition, are extended in time. Moreover, whereas a visual stimulus lasting only 1 ms can provide very rich information, that is not the case for a 1-ms sound. Humans are indeed used to processing much longer acoustic entities. In view of this, it is natural to think that "memory" (in the broadest sense) must play a crucial role in the processing of information provided by sound sources. However, a stronger point can be made: It is reasonable to state that, at least in the auditory domain, "perception" and "memory" are so deeply interrelated that there is no definite boundary between them. Such a view is supported by numerous empirical facts and simple logical considerations. Consider, as a preliminary example, the perception of loudness. The loudness of a short sound, e.g., a burst of white noise, depends on its duration (Scharf 1978). Successive noise bursts equated in acoustic power and increasing in duration from, say, 5 ms to about 200 ms are perceived not only as longer and longer but also as louder and louder. Loudness is thus determined by a temporal integration of acoustic power. This temporal integration implies that a "percept" of loudness is in fact the content of an auditory memory.

A commonsense notion is that memory is a consequence of perception and cannot be a cause of it. In the case of loudness, however, perception appears to be a consequence of memory. This is not a special case: Many other examples of such a relationship between perception and memory can be given. Consider, once more, the perception of white noise. A long sample of white noise, i.e., a completely random signal, is perceived as a static "shhhhh…" in which no event or feature is discernible. But if a 500-ms or 1-s excerpt of the same noise is taken at random and cyclically repeated, the new sound obtained is rapidly perceived as quite different. What is soon heard is a repeating sound pattern filled with perceptual events such as "clanks" and "rasping" (Guttman and Julesz 1963; Warren 1982, Chapter 3; Kaernbach 1993, 2004). It can be said that the perceptual events in question are a creation of memory, since they do not exist in the absence of repetitions. Kubovy and Howard (1976) provided another thought-provoking example. They constructed sequences of binaural "chords"

in which each chord consisted of six simultaneous pure tones with different interaural phase differences (IPDs). The tones had the same frequencies, ranging from 392 to 659 Hz, in all chords. In the first chord, the IPDs were an arbitrary function of frequency. All the subsequent chords were identical to the first chord except for a modification in the IPD of a single tone. The tone with the modified IPD changed from chord to chord, in a sawtooth manner, going gradually from 392 to 659 Hz in some sequences and vice versa in other sequences. Initially, the chords making up such a sequence are perceived as identical stimuli, but after a few iterations an ascending or descending melodic pattern emerges: In each chord, the tone with the modified IPD is perceptually segregated from the other tones, and the listener tracks this tone from chord to chord. The segregation is based on nothing but memory since the segregated tones are, intrinsically, similar to the other tones. The phenomenon is not observable when the chords are separated by very long interstimulus intervals (ISIs), but silent ISIs of 1 s are not too long.

In order to see that perception and memory are deeply interrelated, it is in fact unnecessary to consider specific stimuli. The term "perception" is quite generally used to mean, more precisely, "discrimination" or "identification." In both cases, what is "perceived" is a relation between a stimulus belonging to the present and memory traces of previous stimuli (possibly a single stimulus memory trace in the case of discrimination). When John or Jack says that he perceives the pitch of a newly presented sound as high, he means more precisely that the pitch in question is higher than the average pitch of sounds that he has heard in the past, and memorized. Consider, besides, what psychoacousticians do when they want to assess the "sensation noise" inherent to the perception of some acoustic parameter, that is, the imprecision with which this acoustic parameter is encoded by the auditory system. The only possible method to quantify sensation noise is to measure just-noticeable differences or some other index of discrimination. Thus, one must present successive stimuli and require the listener to compare them. But in such a situation, the internal noise limiting performance may include, in addition to "sensation noise," a "memory noise." Performance will be maximal for some ISI, typically several hundreds of milliseconds if the stimuli are brief. Choosing this optimal ISI does not ensure that the memory noise will be inexistent or even smaller than the sensation noise: For the optimal ISI, the only certainty will be that the memory noise is as small as it can be.

There are multiple forms of auditory memory, and they are certainly based on a variety of neural mechanisms. The present chapter will not consider all of them. For instance, although it has been noted above that one form of auditory memory is involved in the perception of loudness, temporal integrations of this kind (also observable for other auditory attributes) will be ignored in the following. The starting point of the chapter is the general idea that any sound, *once it has ended*, leaves in the brain neural traces that affect the perception of future sounds (and can also, in fact, play a role in the perceptual analysis of the ended sound). The aim of the chapter is to describe a number of interesting psychophysical

phenomena illustrating this general idea, and to relate them, as far as possible, to neurophysiological facts.

2. Neural Adaptation and Its Possible Perceptual Consequences

A very primitive and short-lived form of auditory memory manifests itself in the phenomenon of *forward masking*. The detection threshold of a brief sound is elevated by the presentation of a previous sound if the two sounds have similar or overlapping power spectra and if the time interval separating their offsets does not exceed about 200 ms (Zwislocki et al. 1959). The amount of masking, or in other words, the size of the threshold elevation, is a decreasing function of the time interval in question. The slope of this function increases with masker intensity, so that masking effects last about 200 ms more or less independently of masker intensity. Because a monaural forward masker has at most a very weak masking effect on a probe sound presented to the other ear, it is believed that the physiological substratum of forward masking is located at a relatively peripheral level of the auditory system. What is this substratum? Two hypotheses can be put forth. According to the first one, forward masking is due to a persistence of the neural excitation produced by the masker beyond its physical offset; the detection threshold of the following probe sound is elevated because, in order to be detectable, the probe must produce a detectable increment in the residual excitation produced by the masker; the just-detectable increment is an increasing function of the residual excitation, as predicted by Weber's law. According to the second hypothesis, in contrast, the trace left by the masker is negative rather than positive: forward masking is due to an "adaptation" phenomenon, that is, a decrease in the sensitivity of the neural units stimulated by the masker following its presentation; in order to be detectable, the following probe must overcome this adaptation and thus be more intense than in the masker's absence. Houtgast and van Veen (1980) and Wojtczak and Viemeister (2005) provided psychophysical evidence in support of the adaptation hypothesis. In their experiments, a 10-ms binaural probe was presented shortly after, during, or shortly before a longer and more intense masker presented to only one ear. On the ear stimulated by the masker, the level of the probe was such that the probe was partially masked but detectable. On the other ear (essentially not subjected to the masker influence), the level of the probe was controlled by the listener, who was required to adjust it to the value producing a mid-plane localization of the probe. For this physical level, one could assume that the "internal level" of the probe was the same at the two ears. When the probe was presented shortly after the masker, it appeared that the physical level adjusted at the unmasked ear was lower than the physical level at the masked ear, as if the masker attenuated the probe. This effect was smaller or absent when the probe was presented during the masker or before it. Such an outcome is consistent with the adaptation hypothesis and was not expected under the persistence hypothesis.

Adaptation effects have been observed by physiologists at the level of the auditory nerve. Is it the neural basis of forward masking? In order to answer that question, Relkin and Turner (1988) and Turner et al. (1994) measured the detection thresholds of probe signals preceded by maskers in psychophysical experiments (on human listeners) as well as physiological experiments (on chinchillas), using one and the same two-interval forced-choice procedure in both cases. In the physiological experiments, the relevant neural information was supposed to be the number of spikes appearing in a single auditory nerve fiber within a temporal window corresponding to the period of the probe presentation. The results suggested that the amount of forward masking resulting from adaptation in the auditory nerve is too small to account for the psychophysical phenomenon, and thus that an additional source of masking exists at a higher level of the auditory system. Meddis and O'Mard (2005) proposed a different scenario. They supposed that the detection of an auditory signal is not simply determined by the quantity of spikes conveyed by the auditory nerve, but requires coincidental firing of a number of nerve fibers. Using this assumption in a computer model of the auditory nerve response to probe signals in a forward masking context, they arrived at the conclusion that the model ingredients were sufficient to predict the forward making effects observable psychophysically. Nonetheless, the temporal rules of forward masking seem to be similar for normal listeners and for cochlear implant patients (Shannon 1990), which suggests that the phenomenon mainly takes place beyond the auditory nerve.

Adaptation effects indeed exist throughout the auditory pathway. Ulanovsky et al. (2003, 2004) recently described an interesting form of adaptation in the primary auditory cortex of cats. The animals were presented with long sequences of pure tones separated by silent ISIs of about 500 ms and having two possible frequencies, with different probabilities of occurrence within each sequence (e.g., f1 f1 f1 *f2* f1 f1 f1 f1 f1 *f2* f1 f1 *f2* …). Measures of spike count were made in neurons which, initially, were equally responsive to the two frequencies. What the authors found, in the course of various sequences, is a decrease in spike count in response to the more common frequency, but very little or no concomitant decrease in response to the rarer frequency. This stimulus-specific adaptation could be observed even when the two presented frequencies were only a few percent apart. It did not seem to exist subcortically, in the auditory thalamus. It appeared to have a short-term component, reflecting an effect of one tone on the response to the next tone, but also much slower components, revealing a surprisingly long neural memory: the authors uncovered an exponential trend with a time constant of tens of seconds. It was also found that stimulus-specific adaptation could be elicited by sequences of tones differing in intensity rather than frequency.

The cortical adaptation described by Ulanovsky et al. may not be a source of forward masking. However, it is likely to play a role in another perceptual phenomenon, called *enhancement*. An enhancement effect occurs when, for example, a sum of equal-amplitude pure tones forming a "notched" harmonic series (200, 400, 600, 800, 1200, 1400, 1600 Hz; note that 1000 Hz is missing) is

followed, immediately or after some ISI, by the complete series (200, 400, 600, 800, *1000*, 1200, 1400, 1600 Hz). The second stimulus is not heard as a single sound with a pitch corresponding to 200 Hz, as would happen in the absence of the first stimulus (the "precursor"). Instead, the second stimulus is heard as a sum of two separate sounds: (1) a complex tone similar to the precursor; (2) the 1000-Hz pure tone that was not included in the precursor; this pure tone "pops out,", as if the other tones were weaker due to an adaptation by the precursor. Amazingly, according to Viemeister (1980), enhancement effects are observable even when the precursor and the following stimulus are separated by several minutes or indeed hours (although stronger enhancement is of course obtained for very short ISIs). However, at least for stimulus configurations such as that considered above, enhancement appears to be a monaural effect: it is not observable when the two successive stimuli are presented to opposite ears. Viemeister and Bacon (1982) wondered whether the enhancement of a tone *T* increases its forward masking of a subsequent short probe tone with the same frequency. This was indeed verified in their experiment. Using a precursor that was temporally contiguous to the complex including *T*, they found an increase of forward masking by as much as 8 dB, on average. This increase in forward masking, normally requiring an increase of about 16 dB in masker intensity, could not be ascribed to forward masking of the probe by the precursor itself, because the latter effect was too small. Therefore, the experiment showed that an enhanced tone behaves as if it were increased in intensity. To account for that behavior, the authors noted that in the absence of the precursor, the complex including *T* produced less masking of the probe than *T* alone. This suggested that somewhere in the auditory system, *T* was attenuated by the other components of the complex. [The finding in question was in fact consistent with previous psychophysical studies on the mutual interactions of simultaneous pure tones (Houtgast 1972).] Viemeister and Bacon thus interpreted enhancement as a decrease, caused by adaptation, in the ability of a sound to attenuate other, simultaneous sounds. However, Wright et al. (1993) have cast doubts on the validity of this interpretation.

Enhancement effects are not elicited only when a pure tone is added to a previously presented sum of pure tones. In a white noise presented after a band-reject noise, one can hear clearly, as a separate sound, the band of noise that was rejected in the initial stimulus. Similarly, one can hear a given vowel in a stimulus with a flat spectrum if this stimulus is preceded by a precursor consisting of the "negative" of the vowel in question, i.e., a sound in which the vowel formants—corresponding to spectral peaks—are replaced by "antiformants" corresponding to spectral troughs (Summerfield et al. 1987; see also Wilson 1970). In one of the experimental conditions used by Summerfield et al., the precursor consisted of wideband *noise* with a uniform spectrum while the subsequent stimulus was a *complex tone* with very small spectral "bumps" at frequencies corresponding to the formants of a given vowel. This stimulus configuration still produced significant enhancement: the vowel was identified more accurately than in the absence of the precursor. Moreover, it appeared that the benefit of the noise

precursor for identification was not smaller than the benefit of a comparable *tonal* precursor with the same pitch as the subsequent stimulus. This is important because it suggests that the occurrence of enhancement does not require from the listener the perception of a similarity between the precursor and part of the subsequent stimulus. In turn, this supports the idea that the effect is caused by "low-level" mechanisms such as adaptation. However, since there are actually various forms of neural adaptation in the auditory system, it remains to be determined which one(s) matter(s) for enhancement. In fact, there may be different forms of enhancement, based on different forms of adaptation. Consider again, in this regard, the perceptual phenomenon discovered by Kubovy and Howard (1976) and described at the beginning of the chapter. Essentially, the authors found that a binaural tone can be made to pop out, in a mixture of other binaural tones, by virtue of its relative novelty. This is apparently an enhancement effect. But interestingly, the novelty involved here is neither a new frequency nor a new intensity; it is only a new interaural delay for a given frequency. If enhancement is mediated, in that case again, by adaptation, the corresponding adaptation may well be different from that underlying the enhancement of new energy in some spectral region.

In their papers about stimulus-specific cortical adaptation, Ulanovsky et al. (2003, 2004) do not relate their physiological observations to the auditory phenomenon of enhancement. However, they do state that this form of adaptation "may underlie auditory novelty detection" (Ulanovsky et al. 2003, p. 394). More specifically, they view it as the neural basis of the *mismatch negativity* or MMN. The MMN, initially identified by Näätänen et al. (1978), is a *change-specific* component of the auditory event-related potential recordable on the human scalp. One can also measure it with a brain-imaging tool such as functional magnetic resonance imaging (fMRI). Näätänen and Winkler (1999) and Schröger (1997, 2005) reviewed the enormous literature (about 1000 articles, up to now) devoted to this brain response. An MMN is typically obtained using a stimulus sequence in which a frequent "standard" sound and a rarer "deviant" sound are randomly interleaved. In such conditions, a subtraction of the average potential evoked by the standard from the average potential evoked by the deviant reveals a negative wave peaking at 100–250 ms following stimulus onset. This negative wave is the MMN. A similar wave is obtained by subtracting the response to the deviant when presented in an "alone" condition from the response to the same stimulus in the context of the sequence including more frequent iterations of the standard. In the latter sequence, the ISI between consecutive sounds may be, for instance, 500 ms, but it is not a very critical parameter: An MMN is still recordable for ISIs as long as 7 or 9 s when the standard and deviant stimuli are two tones differing in frequency by 10% (Czigler et al. 1992; Sams et al. 1993). The main source of MMN is located in the auditory cortex (e.g., Kropotov et al. 2000). It seems that any kind of acoustic change can give rise to MMN: The standard and deviant stimuli can differ in frequency, intensity, spectral profile, temporal envelope, duration, or interaural time delay. An increase in the magnitude of the change produces an increase in the amplitude of the MMN and a decrease in its

latency. Giard et al. (1995) reported that the scalp topographies of the MMNs elicited by changes in frequency, intensity, and duration are not identical, which suggests that these three types of change are processed by at least partly distinct neural populations. Analogous results were recently obtained by Molholm et al. (2005) in a study using fMRI rather than electroencephalography. The idea that there are separate and specialized MMN generators is also supported by experiments in which changes occurred on two acoustic dimensions simultaneously: The MMN obtained in response to a two-dimensional change in frequency and interaural relation, or frequency and duration, or duration and intensity, is equal to the sum of the MMNs elicited by its one-dimensional components, exactly as if each of the combined one-dimensional components elicited its own MMN (e.g., Schröger 1995).

A crucial property of the MMN is that it is a largely automatic brain response. It is usually recorded while the subject is required to ignore the stimuli and to read a book or to watch a silent film. The automaticity of the MMN tallies with the suggestion by Ulanovsky et al. (2003) that its neural basis is an adaptation mechanism already taking place in primary auditory cortex. Other authors have also argued that adaptation is the whole explanation (e.g., Jääskeläinen et al. 2004). In this view, the fact that an MMN can be elicited by, for example, a *decrease* in sound intensity, or a change in duration, would mean that certain neurons prone to adaptation are optimally sensitive to particular intensities or durations. However, several experimental results do not fit in with the adaptation hypothesis. For instance, Tervaniemi et al. (1994) recorded a significant MMN in response to occasional *repetitions* of a stimulus in a sequence of "Shepard tones" (sums of pure tones one octave apart) perceived as an endlessly descending melodic line. In the same vein, Paavilainen et al. (2001) report that an MMN can be elicited by the violation of an abstract rule relating the intensity of a pure tone to its frequency ("the higher the frequency, the higher the intensity"). Such findings suggest that even though the MMN is generated pre-attentively, the MMN generator is endowed with some intelligence allowing it to detect novelties more complex than mere modifications of specific sound events. Jacobsen and Schröger (2001) and Opitz et al. (2005) used an ingenious method to identify the respective contributions of adaptation and more "cognitive" operations in the MMN generation process. Consider the two sequences of pure tones displayed in Table 4.1. The "oddball" sequence consists of nine presentations of a 330-Hz standard tone, followed by one presentation of a 300-Hz deviant tone. In the "control" sequence, on the other hand, all tones differ from each other in frequency; however, one tone is matched in both frequency and temporal position to the deviant tone of the oddball sequence, and another tone is matched to the

TABLE 4.1. Frequencies (in Hz) of tones forming an oddball sequence and the corresponding control sequence in the experiment of Opitz et al. (2005)

Oddball	330	330	330	330	330	330(A)	330	330	330	300(B)
Control	585	363	532	399	440	330(C)	484	707	643	300(D)

standard tone of the oddball sequence. A significant difference between brain responses to the tones labeled "A" and "B" may arise from both adaptation or cognitive operations. However, suppose that an MMN is observed when the response to D is subtracted from the response to B. Since both B and D are tones with a novel frequency, one can assume that this MMN is due to cognitive operations rather than to adaptation; a contribution of adaptation is very unlikely, because the frequency difference between B and its predecessors is not larger than the frequency difference between D and any of its predecessors. On the other hand, if the response to A differs from the response to C, the main source of this effect is identifiable as adaptation rather than cognitive operations, because neither A nor C violates a previously established regularity. Following this rationale, Jacobsen and Schröger (2001) and Opitz et al. (2005) obtained evidence that both adaptation and cognitive operations contribute to the "B minus A" MMN. Taking advantage of the fMRI tool, Opitz et al. localized the adaptation component in the primary auditory cortex and the cognitive component in nonprimary auditory areas.

It has been argued that the MMN has a functional value and must be interpreted as a warning signal, drawing the subject's attention toward changes in the acoustic environment. According to Schröger (1997), a change will be detected consciously if the MMN exceeds a variable threshold, the threshold in question being low if the subject pays attention to the relevant auditory stimulation and higher otherwise. Is it clear, however, that the MMN is directly related to the conscious perception of acoustic changes? In support of this idea, Näätänen et al. (1993) and Atienza et al. (2002) found that improvement in the conscious (behavioral) discrimination of two very similar stimuli, following repeated presentations of these stimuli, can be paralleled by the development of an MMN initially not elicited by the same stimuli (see also Tremblay et al. 1997). Besides, according to Tiitinen et al. (1994), an increase in the frequency difference between two tone bursts produces precisely parallel decreases in (1) the subject's behavioral reaction time to the corresponding frequency change and (2) the latency of the MMN recordable with the same stimuli (while the subject is reading a book). The data of Tiitinen et al. suggest in addition that in the vicinity of 1000 Hz, the minimum frequency change able to elicit an MMN (in the absence of attention) is roughly similar to the frequency difference limen measurable behaviorally, under normal conditions. However, Allen et al. (2000) obtained very different results in a study on the discrimination of synthetic syllables. They found that a significant MMN could be measured for stimulus changes that were much too small to be perceived consciously. In itself, this is not inconsistent with Schröger's hypothesis on the relation between the MMN and conscious change detection. But Allen et al. also found essentially identical MMNs in response to inaudible and audible stimulus changes. They were thus led to state that "the neural generators responsible for the MMN are not necessarily linked to conscious perception" (Allen et al. 2000, p. 1389). Another argument against the idea that the MMN-generating mechanism is crucial for the conscious detection of acoustic changes is that according to several authors (see especially Cowan

et al. 1993), a given stimulus elicits a detectable MMN only if it is preceded by *several* presentations of a different stimulus. For the conscious perception of an acoustic change, a sequence of only two sounds is of course sufficient.

In summary, it has been pointed out above that neural adaptation in the auditory system—a "negative" form of auditory memory—is likely to be the cause of forward masking and to play a role in the conscious detection of novelties in the acoustic environment. With respect to the detection of novelty, stimulus-specific adaptation is useful because it makes novel sounds more salient: In a noisy jungle, as pointed out by Jääskeläinen et al. (2004), it is a matter of life or death to detect a novel sound such as that of a twig cracking under the paw of a stalking predator. On the other hand, it may be that adaptation does not help a listener to perceive consciously the *relationship* between a novel sound and a previous sound. Indeed, from a certain point of view, adaptation should impair our ability to judge whether two successive sounds are identical or differ in intensity, because if the two sounds are physically identical, their neural representations will nonetheless be systematically different in consequence of adaptation. In a later section of this chapter, it will be seen that people can consciously detect frequency changes on the basis of automatic neural processes that are apparently unrelated to adaptation as well as to other potential sources of the MMN.

3. Preperceptual Storage

Vision researchers have firmly established that soon after its termination, an optical stimulus has two types of representation in visual memory. Compelling evidence for this duality was provided, in particular, by Phillips (1974). In his experiments, observers had to make same/different judgments on visual patterns produced by randomly filling cells in a square matrix. On each trial, two successive matrices were displayed, both of them for a time of 1 s. These two matrices always had the same number of cells, but the number in question (an index of complexity) was an independent variable, as well as the ISI. In addition, the two matrices could be displayed either exactly in the same position or in slightly different positions. Finally, an irrelevant matrix acting as a mask could be either presented or not presented during the ISI. When the ISI was short (< 100 ms), discrimination performance was strongly dependent on the "position" factor (displacements impaired performance) and strongly affected by the mask; however, in the absence of displacement or mask, performance was excellent regardless of the number of cells. When the ISI was longer (600 ms), opposite results were obtained: the number of cells had a large effect on performance, which was quite poor for 8×8 matrices and still not perfect for 5×5 matrices; however, the position factor had no effect and the mask had only a weak effect. These findings, as well as other results, led to the distinction between: (1) a very short-lived but high-capacity "iconic memory," tied to spatial position and very sensitive to masking; (2) a more enduring but limited-capacity "short-term visual memory," not tied to spatial position and less sensitive to masking.

In the auditory domain, is there a similar duality of memory systems? Cowan (1984) has posited that the answer is yes. In any case, a perceptual phenomenon known as *backward recognition masking* (BRM) seems to imply that one must distinguish a "preperceptual" auditory memory (PPAM) from "postperceptual" short-term auditory memory (STAM). [In the literature, unfortunately, STAM is sometimes referred to as "echoic memory"; this is misleading since STAM is the auditory counterpart of short-term visual memory, not iconic memory.] The phenomenon of BRM was investigated in detail by Massaro (for a short review, see Massaro and Loftus 1996). In his initial experiment, Massaro (1970a) requested listeners to identify as "high" or "low" a 20-ms burst of sinusoidal sound taking two possible frequencies: 870 Hz (correct response: "high") or 770 Hz (correct response: "low"). On each experimental trial, one of the two corresponding test tones was presented and followed by a 500-ms tonal masker of 820 Hz, after an ISI randomly determined among a set of ISI values ranging from 0 to 500 ms. The masker and test tones had the same intensity. Before data collection, the three tested listeners were trained in the task for about 15 hours. The results are displayed in Figure 4.1. It can be seen that identification performance, measured in terms of percent correct, improved markedly and steadily as the ISI increased from about 40 ms to about 250 ms, and then plateaued. Let us stress that for any ISI, the test tones were clearly audible; for small ISIs, the difficulty was not to detect them but only to recognize their pitch. Subsequent experiments indicated that BRM affects, in addition to pitch judgments, judgments of loudness, duration, timbre, spatial position, and speech distinctions.

To account for these results and related ones, Massaro (1972) (see also Massaro and Loftus 1996) essentially argued that "perception takes time." According to his theory, when a short sound S_1 is presented to a listener, an image or

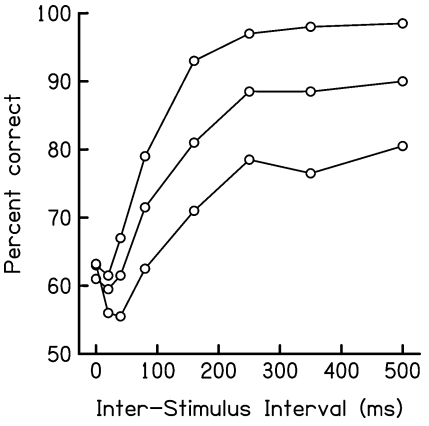

FIGURE 4.1. Results of an experiment on backward recognition masking. Identification performance as a function of the ISI for three different listeners. (Adapted with permission from Massaro DW 1970a; ©American Psychological Association.)

representation of this sound is initially stored in a PPAM system. This storage does not start at the end of the sound but at its onset (or very soon after the onset), and the temporal span of PPAM is about 250 ms, independently of the sound itself. The "perception" (or perceptual analysis) of the sound corresponds to a progressive readout of the information stored in PPAM. If a second sound S_2 is presented less than 250 ms after S_1, the perceptual analysis of S_1 is interrupted because S_2 replaces S_1 in PPAM. Hence, it is not possible to identify S_1 as accurately as in the absence of S_2. Otherwise, the perceptual analysis of S_1 continues until the disappearance of its image in PPAM. Thus, the time available and needed for an optimal perceptual analysis is the fixed temporal span of PPAM. The product of perceptual analysis is progressively transferred (as soon as perceptual analysis begins) into a different and more enduring memory system, STAM (called "synthesized auditory memory" by Massaro).

This theory met with skepticism in the psychoacoustic community. Several research teams obtained results similar to those of Massaro (1970a) in variants of the experiment described above, but it was also found that the type of psychophysical procedure used to study BRM can have a strong influence on the results (Watson et al. 1976; Yost et al. 1976). Another finding was that no BRM occurs when the mask consists of noise and is thus perceptually dissimilar to the test tones. According to Sparks (1976), even a narrowband noise in the spectral region of the test tones is an ineffective masker. From such a finding, it has been inferred that BRM does not reveal the existence of a *pre*perceptual memory and is instead a *post*perceptual effect—an interference effect in STAM. However, the inference in question is unwarranted. It is based on the erroneous idea that "preperceptual" means "not yet processed by the auditory system." On the contrary, if PPAM does exist, its neural substratum is presumably located in the auditory cortex, very far from the cochlea. The absence of BRM of tones by noise may only mean that (contrary to a hypothesis favored by Massaro) PPAM is not a single-channel memory store, completely filled by any type of sound. In this view, tones and noise have separate representations in PPAM.

Hawkins and Presson (1977) reported evidence that, to some extent, BRM depends on attention. In one of their experiments, a two-alternative pitch identification judgment ("high" or "low") had to be made on a monaural tone burst followed by a tonal masker whose frequency varied unpredictably from trial to trial. In separate blocks of trials, the masker was respectively presented (1) to the same ear as the test tone; (2) to the opposite ear; (3) diotically. Since these three conditions were blocked, it could be expected that in conditions 2 and 3, listeners would be able to filter out the masker attentionally and thus to reduce BRM. This did not happen: the results obtained in the three conditions were exactly the same, and similar to those displayed in Figure 4.1. In a second experiment, however, the authors varied the masker frequency *between* blocks of trials rather than within blocks, and this slight change in procedure had spectacular consequences: BRM was now essentially absent in conditions 2 and 3, whereas in condition 1 results similar to those shown in Figure 4.1 were once more obtained. Overall, therefore, Hawkins and Presson's study suggests that selective

attention can largely reduce BRM, but that a purely spatial attentional filter is not sufficient to do so when there is a spatial difference between the masker and test tones. Another suggestion of Hawkins and Presson's study is that a purely spectral attentional filter is also not sufficient to prevent BRM. Bland and Perrott (1978) supported that view: Paradoxically, according to these investigators, a fixed tonal masker has a stronger masking effect when its frequency is far away from the test tones' frequencies than when all stimuli are close in frequency. However, Sparks (1976) found just the opposite.

The fact that BRM seems to be to some extent dependent on attention cannot be taken as an argument against the PPAM concept. On the other hand, as pointed out by Massaro and Idson (1977), it is possible to argue that none of the studies mentioned above provides conclusive evidence for PPAM. In all of them, subjects were required to make absolute judgments: On each trial, the percept evoked by the presented test tone had to be compared with a representation of the two possible test stimuli in a "long-term" memory store. In such conditions, the deleterious effect of the masker may occur while the presented test tone is perceptually analyzed, but also following this perceptual analysis, while the percept (stored in STAM) is compared to the long-term internal representations and a decision is being made. In order to get rid of this ambiguity, Massaro and Idson (1977) replaced the original BRM paradigm by an experimental situation in which listeners simply had to make comparisons between two successive 20-ms tones, differing in frequency and separated by a variable ISI. On each trial, the frequency of the first tone (S_1) was selected at random within a frequency range of several semitones, and the frequency of the second tone (S_2) was, at random, slightly higher or lower. The task was to judge whether S_2 was higher or lower in pitch than S_1. In this situation, again, it appeared that performance increased as the ISI increased, up to at least 250 ms. This could not be explained by assuming that, for short ISIs, S_1 had a deleterious effect on the processing of S_2 because *forward* recognition masking of pitch is nonexistent (Ronken 1972; Turner et al. 1992). A conceivable alternative hypothesis would be that for short ISIs, the main difficulty was not to perceive accurately the frequency of S_1 or S_2 but to identify the temporal order in which the two tones were presented. However, this hypothesis is ruled out by the fact that the temporal order of two spectrally remote short sounds can be reliably identified as soon as the onset-to-onset interval exceeds about 20 ms (Hirsh 1959). Thus, it seems very hard to account for results such as those of Massaro and Idson without admitting the existence of PPAM and the idea that perception takes time, much more time than the stimulus itself if the stimulus is very short. Kallman and Massaro (1979) provided additional evidence that BRM is at least partly due to interference in PPAM rather than STAM.

Massaro and Idson (1977) were actually not the first to report that an increase in ISI can improve performance in a two-interval auditory discrimination task. This had been previously found by several authors (e.g., Tanner 1961). Recently, the present authors also observed such a trend in a study concerned with frequency discrimination (Demany and Semal 2005). In this respect, our data

support the main points of Massaro's theory, in particular the PPAM notion. However, the data are inconsistent with an important detail of Massaro's theory: the idea that the amount of time needed for an optimal perceptual analysis has a fixed value of about 250 ms regardless of the stimulus. In one of our experiments, the two tones presented on each trial (S_1 and S_2) consisted of either 6 or 30 sinusoidal cycles. They were separated by an ISI that varied between blocks of trials, up to 4 s. The frequency of S_1 varied unpredictably within a very wide range: 400–2400 Hz. Performance was assessed in terms of d' (Green and Swets 1974) for relative frequency shifts (from S_1 to S_2) amounting on average to ±5.8% in the "6 cycles" condition and ±1.3% in the "30 cycles" condition. The results are displayed in Figure 4.2A. For each number of cycles, as the ISI increased from 200 ms to 4 s, d' first increased, rapidly, and then decreased, more slowly. The ISI for which d' was maximal (the optimal ISI) provided an estimation of the duration needed to perceive S_1 as accurately as possible. According to Massaro's theory, this optimal ISI should have been nearly the same for the two classes of stimuli. It can be seen, however, that this was not the case: The optimal ISI appeared to be about 400 ms in the "6 cycles" condition and markedly longer, about 1 s, in the "30 cycles" condition. In another experiment, only 30-cycle stimuli were used, but their perceptual uncertainty was manipulated. There were two uncertainty conditions, in which the frequency shifts had the same relative size, on average ±0.8%. In the "high-uncertainty" condition, the frequency of S_1 could take any value from 400 Hz to 2400 Hz on each trial. Of course, S_2 varied in about the same range. In the "low-uncertainty" condition, on the other hand, S_2 could have only three possible frequencies: 400, 980, and 2400 Hz (immediate repetitions of the same S_2 from trial to trial being precluded). Figure 4.2B displays the results. It can be seen, firstly, that performance was globally better in the low-uncertainty condition than in the high-uncertainty condition, and secondly that the optimal ISI was longer in the

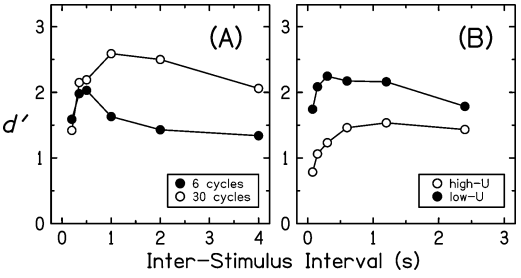

FIGURE 4.2. Results of two experiments by Demany and Semal (2005). In one of them **(A)**, one independent variable was the number of sinusoidal cycles making up each tone: 6 or 30. In the other experiment **(B)**, the number of cycles was fixed (30), but the uncertainty of the stimuli was manipulated; this uncertainty was either high ("high-U") or low ("low-U"). For both experiments, each data point is based on a total of 3000 trials (750 trials for each of the four tested listeners). (Reprinted with permission from Demany L, Semal C 2005; ©Psychonomic Society.)

latter condition. Each of these two experiments shows that frequency information is stored in a PPAM system with a temporal span of at least 1 s. Given its span, the memory store in question is apparently different from that involved in phenomena such as temporal integration of loudness, contrary to a suggestion of Cowan (1984).

4. Short-Term Auditory Memory and the Binding of Successive Sounds

Durlach and Braida (1969) made a distinction between two *memory operating modes*: a "sensory-trace" mode and a "context-coding" mode. In the sensory-trace mode, a percept is directly put in relation with the memory trace of a previous percept. In the context-coding mode, a percept is compared with a *set* of memory traces of previous percepts, including possibly quite ancient traces, and the outcome of the comparison is represented by a more or less precise verbal label (e.g., "halfway between /a/ and /o/" in the domain of timbre). Durlach and Braida pointed out that the context-coding mode is necessarily used in identification tasks (absolute judgments on single stimuli) but may also be used in a discrimination task. If, for instance, one has to make a same/different judgment on two stimuli separated by a 24-hour ISI during which many similar stimuli are presented, then a context coding of the two stimuli to be compared will presumably be the most efficient strategy. It will be so because, in contrast to a sensory trace, a verbal label can be perfectly memorized for a very long time, without any deleterious effect of interfering stimuli. However, the sensory-trace mode is undoubtedly the most efficient mode in a discrimination task such as that used by Massaro and Idson (1977) or the high-uncertainty condition of Demany and Semal (2005).

As one would expect, if two successive tones that differ very slightly in frequency or in intensity are separated by a silent ISI for as long as 5 or 10 s, their behavioral discrimination is definitely poorer than if the ISI is shorter, e.g., 1 s. This is especially true if in the test, the two stimuli presented on each trial vary in a wide range from trial to trial (Harris 1952; Berliner and Durlach 1973). Using such a "roving" procedure, Clément et al. (1999) found that the degradation of discrimination performance as the ISI increases is initially slower for frequency discrimination than for intensity discrimination. The degradation observable for frequency discrimination with silent ISIs and a roving procedure is of special interest, because in that case, the possible influence of context coding on performance is probably minimized and negligible. If so, the data reflect a pure sensory-trace decay. How to account for such a decay? The simplest model that can be thought of in the framework of signal detection theory (Green and Swets 1974) was formulated by Kinchla and Smyzer (1967). According to this model, the discrimination of two successive stimuli is limited by a sum of two independent internal noises: a "sensation noise" corresponding to the imperfect perceptual encoding of the two stimuli, and a "memory noise" resulting from a

random walk of the trace of the first stimulus during the ISI. The random walk assumption implies that the memory noise is a Gaussian variable whose variance is proportional to the ISI. As noted by Demany et al. (2005), one prediction of the model is that when the ISI increases, the relative decay of discrimination performance (d') will be slower if the sensation noise is large than if the sensation noise is small (because sensation noise and memory noise are supposed to be additive). The veracity of this prediction was actually questioned by Demany et al. (2005) on the basis of the data shown in Figure 4.2A and other data. In visual short-term memory, according to Gold et al. (2005), the fate of a trace is deterministic rather than random, contrary to the principal assumption of Kinchla and Smyzer (1967). With regard to physiology, it has been assumed that the maintenance of a sensory trace during a silent ISI is due to a sustained activity of certain neurons within this time interval. In studies on auditory frequency discrimination by monkeys, such neurons have indeed been found, at the level of the auditory cortex (Gottlieb et al. 1989) but also the dorsolateral prefrontal cortex (Bodner et al. 1996).

One might think that in humans, a frequency discrimination task is not appropriate for the study of "pure" sensory-trace decay because a pitch percept is liable to be rehearsed with profit by humming. However, this hypothesis is wrong: In fact, humming during the ISI is not profitable (Massaro 1970b; Kaernbach and Hahn, unpublished data). More generally—not only in the case of pitch—there seems to be a complete *automaticity of retention* in STAM, at least during silent ISIs. In this respect, STAM appears to be very different from short-term *verbal* memory, which is strongly dependent on attention and rehearsal processes.

One experiment suggesting that STAM is automatic was performed by Hafter et al. (1998). Their stimuli were 33-ms audiovisual signals (tone bursts coupled with colored disks). On each trial, two such signals were successively presented, and the subject had to make, in separate blocks of trials, intensity or luminance comparisons between: (1) their auditory components alone; (2) their visual components alone; (3) their auditory components *and* their visual components (which varied independently). The results obtained with a roving procedure showed that in the dual task of the third condition, the division of attention between auditory and visual signals had no deleterious effect on discrimination performance: For each sensory modality, performance was the same in the dual task and the restricted task. In this experiment, however, the ISI was short (301 ms). One can argue that different results might have been found for longer ISIs.

The present authors used longer ISIs in a purely auditory study (Demany et al. 2001). Our stimuli were 500-ms amplitude-modulated tones with three independent parameters, randomly varying from trial to trial: carrier frequency (2000–3500 Hz), modulation frequency (30–100 Hz), and intensity (48–86 dB SPL). The second of the two stimuli presented within a trial (S_2) differed from the first (S_1) with respect to only one of the three parameters. The identity of the parameter in question was selected at random, as well as the direction of the

shift. The task was to indicate whether the shift was positive or negative (without specifying the identity of the shifted parameter). In the experiment proper, the sizes of the shifts were fixed (in percent) for a given listener and parameter. These sizes had been previously adjusted in order to obtain similar levels of performance for the three parameters. Six of the eight experimental conditions are depicted in Figure 4.3. In all but one of these six conditions (the exception was condition D), a visual cue indicating the identity of the shifted parameter was provided on each trial between S_1 and S_2. Thanks to this cue, the listener could attend selectively to the relevant parameter and ignore the remaining perceptual information. Crucially, in conditions B and E, the cue was presented immediately after S_1, and the ISI was long (4 or 6 s). It could thus be expected that in these two conditions, the cue would have a positive effect on the memorization of the relevant parameter of S_1. If so, performance should have been better in condition B than in condition C, since in the latter condition, the cue was presented much later, shortly before S_2 rather than just after S_1 (this was the only difference between conditions B and C). For the same reason, performance should have been better in condition E than in condition F. However, as indicated in Figure 4.3, the average d' values obtained in conditions B and C, or E and F, were very similar. There was not even a trend in the direction expected under the hypothesis that

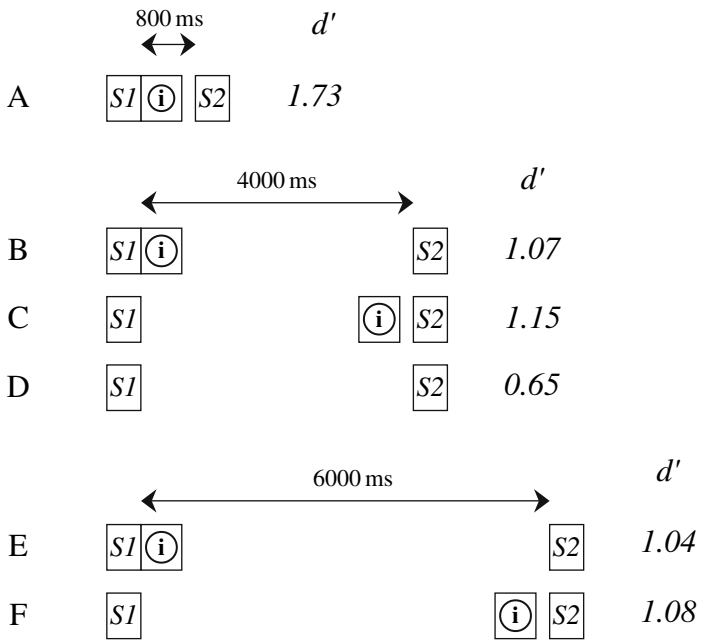

FIGURE 4.3. Results of Demany et al. (2001). "S1" and "S2" are auditory stimuli (amplitude-modulated tones) and each "lock" represents a visual cue. Each d' value is based on at least 4800 trials (at least 1200 trials for each of the four tested listeners). (Adapted with permission from Demany et al. 2001.)

attention aids memory. These data thus support the idea that STAM is automatic. Note that performance was significantly poorer in conditions B and E than in condition A, where the ISI was shorter. This proves that in conditions B and E, performance was limited by memory factors rather than by perceptual factors; a loss of information was taking place during the ISI, and thus one cannot argue that there was no room for a positive effect of attention on memory. It is also important to note that performance was significantly poorer in condition D (no cue) than in conditions B and C. This result, which was predicted by signal detection theory and presumably does *not* reflect an influence of attention on memory (see Demany et al. 2001 for a detailed discussion), rules out a trivial hypothesis according to which the cues were simply ignored.

The just-described study failed to find a benefit of attention for memory while attention was drawn onto one feature of an auditory object among other features of the same auditory object. Is attention more efficacious when, instead, it is drawn onto one auditory object among other auditory objects? This question led Clément (2001) to perform experiments in which, on each trial, the listener was initially presented with a sum of three sinusoidally amplitude-modulated pure tones with distant carrier frequencies (pitches), distant modulation frequencies, and different localizations (one tone was presented to the left ear, another tone to the right ear, and the third tone diotically). These three simultaneous tones, whose carrier frequencies varied from trial to trial, were always perceived as three separate auditory objects. After a silent ISI generally lasting 5 s (sometimes 10 s), their sum was followed by a single tone, identical in every respect to a randomly selected component of the sum except for a slight upward or downward shift in carrier frequency. The task was to identify the direction of this slight frequency shift. In most conditions, a visual cue appearing on the left, middle, or right part of a screen indicated to the listener the relevant component of the tonal complex. As in the study by Demany et al. (2001), this cue occurred either at the very beginning of the ISI or near its end. The outcome was again that the cue's temporal position had no influence on discrimination performance. In one experiment, a cue was always presented at the very beginning of the ISI, but it was invalid on about 20% of trials, unpredictably. On the trials in question, listeners were thus led to rehearse an inappropriate component of the complex. This did not impair performance, to the listeners' own surprise.

The retention of a frequency or pitch trace in STAM is in fact so automatic that paradoxically, it is possible to detect consciously a frequency difference between two tones several seconds apart *in the absence of a conscious perception of the first tone's frequency*. This was shown by Demany and Ramos (2005). In their study, listeners were presented with sums of five synchronous pure tones separated by frequency intervals that varied randomly between 6 and 10 semitones (1 semitone = 1/12 octave). In contrast to the sums of tones used by Clément (2001), these new tonal complexes, or "chords," were perceived as single auditory objects. That was because, among other things, their sinusoidal components did not differ from one another with respect to amplitude envelope or spatial localization. On each trial, a chord was followed, after a silent ISI, by

a single pure tone (*T*). Three conditions, illustrated in Figure 4.4A, were run. In the "up/down" condition, *T* was 1 semitone above or below (at random) one of the three intermediate components of the chord (at random again), and the task was to judge whether *T* was higher or lower in frequency than the closest

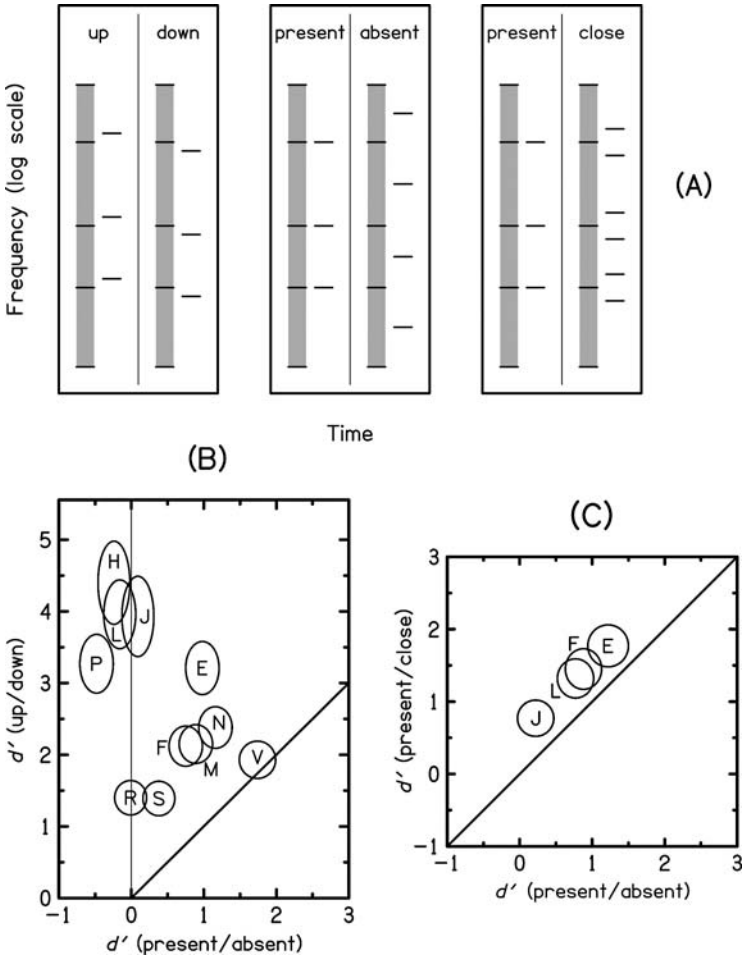

FIGURE 4.4. Experimental conditions and results of Demany and Ramos (2005). **(A)** Stimulus configurations used in the up/down, present/absent, and present/close conditions. Each horizontal segment represents a pure tone, and the shaded areas represent a possible chord. In the experiments, each chord was followed by a single *T* tone, and the ISI always exceeded the duration of both stimuli. **(B)** Results of eleven listeners in the present/absent and up/down conditions. Each ellipse (or circle) is centered on the d' values measured in the two conditions for a given listener, and its surface represents a 95% confidence area. Oblique lines indicate where the ellipses could be centered if d' were identical in the two conditions. **(C)** Results of four listeners in the present/absent and present/close conditions. (Adapted with permission from Demany and Ramos 2005; ©American Institute of Physics.)

component of the chord. In the "present/absent" condition, *T* was either identical to one of the three intermediate components or halfway in (log) frequency between two components, and the task was to judge whether *T* was present in the chord or not. The third condition, "present/close," was identical to the "present/absent" condition except that now, when *T* was not present in the chord, it was 1.5 semitone above or below (at random) one of the three intermediate components. Figure 4.4B shows the results obtained from 11 listeners in the up/down and present/absent conditions. In the present/absent condition, performance was generally quite poor: the average d' was only 0.46. This reflects the fact that although the chords' components were certainly resolved by the listeners' cochleas, it was essentially impossible to hear them individually. The chords' components were "fused" at a central level of the auditory system, and for this reason they produced on each other an "informational masking" effect (see Chapter 6). In the up/down condition, however, overall performance was very good: the average d' was 2.74. Surprisingly, the judgments required in this condition were relatively easy because the one-semitone frequency shifts elicited percepts of pitch shift even while the component of the chord that was one semitone away from *T* could not be consciously heard out. Typically, the listeners perceived *T* as the ending point of a clearly ascending or descending melodic sequence without being able to say anything about the starting point of that sequence. Definite percepts of directional pitch shift were also elicited by the 1.5-semitone frequency shifts occurring on "close" trials in the present/close condition. On "present" or "absent" trials, in contrast, the frequency of *T* was not so strongly placed in relation with a previous frequency. In the present/close condition, therefore, it was to some extent possible to distinguish the two types of trial on the basis of the audibility of a clear pitch shift. Indeed, as shown in Figure 4.4C, four listeners were more successful in that condition (average d': 1.33) than in the present/absent condition (average d': 0.77). All the data presented in Figure 4.4 were obtained for a chord-*T* ISI of 0.5 s. However, four listeners were also tested in the up/down and present/absent conditions using longer ISIs. For a 4-s ISI, performance was still markedly better in the up/down condition (average d': 0.94) than in the present/absent condition (average *d':* 0.33). A subsequent study (Demany and Ramos, in press) showed that performance in the up/down condition was not markedly poorer if the chord was presented to one ear and *T* to the opposite ear rather than all stimuli being presented to the same ear. In another experiment, the five-tone chords described above were replaced by chords of 10 tones with a constant spacing of 5.5 semitones, and in the up/down condition, *T* was positioned one semitone above or below any of the chord's 10 components. A 0.5-s ISI was used. Five listeners were still able to perform the up/down task relatively well (average d' : 1.56). In the present/absent condition, in contrast, their performance was at the chance level (average d' : 0.07).

It should be emphasized that these results are not interpretable in terms of adaptation. If listeners' judgments had been based on the relative adaptation of neurons detecting *T* by the previous chord, then the present/absent condition should have been the easiest condition, since the adaptation of neurons detecting *T* was respectively maximized and minimized on "present" and "absent" trials.

In reality, the present/absent condition was the most difficult one. The results make sense, however, under the hypothesis that the human auditory system is equipped with automatic "frequency-shift detectors" (FSDs) that can operate on memory traces in STAM. More precisely, it is possible to account qualitatively for the relative difficulties of the three experimental conditions by assuming that such detectors exist and that: (1) some of them are activated only by upward shifts, while others are activated only by downward shifts; (2) within each subset, an FSD responds more strongly to small shifts (such as one-semitone shifts) than to larger shifts; (3) when detectors of upward shifts and downward shifts are simultaneously activated—this was presumably the case in each experimental condition—the dominantly perceived shift is in the direction corresponding to the stronger activation. A similar model had been proposed by Allik et al. (1989) to account for the perception of pitch motion in sequences of chords.

The FSDs that were apparently operating on the sound sequences employed by Demany and Ramos might also play a role in the perceptual detection of very small frequency differences between isolated pure tones. This could explain why frequency discrimination and intensity discrimination are not affected in the same manner by the ISI when the ISI increases from 0.5 s to a few seconds (Clément et al. 1999). However, if evolution provided humans with FSDs, that was probably not specifically to allow them to detect minute frequency changes in an automatic way. A more plausible conjecture is that the FSDs' main function is to *bind successive sounds* and as such to serve as a tool for what Bregman (1990) called "auditory scene analysis" (see also Chapter 11 of this volume). Humans feel that they can *perceive* as a whole a succession of sounds such as a short melody. While the last tone is being heard, the first tone, stored in STAM, still belongs to the same "psychological present" (Fraisse 1967). The perceptual coherence of the whole set of tones may be partly due to the existence of FSDs. Physiologically, the FSD hypothesis is not unrealistic. In the auditory cortex of cats, the response of many neurons to a given tone can be greatly increased by previous presentation of another tone with a different frequency (McKenna et al. 1989; Brosch and Schreiner 2000). These neurons are thus particularly sensitive to tone *sequences*, as expected from FSDs. However, in the just-cited studies, increases in firing rate by a previous tone were observed only for ISIs smaller than 1 s. One must also keep in mind, as pointed out earlier in this chapter, that the mere existence of an interaction between two successive tones in the auditory system does not immediately account for the perception of a relation between them. A problem is to dissociate, in the neural activity concomitant to the presentation of the second sound, what is due to the relation between the two sounds from what is due to the intrinsic characteristics of the second sound.

It has been argued above that retention in STAM does not depend on attention and that the brain automatically puts in relation successive sounds separated by nothing but silence. Nevertheless, attention can have positive effects on auditory perception (Hafter, Chapter 5), and the retention of an auditory trace in STAM may be independent of attention only *after the formation of this trace*.

Demany et al. (2004) showed that the retention of a pitch trace is improved if, *during* the presentation of the tone evoking this pitch, attention is focused on the tone in question rather than on another, simultaneous, tone. Presumably, this occurs because attention improves the formation of the memory trace of the focused pitch. Besides, it should also be pointed out that the perception of relations between sounds is probably more attention-dependent when the sounds are *non*consecutive than when they are separated by nothing but silence. Different memory mechanisms are probably involved in these two cases. In a study by Zatorre and Samson (1991), patients who had undergone focal excisions from the temporal or frontal cortex were required to make same/different judgments (pitch comparisons) on pairs of tones separated by a 1650-ms ISI. The ISI was either silent of filled by six interfering tones, to be ignored. When the ISI was silent, the patients' performance was not significantly different from that of normal control listeners. However, when the ISI was filled by tones, a significant deficit was observed in patients with damage in the right temporal lobe or the right frontal lobe. The memory mechanisms involved in the latter case may be similar to those permitting the detection of repetitions in a cyclically repeated white-noise segment of long duration. Interestingly, whereas for humans repetition detection is possible when the repeated noise segment is as long as 10 s (Kaernbach 2004), cats are apparently unable to discriminate repeated noise from nonrepeated noise as soon as the repeated segment exceeds about 500 ms (Frey et al. 2003). For gerbils, the limit is even lower (Kaernbach and Schulze 2002).

The effect of interfering pure tones on the detectability of a pitch difference between two pure tones has been investigated in detail by Deutsch (for a summary of her work, see Deutsch 1999). She showed, among other things, that performance is much more impaired by an interfering tone that is different from the tone to be remembered (S_1) but close to it in pitch than by a remote interfering tone. Deutsch and Feroe (1975) provided evidence that the impairment produced by an interfering tone close in pitch to S_1 is not due to a destruction of the trace of S_1 but rather to an inhibition of this trace: The deleterious effect of an interfering tone I_1 can be reduced by the presentation, following I_1, of another interfering tone I_2, close in pitch to I_1 and less close to S_1; a natural interpretation is that I_2 inhibits the trace of I_1, and in doing so disinhibits the trace of S_1. Using again an experimental paradigm in which same/different judgments had to be made on pure tones possibly different in pitch, Semal and Demany (1991, 1993) wondered whether the effect of interfering tones on performance is exclusively determined by the pitch of the interfering tones or is also dependent on their other characteristics. It could be expected, for instance, that the deleterious effect of an interfering tone close in pitch to S_1 would be reduced if this tone were much less intense than S_1 or consisted of harmonics of a missing fundamental instead of being a pure tone like S_1. However, this was not the case: Pitch appeared to be the only perceptual parameter affecting performance. In the same vein, Semal et al. (1996) found that if the two stimuli to be compared are no longer tones but monosyllabic words, identical in every respect or slightly different in pitch, interfering words are not more deleterious than interfering tones with the

same pitches. These experiments, and a related study by Krumhansl and Iverson (1992), suggested that humans possess a pitch-specific memory module, deaf to loudness and timbre (spectral composition). There are also experimental data suggesting that at least some aspects of timbre are retained in a specific memory module, deaf to pitch (Starr and Pitt 1997).

The pitch-specific memory module, if it does exist, is not likely to play a major role in the auditory phenomenon reported by Demany and Ramos (2005) and described above. In our view, the FSDs hypothesized by Demany and Ramos must be thought of as detectors of *spectral* shifts rather than shifts in pitch per se (i.e., shifts in periodicity regardless of spectral composition). If, as argued above, the raison d'être of the FSDs is to bind successive sounds, then these detectors should clearly operate in the spectral domain rather than an "abstract" pitch/periodicity domain: Listeners do rely on spectral relationships when they analyze a complex auditory scene and make of it a set of temporal streams within which sounds are perceptually bound (van Noorden 1975; Hartmann and Johnson 1991; Darwin and Hukin 2000).

5. Long-Term Traces

For humans, the most important function of the auditory system is to permit *speech* communication. The processing of speech (see in this regard Chapter 10 of this volume) makes use of a long-term auditory memory, at least a rudimentary one: In order to identify an isolated spoken vowel that has just been heard, it is unnecessary to reproduce the sound vocally by trial and error; instead, the percept can be directly matched to an internal auditory template of the vowel, which leads to a rapid identification. Does human auditory long-term memory also include representations of higher-order speech entities, such as words? A majority of authors believe that the answer is negative: It is generally supposed that in the "mental lexicon" used to understand spoken words, meanings are linked with *abstract* representations of words (see, e.g., McClelland and Elman 1986); a given word is supposed to have only one representation, although the actual sound sequence corresponding to this word is not fixed but greatly depends on the speaker and various context factors. Goldinger (1996) has challenged this assumption and argued instead that a given word could be represented as a group of episodic traces retaining "surface details" such as intonation. In one of his experiments, listeners were required to identify monosyllabic words partially masked by noise and produced by several speakers. One week later, the same task was performed again, but the words were repeated either in their original voice or a new voice. The percentage of correct identifications was generally higher in the second session, but the improvement was larger for words repeated in their original voice, which implies that surface details had been kept in memory for one week. In another experiment, however, the retention of surface details appeared to be weaker (significant after one day, but absent after one week). During the second session, the listeners now had to judge

explicitly whether a given word had also been presented in the first session or was new. Therefore, Goldinger's results suggest that surface details of words can be memorized during long periods of time but that the corresponding traces persist mainly in an implicit form of memory (see also, in this respect, Schacter and Church 1992).

Like the perception of speech, the perception of *music* is strongly dependent on long-term memory traces. Shepard and Jordan (1984) performed one of the studies supporting this view. They played to psychology students a sequence of eight pure tones going from middle C (261.6 Hz) to the C one octave above in equal steps corresponding to a frequency ratio of $2^{1/7}$. The task was to judge the relative sizes of the frequency intervals between consecutive tones, from a strictly "physical" point of view and not on any musical basis. It was found that the third and seventh intervals were judged as larger than the other intervals. This can be understood by assuming that despite the instructions, the sequence played was perceptually compared to an internal template corresponding to the major diatonic scale of Western music (do, re, mi, fa, sol, la, ti, do). In this scale, the third and seventh intervals (mi-fa and ti-do) are physically *smaller* than the other intervals. When people listen to music, their knowledge of the diatonic scale and of other general rules followed in what is called "tonal music" generates expectancies about upcoming events, and these expectancies apparently affect the perception of the events themselves. Bigand et al. (2003) recently emphasized that point. In their experiments, they used sequences of chords in a well-defined musical key (e.g., C major). In half of the sequences, the last chord was made dissonant by the addition of an extra tone close in frequency to one of its components. The listeners' task was to detect these dissonances. In the absence of the extra tone, the last chord was either a "tonic" chord, ending the sequence in a normal way according to the rules of tonal music, or a less-expected "subdominant" chord. The harmonic function (tonic or subdominant) of this last chord was independent of its acoustic structure; it was entirely determined by the context of the previous chords. Yet the notes of tonic chords did not appear more frequently among the previous chords than the notes of subdominant chords. Nonetheless, dissonances were better detected in tonic chords than in subdominant chords. Remarkably, this trend was not weaker for nonmusicians than for musicians. In the same vein, Francès (1958/1988) had previously shown that for nonmusicians as well as for musicians, it is easier to perceive a change in a melodic sequence when this sequence is based on the diatonic scale than when that is not the case, all other things being equal as far as possible. Dewar et al. (1977) also found that a change in one note of a diatonic melody is easier to detect when the new melody is no longer diatonic than when the new melody is still diatonic. Such effects are probably due mainly to long-term learning phenomena rather than to intrinsic properties of the diatonic scale, because the scale in question clearly rests, in part, on cultural conventions (Dowling and Harwood 1986; Lynch et al. 1990).

Tillmann et al. (2000) devised an artificial neural network that becomes sensitive to the statistical regularities of tonal music through repeated exposure

to musical material and an unsupervised learning process. This model simulates the acquisition of implicit musical knowledge by individuals without any formal musical education, and can account for their behavior in experiments such as those just mentioned. Note, however, that the knowledge stored in the model is "abstract" in that the model does not explain at all how *specific* pieces of music can be memorized in the long term. Yet it is clear that people possess this kind of memory. Indeed, many persons are able to recognize a tune that they had not heard for years, or even decades. The perception of music is probably affected by episodic traces of this type as well as by internal representations of general, abstract rules.

Although most humans can recognize and identify a large number of tunes, very few can make a precise absolute judgment on the *pitch* of an isolated tone. Even among those who received a substantial musical education, the ability to assign a correct note label to an isolated tone—an ability called "absolute pitch"—is rare (Ward 1999). This ability requires an accurate long-term memory for pitch. It has been claimed, however, that ordinary people exhibit a surprising long-term memory for pitch when they are asked to reproduce vocally a familiar song always heard in the same key (Levitin 1994) or even when they simply speak in a normal way (Braun 2001; Deutsch et al. 2004; see also Deutsch 1991). In his influential theory about pitch perception, Terhardt (1974) has supposed the existence, in every normal adult, of an implicit form of absolute pitch. A periodic complex tone such as a vowel or a violin note is made up of pure tones forming a harmonic series, and some of these pure tones (those having low harmonic ranks) are resolved in the cochlea. When presented in isolation, these resolved pure tones evoke quite different pitch sensations. Yet, when they are summed, one typically hears a single sound and only one pitch, corresponding to the pitch of the fundamental. Why is it the case? According to Terhardt, this is entirely an effect of learning. Because speech sounds play a dominant role in the human acoustic environment, humans learn, according to this theory, to perceive any voiced speech sound, and by extension any sound with a harmonic or quasiharmonic spectrum, as a single sound with one pitch rather than as an aggregate of pure tones with various pitches. More specifically, the pitch evoked by a given pure tone of frequency f is associated in a "learning matrix" with the pitches evoked by its subharmonics ($f/2$, $f/3$, and so on), due to the co-occurrence of pure tones with these frequency relationships in voiced speech sounds. After the formation of the learning matrix (hypothetically taking place in early infancy, or maybe in utero), the associations stored in it would account for a wide range of perceptual phenomena in the domain of pitch.

A fascinating study supporting Terhardt's hypothesis was performed by Hall and Peters (1982). It deserves to be described here in detail. One class of phenomena that Terhardt intended to explain concerned the pitch of quasiharmonic sounds. Consider a sound S consisting of three pure tones at 1250, 1450, and 1650 Hz. The three components of S are not consecutive harmonics of a common fundamental (which would be the case for 1200, 1400, and 1600 Hz). However, their frequencies are close to consecutive integer multiples of 210

Hz: 6 × 210 = 1260, 7 × 210 = 1470, and 8 × 210 = 1680. When listeners are required to match *S* in pitch with a pure tone of adjustable frequency, they normally adjust the pure tone to about 210 Hz. But they do so with some difficulty, because the pitch of *S* is relatively weak. The weakness of its pitch is in part due to the fact that the components of *S* are *high-rank* quasiharmonics of a common fundamental. Indeed, it is well known that sums of low-rank harmonics elicit more salient pitch sensations than sums of high-rank harmonics. This gave to Hall and Peters the following idea. Suppose that *S* is repeatedly presented to listeners together with a simultaneous sound *S′* made up of the first five harmonics of 200 Hz (200, 400, …, 1000 Hz). If, as hypothesized by Terhardt, one can learn to perceive a given pitch simply by virtue of associative exposures, then the pitch initially perceived in *S* (about 210 Hz) might progressively change in the direction of the pitch elicited by *S′* (about 200 Hz). This was indeed found in the study of Hall and Peters. They also observed a trend consistent with Terhardt's hypothesis when the frequencies of the components of *S* were instead 1150, 1350, and 1550 Hz. In this case, the pitch of *S* rose instead of falling. However, Hall and Peters noted that they failed to induce significant changes in the pitch of perfectly harmonic stimuli.

Terhardt claimed that his theory also elucidated various phenomena relating to the perception of musical intervals. A sum of two simultaneous pure tones is perceived as more consonant when the frequency ratio of these tones is equal to 2:1 (an octave interval) or 3:2 (a fifth) than when the two tones form slightly smaller or larger intervals. In addition, people perceive a stronger consonance or "affinity" between two *sequentially* presented pure tones when they form approximately an octave or a fifth than when they form other musical intervals. Curiously, however, in order to form an optimally tuned melodic octave, two sequentially presented pure tones must generally have a frequency ratio not exactly equal to 2:1 but slightly larger, by an amount that depends on the absolute frequencies of the tones (Ward 1954); this has been called the "octave enlargement phenomenon." According to Terhardt, all these phenomena originate from the learning process that he hypothesized. The rationale is obvious for the origin of consonance, but needs to be explained with regard to the octave enlargement phenomenon. Terhardt's learning hypothesis is a little more complex than suggested above. He actually argued that the pitch elicited by a given harmonic of a voiced speech sound (before the formation of the learning matrix) is in general slightly different from the pitch of an identical pure tone presented alone, due to mutual interactions of the harmonics at a peripheral stage of the auditory system (Terhardt 1971). Consequently, the pitch intervals stored in the learning matrix should correspond to simple frequency ratios (e.g., 2:1) for pure tones presented simultaneously, but to slightly different frequency ratios for pure tones presented sequentially. This would account for the octave enlargement phenomenon.

Terhardt's conjecture on the origin of the octave enlargement phenomenon has been brought into question by Peters et al. (1983), who found that for adult listeners, the pitch of individual harmonics of complex tones does not

differ significantly from the pitch of identical pure tones presented in isolation. Moreover, Demany and Semal (1990) and Demany et al. (1991) reported other results challenging the idea that the internal template of a melodic octave originates from the perception of simultaneous pure tones forming an octave. They found that at relatively high frequencies, mistunings of simultaneous octaves are poorly detected, whereas the internal template of a melodic octave can still be very precise. Burns and Ward (1978) have argued that in musicians, the perception of melodic intervals is "categorical" due to a learning process that is determined by cultural conventions and has nothing to do with Terhardt's hypothetical learning matrix. These authors assessed the discriminability of melodic intervals differing in size by a given amount (e.g., 0.5 semitone) as a function of the size of the standard interval. In musicians, they observed markedly nonmonotonic variations of performance; performance was poorer when the two intervals to be discriminated were identified as members of one and the same interval category (e.g., 3.75 and 4.25 semitones, two intervals identified as major thirds) than when the two intervals were labeled differently (e.g., 4.25 and 4.75 semitones, the latter interval being identified as a fourth). In nonmusicians, on the other hand, performance did not depend on the size of the standard interval. The latter observation suggested that musicians' categorical perception of intervals stems from their musical training itself rather than from some "natural" source. However, the results obtained in nonmusicians by Burns and Ward are at odds with those of Schellenberg and Trehub (1996) on infants. According to Schellenberg and Trehub, 6-month-old infants detect more readily a one-semitone change in a melodic interval formed by pure tones when the first interval corresponds to a simple frequency ratio (3:2 or 4:3) than when the first interval is the medieval *diabolus in musica*, i.e., a tritone (45:32). This makes sense in the framework of Terhardt's theory. So, the roots of our perception of melodic intervals continue to be a subject of controversy (recently nourished by Burns and Houtsma 1999 and Schwartz et al. 2003).

It should be noted that Terhardt's theory is not the only pitch theory assuming the existence, in every normal adult listener, of internal templates of harmonic frequency ratios. In most of the other theories, the question of the templates' origin is avoided. However, Shamma and Klein (2000) have proposed a learning scenario that differs from Terhardt's scenario. In their theory, it is unnecessary to hear, as supposed by Terhardt, voiced speech sounds or periodic complex tones of some other family in order to acquire harmonic templates. Broadband noise is sufficient. The templates emerge due to: (1) the temporal representation of frequency in individual fibers of the auditory nerve; (2) phase dispersion by the basilar membrane; and (3) a hypothetical mechanism detecting temporal coincidences of spikes in remote auditory nerve fibers.

In the domain of *spatial hearing*, the question of the influence of learning on perception began to be asked experimentally very long ago. In a courageous study on himself, Young (1928) wore the apparatus depicted in Figure 4.5 for 18 days. This apparatus inverted the interaural time differences that are used by the binaural system to localize sounds in the horizontal plane, and Young's aim

FIGURE 4.5. The "pseudophone" of Young (1928), after Vurpillot (1975). (Reprinted with permission from Vurpillot 1975.)

was to determine whether his perception of sound localization could be reshaped accordingly. This was not the case: After 18 days, in the absence of visual cues, he was still perceiving on the right a sound coming from the left and vice versa. However, less radical alterations of the correspondence between the azimuthal position of a sound source and the information received by the binaural system are able to induce, in appropriately trained adult listeners, real changes in perceptual localization (Shinn-Cunningham 2000). Localization in the vertical plane, which depends on spectral cues related to the geometry of the external ear (pinna), may be even more malleable. If the usual spectral cues are suddenly modified by the insertion of molds in the two conchae, auditory judgments of elevation are at first seriously disrupted but become accurate again after a few weeks (Hofman et al. 1998). Interestingly, according to Hofman et al., the subject is not aware of these changes in daily life. Moreover, the new spectral code learned does not erase the previous one: When the molds are removed after having been worn permanently for several weeks, the subject immediately localizes sounds accurately and doesn't need a readaptation period; yet, a memory of the spectral code created by the molds appears to persist for several days. Learning-induced modifications of perceptual sound localization have been observed not only in humans but also in barn owls (Knudsen and Knudsen 1985; Linkenhoker and Knudsen 2002). The behavioral changes observed in barn owls were found to be associated with changes in the tuning characteristics of individual auditory

neurons (Zheng and Knudsen 1999). It is not clear, however, that psychophysical findings such as those of Hofman et al. (1998) have a similar neural basis. Actually, because their subjects did not need a readaptation period when the molds were removed, Hofman et al. avoided interpreting their findings in terms of "neural plasticity." In their view, the changes that they observed were comparable to the acquisition of a new language and were not based on functional modifications of the auditory system itself.

In humans, nonetheless, influences of auditory experience on auditory perception may well be based in part on functional modifications of the auditory system. This has been suggested to account for data concerning, in particular, *discrimination learning*. When some auditory discrimination threshold is repeatedly measured in an initially naive subject, it is generally found that the subject's performance gets better and better before reaching a plateau: The measured thresholds decrease, at first rapidly and then more and more slowly. Thousands of trials may be necessary to reach the plateau, even though the standard stimulus does not change and the challenge is only to perceive a difference between this stimulus and comparison stimuli varying in a single acoustic dimension. Interestingly, however, the amount of practice needed to reach the plateau appears to depend on the nature of the acoustic difference to be detected: A larger number of trials is required for the detection of interaural intensity differences than for the detection of interaural time differences (Wright and Fitzgerald 2001); also, the number of trials needed to optimally detect differences in (fundamental) frequency depends on the spectrum of the stimuli and is apparently longer for complex tones made up of resolvable harmonics than for pure tones or for complex tones made up of unresolvable harmonics (Demany and Semal 2002; Grimault et al. 2002). Even more interestingly, it has been found in several studies that the perceptual gain due to a long practice period with a restricted set of stimuli is, to some extent, stimulus-specific or dimension-specific. For example, after 10 one-hour sessions of practice in the discrimination of a 100-ms temporal interval (between tone bursts) from slightly different intervals, discrimination of temporal intervals is significantly improved for a 100-ms standard but not for a 50-ms or 200-ms standard (Wright et al. 1997). Analogous phenomena have been reported in the domain of pitch (Demany and Semal 2002; Grimault et al. 2002; Fitzgerald and Wright 2005) and binaural lateralization (Wright and Fitzgerald 2001). This specificity of learning clearly indicates that what is learned is genuinely *perceptual*. The rapid improvements of performance observable at the onset of practice are apparently less stimulus-specific. For this reason, they have been interpreted as due mainly to "procedural" or "task" learning (Robinson and Summerfield 1996). However, Hawkey et al. (2004) recently suggested that their main source is in fact a true perceptual change as well.

One way to account for improvements in perceptual discrimination is to assume that the subject learns to process the neural correlates of the stimuli in a more and more efficient manner. For instance, irrelevant or unreliable neural activity could be used initially in the task and then progressively discarded

(Puroshotaman and Bradley 2005). Alternatively, or in addition, the neural correlates of the stimuli could be by themselves modified during the training process, with beneficial consequences in the task; this would mean that auditory experience can be embodied by changes in the functional properties of the auditory system (Weinberger 1995, 2004; Edeline 1999). Fritz et al. (2003) and Dean et al. (2005), among others, reported results supporting the latter hypothesis. Fritz et al. (2003) trained ferrets to detect a pure tone with a specific frequency among sounds with broadband spectra, and they assessed simultaneously the spectrotemporal response field of neurons in the animals' primary auditory cortex. They found that the behavioral task swiftly modified neural response fields, in such a way as to facilitate perceptual detection of the target tone. Some of the changes in receptive fields persisted for hours after the end of the task. According to Dean et al. (2005), the neurometric function (spike rate as a function of stimulus intensity) of neurons in the inferior colliculus of the guinea pig is also plastic; it depends on the statistical distribution of the stimulus intensities previously presented to the animal. In these neurons, apparently, there is an adjustment of gain that optimizes the accuracy of intensity encoding by the neural population as a whole for the intensity range of the recently heard sounds. Recanzone et al. (1993) provided further support to the idea that auditory experience can change the properties of the auditory system itself. For several weeks, they trained adult owl monkeys in a frequency-discrimination task using pure tones and an almost invariable (but subject-dependent) standard frequency. Behavioral discrimination performance improved, and this improvement was limited to the frequency region of the standard. The authors then found that in the monkeys' primary auditory cortices, neural responses to pure tones had been affected by the training. In particular, the cortical area responding to the standard frequency was abnormally large. Remarkably, this widening was not observed in a control monkey that was passively exposed to the same acoustic stimulation as that received by one of the trained monkeys, without being required to attend to the stimuli. In a similar study on cats rather than owl monkeys, Brown et al. (2004) completely failed to replicate the main results of Recanzone et al.; but cats are poor frequency discriminators in comparison with monkeys and humans. In humans, discrimination learning has been shown to induce modifications in neuromagnetic responses of the brain to the stimuli used in the training sessions (Cansino and Williamson 1997; Menning et al. 2000). Moreover, several recent studies suggest that auditory experience is liable to induce functional changes at a subcortical level of the human auditory system (Krishnan et al. 2005; Philibert et al. 2005; Russo et al. 2005). In the study by Philibert et al., for instance, the tested subjects were hearing-impaired persons who were being fitted with binaural hearing aids, thanks to which they became able to perceive high-frequency sounds at medium or high loudness levels. Before and during the rehabilitation period, electrophysiological recordings of their auditory brainstem responses to clicks were made in the absence of the hearing aids. These recordings revealed a progressive shortening of wave-V latency, hypothetically interpreted by the authors as reflecting a neural reorganization in the inferior colliculus.

In the future, undoubtedly, further investigations of the neural correlates of perceptual changes due to experience will be carried out using more and more refined and powerful techniques.

6. Concluding Remarks

The paramount goal of perception is identification, and in the case of auditory perception it is the identification of *sound sources* and *sound sequences* (because, at least for humans, meanings are typically associated with combinations of successive sounds rather than with static acoustic features). In themselves, the mechanisms of auditory identification are as yet unclear (see McAdams 1993 for a review of models). However, it is clear that identification would be impossible in the absence of a long-term auditory memory. Another form of auditory memory, STAM, is essential for the processing of sound sequences, because without STAM, as emphasized above, the successive components of a sequence could not be connected and therefore the sequence could not be perceived as a single auditory object. These considerations suggest that psychoacousticians should intensively work on the various forms of auditory memory in order to clarify their functioning at the behavioral level. Indeed, basic questions remain unanswered in this field. Yet, during the last two decades, there has been little psychophysical research on auditory memory. Most of the experimental studies intended to provide information on auditory memory have been focused on an evoked electric potential (the MMN) which is only an indirect clue, perhaps not tightly related to psychological reality (Allen et al. 2000). The main cause of psychoacousticians' reluctance may have been the belief that auditory memory is strongly dependent on attentional factors and neural structures that are not part of the auditory system per se. Contrary to such a view, the present authors believe that auditory memory (STAM, at least) is largely automatic and that what people consciously *hear* is to a large extent determined by mnemonic machineries. What people consciously *see* may not be less affected by visual memory. However, STAM is possibly more automatic than its visual counterpart: The perceptual phenomenon reported by Demany and Ramos (2005) and described in Section 4 does not seem to have a visual counterpart. Indeed, the paradoxical sensitivity to change demonstrated by this phenomenon stands in sharp contrast to the surprising "change blindness" of vision in many situations (Rensink et al. 1997; Simons and Levin 1997). Nowadays, a variety of visual phenomena crucially involving visual memory are thoroughly investigated by psychophysicists. It is hoped that the present chapter will encourage some of its readers to undertake research of this type in the auditory domain.

Acknowledgments. The authors thank Emmanuel Bigand, Juan Segui, and especially Christophe Micheyl for helpful discussions.

References

Allen J, Kraus N, Bradlow A (2000) Neural representation of consciously imperceptible speech sound differences. Percept Psychophys 62:1383–1393.

Allik J, Dzhafarov EN, Houtsma AJM, Ross J, Versfeld HJ (1989) Pitch motion with random chord sequences. Percept Psychophys 46:513–527.

Atienza M, Cantero JL, Dominguez-Marin E (2002) The time course of neural changes underlying auditory perceptual learning. Learn Mem 9:138–150.

Berliner JE, Durlach NI (1973) Intensity perception. IV. Resolution in roving-level discrimination. J Acoust Soc Am 53:1270–1287.

Bigand E, Poulin B, Tillmann B, Madurell F, d'Adamo DA (2003) Sensory versus cognitive components in harmonic priming. J Exp Psychol [Hum Percept Perform] 29:159–171.

Bland DE, Perrott DR (1978) Backward masking: Detection versus recognition. J Acoust Soc Am 63:1215–1217.

Bodner M, Kroger J, Fuster JM (1996) Auditory memory cells in dorsolateral prefrontal cortex. NeuroReport 7:1905–1908.

Braun M (2001) Speech mirrors norm-tones: Absolute pitch as a normal but precognitive trait. Acoust Res Let Online 2:85–90.

Bregman AS (1990) Auditory Scene Analysis. Cambridge, MA: MIT Press.

Brosch M, Schreiner CE (2000) Sequence sensitivity of neurons in cat primary auditory cortex. Cereb Cortex 10:1155–1167.

Brown M, Irvine DRF, Park VN (2004) Perceptual learning on an auditory frequency discrimination task by cats: Association with changes in primary auditory cortex. Cereb Cortex 14:952–965.

Burns EM, Houtsma AJM (1999) The influence of musical training on the perception of sequentially presented mistuned harmonics. J Acoust Soc Am 106:3564–3570.

Burns EM, Ward WD (1978) Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. J Acoust Soc Am 63:456–468.

Cansino S, Williamson SJ (1997) Neuromagnetic fields reveal cortical plasticity when learning an auditory discrimination task. Brain Res 764:53–66.

Clément S (2001) La mémoire auditive humaine: Psychophysique et neuroimagerie fonctionnelle. PhD thesis, Université Victor Segalen, Bordeaux, France.

Clément S, Demany L, Semal C (1999) Memory for pitch versus memory for loudness. J Acoust Soc Am 106:2805–2811.

Cowan N (1984) On short and long auditory stores. Psychol Bull 96:341–370.

Cowan N, Winkler I, Teder W, Näätänen R (1993) Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). J Exp Psychol [Learn Mem Cogn] 19:909–921.

Czigler I, Csibra G, Csontos A (1992) Age and inter-stimulus interval effects on event-related potentials to frequent and infrequent auditory stimuli. Biol Psychol 33: 195–206.

Darwin CJ, Hukin RW (2000) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. J Acoust Soc Am 107:970–977.

Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci 8:1684–1689.

Demany L, Ramos C (2005) On the binding of successive sounds: Perceiving shifts in nonperceived pitches. J Acoust Soc Am 117:833–841.

Demany L, Ramos C (2007) A paradoxical aspect of auditory change detection. In: Kollmeier B, Klump G, Hohmann V, Mauermann M, Uppenkamp S, Verhey J (eds) Hearing: From Basic Research to Applications. Heidelberg: Springer, pp. 313–321.
Demany L, Semal C (1990) Harmonic and melodic octave templates. J Acoust Soc Am 88:2126–2135.
Demany L, Semal C (2002) Learning to perceive pitch differences. J Acoust Soc Am 111:1377–1388.
Demany L, Semal C (2005) The slow formation of a pitch percept beyond the ending time of a short tone burst. Percept Psychophys 67:1376–1383.
Demany L, Semal C, Carlyon RP (1991) On the perceptual limits of octave harmony and their origin. J Acoust Soc Am 90:3019–3027.
Demany L, Clément S, Semal C (2001) Does auditory memory depend on attention? In: Breebart DJ, Houtsma AJM, Kohlrausch A, Prijs VF, Schoonhoven R (eds) Physiological and Psychophysical Bases of Auditory Function. Maastricht, the Netherlands: Shaker, pp. 461–467.
Demany L, Montandon G, Semal C (2004) Pitch perception and retention: Two cumulative benefits of selective attention. Percept Psychophys 66:609–617.
Demany L, Montandon G, Semal C (2005) Internal noise and memory for pitch. In: Pressnitzer D, de Cheveigné A, McAdams S, Collet L (eds) Auditory Signal Processing: Physiology, Psychoacoustics, and Models. New York: Springer, pp. 230–236.
Deutsch D (1991) The tritone paradox: An influence of language on music perception. Music Percept 8:335–347.
Deutsch D (1999) The processing of pitch combinations. In: Deutsch D (ed) The Psychology of Music. New York: Academic Press, pp. 349–411.
Deutsch D, Feroe J (1975) Disinhibition in pitch memory. Percept Psychophys 17:320–324.
Deutsch D, Henthorn T, Dolson M (2004) Absolute pitch, speech, and tone language: Some experiments and a proposed framework. Music Percept 21:339–356.
Dewar K, Cuddy L, Mewhort J (1977) Recognition memory for single tones with and without context. J Exp Psychol [Hum Learn Mem Cogn] 3:60–67.
Dowling WJ, Harwood DL (1986) Music Cognition. Orlando: Academic Press.
Durlach NI, Braida LD (1969) Intensity perception. I. Preliminary theory of intensity resolution. J Acoust Soc Am 46:372–383.
Edeline JM (1999) Learning-induced physiological plasticity in the thalamo-cortical sensory systems: A critical evaluation of receptive field plasticity, map changes and their potential mechanisms. Prog Neurobiol 57:165–224.
Fitzgerald MB, Wright BA (2005) A perceptual learning investigation of the pitch elicited by amplitude-modulated noise. J Acoust Soc Am 118:3794–3803.
Fraisse P (1967) Psychologie du Temps. Paris: Presses Universitaires de France.
Francès R (1988) The Perception of Music (JW Dowling, trans). Hillsdale, NJ: Lawrence Erlbaum. Original work: La Perception de la Musique. Paris: Vrin, 1958.
Frey HP, Kaernbach C, König P (2003) Cats can detect repeated noise stimuli. Neurosci Lett 346:45–48.
Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6:1216–1223.
Giard MH, Lavikainen J, Reinikainen K, Perrin F, Bertrand O, Pernier J, Näätänen R (1995) Separate representations of stimulus frequency, intensity, and duration in auditory sensory memory. J Cogn Neurosci 7:133–143.
Gold JM, Murray RF, Sekuler AB, Bennett PJ, Sekuler R (2005) Visual memory decay is deterministic. Psychol Sci 16:769–774.

Goldinger SD (1996) Words and voices: Episodic traces in spoken word identification and recognition memory. J Exp Psychol [Learn Mem Cogn] 22:1166–1183.
Gottlieb Y, Vaadia E, Abeles M (1989) Single unit activity in the auditory cortex of a monkey performing a short term memory task. Exp Brain Res 74:139–148.
Green DM, Swets JA (1974) Signal Detection Theory and Psychophysics. Huntington, NY: Krieger.
Grimault N, Micheyl C, Carlyon RP, Collet L (2002) Evidence for two pitch encoding mechanisms using a selective auditory training paradigm. Percept Psychophys 64:189–197.
Guttman N, Julesz B (1963) Lower limits of auditory periodicity analysis. J Acoust Soc Am 35:610.
Hafter ER, Bonnel AM, Gallun E (1998) A role for memory in divided attention between two independent stimuli. In: Palmer AR, Rees A, Summerfield AQ, Meddis R (eds) Psychophysical and Physiological Advances in Hearing. London: Whurr, pp. 228–236.
Hall JW, Peters RW (1982) Change in the pitch of a complex tone following its association with a second complex tone. J Acoust Soc Am 71:142–146.
Harris JD (1952) The decline of pitch discrimination with time. J Exp Psychol 43:96–99.
Hartmann WM, Johnson D (1991) Stream segregation and peripheral channeling. Music Percept 9:155–184.
Hawkey DJC, Amitay S, Moore DR (2004) Early and rapid perceptual learning. Nat Neurosci 7:1055–1056.
Hawkins HL, Presson JC (1977) Masking and preperceptual selectivity in auditory recognition. In: Dornic S (ed) Attention and Performance VI. Hillsdale: Lawrence Erlbaum, pp. 195–211.
Hirsh IJ (1959) Auditory perception of temporal order. J Acoust Soc Am 31:759–767.
Hofman PM, van Riswick JGA, van Opstal AJ (1998) Relearning sound localization with new ears. Nat Neurosci 1:417–421.
Houtgast T (1972) Psychophysical evidence for lateral inhibition in hearing. J Acoust Soc Am 51:1885–1894.
Houtgast T, van Veen TM (1980) Suppression in the time domain. In: van den Brink G, Bilsen FA (eds) Psychophysical, Physiological and Behavioural Studies in Hearing. Delft: Delft University Press, pp. 183–188.
Jääskeläinen IP, Ahveninen J, Bonmassar G, Dale AM, Ilmoniemi RJ, Levänen S, Lin FH, May P, Melcher J, Stufflebeam S, Tiitinen H, Belliveau JW (2004) Human posterior auditory cortex gates novel sounds to consciousness. Proc Natl Acad Sci USA 101:6809–6814.
Jacobsen T, Schröger E (2001) Is there pre-attentive memory-based comparison of pitch? Psychophysiology 38:723–727.
Kaernbach C (1993) Temporal and spectral basis of the features perceived in repeated noise. J Acoust Soc Am 94:91–97.
Kaernbach C (2004) The memory of noise. Exp Psychol 51:240–248.
Kaernbach C, Schulze H (2002) Auditory sensory memory for random waveforms in the Mongolian gerbil. Neurosci Lett 329:37–40.
Kallman HJ, Massaro DW (1979) Similarity effects in backward recognition masking. J Exp Psychol [Hum Percept Perform] 5:110–128.
Kinchla RA, Smyzer F (1967) A diffusion model of perceptual memory. Percept Psychophys 2:219–229.
Knudsen EI, Knudsen PF (1985) Vision guides the adjustment of auditory localization in young barn owls. Science 230:545–548.

Krishnan A, Xu Y, Gandour J, Cariani P (2005) Encoding of pitch in the human brainstem is sensitive to language experience. Cogn Brain Res 25:161–168.
Kropotov JD, Alho K, Näätänen R, Ponomarev VA, Kropotova OV, Anichkov AD, Nechaev VB (2000) Human auditory-cortex mechanisms of preattentive sound discrimination. Neurosci Lett 280:87–90.
Krumhansl CL, Iverson P (1992) Perceptual interactions between musical pitch and timbre. J Exp Psychol [Hum Percept Perform] 18:739–751.
Kubovy M, Howard FP (1976) Persistence of a pitch–segregating echoic memory. J Exp Psychol [Hum Percept Perform] 2:531–537.
Levitin DJ (1994) Absolute memory for musical pitch: Evidence from the production of learned melodies. Percept Psychophys 56:414–423.
Linkenhoker BA, Knudsen EI (2002) Incremental training increases the plasticity of the auditory space map in adult barn owls. Nature 419:293–296.
Lynch MP, Eilers RE, Oller DK, Urbano RC (1990) Innateness, experience, and music perception. Psychol Sci 1:272–276.
Massaro DW (1970a) Preperceptual auditory images. J Exp Psychol 85:411–417.
Massaro DW (1970b) Retroactive interference in short-term recognition memory for pitch. J Exp Psychol 83:32–39.
Massaro DW (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol Rev 79:124–145.
Massaro DW, Idson WL (1977) Backward recognition masking in relative pitch judgments. Percept Mot Skills 45:87–97.
Massaro DW, Loftus GR (1996) Sensory and perceptual storage. In: Bjork EL, Bjork RA (eds) Memory. San Diego: Academic Press, pp. 67–99.
McAdams S (1993) Recognition of sound sources and events. In McAdams S, Bigand E (eds) Thinking in Sound: The Cognitive Psychology of Human Audition. Oxford: Clarendon Press, pp. 146–198.
McClelland JL, Elman, JL (1986) The TRACE model of speech perception. Cogn Psychol 18:1–86.
McKenna TM, Weinberger NM, Diamond DM (1989) Responses of single auditory cortical neurons to tone sequences. Brain Res 481:142–153.
Meddis R, O'Mard LO (2005) A computer model of the auditory-nerve response to forward-masking stimuli. J Acoust Soc Am 117:3787–3798.
Menning H, Roberts LE, Pantev C (2000) Plastic changes in the auditory cortex induced by intensive frequency discrimination training. NeuroReport 11:817–822.
Molholm S, Martinez A, Ritter W, Javitt DC, Foxe JJ (2005) The neural circuitry of pre-attentive auditory change-detection: An fMRI study of pitch and duration mismatch negativity generators. Cereb Cortex 15:545–551.
Näätänen R, Winkler I (1999) The concept of auditory stimulus representation in cognitive neuroscience. Psychol Rev 125:826–859.
Näätänen R, Gaillard AW, Mäntysalo S (1978) Early selective-attention effect on evoked potential reinterpreted. Acta Psychol 42:313–329.
Näätänen R, Schröger E, Karakas S, Tervaniemi M, Paavilainen P (1993) Development of a memory trace for a complex sound in the human brain. NeuroReport 4:503–506.
Opitz B, Schröger E, von Cramon DY (2005) Sensory and cognitive mechanisms for preattentive change detection in auditory cortex. Eur J Neurosci 21:531–535.
Paavilainen P, Simola J, Jaramillo M, Näätänen R, Winkler I (2001) Preattentive extraction of abstract feature conjunctions from auditory stimulation as reflected by the mismatch negativity (MMN). Psychophysiology 38:359–365.

Peters RW, Moore BCJ, Glasberg BR (1983) Pitch of components of complex tones. J Acoust Soc Am 73:924–929.
Philibert B, Collet L, Vesson JF, Veuillet E (2005) The auditory acclimatization effect in sensorineural hearing-impaired listeners: Evidence for functional plasticity. Hear Res 205:131–142.
Phillips WA (1974) On the distinction between sensory storage and short-term visual memory. Percept Psychophys 16:283–290.
Purushotaman G, Bradley DC (2005) Neural population code for fine perceptual decisions in area MT. Nat Neurosci 8:99–106.
Recanzone GH, Schreiner CE, Merzenich MM (1993) Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. J Neurosci 13:87–103.
Relkin EM, Turner CW (1988) A reexamination of forward masking in the auditory nerve. J Acoust Soc Am 84:584–591.
Rensink RA, O'Regan JK, Clark JJ (1997) To see or not to see: The need for attention to perceive changes in scenes. Psychol Sci 8:368–373.
Robinson K, Summerfield AQ (1996) Adult auditory learning and training. Ear Hear 17: 51S–65S.
Ronken DA (1972) Changes in frequency discrimination caused by leading and trailing tones. J Acoust Soc Am 51:1947–1950.
Russo NM, Nicol TG, Zecker SG, Hayes EA, Kraus N (2005) Auditory training improves neural timing in the human brainstem. Behav Brain Res 156:95–103.
Sams M, Hari R, Rif J, Knuutila J (1993) The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. J Cogn Neurosci 5:363–370.
Schacter DL, Church B (1992) Auditory priming: Implicit and explicit memory for words and voices. J Exp Psychol [Learn Mem Cogn] 18:915–930.
Scharf B (1978) Loudness. In: Carterette EC, Friedman MP (eds) Handbook of Perception, IV: Hearing. New York: Academic Press, pp. 187–242.
Schellenberg EG, Trehub SE (1996) Natural musical intervals: Evidence from infant listeners. Psychol Sci 7:272–277.
Schröger E (1995) Processing of auditory deviants with changes in one vs. two stimulus dimensions. Psychophysiology 32:55–65.
Schröger E (1997) On the detection of auditory deviations: A pre-attentive activation model. Psychophysiology 34:245–257.
Schröger E (2005) The mismatch negativity as a tool to study auditory processing. Acta Acust 91:490–501.
Schwartz DA, Howe CQ, Purves D (2003) The statistical structure of human speech sounds predicts musical universals. J Neurosci 23:7160–7168.
Semal C, Demany L (1991) Dissociation of pitch from timbre in auditory short-term memory. J Acoust Soc Am 89:2404–2410.
Semal C, Demany L (1993) Further evidence for an autonomous processing of pitch in auditory short-term memory. J Acoust Soc Am 94:1315–1322.
Semal C, Demany L, Ueda K, Hallé PA (1996) Speech versus nonspeech in pitch memory. J Acoust Soc Am 100:1132–1140.
Shamma S, Klein D (2000) The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. J Acoust Soc Am 107:2631–2644.
Shannon RV (1990) Forward masking in patients with cochlear implants. J Acoust Soc Am 88:741–744.

Shepard RN, Jordan DS (1984) Auditory illusions demonstrating that tones are assimilated to an internalized musical scale. Science 226:1333–1334.
Shinn-Cunningham BG (2000) Adapting to remapped auditory localization cues: A decision-theory model. Percept Psychophys 62:33–47.
Simons DJ, Levin DT (1997) Change blindness. Trends Cogn Sci 1:261–267.
Sparks DW (1976) Temporal recognition masking—or interference? J Acoust Soc Am 60:1347–1353.
Starr GE, Pitt MA (1997) Interference effects in short-term memory for timbre. J Acoust Soc Am 102:486–494.
Summerfield Q, Sidwell A, Nelson T (1987) Auditory enhancement of changes in spectral amplitude. J Acoust Soc Am 81:700–708.
Tanner WP (1961) Physiological implications of psychophysical data. Ann NY Acad Sci 89:752–765.
Terhardt E (1971) Pitch shifts of harmonics, an explanation of the octave enlargement phenomenon. In: Proceedings of the 7th International Congress on Acoustics (Budapest) 3:621–624.
Terhardt E (1974) Pitch, consonance, and harmony. J Acoust Soc Am 55:1061–1069.
Tervaniemi M, Maury S, Näätänen R (1994) Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. NeuroReport 5:844–846.
Tiitinen H, May P, Reinikainen K, Näätänen R (1994) Attentive novelty detection in humans is governed by pre-attentive sensory memory. Nature 372:90–92.
Tillmann B, Bharucha JJ, Bigand E (2000) Implicit learning of tonality: A self-organizing approach. Psychol Rev 107:885–913.
Tremblay K, Kraus N, Carrell TD, McGee T (1997) Central auditory system plasticity: Generalization to novel stimuli following listening training. J Acoust Soc Am 102:3762–3773.
Turner CW, Zeng FG, Relkin EM, Horwitz AR (1992) Frequency discrimination in forward and backward masking. J Acoust Soc Am 92:3102–3108.
Turner CW, Relkin EM, Doucet J (1994) Psychophysical and physiological forward masking studies: Probe duration and rise-time effects. J Acoust Soc Am 96:795–800.
Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. Nat Neurosci 6:391–398.
Ulanovsky N, Las L, Farkas D, Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. J Neurosci 24:10440–10453.
van Noorden LPAS (1975) Temporal coherence in the perception of tone sequences. PhD Thesis, Technische Hogeschool Eindhoven, the Netherlands.
Viemeister NF (1980) Adaptation of masking. In: van den Brink G, Bilsen FA (eds) Psychophysical, Physiological and Behavioural Studies in Hearing. Delft: Delft University Press, pp. 190–198.
Viemeister NF, Bacon SP (1982) Forward masking by enhanced components in harmonic complexes. J Acoust Soc Am 71:1502–1507.
Vurpillot E (1975) La perception de l'espace. In: Fraisse P, Piaget J (eds) Traité de Psychologie Expérimentale, vol. VI: La Perception. Paris: Presses Universitaires de France, pp. 113–198.
Ward WD (1954) Subjective musical pitch. J Acoust Soc Am 26:369–380.
Ward WD (1999) Absolute pitch. In: Deutsch D (ed) The Psychology of Music. San Diego: Academic Press, pp. 265–298.
Warren RM (1982) Auditory Perception: A New Synthesis. New York: Pergamon.

Watson CS, Kelly WJ, Wroton HW (1976) Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty. J Acoust Soc Am 60:1176–1186.

Weinberger NM (1995) Dynamic regulation of receptive fields and maps in the adult sensory cortex. Annu Rev Neurosci 18:129–158.

Weinberger NM (2004) Specific long-term memory traces in primary auditory cortex. Nat Rev Neurosci 5:279–290.

Wilson JP (1970) An auditory after-image. In: Plomp R, Smoorenburg GF (eds) Frequency Analysis and Periodicity Detection in Hearing. Leiden, the Netherlands: Sijthoff, pp. 303–318.

Wojtczak M, Viemeister NF (2005) Mechanisms of forward masking. J Acoust Soc Am 115: 2599.

Wright BA, Fitzgerald MB (2001) Different patterns of human discrimination learning for two interaural cues to sound-source location. Proc Natl Acad Sci USA 98:2307–12312.

Wright BA, McFadden D, Champlin CA (1993) Adaptation of suppression as an explanation of enhancement effects. J Acoust Soc Am 94:72–82.

Wright BA, Buonomano DV, Mahncke HW, Merzenich MM (1997) Learning and generalization of auditory temporal-interval discrimination in humans. J Neurosci 17:3956–3963.

Yost WA, Berg K, Thomas GB (1976) Frequency recognition in temporal interference tasks: A comparison among four psychophysical procedures. Percept Psychophys 20:353–359.

Young PT (1928) Auditory localization with acoustical transposition of the ears. J Exp Psychol 11:399–429.

Zatorre RJ, Samson S (1991) Role of the right temporal neocortex in retention of pitch in auditory short-term memory. Brain 114:2403–2417.

Zheng W, Knudsen EI (1999) Functional selection of adaptive auditory space map by $GABA_A$–mediated inhibition. Science 284:962–965.

Zwislocki J, Pirodda E, Rubin H (1959) On some poststimulatory effects at the threshold of audibility. J Acoust Soc Am 31:9–14.