# 5
# Auditory Attention and Filters

Ervin R. Hafter, Anastasios Sarampalis, and Psyche Loui

## 1. Introduction

The traditional approach to the study of sound sources emphasized bottom-up analysis of acoustic stimuli, whether they were more primitive features such as frequency, sound level, and source direction, or complexes made by combinations of primitives. However, a simple scan of the table of contents of this volume shows that the field has evolved considerably toward a realization of important top-down processes that modulate the perception of sounds as well as control how we derive and interpret the natural acoustical events of everyday life. Probably the most commonly used word in this regard is *attention*, a term whose meaning is "understood" by everyone, but whose scientific description encompasses a variety of operational definitions. The current chapter does not attempt to address all of these approaches. Rather, it concentrates on the listener's ability to extract relevant features of the auditory scene and seeks to understand the seeming ability to focus on some parts of the auditory stream at the expense of others. While perceptual attention is typically defined in terms of internal processes that help us extract relevant information from a complex environment, such a broad view does not tell us about the nature or locus of the selection process. The focus of attentional theories ranges from stimulus cohesion, whereby attention binds sensory features into higher-order percepts (Treisman and Gelade 1980), to segregation, which breaks stimuli into parts so that selective processing can be allotted to the attended element. The concentration here is on the latter, though issues of feature binding become important when segregation is based on a combination of auditory features such as harmonicity and timbre, or on the collective effects of a string of tones in a melody.

Intensive study of attention to specific portions of auditory messages began in the 1950s through seminal experiments and commentary by a group of individuals (see Cherry 1953; Broadbent 1958; Deutsch and Deutsch 1963; Norman 1969; Treisman 1969; Moray 1970) whose focus was on what happens when we are confronted with sounds from multiple auditory sources. Cherry's colorful description of the "cocktail party effect" was seen as an example of the listener's ability to derive information from one stream of speech in the presence of others, and from this grew a theoretical discussion of the presence of specific filters (or channels) that select information that matches their pass bands for interpretation by more central processes. A common methodology in this early work on spatially defined attentional channels was dichotic listening, in which two different streams of spoken material were presented simultaneously, one to each ear, and the subject was instructed to "shadow" the attended speech at one ear by repeating it as it was heard. Inability to recall information from the unattended ear fitted well with Broadbent's (1958) filter theory. In order to model the perceptual bottleneck caused by two messages reaching a sensory buffer at the same time, he proposed a selective filtering process that would allow one message to go forward while leaving the other to decay in short-term memory. In this way, a more robust feature of the unattended stimulus such as its fundamental frequency might persist until after analysis of the selected message was complete. Based on subsequent evidence that some higher-level information, such as the listener's name, might break through to awareness from the unattended source (e.g., Moray 1960), Treisman (1964, 1969) abandoned the notion of all-or-none filters in favor of attentional attenuators that selectively reduce the effectiveness of a stimulus without totally blocking it. From a more cognitive perspective, Deutsch and Deutsch (1963) eschewed the idea of peripheral filtering in favor of late selection that could include semantic factors as well. In the years since, there have been no clear winners in the debate about early and late sites of selection, with evidence for attentional filtering ranging from hard-wired frequency analyzers to segregation of simultaneous auditory streams. One potential reason for this ambiguity is described by Broadbent's (1958) notion that information blocked by an early bottleneck may be later processed in short-term memory. In this regard, Norman (1969) demonstrated that when subjects are asked about what they just heard on the unattended side when a shadowing task is suddenly terminated, they can produce up to 30 seconds of recall.

The more specific a definition of attention, the greater its reliance on the experimental operations used to define it. Traditionally, work on visual attention has relied heavily on measures of reaction times (RTs). Conversely, studies of auditory attention have concentrated on paradigms that measure the ability to extract signals from a noisy background for detection and/or discrimination. In this case, attention is often thought of in terms of the listener's focus on the expected location of a signal along a monitored dimension. While this is the major concern of the current chapter, also noted are cases in which the use of RT in audition has been useful for highlighting the distinction between endogenous cues, whose information directs attention to the predicted locations of meaningful stimuli, and exogenous cues, which provoke a reflexive attraction to a location, regardless of its relation to the task. Finally, it is obvious that auditory attention affects more than just simple acoustical features, having profound influences on such complex processes as informational masking, spatial hearing, and speech understanding. However, because those topics appear elsewhere in this volume (Chapters 6, 8, and 10), the stress here is on the effects of cueing in activating attentional filters based on expected features of the auditory signal.

# 2. Signal Detection

## 2.1 Detecting a Single Signal in Noise

Sensory coding generally begins with the breakdown of complex stimuli into more narrowly defined regions along fundamental dimensions. In audition, this is done by processes in the cochlea that separate the acoustic input into the separate frequency channels leading to frequency-specific activity in the auditory nerve. In the history of psychoacoustics, these frequency channels have taken on various names including “critical bands” and “auditory filters” (see Moore 2003 for a review). Division of this kind provides a distinct advantage for signal detection by increasing the signal-to-noise ratio (S/N) for narrowband signals through the rejection of interference from more distant regions of the spectrum.

An early demonstration of exclusive processing of a single band was evident from the effects of reducing the width of a noise masker on detection of a tonal signal. At first, limiting the masker had no effect on signal detection, but performance rose when the bandwidth was reduced to the point that its edges fell into the so-called critical band surrounding the signal (Fletcher 1940). The ability to respond to the stimulus within a selected region of the spectrum is reminiscent of one of the oldest theories of attention, in which it is pictured as focusing of a searchlight on the relevant information. A demonstration is seen in measures of the “critical ratio” (Fletcher 1940). Here, the bandwidth of the masker remains wide, but the subject is informed about the frequency to be detected by presentation of the signal before the experiment begins. The primary assumption of this method is that the band level of the masker (BL) of the noise within the listening band is proportional to the level of the signal at threshold. Given that the total wideband noise has a spectrum level of $N_0$ (level/Hz), the width of a rectangular equivalent of the listening band is computed by dividing BL by $N_0$. Hence, the term *critical ratio* (see Hartmann 1997). The high correlation between critical ratios and other measures of the bandwidths (Scharf 1970; Houtsma 2004) lends credence to the view that attention can focus on a specific region of an auditory dimension, even when the stimulus covers a much wider range. Throughout this chapter, we will use the term *filter* to describe this kind of attentional selection in a variety of auditory dimensions.

The important attentional assumption of the critical ratio is that the listener is accurately informed about where to listen. As noted above, this is typically done by playing a sample of the signal before data collection begins, and it is often enhanced by feedback presented after each trial. A potential weakness of feedback is that the subject must hold a representation of the stimulus in the signal interval in memory for comparison to the feedback, which may explain improvements in performance at the beginning of tests with weak signals (Gundy 1961). Some experiments have attempted to focus the subject with simultaneous cueing such as by adding the signal to a continuous tonal pedestal set to the same frequency as the signal (see Green 1960) or, for detection of a monaural signal in noise, by a sample of the signal presented to the contralateral ear (Taylor and Forbes 1969; Yost et al. 1972). However, one must be cautious about the possibility that these cues introduce detectable changes in dimensions other than energy. In the former case, this might due to introduction of transients when signals are added to the pedestal (e.g., Macmillan 1971; Leshowitz and Wightman 1972; Bonnel and Hafter 1999), while in the latter it might reflect introduction of binaural effects of the kind responsible for binaural masking level differences (BMLDs) (Jeffress et al. 1956). Perhaps the most efficient way to inform the subject about the signal is to begin each trial with an iconic cue, i.e., one that matches the signal in every respect except level. Typically, the level of this cue is set high enough to make it clearly audible but not so much as to prevent the qualitative impression that both cue and signal are heard as tones in noise. A special advantage of this kind of cueing is that it can provide control data for quantification of the effects of uncertainty when signals are drawn at random from a range of frequencies on a trial-by-trial basis.

## *2.2 Detection with Frequency Uncertainty*

Obviously, focusing attention on a single filter is a weak strategy for detecting a tonal signal when there is uncertainty about its frequency. Traditionally, this has been studied by choosing the signal on each trial at random from a set of *M* frequencies. An assumption of independence between maskers at the outputs of the "auditory filters" centered on the *M* tones has led to the term M-orthogonal bands (MOB) to describe models couched in signal detection theory (SDT) (Green and Swets 1966). When *M* is relatively small, it is assumed that the subjects are able to monitor the appropriate bands through repeated testing with feedback or through trial-by-trial iconic cues. In a more complex situation from the subject's perspective, effects of uncertainty have been tested by presenting signals drawn completely at random from a wide range of frequencies. In this case, an MOB model predicts that detection must fall with increasing *M* due to the increased probability of a false alarm produced by noise alone in one of the nonsignal filters, while views based on issues of shared attention point to such factors as inaccuracy in choosing which filters to monitor as well as higher demands on a limited attentional resource. These will be discussed later, in Section 3.1.

Green (1960) used SDT to examine the MOB model through comparisons between the effects of uncertainty on human behavior and that of a hypothetical ideal observer whose knowledge of the signal is exact. For the "ideal," increasing *M* has three effects on the psychometric functions. These functions, which relate performance—in units of the percentage of correct decisions in a two-alternative forced-choice (2AFC) task—to signal level are depicted in Figure 5.1. First is a shift to the right, indicative of the need for higher signal levels to maintain constant performance with increased uncertainty. The second is an increase in the slopes of the functions, making the rise in performance from chance to perfect happen over a smaller change of level. The third is a deceleration of the other two effects, with each successive increment in *M* having less of an effect than the one before. In comparison to the ideal observer with $M = 1$, the

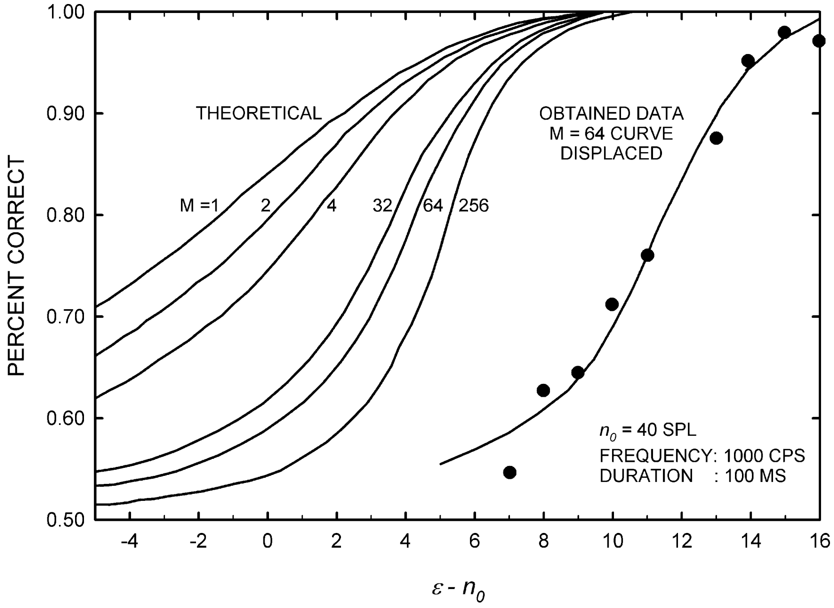


Figure 5.1. Predictions of an M-orthogonal band model of the effects of stimulus uncertainty on signal detection for an ideal energy detector with signal known exactly. (Reprinted with permission from Figure 6 in Green 1960).

empirically measured function from human subjects listening for only a single signal, that is, without frequency uncertainty, showed both the rightward shift indicative of reduced performance and a steeper slope. Rather, the human data closely resemble the ideal observer with an *M* of 64. Green (1961) argued that this steeper slope explains why increasing *M* from 1 to 51 (with 51 possible signal frequencies spaced evenly over a range of 3000 Hz) produced a rise in threshold of only 3 dB compared to the 6 dB change expected for the ideal observer. From these results, Green (1960, 1961) concluded that listeners, even in the simplest of cases, have a high degree of signal uncertainty about such basic parameters as frequency, phase, duration, and time of occurrence. Lumping uncertainty along multiple parameters produces a single variable that we call "unavoidable uncertainty" (UU). From this perspective, when the experimenter varies one signal parameter from 1 to *M* and holds all other parameters constant, uncertainty for the human observer is best described as a rise in uncertainty from UU to *M* + UU. Green (1960) applies this idea to a feature other than frequency, noting that variation of the moment of the signal's presentation over a range of 8 s produced a rise in threshold of less than 2 dB (Egan et al. 1961). Still more support for the idea that human performance is plagued by uncertainty along multiple dimensions is beautifully seen in classic papers by Jeffress (1964, 1970), who showed that receiver operating characteristics (ROCs) based on human performance in a tone-detection task fitted well to comparable ROCs for an ideal observer who has no knowledge of the signal's phase. In summary, this view, based on the MOB model of SDT, sees the human subject as an optimal detector who must rely on noisy data for making judgments. While this seems true, it leaves open questions about top-down processes and questions about how shared attention might interact with stimulus uncertainty. This will be discussed more fully in Section 3.2.

## 2.3 Use of a Probe-Signal Method to Measure the Shapes of the Listening Bands

While the critical ratio offers a way for getting at the width of a listening band employed in wideband noise, it is limited to a single feature of that band, revealing nothing about the structure of the filter. In order to address that problem, Greenberg and Larkin (1968) described a probe-signal method designed to provide a direct description of both the width and shape of the listening band. Their subjects were trained to detect 1000-Hz signals presented in wideband noise. During probe-signal conditions, the subject's expectancy of 1000 Hz was maintained by using that frequency on the majority of trials, while for the remaining trials, frequencies of the "probes" were chosen from a set symmetrically spaced around 1000 Hz. Detection probabilities were lower with probes than with expected signals by an amount that grew with their distances from 1000 Hz. The explanation given for this result was that subjects had responded to stimuli within an internal filter centered on 1000 Hz, thus reducing the levels of probes via attenuation by the skirts of that filter. In support of this proposal, the estimated width of the internal filter was similar to that of the "critical band" measured by other means, a result later confirmed by Dai et al. (1991), who showed that bandwidths obtained with probe signals from 250 to 4000 Hz resembled those found with notched-noise masking (Patterson and Moore 1986). Obviously, listening for an expected frequency does not make the rest of the spectrum inaudible. In a single-interval (yes/no) task, Dai et al. (1991) inserted some probes well outside the "auditory filter" of the expected frequency. Presenting these tones at several levels allowed them to plot psychometric functions that then were used to evaluate the salience of each probe. Results showed a maximum attenuation of only 7 dB for all probes, regardless of their distance from expectation.

Because the most intense probes were higher in level than the expected signals, the authors suggested that subjects might have changed their response strategies, accepting even widely divergent frequencies as valid signals. This addresses a potential criticism of the probe-signal method first raised in Greenberg and Larkin (1968) and referred to in Scharf et al. (1987) as a "heard but not heeded" strategy, which posits that filter-like results could also appear in the normal probe method if subjects chose not to respond to sounds that were not like the expected signal. In order to address this, Scharf et al. (1987) included a condition in which trial-by-trial feedback was provided so as to encourage the subject to respond to probes that differed in perceived quality from the expected signal. When this, as well as other conditions intended to test the idea, had little effect, Scharf et al. (1987) concluded that although the “heard but not heeded” hypothesis might have had some effect on results from the probe-signal method, the width of the listening band was based primarily on sensory filtering. This does not, however, preclude individual differences among listeners based on where they place their attentional filters or how they interpret sounds that differ from the expected signals in qualitative as well as quantitative ways. In this regard, Penner (1972) varied payoffs in a probe-signal method and showed that different subjects used different subjective strategies. This led some to act as if listening through narrower, auditory-filter-like bandwidths and others to show wider filters that were more inclusive of distant probes.

### *2.4 Probe-Signal Measures of Selectivity in Domains Other Than Frequency*

The probe-signal method has been used to study other situations in which dimensions might be represented by their own internal filters. For example, Wright and Dai (1994) studied listening bands found with off-frequency probes using signals that could have one of two durations: 5 or 295 ms. In mixed conditions, where the durations of the expected signals and probes did not always match, probes were more poorly detected if the durations of the probe and signal were different, a result interpreted as an indication of attention to distinct locations in the time–frequency plane. In a follow-up experiment, Dai and Wright (1995) looked for evidence of attention to specific filters in the signal-duration domain by measuring performance in a probe task in which the expected signals and probes differed only in their durations. In a condition in which the duration was 4, 7, 24, 86, 161, or 299 ms chosen at random, performance was only slightly less than when each was tested alone, suggesting little effect of duration uncertainty. Then, in separate blocks, with expected signals whose durations were either 4 or 299 ms, the durations of occasional probes were 7, 24, 86, 161 ms. There, seemingly in keeping with the idea of tuning in the duration domain, detection of probes declined to chance as their durations differed from expectation. While the first result shows that subjects could monitor multiple durations with ease, the second seems to demonstrate focus on a specific listening band defined by duration. Some element of the seeming focus on an expected duration may stem from a “heard-but-not-heeded” strategy (Scharf et al. 1987), where all signals are heard but some are ignored because they do not fit the description of an expected signal.

## 3. Attention and Effort

The National Aeronautics and Space Agency (NASA) once asked this lab why trained airline pilots using one of their new flight simulators were crashing on landing at an uncharacteristically high rate. In response, we proposed that landing an airplane requires repeated answers to the yes/no question, “Is it safe to continue?" Based on a traditional SDT perspective, we cited Norman and Bobrow's (1975) argument that $d'$ in basic signal detection is "data-limited," that is, affected by the S/N but not by attention, and argued it was probably a more relaxed response criterion ($\beta$) in the simulator that led to more false alarms, i.e., crashes. When asked by NASA to prove this, we proposed a task in which subjects would be paid either by the hour or by a pay-for-performance scheme meant to be more like real flying, with a false alarm on a rare percentage (2%) of the noise-only trials producing a "crash," sending the subjects home without pay. Subjects in this condition reported heightened attention and, not surprising in the light of Kahneman's (1973) discussion of the relation between attention and effort, a higher level of stress. In order to increase the overall cognitive load of the task, the signal's frequency was varied from trial to trial, drawn at random from a wide range of possibilities. For maximum uncertainty, these tones were presented without cues; for minimum uncertainty; each was preceded by an iconic cue.

Findings reported to NASA (Hafter and Kaplan 1976) showed that the nonstressful (hourly-pay) condition produced a typical uncertainty effect, that is, a difference between thresholds, with and without cues, of about 3 dB. Furthermore, with uncertainty held to a minimum, varying the payoff from easy to stressful had no effect, in accord with Norman and Bobrow (1975). The most interesting result was an interaction between uncertainty and payoff, with the risky scheme reducing the effect of uncertainty by half. Evidence that payoff could improve performance led us to postulate that the widths of the effective listening bands were subject to cognitive processes, with a high cost of shared attention produced associated with uncertainty being somewhat relieved by subjects using fewer, albeit wider, filters. From this perspective, the threat of a potential crash led observers to attend more closely, responding to more, albeit narrower, effective listening bands.

## *3.1 Effects of Signal Uncertainty on the Bandwidths of the Effective Filters*

The idea that frequency uncertainty might produce changes in the listening bands is not new, with suggestions varying from a single auditory filter switched to each possible frequency in accord with its probability to the proposition of a single wideband filter that encompasses all possibilities (Swets 1984). However, all such models predict large losses in detectability when the signal is drawn from a large range of possible frequencies, and that is simply not the case. What is more, the idea of locally wider bands seems antithetical to the traditional view that auditory filters are immutably constrained by the mechanics of the cochlea. Scharf and his colleagues (e.g., Scharf et al. 1997) have argued for top-down control of the peripheral filters through efferent innervation of the cochlea by the olivocochlear bundle (OCB). This was based on observations of wider bands found using a probe-signal method in patients whose OCBs had been severed during surgery. In rebuttal to the view of top-down control, however, Ison et al.

(2002) cite reduced OCB function due to aging when noting that probe-signal measures showed only small differences in width between young and old patients. In order to address the idea in Kaplan and Hafter (1976) that uncertainty and, presumably, the increased cost of shared attention had widened the effective bandwidths in Hafter and Kaplan (1976), Schlauch and Hafter (1991) devised a means for use of the probe-signal method to examine the filters as a function of a controlled amount of uncertainty in a way that would be unbiased by the unavoidable uncertainty discussed earlier in conjunction with the M-orthogonal band model (Green 1960).

The probe-signal method is based on a primary assumption that the subject responds only to sounds within the filter centered on the expected frequency. In this case, signals to be detected in wideband noise would be drawn at random from the range 600–3750 Hz, but uncertainty would be erased by beginning each trial with a clearly audible cue that told the subject what frequency to expect. Within-subject comparisons showed that performance with these iconic cues was as good as that in a long block that presented only a single frequency. Expectancy was established by presenting iconic cues (with cue-to-signal ratios $f_c/f_s$ of 1.00) on 76% of the trials. However, in the remaining 24% of the trials, probe signals differed from expectancy values by small distances. In line with the quasi-logarithmic distribution of frequency in the cochlea (e.g., Greenwood 1961; Moore and Glasberg 1983), it was assumed that auditory filters are well characterized by a single quality factor, Q (the ratio of a filter's center frequency to its bandwidth). For analysis of the hypothetical constant-Q filter, probes were set to one of four log-distances from expected frequencies, with $f_c/f_s$ of 0.95, 0.975, 1.025, or 1.05, and data were averaged for each value of $f_c/f_s$, regardless of $f_c$. Preliminary measures of performance taken at several points across the range of possibilities were used to derive a function describing the signal levels needed for equal detectability. Once the experiment began, all signals were set accordingly. Performance was measured as a percentage of correct responses [P(C)] in a 2AFC task.

Ordinarily, filters are plotted as decibels of loss relative to a value of zero dB assigned to their center frequencies. In order to convert P(C) into dB, psychometric functions were derived from additional tests conducted with several overall signal levels. A filter constructed in this way is illustrated in the left panel of Figure 5.2. Data points are averages from three subjects, and the fitted line is a rounded exponential (ROEX) model of the auditory filter (Patterson and Nimmo-Smith 1980) plotted in terms of a single bandwidth parameter, $p$, as in Patterson and Moore (1986). The close fit between this derived filter and that in Patterson and Moore (1986) in both shape and bandwidth suggests that subjects here attended to single auditory filters at the cued locations, and were unaffected when the signal was roved from trial to trial.

In order to study the effects of uncertainty on bandwidth, Schlauch and Hafter (1991) planned to use the same method while increasing the number of bands that the subject must monitor on each trial. For this, they eschewed comparisons to the ideal observer of SDT as in Green 1960, and chose instead to define the

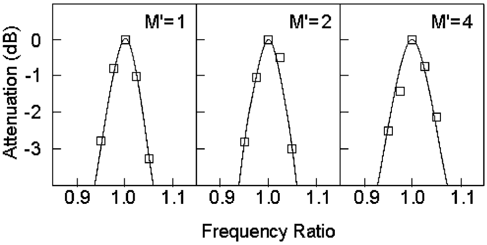


FIGURE 5.2. Listening bands derived for using a probe-signal method for detection of signal frequencies that varied the range 600–3750 Hz (adapted from Schlauch and Hafter 1991). Each trial began with a cue containing $M'$=1, 2, or 4 tones, one of which defined the expected frequency of the signal. The abscissa is plotted in terms of the ratio of cued frequency to probe frequency ($f_c/f_s$). The ordinate shows the loss in effective level of probe signals ($f_c/f_s \neq 1.00$) in dB relative to that found with the signals at the expected value ($f_c/f_s = 1$). These were found using separately obtained psychometric functions to convert performance in P(C) effective signal levels in dB. The fitted curves are ROEX filters (Patterson and Moore 1986) (see text).

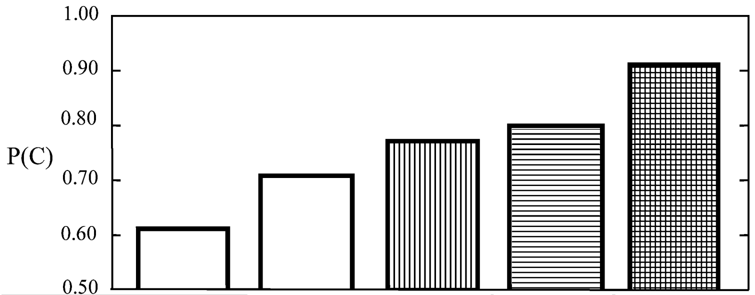
case of minimal uncertainty in terms of the subject's own performance. This is described in the leftmost panel of the figure by the label $M' = 1$ to indicate that these data represent the best that a subject could do when required to monitor only a single band on each trial. Uncertainty was defined in terms of the number of tones in each cue, be they one, two or four tones ($M' = 1$, 2, or 4). In all cases, the $M'$ tones were chosen at random from the range of possible frequencies. While only one of the cue tones on a trial matched the signal, the subject would have to monitor filters at all of their frequencies. Results from $M' = 2$ and 4 are plotted in the center and rightmost panels of Figure 5.2, again normalized in order to set performance with a ratio $f_c/f_s = 1.00$ to zero dB. While the fitted curves are ROEX filters (Patterson and Moore 1986), there is a small but consistent increase in bandwidth parameter with increasing $M'$, lending support for the view that listening bands, as measured in auditory masking, are labile in ways that allow them to be affected by attentional factors.

## *3.2 Relative Cueing for Detection at Emergent Levels of Processing*

Just as iconic cues can alleviate the loss of detectability due to frequency uncertainty, so too can a variety of other cues that bear a more distant relation to the signal. Although these are typically less effective than iconic cues, we have seen improvement in the detection of randomly chosen tones cued by tones related to the signal by a musical interval (Hafter et al. 1993), by a five-tone harmonic sequence for which the signal is the missing fundamental (Hafter and Schlauch 1989), and with musically trained subjects who have absolute pitch and are cued with a visual description of the signal on a musical score (Plamondon and Hafter 1990). From this, it would seem that successful cueing requires only that the cue and signal sound alike in the "mind's ear." However, an interesting result brought this into question. Whereas a harmonic sequence helped a subject listen for its missing fundamental, the reverse was not true. That is, when the signal to be detected was a randomly chosen harmonic sequence, cueing with its missing fundamental did not improve detection. A potential explanation for this begins with the suggestion that the harmonic sequences were detected on the basis of their emergent property: their complex pitch. The hierarchy of processing in the auditory nervous system means that signals are represented in multiple sites along the auditory pathway. If one assumes that attention can be focused on any level of processing, it follows that while detection of a five-tone complex might be based on activity in five distinct locations in a neural representation of frequency, it could also reflect activity in a single location in a representation of complex pitch. As discussed in Hafter and Saberi (2001), for optimal performance, attention should be directed toward the level with the best signal-to-noise ratio (S/N) and, in the case of a complex signal, this is on the complex feature rather than the primitives. This can be understood in terms of the probability of false positives in detecting a five-tone complex. If done on the basis of individual frequencies, a false alarm should happen in response to multiple peaks in the spectrum of noise-alone trials, regardless of their frequencies. Conversely, for detections based on complex pitch, a false alarm should happen only if peaks in the noise spectrum happen to be related by a common fundamental.

Why, then, the asymmetry in cueing with complexes and their fundamentals? An alternative to the simple sounds-like hypothesis says that a successful cue specifies a unique location at the level of processing where detection takes place. From this perspective, a complex pitch would be able to specify the location of its fundamental in a representation organized by frequency, but because a pure tone does not belong to a single harmonic complex, a single frequency cannot specify a unique location in a representation organized by complex pitch.

In order to test this hypothesis, Hafter and Saberi (2001) compared performance from five conditions in which stimuli might be cued and detected at more than one level of processing. In all cases, signals and cues were made up of three tones and presented in continuous background noise. Individual tones were preset to be equally detectable in accord with the relation between thresholds and frequency. Specifics of the five conditions as well as results from the 2AFC detection task are described in Figure 5.3. In Condition 1, signals drawn at random from the frequency range 400–4725 Hz were presented without cues. The average level of these tones was set to produce extremely low performance [P(C) = 0.60] in order to leave room for improvement in other conditions. This level was then used throughout the experiment. Signals in Condition 2 were missing-fundamental harmonic complexes. Each was created by first selecting a frequency designated as a fundamental ($f_0$) from the range 200–675 Hz. Its next six harmonics ($f_1$ to $f_6$) were computed, and three of them were chosen at random to be a signal. For example, if the randomly chosen $f_0$ was 310 Hz and the harmonics chosen for the signal were $f_1$, $f_4$, and $f_6$, the signal would consist



| Condition | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 3-tone signals | Random | Harmonic e.g. $f_2 f_4 f_5$ | Random | Harmonic e.g. $f_1 f_3 f_5$ | Harmonic e.g. $f_3 f_4 f_6$ |
| 3-tone cues | | | Same as signals | Other 3 $f_2 f_4 f_6$ | Same as signals |
| Level of processing | Frequency | Pitch | Frequency | Pitch | Pitch |

FIGURE 5.3. Results (adapted from Hafter and Saberi 2001) demonstrating cueing and signal detection based on two different levels of processing, one tied to analysis of individual frequencies and one tied to the emergent property complex pitch (see text). Signals and cues in all five conditions were three-tone complexes.

of 620, 1550, and 2170 Hz. These were also without cues. Although levels here were the same as those in Condition 1, improved performance [P(C) = 0.70)] confirms the prediction that detection should be better if based on a complex pitch than if on a comparable set of independent frequencies. In cued conditions, 3, 4, and 5, cues were set 7 dB above the signals. Signals in Condition 3 were random tones selected as in Condition 1. Cues matched the signals in frequency, thus informing the subject about where to listen in the frequency domain. Comparing results [P(C) = 0.76] to those in Condition 1 shows that cues ameliorated uncertainty about frequencies in the signals. Signals in Condition 4 were selected in the same way as in Condition 2. However, each cue was chosen to match the signal in the complex pitch domain without highlighting its frequencies. For this, after three of the six harmonics had been chosen to be a signal, the remaining three were used as the cue. In terms of the example above for Condition 2, if $f_0$ was 310 Hz and $f_1$, $f_4$, and $f_6$ made up the signal, the cue would be $f_2$, $f_3$, and $f_5$, or 930, 1240, and 1860 Hz. Comparing results [P(C) = 0.79] to those in Condition 2 shows that the cues ameliorated uncertainty about the complex pitch of the signal. Finally, Condition 5 used cues to inform the subject both about frequencies in a signal and its complex pitch. For this, signals were harmonic complexes chosen as in Conditions 2 and 4, but frequencies in each cue were identical to those in the signal. Comparisons of these results [P(C) = 0.91] to those in Conditions 3 and 4 shows the added effectiveness of cueing both signal features frequency and complex pitch. A prediction for the optimal summation of information in the two domains obtained by summing the two $(d')^2$ values from Conditions 3 and 4 finds that performance in Condition 5 was slightly higher than the predicted value, perhaps indicating a form of useful crosstalk whereby cues in one dimension enhanced the efficacy of cues in the other. In a more general sense, these data strongly support the idea that detection of a complex signal can be based on multiple dimensions and that attention can be focused on these dimensions through cues that share a unique level of processing with the signals.

## 4. Expectancy and the Analysis of Clearly Audible Signals

To this point, we have talked about how cueing can improve detection by indicating to the listener where to expect a weak signal along some dimension. Related effects with suprathreshold stimuli show that cues can also change the way that an audible target is perceived. This is especially obvious with running speech, where syntax and semantics affect how speechlike sounds are heard. The focus here will be on nonspeech cueing of audible signals based on connections between cues and signals that range from simple relations such as harmony and interstimulus intervals to more-complex patterns established by presenting stimuli in an auditory stream.

### *4.1 Attention Focused by Musical Expectancy*

Attention to sequential information is especially important in music, where the basic nature of the stimulus is represented in the relation between auditory events over time. From this perspective, a musical context can act as a cue to establish expectations of future events in the stream. Many have considered expectancy to be a major feature of music, particularly Meyer (1956), who has postulated that the systematic violation of expectation is a primary factor in the elicitation of emotion by music. Much of the experimental work on musical expectation has used a variant of the probe-tone paradigm of Krumhansl (e.g., Krumhansl and Kessler 1982), which presents a melodic scale followed by a probe tone that is rated by the listener for its goodness of fit to the preceding notes. Ratings are highest for probes that match the melody in tonal and harmonic contexts. For instance, if the melody is in C major, the most highly rated pitch class is C, followed by G, E, and F. Similarly, probe-tone profiles have been shown to reflect the statistics underlying a musical composition, as expectancies are established by melodies composed in a single key (Krumhansl 1990). Reaction time (RT) has also been employed in the study of musical expectation. As an example, Bharucha and Stoeckig (1986) presented pairs of chords and asked subjects to say whether the second chord was consonant or dissonant. Results showed that if the second chord was expected (based on harmonic relations to the first), RTs were faster for consonance, but if the second chord was unexpected, RTs were faster for dissonance. Interestingly, while musically trained subjects were faster on average, the main effect held for those without musical training. It is tempting to speculate on whether musical similarity ratings, especially those tied to harmonicity, derive from fundamental auditory processes as might be predicted by models of frequency discrimination and pitch, or are reflective of musical experience. The learning hypothesis gains credence from the observation that the modern musical interval of the fifth is different from that used in sixteenth-century music but alas, we have no experimental data on similarity ratings from the sixtenth century. As always, the likely answer is that both nature and nurture are probably involved.

## *4.2 Attention Focused Through Internal Oscillations Entrained to Temporal Sequences*

Another auditory factor implicated in sequential cueing is the timing of events or rhythm. As with other factors such as pitch contour, regularity of timing can cause the listener to look ahead, prescribing the appropriate moment for evaluation of an upcoming target. Evidence for this kind of entrainment to rhythm is found in the work of Jones and her colleagues (e.g., Jones and Boltz 1989; Large and Jones 1999; Jones et al. 2002), who postulate that consistent timing in a sequence of events produces an anticipatory attentional focus based on the temporal structure of the sequence. Support for this has come from a paradigm in which the subject hears a sequence of tones of different frequencies played at regular intervals. The sequence begins with a reference tone and ends with a target tone that follows the penultimate tone by a variable inter-onset interval (IOI). The listener's task is to judge whether the pitch of the target is the same or different from that of the reference. For IOIs of up to 1200 ms, Jones et al. (2002) found that performance on the discrimination task was maximal when the IOI matched expectations established by the rhythm, but fell off as a function of the difference between the actual IOI and expectancy. As shown in Figure 5.4 (from Jones et al. 2002), this fits with the notion of attentional filters discussed above, implying that regularity in a sequence can be used to select a temporal filter that focuses attention at a specific time. In keeping with that interpretation, the width of the filter was at a minimum when tones in the sequence were presented with a regular rhythm, but grew wider for cases in which the context was less regular. Comparing these results with experiments discussed above, in which subjects listened for a tone at a cued frequency, one might say that just as the earlier study showed sensitivity to a change in level at a specific place in acoustic frequency, Jones et al. (2002) showed sensitivity to a change in frequency at a specific instant in time. These kinds of multifilter interactions, often with separate dimensions examined in tandem, possibly represent a significant part of analyses making up complex perception. A model for how we focus attention in time is proposed by Large and Jones (1999), who posit that the allocation of attention is controlled by a set of nonlinear internal oscillators that can entrain to events in the acoustic stream while tracking complex rhythms.

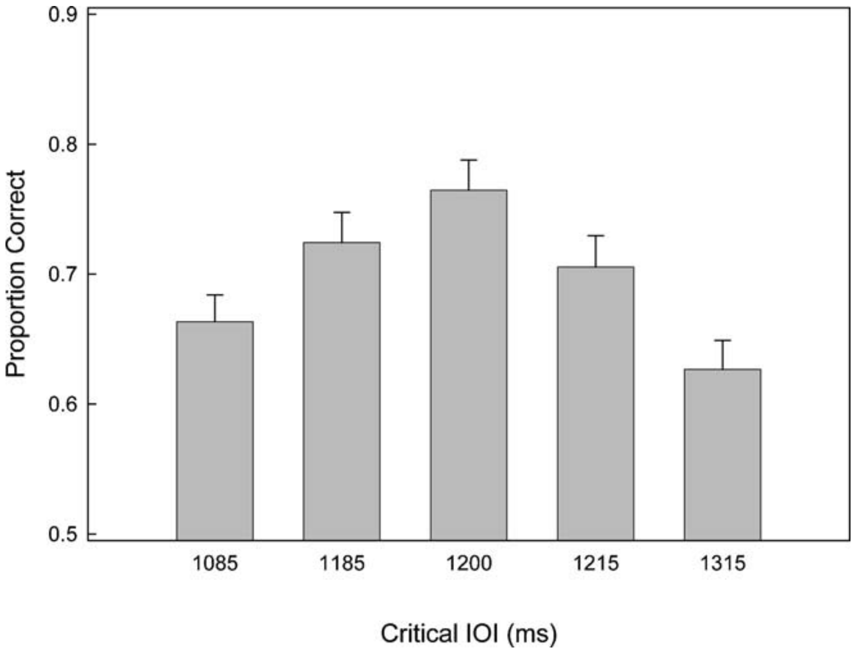


FIGURE 5.4. Results adapted from Jones et al. (2002) in which subjects compared the frequency of the first tone in a rhythmic sequence to the last. They suggest that stimulus regularity can act as a cue, focusing attention on the moment when it is most needed. In this way, the peaked function can be thought of as a kind of filtering in the time domain that diminishes processing before and after the expected time (see text).

## *4.3 Segregation into Multiple Auditory Streams*

Most auditory communication relies on information carried in acoustic sequences. It is impossible to reference all of the important work on streaming by Bregman and his colleagues, but for a remarkable compendium of knowledge about streaming, how it works and how it interacts with other features of audition, the book by Bregman (1990) is highly recommended. From the perspective of auditory attention and selective filtering, one can argue that the pattern of acoustical features in a sequence establishes expectancies for higher-order structured relations such as those found in melodies.

A major approach to these issues has been through examination of stream segregation, whereby a sequence may be perceived as two streams that are essentially coexistent in time. An example from music in which a sequence is not parsed into separate streams is called hocketing. It occurs when a melodic line carried by interleaved sequences from different instruments or voices is heard as a single stream. However, the long history of polyphonic music shows that with a greater separation of the notes in a sequence, the percept can be one of two separate streams. Composers from Bach to Moby have utilized this to play separate melodies on alternating notes, even when both are produced by a single-voiced instrument such as the recorder. In the laboratory, one way to see whether a mixture of alternating sound sequences is segregated is simply to ask subjects whether they hear one stream or two. When the sequence is a simple alternation between tones, A and B, segregation depends on the physical difference between A and B as well as the speed of presentation, with tones that are more similar typically reported as a single stream (Bregman 1990). A popular approach, suggested by van Noorden (1975), is to present a sequence of alternating tones whose perceived rhythm is ambiguous, depending upon how the elements are grouped. An example of this paradigm is illustrated in Figure 5.5, where two frequencies, A and B, are presented in ABA triplets. When the frequency separation between alternating tones is small, or when the tempo is not too fast, the listener reports hearing a single stream that resembles the "galloping" rhythm of a horse. Conversely, when the separation is large (typically three semitones or more) (Bregman 1990) or the sequence is played at a brisk tempo, the percept changes to that of two simultaneously occurring rhythms corresponding to the separate A and B patterns in a way that has been likened to Morse code. Thus, Carlyon and his colleagues (e.g., Carlyon et al. 2003) have found it convenient to instruct subjects to refer to the two kinds of percepts with the terms "Horse" and "Morse." A key point in stream segregation

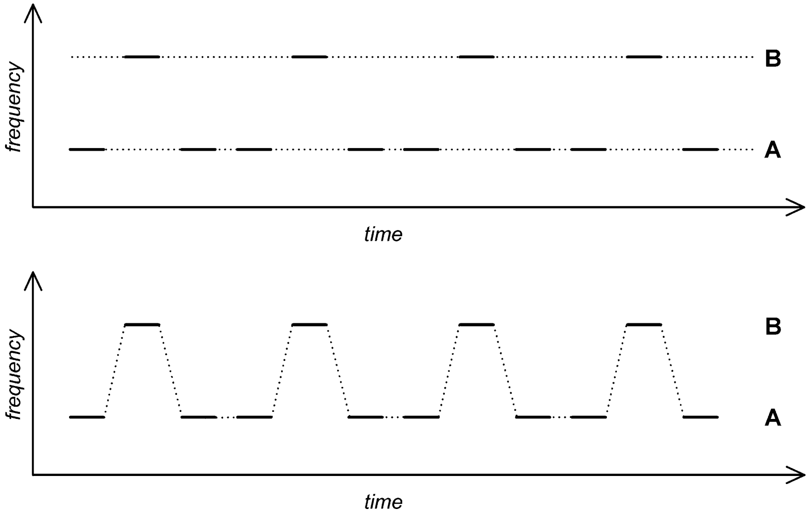


FIGURE 5.5. Schematic description of a common paradigm in which two auditory components alternate in triads. The ordinate plots frequency, and the abscissa, time. Two frequencies labeled A and B are represented by the darker bars (see text). Lighter lines in the two panels portray two different ways in which the stimuli might be perceptually grouped by the listener. The lower panel is heard as a galloping sound produced by grouping the individual triads; the upper panel is heard as two pulsing streams at the two frequencies.

is that while listeners can switch the focus of attention to either percept at will, they generally do not report hearing both at the same time.

A longstanding discussion about the role of attention in auditory grouping concerns whether it is preattentive or is representative of a top-down, more schematic analysis (Bregman 1990). An interesting feature of stream segregation is that it is not instantaneous, but rather builds up over time, sometimes not reaching a peak until many seconds after the onset of the sequence. On the grounds that segregation is a form of grouping, Carlyon and his associates have used the buildup to examine the role of attention when the subject hears the ambiguous stream while responding in a second, independent task. A study by Carlyon et al. (2001) presented 21 s of ABA repetitions, like those shown in Figure 5.5, to the left ear of a subject who used a pair of buttons to report whether the perception was of one stream ("Horse") or two streams ("Morse"). This generally took a few seconds. In a potentially competing task, the 21-s period also began with a presentation to subject's right ear of 10 s of 400-ms bursts of noise whose levels rose quickly to a base amplitude and then, more slowly, either increased or decreased. When the subject was told to ignore the noises, the switch from "Horse" to "Morse" was at about the same time as when sounds were in the left ear alone. However, when the instructions were to spend the first 10 s responding to noises in the right ear as either "approaching" or "withdrawing" and then switch to applying streaming instructions to the tones in the left ear, a buildup of segregation like that seen with no distraction began at the end of the tenth second. From this, the authors concluded that attending to the sequence of tones is important for streaming to build up.

In a later study, Carlyon et al. (2003) looked more deeply into the effects of different kinds of secondary tasks. Here, 13 s of ABA sequences were presented for streaming, but during the first 10 s, subjects were instructed to perform one of three additional tasks that were auditory, visual, or purely cognitive. For auditory distraction, subjects counted the number (3, 4, or 5) of tones in the sequence that had been chosen at random to be perturbed by addition of clearly audible 16-Hz amplitude modulation. Simultaneously, they saw a movie displaying a sequence of 125-ms ovals, most of which were filled by solid lines drawn on an angle from side to side, and for visual distraction, counted the number (3, 4, or 5) of ovals with short line segments. In the nonperceptual cognitive task, subjects were instructed to count backward from a random number presented before each the trial. At the end of the 10 s, the display told subjects to respond to the ABA sequence as "Horse" or "Morse." Results showed considerable stream segregation in all conditions, suggesting that segregation could, to some extent, take place while one was attending to another task. However, support for the idea of special demands required by perceptual attention to another sensory domain or to a competing cognitive task was seen in the fact that the number of "Morse" (stream-segregated) responses was least for the condition with mental arithmetic and most when the tokens to be monitored were superimposed on the auditory stream.

As discussed above, the importance of simple auditory dimensions such as frequency in stream segregation has led to discussion of whether it is a process that relies primarily on lower-level attentional or even preattentional processes. However, it is clear that segregation can also occur with streams defined by differences in emergent perceptual properties such as timbre. In this regard, studies of multidimensional scaling in timbre have identified two axes of sound quality defined by spectral and temporal envelope distributions. Using a streaming paradigm like the one described in Figure 5.6, Wessel (1979) showed streaming based on differences of timbre. For the creation of this "Wessel illusion," sequences were made up of repeated sets of three tones, shown in Figure 5.6, played with a constant intertone interval. If the timbres of the tones are identical, the listener hears repeated three-tone sequences that rise in pitch. However, if the alternate tones are set to one of two timbres, say flutes and trumpets, as described by solid and open symbols in the figure, and if the timbres are sufficiently different along a relevant feature such as the spectral centroid or attack time, the stream breaks into two timbrally defined melodies, or *Klangfarbenmelodien.* Now the percept is of two three-tone sequences, with one characterized by the falling sequence of solid symbols and one by the sequence of open symbols.

Another demonstration of higher-order attentional processes in stream segregation can be found in Dowling (1973), who interleaved familiar melodies with sequences of randomly chosen tones. He found the usual result that segregation was easier when the frequency range of the melody and interfering tones did not overlap. However, stream segregation was strengthened by cues that specified in advance the melody to be heard. Dowling et al. (1987) later showed that when the two streams overlapped in frequency, segregation was better if the "on-beat" stream (the one beginning the stimulus) was the target melody. However, when the streams did not overlap in frequency, it made no difference which stream came first.

It is, perhaps, stretching things to say that cueing the positions along an auditory dimension for detection or identification of a weak signal is the same as preparing the listener to respond to qualitative features of clearly audible tones and streams, but a common thread that leads us to think in terms of an umbrella of auditory attention is that detection, identification, and interpretation are all affected by expectancies of the signal to come, be they from preceding stimuli or from long-term memory. Thought of in this way, streaming suggests

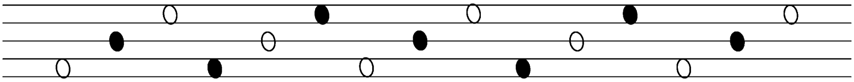


FIGURE 5.6. The "Wessel illusion" (Wessel 1979) in which triplets are presented that when of the same timbre, sound like repeated three-tone melodies that rise in pitch. When alternate tones are set to different timbres as described by the solid and open symbols, and the spectral centroids of the two sounds are sufficiently different, the percept is of two three-tone melodies that fall in pitch (see text).

that features in an ongoing auditory sequence can cue the listener to expect what is next in a way that allows the successive stimulus to be accepted into the stream. This is especially interesting with simultaneous or interleaved sequences, for it describes a role for focused attention in separating auditory objects or events on the basis of shared commonalities. Clearly, Wessel's (1979) use of timbre as the dimension on which the streaming illusion is based goes beyond the suggestion (Hartmann and Johnson 1991) that stream segregation is based simply on channeling in the auditory periphery through differences between tones in fundamental dimensions such as frequency, spatial separation, and duration. In this regard, Moore and Gockel (2002) have examined evidence that suggests that any "sufficiently salient perceptual difference may lead to stream segregation."

## 5. Reflexive Attraction of Attention

To this point we have discussed informational cues that tell the listener about what or where a signal may be. A distinction has been made between these and another kind of cue, such as an unexpected shout, that seems to pull attention to a place in space, regardless of its importance.

### *5.1 Comparisons Between Endogenous and Exogenous Cueing*

Posner (1980) posited a distinction between two kinds of attentional orienting mechanisms, endogenous and exogenous. Up to this point, we have discussed primarily the former. Endogenous cue essentially "push" (Jonides 1981) attention to a place in an auditory dimension where a signal is likely to be, thus telling the subject where to listen. As implied by the word endogenous, the subject must extract relevant information in the cue and make the connection to the expected location of the signal. An endogenous cue need not be identical in form or place to the signal, the requirement being only that it supply useful information that can be deciphered by the listener. That is demonstrated in the case discussed above in which a successful cue tone was related to the signal by a known frequency distance, the musical fifth. For an endogenous cue to be meaningful, the location to which it points must be significantly correlated with the actual position of the signal. Conversely, exogenous cues carry no information about where the signal will be. Rather, they "pull" (Jonides 1981) attention to a location in a reflexive manner without regard for where the signal will actually appear. Such cues can affect performance, positively if they happen to match the signal and negatively if they do not. When a cue of either type correctly defines the location of a signal, it is called "valid"; when it does not, it is called "invalid."

The most common experimental demonstration of the difference between these two kinds of cues is through measurement of reaction times (RTs). They are easily described in the classic visual spatial-cueing paradigm developed by

Posner (1980), whereby the subject fixates on a point between potential signal locations placed, say, on either side of fixation, and the instruction is to respond as quickly as possible to a signal presented at either of these locations. Since an endogenous cue must, by definition, match the signal with a probability greater than chance, subjects who use the cue will show a faster RT on valid trials and a slower RT on invalid trials. In this situation, purely endogenous cues might be arrows at the point of fixation pointing to the left or right, or a written script on the screen with the words "left" or "right," or even an auditory cue with the subject trained to expect a signal on the left after a high pitch and on the right after a low pitch. Exogenous cueing needs no training, acting as it does at a more primitive level in much the way that an unexpected shout might draw one's attention without consideration for what was said. In the visual task described above, a successful exogenous cue might be a presignal flash presented at random at one of the two potential signal locations. Although totally unrelated to where the signal will be, a flash that happens, by chance, to be valid will speed RT, and one that is invalid will slow it. Because an endogenous cue relies on the use of information, it is thought to elicit top-down control, while the more primitive response to an exogenous cue is thought to be based on bottom-up processes. The time between the cue and signal, often called the stimulus-onset asynchrony (SOA), is critical for RT, in that the reflexive effect of an exogenous cue is gone after a relatively short SOA, i.e., generally less than a second, while the effect of an endogenous cue can last much longer, covering SOAs on the order of seconds.

In the auditory modality, the first conclusive evidence of a difference between endogenous and exogenous cueing was observed by Spence and Driver (1994), who used RT to measure auditory discrimination in a spatial cueing task. For these experiments, several speakers were arranged spatially around the listener. Each trial consisted of two sounds—a cue and a target—separated by a variable SOA. In the endogenous condition, subjects were told that the location of the cue predicted the location of the target in 75% of the trials, while in the exogenous condition, subjects were told to ignore the cue because if offered no information about the target. The cue was a pure tone played from one of the speakers; the target, a tone or burst of noise. The task was either to identify the spatial location of the target (front/back or up/down) or, with tonal signals, to identify its pitch (high/low). Effective cueing was defined as RTs that were faster with valid cues than with invalid cues. Effects with exogenous cues were small, short-lived, and observed only for spatial localization, while those with endogenous cues were larger, persisted over longer SOAs and occurred both in frequency discrimination and localization. In subsequent work (Spence and Driver 1998, 2000), in which the focus was on visual and tactile as well as auditory cues, valid exogenous cues again produced faster RTs at short SOAs, unlike endogenous cues, for which the RT advantage lasted over longer delays. A difference, however, was that the effects of both kinds of cueing were observed in discriminations of differences in frequency as well as in spatial location.

Differential roles for endogenous and exogenous cueing in detection have been discussed by Green and McKeown (2001). They note that the use of the probe-signal method to measure the shapes of listening bands probably reflects elements of both kinds of cueing. The idea is that the cue is endogenous in the sense that it predicts where, in frequency, the majority of signals will be. It is exogenous because of a possible reflexive pull of attention to the frequency of the cue, regardless of its predictability. Arguing that filters should be wider without the exogenous component, they note that data from Hafter et al. (1993), where bandwidths found when the frequency of the expected signal was a musical fifth above that of the cue, were wider than those seen when the cue and signal had the same frequency.

Another difference between endogenous and exogenous cueing in masking is also seen in Johnson and Hafter 1980, where the two are put into opposition. In a yes/no detection task, the signal on any trial could be one of two widely spaced frequencies (500 and 1200 Hz). Tonal cues were presented in alternation, with 500 Hz on odd trials and a 1200-Hz cue on even trials. Given that endogenous cueing acts to reduce inaccuracy in the filters to be monitored, one would expect these cues to provide cumulative information over the course of a session, constantly reminding the subject about where to listen. Not surprisingly, on valid trials, this accumulated knowledge plus any exogenous effects of the cues produces a 1.5-dB increase in gain relative to an uncued control condition. Conversely, on invalid trials, one might expect the endogenous benefit, but it would be in conflict with the reflexive pull of the exogenous cue to the wrong filter. This is seen in a loss in performance relative to the control conditions of 1.0 dB.

## *5.2 Inhibition of Return in Exogenous Cueing*

Mondor and Lacey (2001) compared the effects of exogenous auditory cueing on RT in tasks based on discrimination of auditory properties of targets other than their spatial direction. In three separately tested conditions, the subject was asked to make a rapid discrimination between two stimulus values in one of three dimensions: level, frequency, or timbre. Cues in each condition presented one of the same two values used as targets, but this was done randomly in such a way that the cue matched the target in one-half of the trials and not in the other half. Thus, the cues were exogenous, offering no information about the correct response. Targets followed cues by SOAs of 150, 450, or 750 ms. All sounds came from a single speaker, so stimulus direction was not a factor. Of interest was the difference in RT between invalid and valid cues, where exogenous cueing generally works only for short SOAs. The RT difference was positive, with an SOA of 150 ms, indicating of a reflexive pull to the appropriate place in the stimulus dimension being tested. The RT difference was zero when the SOA was 450 ms, indicative of no effect of cueing, and was negative for the SOA of 750 ms. The latter effect, in which responses were actually faster to an invalid cue, has been likened to a kind of late rebound. Generally called inhibition of return (IOR) (Posner and Cohen 1984), it is thought to represent an internal inhibition of responses to the cued location.

In summary, comparisons between endogenous and exogenous cueing in audition show similar results to those found in vision. With endogenous cueing, valid (informative) cues can speed responses for SOAs of several seconds, while invalid (counterinformative) cues tend to slow them down. With exogenous cueing, results are more temporally dependent; valid cues produce faster RTs for short SOAs but slower RTs with longer SOAs, as observed in IOR.

## 6. Auditory Attention and Cognitive Neuroscience

The surging interest in cognitive neuroscience has led to numerous methods for observing neural activity while the brain is engaged in auditory attention. Two of the most common neuroimaging techniques are functional magnetic resonance imaging (fMRI) and positron-emission tomography (PET). Measures of fMRI estimate neural activity through measures of the blood oxygenation level dependent response (BOLD), which represents a coupling between hemodynamic and neural activity. Neural activity in PET relies on radioactive tracers taken up by the most active neurons. In an example of the use of fMRI for the study of auditory streaming, Janata et al. (2002) used polyphonic music in which the subject listened to a duet whose melodies differed in their respective timbres. In one condition, a subject was asked to attend to and track one of the melodies. In separate comparison tasks, subjects either listened to the duets passively or rested in silence. Results showed that active listening produced a greater BOLD response in the superior temporal gyrus (STG) and some frontal and parietal areas including precentral gyrus, supplementary and presupplementary motor areas (SMA), and intraparietal sulcus (IPS). Because the STG is primarily involved in auditory processing (e.g., Zatorre et al. 2002), while the IPS, precentral gyri, SMA, and pre-SMA are implicated in more general working-memory and attentional tasks, results from this experiment are in agreement with other studies using fMRI (Petkov et al. 2004) and PET (e.g., Zatorre et al. 1999) in implicating a frontoparietal network that couples with domain-specific sensory cortices during sustained attention to auditory and musical stimuli.

Perhaps the most developed approach to the study of human neuroscience and auditory attention has been in the domain of electrophysiology using event-related potentials (ERPs), a technique that allows detailed study of the time course of attention. Several components of the ERP have been specifically linked to auditory attention. The earliest known sound-evoked potential is the auditory brainstem response, or ABR, which is a complex of seven individual waveform components starting at around 10 ms after the onset of a brief sound. The ABR complex is thought to be independent of attentional modulation, and is therefore described by Hackley (1993) and Woldorff et al. (1998) as being strongly automatic, differentiating neural processes in the brainstem from the brain's attentional network. The first known cortical ERP shown to be related to hearing is the N1, a negative waveform that is largest at around 100 ms after the onset of any auditory stimulus. Hillyard et al. (1973) found the N1 component to be sensitive to modulation by attention in a study using dichotic listening, where a rapid stream of tone pips was sent randomly to each ear. Listeners were required to perform an attention-demanding fine pitch discrimination task on tones coming into one of their ears. Comparisons between ERPs evoked by pips to the left and right ears showed a larger N1 for stimuli in the attended ear than for those in the unattended ear. Thus, the neural generators of the N1 component were said to be partially automatic, because while sensory stimulation without attention was sufficient for its elicitation, it was sensitive to the enhancement or modulation of attention. Woldorff and Hillyard (1991) later localized the N1 to the primary auditory cortex through the use of magnetoencephalography (MEG).

Another interesting component of the ERP that has been implicated in attention is called the mismatch negativity (MMN). It is elicited in the so-called oddball paradigm, whereby a standard stimulus is presented repeatedly with high probability, but an occasional deviant stimulus is introduced at random locations in the stimulus stream. When comparing ERPs evoked by the standard and deviant stimuli, the MMN is a negative waveform in the deviant response at about 150–200 ms poststimulus. Because this occurs in all senses that have been tested and reflects any stimulus that differs from its surrounding context, Näätänen (1988) claimed that the MMN is independent of attention. However, in a subsequent study using the dichotic oddball paradigm with stimuli presented at fast rates, Woldorff and Hillyard (1991) found that the MMN was larger on the attended side than on the unattended side, suggesting that while mismatch detection, like sensory processing, is partially automatic in its elicitation, it can be modulated by attention. Using MEG, Woldorff et al. (1998) also found attention modulation of a magnetic analogue of the MMN localized in the primary auditory cortex. Based on these findings, Hackley (1993) and Woldorff and Hillyard (1991) proposed that, unlike the case with other evoked potentials such as the auditory brainstem reflexes, the N1 and MMN in early sensory cortices are "partially automatic," such that attention is not required for cortical activity but can strongly enhance or modulate cortical processing. This is supported by Sussman et al. (1998, 1999), who linked the MMN to auditory attention and auditory streaming through tasks in which subjects heard a sequence of tones that alternated between high and low frequencies. If segregated, each of the pitch streams was organized into its own simple melody. On a small number of trials, the order of tones was changed in the low-frequency range in order to test the hypothesis that an MMN would appear only if the two melodies were segregated. As expected, this was true when the tempo was fast, both when the subjects were instructed to attend to the tones and when they were asked to perform the secondary task of reading a book. However, when the tempo was slow, in which case typically there was less streaming, an MMN was seen only when the subjects were instructed to attend to the auditory stimuli.

## 7. Summary

There are many manifestations of top-down control of sensory processing that fall under the great heading of attention. Internal states, emotions, social constraints, deep knowledge, all can affect what we hear and how we interpret it. In this chapter we have concentrated on one specific role of attention, the ability to extract signals from a background. From this perspective, we have looked at some of the putative internal filters that constrain the processing of a signal, especially when there is subjective uncertainty about where the signal may fall along a relevant dimension. This has been done through examination of informational cues that tell the listener where to listen, be it in frequency, at a higher level of processing such as complex pitch, at a specific moment in time marked by temporal rhythms, or at places in pitch prescribed by musical melodies. In addition, we have discussed cases in which the relation between sequential sounds can affect whether the sequence is heard as perceptual whole or divided into separate streams. We have avoided the excellent literature on attention with speech and speechlike sounds as well as "informational masking," because those topics appear elsewhere in this volume. Also absent is mention of a proposed mechanism for attention, though as briefly noted in the final section, there is a rapidly growing movement in human neuroscience to look for the neural networks that produce results seen in the behavior. From our perspective, the proposal is that the kinds of psychoacoustic measures discussed here are of value for understanding the role of attention in the perception of more-complex arrays of sound in the natural world.

*Acknowledgments.* Great thanks are in order for Anne-Marie Bonnel, whose many valuable ideas added immeasurably to this chapter. Also, we thank Brian Moore, whose tough and insightful comments of an earlier draft are greatly appreciated.

## *References*

Bharucha JJ, Stoeckig K (1986) Reaction-time and musical expectancy—priming of chords. J Exp Psychol [Hum Percept Perform] 12:403–410.

Bonnel AM, and Hafter ER (1998) Divided attention between simultaneous auditory and visual signals. Perception and Psychophysics 60(2), 179–190.

Bregman AS (1990) Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: Bradford Books, MIT Press.

Broadbent DE (1958) Perception and Communication. New York: Pergamon Press.

Carlyon RP, Cusack R, Foxton JM, Robertson IH (2001) Effects of attention and unilateral neglect on auditory stream segregation. J Exp Psychol [Hum Percept Perform] 27: 115–127.

Carlyon RP, Plack CJ, Fantini DA, Cusack R (2003) Cross-modal and non-sensory influences on auditory streaming. Perception 32:1393–1402.

Cherry EC (1953) Some experiments on the recognition of speech with one and with two ears. J Acoust Soc Am 24:975–979.

Dai H, Wright BA (1995) Detecting signals of unexpected or uncertain durations. J Acoust Soc Am 98:798–806.

Dai HP, Scharf B, Buus S (1991) Effective attenuation of signals in noise under focused attention. J Acoust Soc Am 89:2837–2842.

Deutsch JA, Deutsch D (1963) Attention: Some theoretical considerations. Psychol Rev 70:80–90.

Dowling WJ (1973) The perception of interleaved melodies. Cogn Psych 5:322–337.

Dowling WJ, Lung KM-T, Herrbold S (1987) Aiming attention in pitch and time in the perception of interleaved melodies. Percept Psychophys 41:642–656.

Egan JP, Greenberg GZ, Schulman AI (1961) Intervals of time uncertainty in auditory detection. J Acoust Soc Am 33:771–778.

Fletcher H (1940) Auditory patterns. Rev Mod Phys 12:47–65.

Green DM (1960) Psychoacoustics and detection theory. J Acoust Soc Am 32:1189–1203.

Green DM (1961) Detection of auditory sinusoids of uncertainty frequency. J Acoust Soc Am 33:897–903.

Green DM, Swets JA (1966) Signal Detection Theory and Psychophysics. New York: John Wiley & Sons.

Green TJ, McKeown JD (2001) Capture of attention in selective frequency listening. J Exp Psychol [Hum Percept Perform] 27:1197–1210.

Greenberg GZ, Larkin WD (1968) Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. J Acoust Soc Am 44:1513–1523.

Greenwood DD (1961) Critical bandwidth and the frequency coordinates of the basilar membrane. J Acoust Soc Am 33:1344–1356.

Gundy RF (1961) Auditory detection of an unspecified signal. J Acoust Soc Am 33: 1008–1012.

Hackley SA (1993) An evaluation of the automaticity of sensory processing using event-related potentials and brainstem reflexes. Psychophysiology 30:415–428.

Hafter ER, Kaplan R (1976) The interaction between motivation and uncertainty as a factor in detection. NASA project report, Ames Research Center, Moffit Field, CA.

Hafter ER, Saberi K (2001) A level of stimulus representation model for auditory detection and attention. J Acoust Soc Am 110:1489–1497.

Hafter ER, Schlauch RS (1989) Factors in detection under uncertainty. J Acoust Soc Am 86:S112.

Hafter ER, Schlauch RS, Tang J (1993) Attending to auditory filters that were not stimulated directly. J Acoust Soc Am 94:743–747.

Hartmann WM (1997) Signals, Sounds, and Sensation. New York: AIP Press.

Hartmann WM, Johnson D (1991) Stream segregation and peripheral channeling. Music Percept 9:155–184.

Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical signs of selective attention in the human brain. Science 182:177–80.

Houtsma AJM (2004) Hawkins and Stevens revisited with insert earphones (L). J Acoust Soc Am 115:967–970.

Ison JR, Virag TM, Allen PD, Hammond GR (2002) The attention filter for tones in noise has the same shape and effective bandwidth in the elderly as it has in young listeners. J Acoust Soc Am 112:238–246.

Janata P, Tillmann B, Bharucha JJ (2002) Listening to polyphonic music recruits domain-general attention and working memory circuits. Cogn Affect Behav Neurosci 2: 121–140.

Jeffress LA (1964) Stimulus-oriented approach to detection. J Acoust Soc Am 36: 766–774.

Jeffress LA (1970) Masking. In: Tobias J (ed) Foundations of Modern Auditory Theory 1. New York: Academic Press, pp. 87–114.

Jeffress LA, Blodgett HC, Sandel TT, Wood CL III (1956) Masking of tonal signals. J Acoust Soc Am 3:416–426.

Johnson DM, Hafter ER (1980) Uncertain-frequency detection——Cueing and condition of observation. Percept Psychophys 28:143–149.

Jones MR, Boltz M (1989) Dynamic attending and responses to time. Psychol Rev 96: 459–491.

Jones MR, Moynihan H, MacKenzie N, Puente J (2002) Temporal aspects of stimulus-driven attending in dynamic arrays. Psychol Sci 13:313–319.

Jonides J (1981) Voluntary versus automatic control over the mind's eye. In: Long J, Baddeley AD (eds) Attention and Performance IX. Hillsdale, NJ: Lawrence Erlbaum.

Kahneman D (1973) Attention and Effort. Upper Saddle River, NJ: Prentice-Hall.

Krumhansl CL (1990) Tonal hierarchies and rare intervals in music cognition. Music Percept 7:309–324.

Krumhansl CL, Kessler EJ (1982) Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. Psychol Rev 89:334–368.

Large EW, Jones MR (1999) The dynamics of attending: How people track time-varying events. Psychol Rev 106:119–159.

Leshowitz B, Wightman FL (1972) On the importance of considering the signal's frequency spectrum: Some comments on Macmillan's Detection and recognition of increments and decrements in auditory intensity experiment. Percept Psychophys 12:209–212.

Macmillan NA (1971) Detection and recognition of increments and decrements in auditory intensity. Perception and Psychophysics, 10:233–238.

Meyer L (1956) Emotion and Meaning in Music. Chicago: University of Chicago Press.

Mondor TA and Lacey TE (2001) Facilitative and inhibitory effects of cuing, sound duration, intensity, and timbre. Percept Psychophys 63:726–736.

Moore BCJ (2003) An Introduction to the Psychology of Hearing, 5th ed. San Diego: Academic Press.

Moore BCJ, Glasberg BR (1983) Suggested formulas for calculating auditory-filter bandwidths and excitation patterns. J Acoust Soc Am 74:750–753.

Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. Acta Acust 88:320–333.

Moray N (1960) Broadbent's filter theory—postulate H and the problem of switching time. Q J Exp Psychol 12:214–220.

Moray N (1970) Attention: Selective Processes in Vision and Hearing. New York: Academic Press,.

Näätänen R. 1988. Implications of ERP data for psychological theories of attention. Biological Psychology, 26(1–3):117–63

Norman DA (1969) Memory while shadowing. Q J Exp Psychol 21:85–93.

Norman DA, Bobrow DG (1975) On data-limited and resource-limited processes. Cogn Psych **7**:44–64.

Patterson RD, Moore BCJ (1986) Auditory filters and excitation patterns as representations of frequency resolution. In: Moore BCJ (ed) Frequency Selectivity in Hearing. New York: Academic Press, pp. 123–177.

Patterson RD, Nimmo-Smith I (1980) Off-frequency listening and auditory-filter asymmetry. J Acoust Soc Am 67:229–245.

Penner MJ (1972) Effects of payoffs and cue tones on detection of sinusoids of uncertain frequency. Percept Psychophys 11:198–202.

Petkov CI, Kang X, Alho K, Bertrand O, Yund EW, and Woods DL (2004) Attentional modulation of human auditory cortex. Nat Neurosci 7:658–663.

Plamondon L, Hafter ER (1990) Selective attention in absolute pitch listeners. J Acoust Soc Am Suppl 1 88:S49.

Posner MI (1980) Orienting of attention. Q J Exp Psychol 32:3–25.

Posner MI, Cohen Y (1984) Components of visual orienting. In: Bouma H, Bouwhuis DG (eds) Attention & Performance X. Cambridge, MA: MIT Press, pp. 531–555.

Scharf B (1970) Critical Bands. In: Tobias J (ed) Foundations of Modern Auditory Theory 1. New York: Academic Press, pp. 159–202.

Scharf B, Quigly S, Aoki C, Peachy N, Reeves A (1987) Focused auditory attention and frequency selectivity. Percept Psychophys 42:215–223.

Scharf B, Magnan J, Chays A (1997) On the role of the olivocochlear bundle in hearing: 16 case studies. Hear Res 103:101–122.

Schlauch RS, Hafter ER (1991) Listening bandwidths and frequency uncertainty in pure-tone signal detection. J Acoust Soc Am 90:1332–1339.

Spence CJ, Driver J (1994) Covert spatial orienting in audition: Exogenous and endogenous mechanisms. J Exp Psychol [Hum Percept Perform] 20:555–574.

Spence C, Driver J (1998) Auditory and audiovisual inhibition of return. Percept Psychophys 60:125–139.

Spence C, Driver J (2000) Attracting attention to the illusory location of a sound: Reflexive crossmodal orienting and ventriloquism. NeuroReport 11:2057–2061.

Sussman E, Ritter W, Vaughan HG Jr (1998) Attention affects the organization of auditory input associated with the mismatch negativity system. Brain Res 789:130–138

Sussman E, Ritter W, Vaughan HG Jr (1999) An investigation of the auditory streaming effect using event-related potentials. Psychophysiology 36:22–34

Swets JA (1984) Mathematical models of attention. In: Parasuraman R, Davies DR (eds) Varieties of Attention. London: Academic Press, pp. 183–242.

Taylor MM, Forbes SM (1969) Monaural detection with contralateral cue (MDCC). I. Better than energy detector performance by human observers. J Acoust Soc Am 46:1519–1526.

Treisman AM (1964) Monitoring and storage of irrelevant messages in selective attention. J Verb Learn Verb Behav 3:449–459.

Treisman AM (1969) Strategies and models of selective attention. Psychol Rev 76: 282–299.

Treisman AM, Gelade G (1980) Feature-integration theory of attention. Cogn Psycho 12:97–136.

van Noorden LPAS (1975) Temporal coherence in the perception of tone sequences. PhD thesis, Eindhoven University of Technology.

Wessel DL (1979) Timbre space as a musical control structure. Comp Music J 3:45–52.

Woldorff MG, Hillyard SA (1991) Modulation of early auditory processing during selective listening to rapidly presented tones. Electroencephalogr Clin Neurophysiol 79:170–191.

Woldorff MG, Hillyard SA, Gallen CC, Hampson SR, Bloom FE (1998) Magnetoencephalographic recordings demonstrate attentional modulation of mismatch-related neural activity in human auditory cortex. Psychophysiology 35:283– 292.

Wright BA, Dai H (1994) Detection of unexpected tones with short and long durations. J Acoust Soc Am 95:931–938.

Yost WA, Penner MJ, Feth LL (1972) Signal detection as a function of contralateral signal-to-noise ratio. J Acoust Soc Am 51:1966–1970.

Zatorre RJ, Mondor TA, Evans AC (1999) Auditory attention to space and frequency activates similar cerebral systems. NeuroImage 10:544–554.

Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: Music and speech. Trends Cogn Sci 6:37–46.