Chapter 6
Spatial Stream Segregation

John C. Middlebrooks

Abstract "Stream segregation" refers to a listener's ability to disentangle inter-leaved sequences of sounds, such as the ability to string together syllables from one talker in the presence of competing talkers. Spatial separation of sound sources is a key factor that enables the task of segregation. Psychophysical tasks that require listeners to integrate sounds across locations demonstrate that listeners can over-come spatial separation of sources, suggesting that space is a relatively weak segregating factor. Contrary to that suggestion tasks that require listeners to isolate a sound sequence within a complex background demonstrate robust benefits of spatial separation of the target from other sources. This chapter reviews psy-chophysical studies that show weak versus strong spatial effects on streaming and shows that the spatial acuity of stream segregation can approach the limits of acuity of spatial hearing. Responses from auditory cortex in anesthetized animals are presented demonstrating that single neurons can exhibit spatial stream segregation by synchronizing selectively to one or the other of two interleaved sound sequences. The results from animals imply that perceptually segregated sound sequences are represented in auditory cortex by discrete mutually synchronized neural populations. Human magneto- and electroencephalographic results then are described showing selective enhancement of cortical responses to attended versus unattended sounds. Available results lead to a picture showing bottom-up segre-gation of sound sources by brainstem mechanisms on the basis of spatial and other cues, followed by top-down selection of particular neural populations that could underlie perceptual auditory objects of attention.

Keywords Auditory cortex · Rhythmic masking release · Scene analysis · Spatial hearing · Spatial streaming · Stream integration · Spatial release from masking

J.C. Middlebrooks (✉)
Department of Otolaryngology, Department of Neurobiology & Behavior,
Department of Cognitive Sciences, Department of Biomedical Engineering,
Center for Hearing Research, University of California, Irvine, CA 92697-5310, USA
e-mail: j.midd@uci.edu

6.1   Introduction

"Stream segregation" refers to a listener's ability to disentangle temporally inter-leaved sequences of sounds from multiple sources. It may be regarded as an ele-ment of auditory scene analysis (Bregman 1990) and/or as a part of the solution to the "cocktail party problem" (Cherry 1953). In speech, normal-hearing listeners do this when they attend to sequences of syllables from one talker in the presence of a crowd of other talkers. In music, a listener can pick out a single musical line from an ensemble of multiple instruments, or a composer can exploit tricks of pitch and rhythm to create from a single instrument the impression of multiple segregated lines. As described in Chap. 2 by Shinn-Cunningham, Best, and Lee, the perceptual correlate of utterances from a specific talker or a musical line or any other distinct sound source can be referred to as an "auditory object" (Woods and Colburn 1992; Griffiths and Warren 2004). Stream segregation is a major element of auditory object formation.

The individual elements of sound sequences from multiple sources might overlap partially or completely in time, or the elements might interleave with no temporal overlap. Temporal and/or spectral overlap of sounds from multiple sources can result in energetic or informational masking, which are the topics of Chaps. 3 (Culling and Stone) and 4 (Kidd and Colburn). Even in the case of sequential interleaving of sound elements, in which there is no temporal overlap, it is a challenge for a listener to construct one or more discrete auditory objects when exposed to multiple competing sequences of sounds.

Sound features that enable stream segregation include differences in fundamental frequencies (corresponding to pitches), spectra (corresponding to timbre), and temporal envelopes, particularly differences in onset times (reviewed by Moore and Gockel 2002, 2012). The present chapter focuses on another key factor in stream segregation, the *spatial* differences among multiple sources. Spatial separation of sound sources has long been appreciated to aid in formation of auditory objects (Cherry 1953). Cherry, for example, wrote that "the voices come from different directions" (p. 976) was a key factor in segregating competing talkers. He simulated "different directions" by presenting two spoken messages dichotically, one to each ear, and noted that recognition of one or the other message improved dramatically compared to a condition in which the messages were presented diotically (i.e., mixed and both presented to both ears) (Cherry 1953). Surprisingly, objective measures of spatial effects on stream segregation have yielded a wide variety of results, ranging from "weak-to-no" effect of space to "robust spatial streaming." Those conflicting results seemingly can be reconciled by considering the require-ments of specific psychophysical tasks, as discussed in Sect. 6.2.

Regardless of the particular sound feature, the brain substrates of stream seg-regation likely involve brainstem and thalamocortical mechanisms for bottom-up formation of auditory objects combined with cortical mechanism for top-down selection among those objects. At least one study has suggested that neuronal stream segregation based on tonal frequencies is accomplished as early in the

auditory pathway as the cochlear nucleus (Pressnitzer et al. 2008). Most other physiological studies, however, have focused on forebrain levels. Correlates of stream segregation based on frequencies of pure tones have been demonstrated in neural recordings in primary auditory cortex (area A1) of macaque monkeys (*Macaca fascicularis*: Fishman et al. 2001; *Macaca mulatta*: Micheyl et al. 2005) and ferrets (*Mustela putorius furo*: Elhilali et al. 2009). In humans, correlates of streaming based on fundamental frequencies or interaural time differences (ITDs) have been demonstrated in nonprimary auditory cortex using event-related potentials, magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) (Snyder and Alain 2007; Schadwinkel and Gutschalk 2010; Carl and Gutschalk 2012). Evidence of cortical streaming of high-level auditory objects, including speech streams, has been observed in humans with MEG techniques (Ding and Simon 2012a, b) and with recordings from the cortical surface (Mesgarani and Chang 2012); these topics are reviewed in Chap. 7 by Simon. Correlates of spatial stream segregation by single neurons in cortical area A1 are demonstrated by the study reviewed in Sect. 6.3. A model resulting from that study posits that spatial stream segregation arises in auditory cortex as a product of brainstem spatial processing that is then sharpened by forward suppression in the thalamocortical projection. Results from other studies suggest that spatially sensitive neurons are ubiquitous in auditory cortex but that various auditory cortical areas differ in their relative contributions to spatial stream segregation and to other aspects of spatial hearing.

6.2   Psychophysics of Spatial Stream Segregation

Psychophysical studies have evaluated the conditions under which interleaved sequences of sounds elicit perceptions either of single integrated streams or of two or more segregated streams. An oft-cited example is the dissertation work by van Noorden (1975). van Noorden presented listeners with sequences of tones, denoted here by A and B, that differed in frequency. When sequence ABA_ABA_ABA… was presented at a slow rate or with closely spaced frequencies, listeners reported hearing a succession of "gallops" consisting of the ABA triplets. At a higher rate or with wider frequency separation, however, two perceptually distinct streams emerged, one consisting of a rapid sequence of the A tones and the other consisting of a slower sequence of the B tones. van Noorden wrote of "fusion" (i.e., integration) of the A and B tones into a single stream of gallops and "fission" (segregation) of two segregated A and B streams.

Psychophysical measures of the importance of spatial cues for stream segregation have yielded quite disparate results depending on the design of the experiment. Studies reviewed in Sect. 6.2.1 have required listeners to *integrate* information across multiple source locations. Listeners' generally good performance in such tasks seems to show that location is a weak segregation cue that can be defeated easily when task performance demands integration. In contrast, studies reviewed in

Sects. 6.2.2 and 6.2.3 required listeners to *segregate* multiple competing sounds. Those studies demonstrate that spatial separations of target and masker(s) are potent cues that a listener may exploit when attempting to segregate a particular target from other distracters, like the task of hearing out a particular talker amid a background of other voices.

6.2.1 Weak Disruption of Stream Integration by Spatial Cues

A number of psychophysical studies have tested the ability of a listener to integrate sequences of sounds that vary in spatial or other parameters. Such studies have been referred to variously as measures of fusion (van Noorden 1975), primitive streaming (Bregman 1990), obligatory or involuntary streaming (Vliegen et al. 1999), or integration (Micheyl and Oxenham 2010). Information needed for performance of integrative streaming tasks is distributed among two or more potentially segregated streams that the listener must fuse in order to make a correct judgment. In integrative tasks, the magnitude of stream segregation can be inferred by the degree to which a putative streaming factor *impairs* task performance by forcing signal components into differing perceptual streams. One commonly used test of stream integration is a so-called temporal asymmetry task. Sequences of sounds differing in spectral or spatial parameters, denoted here as A and B, are presented as sequences of ABA_ABA_…, and listeners are required to detect sequences in which the B sound falls asymmetrically between the two A markers; that is, when the AB time interval differs from the BA interval. Performance on such a task is impaired when the A and B sounds diverge into differing perceptual streams as, for example, when the A and B sounds differ in spectrum (Vliegen et al. 1999) or ear of entry (Boehnke and Phillips 2005); those are conditions in which the A and B sounds are assumed to activate distinct neural populations. van Noorden (1975) noted that it was easy to achieve a subjective experience of fission (i.e., segregation) when sounds were presented to opposite ears.

Surprisingly, differences in spatial cues in the A and B sounds, specifically ITDs or interaural level differences (ILDs), result in little or no impairment of temporal asymmetry detection when the ITDs and ILDs fall within the ranges produced by natural sound sources. Boehnke and Phillips (2005) found no significant effect of ITD on temporal asymmetry detection when A and B noise bursts had ITDs of 600 μs, opposite in sign between A and B; ±600 μs correspond approximately to the ITDs produced by free-field sound sources located to the extreme right and left of the listener (Kuhn 1977; Middlebrooks and Green 1990). Similarly, Füllgrabe and Moore (2012) found that ITDs up to 500 μs in tonal stimuli had only weak effects on temporal asymmetry detection. Both of those groups reported little or no subjective experience of stream segregation based on ITD using a procedure similar to that employed by van Noorden (1975) to evaluate tonal stream segregation.

Boehnke and Phillips (2005) also tested stream segregation based on ILDs, presenting stimuli with 12-dB ILDs differing in sign between the A and B noise bursts (corresponding roughly to free-field sound sources located >30° to the right and left of the frontal midline; Shaw 1974). That condition produced statistically significant but weak disruption of temporal asymmetry detection although there was a clear subjective experience of stream segregation.

Fusion of sounds sharing a common onset can be highly resistant to degradation by conflicting spatial cues. Several groups have tested perception of components of speech sounds presented dichotically. Cutting (1976) constructed two-formant syllables, /ba/ and /ga/, and presented various mismatched pairs of formants to listeners, one formant to each ear. In many instances, listeners presented with a lower formant from /ba/ in one ear and an upper /ga/ formant in the other reported hearing a single fused /da/ sound, even though there were no /da/ components in the stimulus. Broadbent and Ladefoged (1957) constructed stimuli consisting of brief sentences in which odd-numbered formants were presented to one ear and even-numbered formants to the other ear. Nearly all of the listeners experienced fusion of the dichotic stimuli such that they reported hearing only a single voice from a single (midline) location. Hukin and Darwin (1995) used vowel formant boundaries as a measure of listeners' ability to fuse vowel components that differed in spatial cues. Presentation of the 500-Hz component of a vowel at the ear opposite from the other components was equivalent to reducing the level of the 500-Hz component by only 5 dB. Displacement of the 500-Hz component from the other vowel components with 666-µs ITDs differing in sign had an even smaller effect on the formant boundary. In a study using free-field stimuli from multiple loudspeakers, Takanen and colleagues (2013) studied recognition of concurrent vowels, with odd-numbered formants presented from one location and even-numbered formants from another. Vowel recognition by their listeners showed essentially no influence of spatial separation of sources of odd and even formants. All of these split-formant speech tasks show the capacity of common onsets to bind together elements of auditory objects; see Chap. 5 by Elhilali and Elhilali et al. (2009) for related animal physiological studies. Also, fusion was experienced only when the various components shared a common fundamental frequency or when the formants were excited by noise (i.e., were aperiodic; Takanen et al. 2013). Introduction of differences in the fundamental frequencies at the two ears could disrupt fusion (Broadbent and Ladefoged 1957).

In summary, published studies of spatial stream segregation measured with tasks that demanded integration have demonstrated minimal disruption of integration by spatial cues and only for cues corresponding to extreme spatial separation, as in the opposite-ear condition. More realistic differences in spatial cues between two sounds apparently are insufficient to disrupt fusion, especially when sounds are bound by common onset and/or common fundamental frequency. The failure of integrative tasks to demonstrate a strong effect of spatial cues may seem to conflict with the classic result by Cherry (1953), who showed that listeners readily segregated conflicting speech streams presented to opposite ears. In Cherry's study, however, listeners were encouraged to segregate the messages at the two ears, and

segregation *enhanced* performance, as in the stream segregation tasks described in the following section.

6.2.2   Robust Stream Segregation by Spatial Cues

Stream segregation can be measured directly by requiring a listener to segregate two or more sound streams and to make a judgment based on information in one stream while rejecting distraction by the other streams. Such tasks have been referred to as measures of fission (van Noorden 1975), segregation (Micheyl and Oxenham 2010), and voluntary segregation (Stainsby et al. 2011). This is the task of a person attempting to follow a particular conversation amid a crowd of other talkers. The magnitude of stream segregation is quantified by the degree to which it *improves* performance. Direct measures of stream segregation have demonstrated robust effects of space or of spatial cues. Hartmann and Johnson (1991) demonstrated that two interleaved melodic lines could be segregated when the two melodies were presented to the two ears with ITDs of ± 500 µs. Identification of the melodies in that condition was nearly as accurate as when the signals were presented to opposite ears. In a study by Saupe and colleagues (Saupe et al. 2010), listeners heard musical phrases played by synthesized instruments differing in timbre and were instructed to report a large descending pitch interval played by a particular target instrument. Performance was enhanced substantially when the sources were separated in location in the free field by 28° compared to a co-located condition. Sach and Bailey (2004) asked listeners to distinguish rhythmic patterns of 500-Hz tone pips localized to the perceptual midline in the presence of interleaved masker pulses. Performance improved significantly when the masker was lateralized by introduction of a 100- to 200-µs ITD or a 4-dB interaural level difference (ILD). The preceding three studies demonstrate that spatial features of sounds can enhance perceptual segregation of target and masker and can thereby enhance target recognition.

Spatial stream segregation can contribute substantially to recognition of speech in the presence of competing speech or other sounds. Most real-world efforts to recognize speech in the presence of other sounds are confounded by some combination of energetic masking, in which signal and masker overlap in time and spectrum, and sequential masking, in which there is no spectrotemporal overlap. Spatial cues are particularly important for segregating interleaved sequences of sounds from competing talkers and for linking together sequential sounds from the same talker. That phenomenon was illustrated by work by Ihlefeld and Shinn-Cunningham (2008a, b). In their experiments, energetic masking was minimized by restricting target and masker to multiple interleaved nonoverlapping spectral bands. A 90° separation of target and masker sources substantially improved the rate of correct identification of words, particularly by reducing the instances in which target words were replaced by words from the masker string. A cue to the location (but not the timbre) of the target enhanced the spatial effect. Kidd and colleagues evaluated the importance of spatial cues for linkage of

successive words in an utterance (Kidd et al. 2008). Listeners heard pairs of five-word sentences spoken by two talkers in which successive words alternated between the two talkers. The speech sources could be co-located or could be separated in perceived interaural location by introduction of ITDs. Along with talker identity and correct syntactic structure, interaural location improved word recognition by linking together words from the target talker.

6.2.3   Spatial Acuity of Stream Segregation

A study of the ability of human listeners to form perceptual streams based on source location utilized interleaved sequences of target and masker noise bursts having identical spectral envelopes and differing only in source location (Middlebrooks and Onsan 2012). A nonverbal objective task was adopted to facilitate comparison of human psychophysical results with animal psychophysical and physiological results. The sound bursts had no temporal overlap, thereby isolating the phenomenon of stream segregation and eliminating any energetic masking. Success in the task required a listener to segregate otherwise identical sequences of sounds into distinct streams on the basis of source location and to discriminate rhythmic patterns within one of those streams. The schematic in Fig. 6.1 shows, in solid bars,
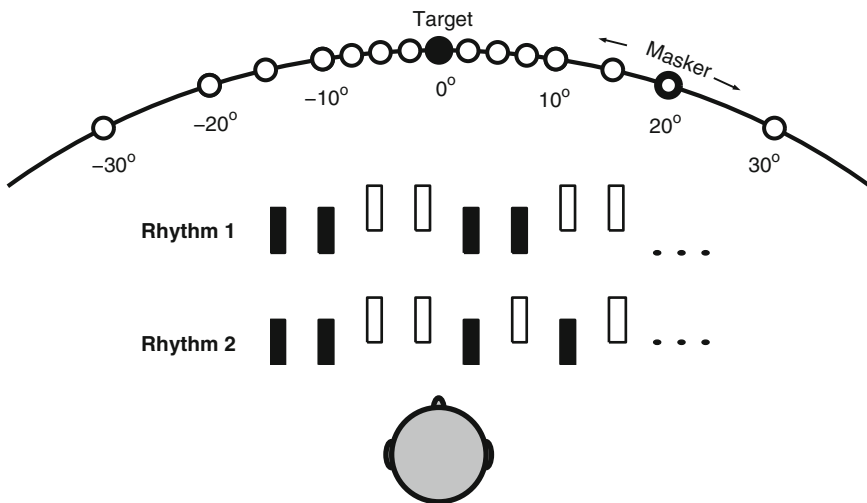

Fig. 6.1 Schematic of a psychophysical measure of spatial stream segregation using rhythmic masking release (RMR). The listener heard sequences of noise bursts presented from loudspeakers positioned in the horizontal plane. The target source was fixed at 0 or 40°, and the masker source, shown here at 20°, varied in location between trials. Target and masker noise bursts (indicated by solid and open bars, respectively) were interleaved in time and were identical except for their source locations. On each trial, one or the other illustrated rhythm was repeated four times without interruption, and the listener indicated whether the target sequence was rhythm 1 or 2

the two target rhythms that were to be discriminated along with, in open bars, the complementary masking sequences; the component broadband or band-passed noise bursts were 20 ms in duration and presented at an aggregate rate of 10/s. In this single-interval design, listeners reported by button press whether they heard rhythm 1 or rhythm 2. When target and masker sources were co-located, the stimulus was by design an undifferentiated sequence of noise bursts. In that condition, target and masker were heard as a single stream, and discrimination of the target rhythm was impossible. Hypothetically, spatial separation of target and masker sources could lead to perception of target and maskers sequences as distinct streams, thereby permitting analysis of the temporal pattern within the target stream and recognition of the target rhythm. This is referred to as rhythmic masking release.

The performance of one listener is shown in Fig. 6.2, with 6.2A and B representing results for target sources fixed respectively at 0 and 40° azimuth in the horizontal plane; the locations of masker sources are plotted as the horizontal axes. The accuracy in performance of the task is given by the discrimination index, d', for discrimination of rhythm 1 from rhythm 2, where d' near zero indicates random-chance performance and $d' = 1$ was taken as the criterion for threshold rhythmic masking release. The expected near-zero d' values were obtained when the masker source location coincided with the target location. Even small displacements of the masker source, however, resulted in emergence of perceptually segregated streams, which resulted in unmasking of the target sequence and rapid improvement in rhythm discrimination. Dashed lines in the figure indicate the crossings of $d' = 1$ that indicate threshold target-masker displacements. In the broadband stimulus condition shown in Fig. 6.2, the median threshold for rhythmic masking release across seven listeners was 8.1° when the target source was at 0° and was 11.2° when the target was at 40°. When asked to report their subjective experiences, these listeners tended to report hearing two distinct streams when the target–masker separation was at or wider than the listeners' masking release thresholds and single streams when the separation was narrower.

Rhythmic masking release thresholds were significantly wider for the 40° than for the 0° target location, although the difference in medians was only 3.1°. A previous physiological study in auditory cortex of domestic cats (*Felis catus*) demonstrated proof of concept of a model of spatial hearing based on comparison of summed activity of left- and right-tuned neural populations in auditory cortex (Stecker et al. 2005). Models invoking only two or three spatial channels have gained some favor in regard to human psychophysics (Phillips 2008; Dingle et al. 2010) and to human neurophysiology using far-field magnetic and electric recordings (Salminen et al. 2009; Magezi and Krumbholz 2010; Briley et al. 2013). Those models, however, predict a rapid fall-off in spatial acuity with increasing distance of the target to the left or right of the midline, contrary to the spatial stream segregation results presented here. That is, in rhythmic masking release tested with a target at 40°, the target and all the tested masker locations would have been within the receptive fields of putative right-tuned neuronal populations, and a left-vs-right channel model would have predicted low (i.e., poor) spatial acuity. The observed
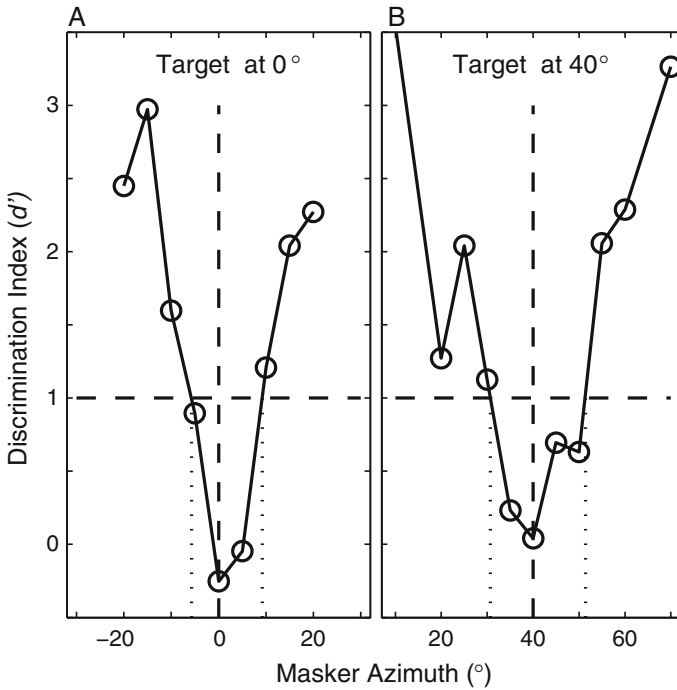
Fig. 6.2 Spatial stream segregation by one listener. Performance in the rhythmic masking release (RMR) task is represented by an index of the discrimination of rhythm 1 versus rhythm 2 (d') as a function of location of the masker source in the horizontal plane. (**A**, **B**) Conditions in which the target was fixed at 0 and 40°, respectively. RMR thresholds were given by the minimum interpolated target/masker separations at which performance exceeded a criterion of $d' = 1$, indicated by dashed lines. Two thresholds, indicated by dotted lines for masker locations to the left and right of the target, were determined for each listener and condition. (From Middlebrooks and Onsan 2012)

high acuity for the 40° target conflicts with that prediction and is more consistent with spatial-hearing models that incorporate the spatial sensitivity of single neurons (Middlebrooks et al. 1994; Lee and Middlebrooks 2013).

Thresholds for rhythmic masking release approached the thresholds measured in the same listeners for discrimination of a right-to-left from a left-to-right sequence of two sounds, their minimum audible angles (MAAs). The distributions of masking release thresholds overlapped with those of MAAs, but masking release thresholds of individual listeners generally were somewhat wider than their MAAs.

The high level of performance and fine spatial acuity obtained in this test of spatial stream segregation using rhythmic masking release contrast markedly with the weak, low-acuity spatial effects observed with tests of stream integration. Again, it appears that a listener can overcome spatial separation in tasks in which segregation is a liability for integrating information from multiple sources but,

alternatively, can take advantage of the spatial arrangement of an auditory scene when the goal is to attend to one of several interleaved sound streams.

6.2.4   Acoustic Cues for Spatial Stream Segregation

The locations of sound sources in space are computed within the central auditory system from acoustical cues that result from interaction of the incident sound wave with the head and external ears (Middlebrooks and Green 1991). The dominant cues for localization of broadband or low-pass sounds in the horizontal dimension (i.e., in azimuth) are ITDs in the ongoing temporal fine structure of sounds (Wightman and Kistler 1992), and the dominant cues for horizontal localization of high-pass sounds are ILDs (Macpherson and Middlebrooks 2002). One can distinguish the relative contribution of fine-structure ITDs or of ILDs to spatial stream segregation in the horizontal dimension by testing with low- or high-pass sounds. In the vertical dimension, the primary spatial cues are spectral shapes that result from the direction-depending filtering properties of the external ears. One can isolate spectral shape cues by testing locations in the vertical midline, where ITDs and ILDs are essentially uninformative.

Spatial stream segregation in the horizontal dimension was tested for broadband (0.4–16 kHz), low-band (0.4–1.6 kHz), and high-band (4–16 kHz) stimuli using the rhythmic masking release task (Middlebrooks and Onsan 2012); in each condition, pass-bands for target and masker stimuli were identical. Performance in the low-band condition was not significantly different from that in the broadband condition (Fig. 6.3). In contrast, performance was substantially worse in the high-band condition in which low-frequency cues were eliminated. Those results suggest that low-frequency ITD cues provided the highest spatial acuity for stream segregation in the horizontal dimension. A separate test demonstrated that the spatial stream segregation in the absence of low-frequency ITD cues (i.e., high-pass sounds) was derived primarily from ILD cues, with little or no contribution from better-ear level cues or from ITDs in envelopes of high-frequency sounds (Middlebrooks and Onsan 2012).

The demonstration that spatial stream segregation in the horizontal dimension relies on ITD and (with lesser spatial acuity) on ILD cues raises the question of whether the observed stream segregation is a property of spatial hearing in general or whether it is specifically a binaural process. That question was addressed by testing for spatial stream segregation with target and masker sources both located in the vertical midline, where binaural cues to target and masker separation are negligible and where spectral shapes are the primary spatial cue (Middlebrooks and Onsan 2012). The performance of one listener in such a task is shown in Fig. 6.4; in this figure, the horizontal axis depicts the vertical location of the masker above or below the horizontal plane (at 0° elevation). An unanticipated result was that sensitivity to target–masker separation depended rather strongly on the durations of the broadband sound bursts that formed the target and masker sequences. When
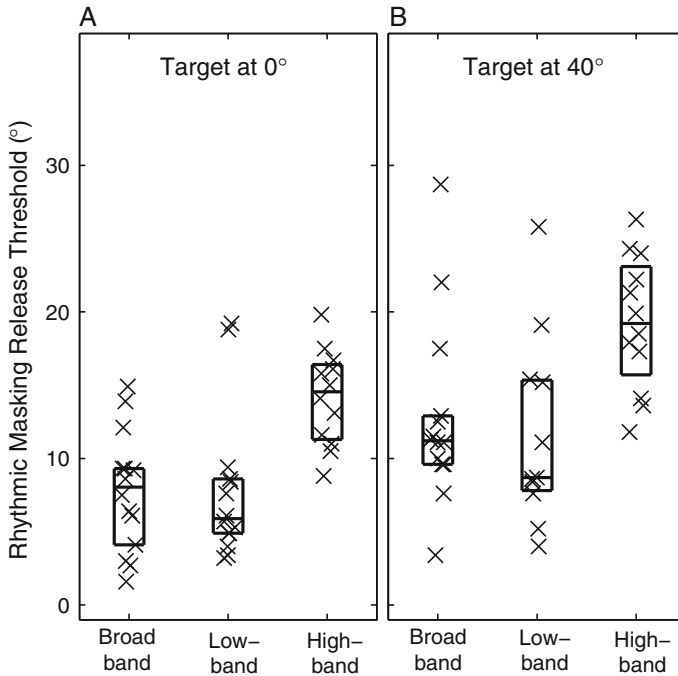
Fig. 6.3 Distributions of rhythmic masking release (RMR) thresholds as a function of stimulus band. Boxes indicate 25th, 50th, and 75th percentiles of distributions across seven listeners and across maskers to left and right of targets. Broadband pass bands were 0.4–16 kHz, low-pass bands were 0.4–1.6 kHz, and high-pass bands were 4.0–16 kHz. Thresholds were significantly wider for 40° (**right**) compared to 0° (**left**) target locations ($p < 0.0005$, paired signed rank test, in the broadband condition) and were significantly wider in the high-band condition than in the broadband or low-band conditions ($p < 0.005$ at 0° and $p < 0.05$ at 40°, Bonferroni-adjusted paired comparisons) (Middlebrooks and Onsan 2012)

burst durations were 10 ms (Fig. 6.4A), this listener and most others were unable to reach criterion sensitivity at any tested target–masker separation. Sensitivity improved for 20-ms bursts (Fig. 6.4B), although in this example the sensitivity hovered around $d' = 1$ for most of the masker locations below the horizontal plane. When the bursts were lengthened to 40 ms, however, sensitivity improved to levels comparable with those observed in the horizontal dimensions; the median of thresholds across listeners was 7.1° for 40-ms bursts in the vertical dimension compared with a median of approximately 4° for 40-ms bursts in the horizontal dimension. The observation that spatial stream segregation sensitivity varied with sound-burst duration is somewhat parallel to observations of impaired vertical localization of brief noise bursts (Hartmann and Rakerd 1993; Hofman and Van Opstal 1998; Macpherson and Middlebrooks 2000), although the impaired vertical localization in the previous studies was associated particularly with high sound
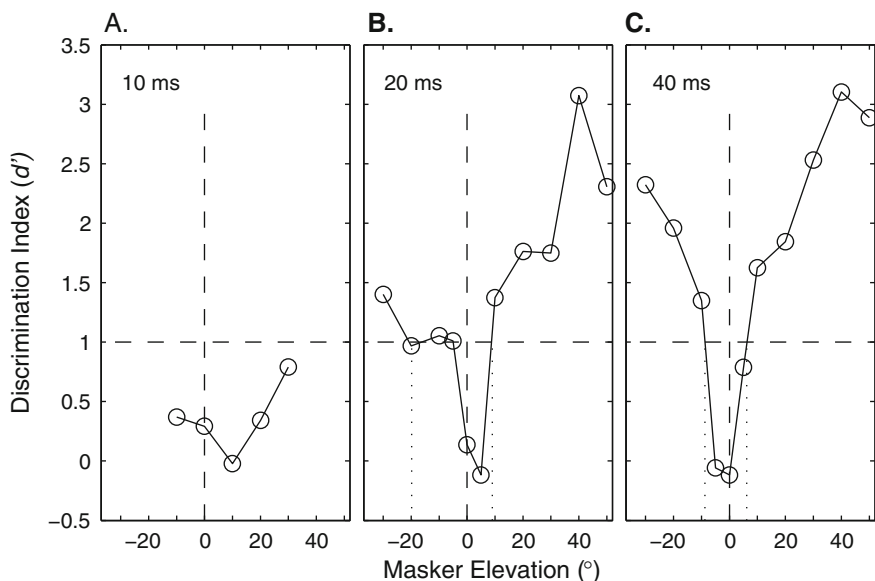
Fig. 6.4 Spatial stream segregation in the vertical midline by one listener. Target sources were fixed at 0° (i.e., in the horizontal plane containing the listener's ears), and the horizontal axes plot the location of the masker above or below the horizontal plane. The three panels show conditions in which the broadband noise bursts constituting the sound sequences were 10, 20, or 40 ms in duration (Middlebrooks and Onsan 2012)

levels whereas sounds were presented at moderate levels in the stream segregation experiments.

The distributions of rhythmic masking release thresholds observed in various experimental conditions are summarized in Fig. 6.5A. Minimum audible angles (Fig. 6.5B) also were measured for the same listeners as an indicator of their localization acuity independent of the complexity of the rhythmic masking release task; note the difference in the vertical scale between Fig. 6.5A and B. There is a striking difference between the two panels. Rhythmic masking release thresholds varied markedly in the horizontal dimension as a function of pass band and in the vertical dimension as a function of burst duration. In contrast, there was little variation in MAAs across those stimulus conditions. The difference in stimulus dependence between spatial stream segregation and MAAs suggests that the two spatial phenomena might result from differing brain structures or mechanisms. That issue is considered further in Sect. 6.4. That spatial stream segregation is observed in both the horizontal and vertical planes confirms, however, that spatial differences among competing sounds can support stream segregation irrespective of whether the spatial differences are processed by binaural or by other spatial mechanisms.
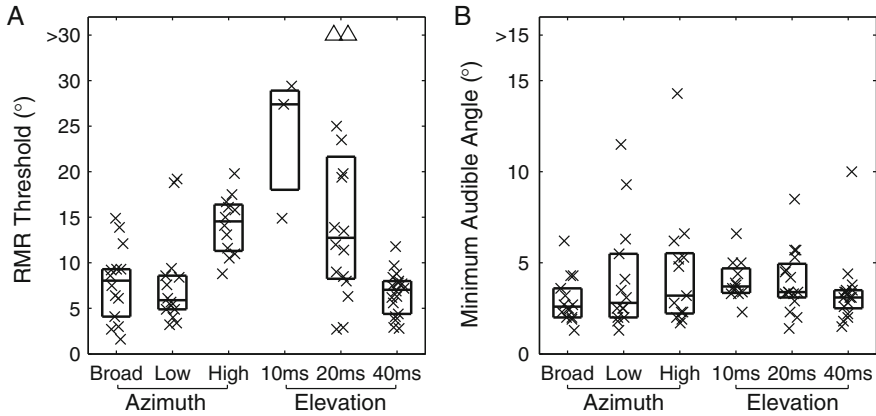
Fig. 6.5 Distributions of rhythmic masking release (RMR, **A**) and minimum audible angles (MAAs, **B**) thresholds in various conditions of stimulus pass-band and spatial dimension. The vertical scales differ between the two panels. Stimulus locations in the horizontal plane (azimuth) were tested in broadband, low-band, and high-band conditions, all with 20-ms sound bursts and with the target at 0° azimuth. RMR thresholds in the horizontal dimension varied significantly with stimulus pass-band ($p < 0.0005$, Kruskal–Wallis test), whereas there was no significant variation in MAAs across those conditions ($p > 0.05$). Locations in the vertical midline were tested with broadband sounds having 10 ms, 20 ms, and 40 ms durations, with the target at elevation 0°. RMR thresholds in the vertical dimension varied significantly with burst duration ($p < 0.0001$), and all pairwise differences between durations were significant ($p < 0.05$, Bonferroni-adjusted paired comparisons). MAA thresholds also varied significantly with burst duration ($p < 0.005$) but that was due entirely to the difference between 10- and 40-ms conditions; no other paired comparisons were statistically significant (Middlebrooks and Onsan 2012)

6.3 A Bottom-Up Substrate for Spatial Stream Segregation

Neural correlates of stream segregation based on tone frequency have been demonstrated in cortical area A1 of the macaque monkey (Fishman et al. 2001; Micheyl et al. 2005). When presented with sequences of tone pips that alternate in frequency, cortical neurons tend to synchronize to tones of one or the other frequency. Tonal stream segregation operates within a substrate of tonotopic organization in which frequency-selective single neurons are organized into orderly maps of tone frequency onto cortical place. Proposed mechanisms of tonal stream segregation have invoked inhibitory interactions among loci along the cortical tonotopic axis (Fishman et al. 2001).

Neural pathways for spatial hearing begin with analysis of acoustic spatial cues in the auditory brainstem. Results of that analysis are conveyed to the level of auditory cortex, where responses of single neurons vary in magnitude and timing according to sound-source location (Middlebrooks et al. 1994). The spatial sensitivity of single neurons, however, is far less precise than is the ability of an animal

to localize a sound by, for instance, orienting to a sound source to receive a food reward (May and Huang 1996; Tollin et al. 2005). Neural localization performance comparable in precision to behavior has been demonstrated only in the coordinated activity of populations of neurons (Furukawa et al. 2000; Miller and Recanzone 2009). There is substantial evidence *contrary* to the presence of orderly maps of sound-source location onto cortical place (King and Middlebrooks 2011). That raises the question of whether or not single cortical neurons could exhibit spatial stream segregation analogous to the tonal stream segregation shown by Fishman and colleagues (2001). A recent study addressed that question and showed that, indeed, responses of single cortical neurons can segregate competing streams of sounds from differing locations (Middlebrooks and Bremen 2013). Section 6.3.1 reviews that study, showing that the acuity of spatial stream segregation by single cortical neurons is substantially greater than the acuity for locations of single sound sources and approaches that of humans in psychophysical tests. Section 6.3.2 considers the evidence that spatial stream segregation reflects bottom-up processing within the auditory pathway at or below the level of the thalamocortical projection.

6.3.1 Spatial Stream Segregation in Primary Auditory Cortex

A putative substrate of spatial stream segregation was studied in primary auditory cortex of anesthetized cats (Middlebrooks and Bremen 2013). The use of general anesthesia almost certainly influenced cortical responses, and any failure to find evidence of spatial stream segregation might have been blamed on anesthetic effects. Contrary to that concern, however, spatial stream segregation was observed in that anesthetized preparation, suggesting that at least some basal level of segregation arises from bottom-up processes that do not require an animal's attention.

Stimuli consisted of sequences of brief noise bursts alternating in location from two sources in the horizontal plane in a free sound field; the source locations were varied parametrically. As discussed in Sect. 6.2.3, human listeners report hearing such stimuli as two distinct streams when target–masker source separations are approximately 10° or wider. In the cat cortical experiment, the base rate of noise-burst presentation (i.e., the aggregate of both sources) was 5 or 10/s. Cortical neurons synchronized closely to noise burst presented at half the base rate from only one of the sources, as shown in the representative post–stimulus time (PST) histograms in the left column of panels in Fig. 6.6. The example neuron showed little sensitivity to the location of a single source, with essentially equal responses to sounds presented from straight ahead (0°, Fig. 6.6C) or from 40° contralateral (Fig. 6.6A) or ipsilateral (Fig. 6.6E) with respect to the side of the recording site. When one source was held at 0° and a second source was added from the same location (equivalent to simply raising the rate to the full aggregate rate), there was a reliable response to the first noise burst, but responses to later bursts

were sparse and irregular (Fig. 6.6D). In the figure, red and blue bars indicate spikes that were synchronized to the A or B source, respectively, although the A and B designation is arbitrary in the co-located condition shown in Fig. 6.6D. When the A source was held at 0° and the B source was shifted to ipsilateral 40° (Fig. 6.6F), the response to the B source largely disappeared and the neuron responded reliably to the A source. In that configuration, the response of the neuron could be said to segregate the A sound sequence from the B sequence. A largely symmetrical response was observed when the A source was held at 0° and the B source was shifted to contralateral 40° (Fig. 6.6B). In that configuration, the B sound sequence dominated the response of the neuron.
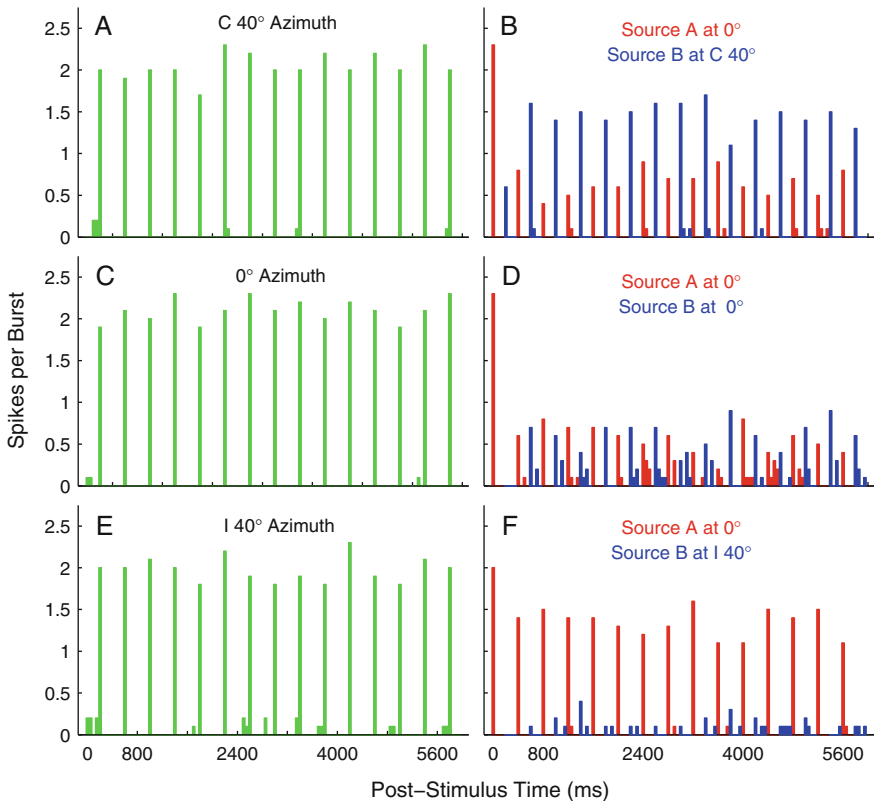

Fig. 6.6 Post–stimulus time histograms of a single neuron in cortical area A1 of an anesthetized cat. Bars indicate the neural spikes per stimulus burst in 50-ms bins. In the left panels, the stimuli were sequences of 5-ms noise bursts at a rate of 2.5/s presented from 40° contralateral with respect to the side of the recording site (**A**), straight ahead (**C**), or 40° ipsilateral (**E**). In the right panels, red bars indicate responses synchronized to source A, fixed at 0° azimuth, and blue bars indicate responses synchronized to source B located at contralateral 40° (**B**), straight ahead (**D**), or ipsilateral 40° (**F**). The aggregate rate of A and B sound bursts in the right panels was 5/s (Middlebrooks and Bremen 2013)

Responses of the unit represented in Fig. 6.6 are plotted in the left column of Fig. 6.7 as spike counts synchronized to A or B sound sources as a function of the location of the B source; Fig. 6.7A, C, and E represent conditions in which the A source was fixed in location at contralateral 40°, 0°, or ipsilateral 40°, respectively. As noted in the preceding text, the response to the B source alone (green line, duplicated in each panel) showed little sensitivity to the source location. Spatial sensitivity sharpened, however, in conditions of competing sources. In panels A, C, and E the neural response was suppressed, compared to the B-alone condition, in configurations in which the B source location coincided with the A location. In each case, however, the response that was synchronized to one or the other source, and the difference between the responses to the two sources, increased dramatically as the two sources were moved apart. In the right column of panels in Fig. 6.7, the blue line plots a measure of discrimination of spike counts synchronized to the B versus the A source; the discrimination index, d', was computed from a receiver operating characteristic (ROC) analysis of trial-by-trial spike counts. The dashed black lines indicate criteria of $d' = \pm 1$. In nearly every case in this illustration, the neural spikes segregated the A and B sources with d' larger than 1 when they were separated by the minimum tested distance, either 10° or 20°. In contrast, a comparison of spike counts elicited by a single source at varying locations compared to a single source fixed at contralateral 40°, 0°, or ipsilateral 40° (green line) indicates minimal sensitivity of the neuron to the location of a single source.

The responses of the single neuron shown in Figs. 6.6 and 6.7 were representative of the sample from primary auditory cortex in the Middlebrooks and Bremen (2013) study in several regards. First, spike rates of the majority of neurons could reliably segregate two interleaved sequences of noise bursts. When the base stimulus rate was 10/s, for instance, spike rates of 78% of single- and multiple-unit recordings could segregate with $d' \geq 1$ sequences from one or more pairs of source locations separated by only 20°. Second, like the illustrated responses, the majority of neurons tended to synchronize preferentially to the more contralateral of the two sound sources. Nevertheless, a sizeable minority of neurons (not illustrated) preferred the more ipsilateral source. Third, nearly every neuron showed greater spatial sensitivity in conditions of two competing sound sources compared to conditions of a single source. The majority of spike rates were modulated by no more than 50% as a function of the location of a single source, as shown by the green line in Fig. 6.7. Other neurons showed contralateral hemifield tuning to single sources in that they responded strongly for a single source in the spatial hemifield contralateral to the recording site; there also were a few examples of ipsilateral or frontal spatial tuning for single sources. In nearly every case, however, tuning widths were narrower, modulation of spike rates by source location was deeper, and discrimination of locations by trial-by-trial spike counts was greater in competing-source conditions compared to the single-source condition ($p < 10^{-6}$, all pairwise comparisons).

Neurons that synchronized preferentially to the more contra- or ipsilateral of two competing sources tended to occupy distinct modules within the cortex such that an electrode track through the cortex would encounter uninterrupted sequences of neurons showing only one laterality followed by sequences showing the other
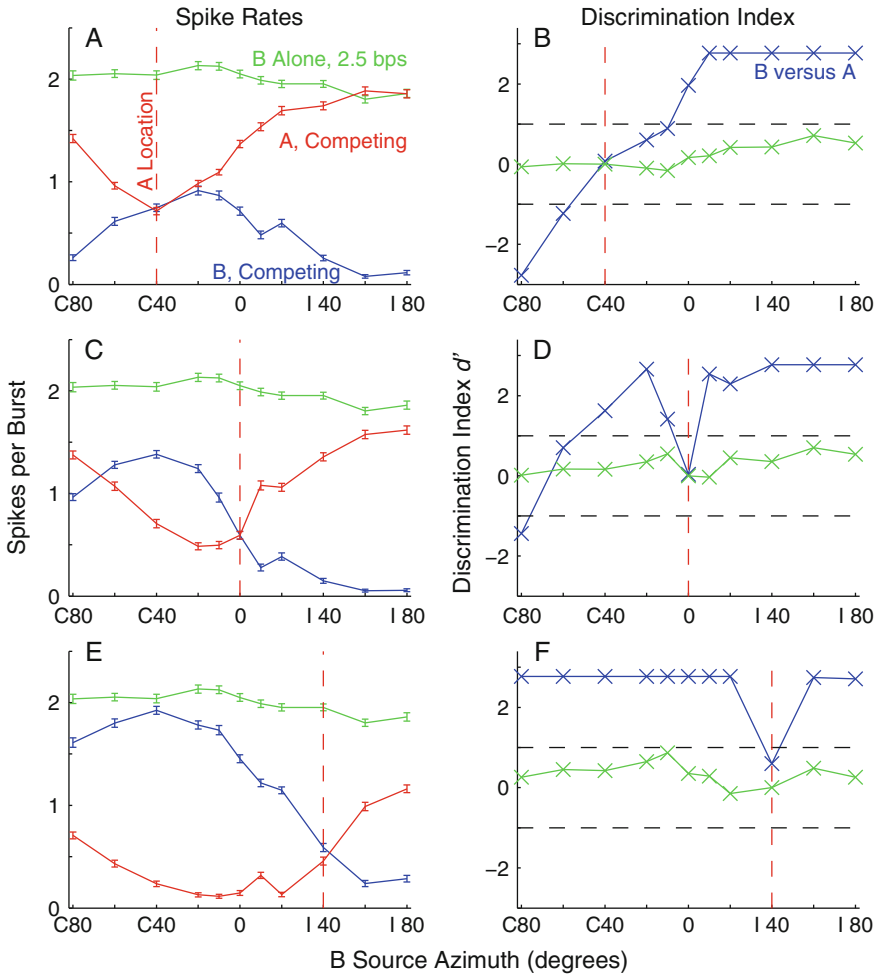
Fig. 6.7 Spike rates and stream segregation by the neuron shown in Fig. 6.6. In the left panels, lines indicate mean spikes per sound burst for a single source (green) or for spikes synchronized to the A (red) or B (blue) source when A and B sound sequences were interleaved. In each panel the location of the A source was fixed at the location indicated by the vertical dashed red line. The right panels show the discrimination index, d', for discrimination of spike rates synchronized to A versus B sources (blue) or for discrimination of spike rates elicited by a single source varied in location compared to the source fixed at the fixed location (green). Positive values of d' indicate stronger responses to the more contralateral of two sound sources (Middlebrooks and Bremen 2013)

laterality; a permutation test showed a probability $<10^{-5}$ that such nonrandom distribution of laterality preferences could have arisen by chance. Sequences of neurons having constant laterality preference often were elongated along cortical columns, but there also were examples of such sequences extending across

columns, that is, spanning a range of characteristic frequencies. In anesthetized conditions, A or B sound sequences have no special significance as target or masking sounds. In awake conditions, however, the listener might identify one or the other sequence as an auditory object of interest. In such conditions, one might hypothesize that some as-yet-unidentified top-down mechanism facilitates activity of cortical modules synchronized to the target and/or suppresses activity of modules tuned to the masker.

6.3.2 Spatial Rhythmic Masking Release by Cortical Neurons

The degree of spatial stream segregation shown by single neurons in auditory cortex in anesthetized cats was sufficient to support spatial rhythmic masking release comparable to that demonstrated in human psychophysics (Middlebrooks and Onsan 2012). Rhythmic sequences of noise bursts such as those used in the human study (Fig. 6.1) were presented, and responses of single neurons were studied in primary auditory cortex of anesthetized cats (Middlebrooks and Bremen 2013). Responses of a single neuron are represented by PST histograms in Fig. 6.8. The sequence of open and filled squares across the top of each panel represents the rhythms of target (open) and masker (filled) noise bursts; top and bottom rows of panels represent rhythm 1 and rhythm 2. When target and masker sources were co-located, as in Fig. 6.8B and E, human listeners reported hearing a single stream and the cat cortical neuron synchronized equally to both sources. When the masker was shifted to contralateral 80° (Fig. 6.8A and D) or ipsilateral 80° (Fig. 6.8C and F), however, human listeners reported hearing two segregated streams, and the neuron responded with distinctive PST patterns. Within each stimulus sequence, the neuron tended to respond strongly to a change from target to masker location or vice versa. That resulted in two strong responses to target bursts for each repetition of rhythm 1 (Fig. 6.8A and C) and three strong responses per repetition of rhythm 2 (Fig. 6.8D and F).

A linear-classifier analysis was used to test whether distinctive PST histogram patterns of single neurons such as that shown in Fig. 6.8 could reliably distinguish target rhythms and thereby perform the rhythmic masking release task. That analysis used linear regression with terms given by spike counts in 50-ms time bins and coefficients optimized to yield outputs of 1 or 2 depending on the stimulus rhythm. An ROC analysis of the distribution of outputs resulting from stimulus rhythms 1 or 2 yielded d' for discrimination of the rhythms by single neurons; a one-out procedure was used so that test trials differed from the trials that were used to compute regression coefficients. Discrimination results are shown in Fig. 6.9A for the neuron represented in Fig. 6.8, and Fig. 6.9B shows the distribution of discrimination results across the neuron sample. More than 25% of the isolated single neurons could isolate target and masker streams adequately at a 10°
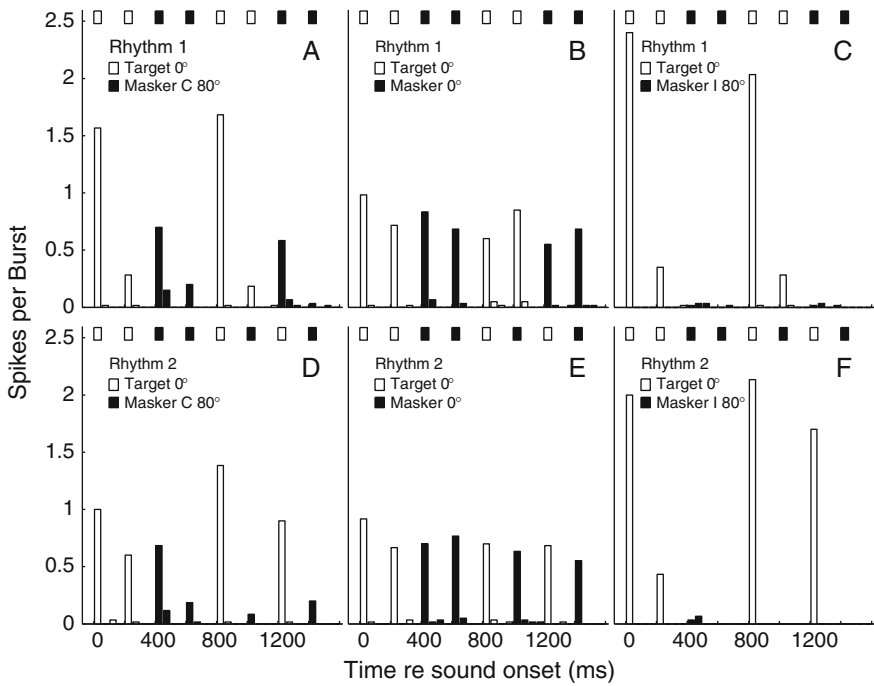
Fig. 6.8 Synchrony of a cortical neuron to rhythmic masking release stimuli. Top and bottom rows of panels represent responses to rhythms 1 and 2, respectively. Stimulus rhythms are represented by the row of squares at the top of each panel. The time axis is folded on the 1600-ms rhythm duration, so that each panel represents means of three repetitions per trial times 20 trials. Open and solid response bars indicate mean spike rates per target (open) and masker (solid) burst in 50-ms bins. Target and masker source locations were as indicated in each panel (Middlebrooks and Bremen 2013)

separation to achieve $d' \geq 1$. That compares favorably with median rhythmic masking release thresholds of 8.1° in human psychophysics reviewed in Sect. 6.2.3 (see also Middlebrooks and Onsan 2012). Task performance by the human listeners required judgments of the perceived rhythmic patterns, whereas in the neurophysiological study the rule for identifying the two rhythms was effectively programmed into the computer as a result of the feedback inherent in the regression analysis. For that reason, the results do not address discrimination of rhythms by neurons. Nevertheless, one can conclude that neural segregation of A and B sound sequences was sufficient to discriminate the rhythms and that segregation could be accomplished by brain mechanisms that were active even under anesthesia.
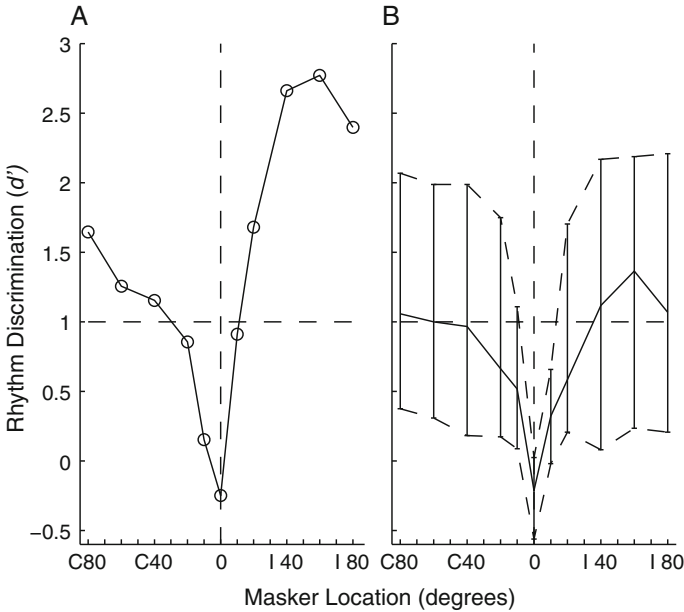
Fig. 6.9 Identification of masked rhythms based on synchronized neural responses. (**A**) The d' for discrimination of rhythm 1 versus 2 by a single neuron as a function of masker source location; the target was fixed at 0°. (**B**) The average d' across 57 single units. The curves indicated 25th, 50th, and 75th percentiles of the distribution (Middlebrooks and Bremen 2013)

6.3.3 A Mechanism for Bottom-Up Spatial Stream Segregation

The spatial stream segregation by single neurons observed in auditory cortex of anesthetized cats (Middlebrooks and Bremen 2013) could be predicted quantitatively by a model invoking (1) relatively weak spatial sensitivity inherited from the ascending brainstem auditory pathway; and (2) forward suppression tentatively localized somewhere in the projection from the medial geniculate body (MGB) and primary auditory cortex. Figure 6.10 shows recorded spike counts (symbols) and model predictions (lines) for one neuron tested with a 5/s stimulus rate (top row of panels) and another neuron tested at a 10/s rate (bottom row). Components of the model are discussed in the following.

The spatial sensitivity of cortical neurons for single sound sources presumably reflects sensitivity inherited from brainstem inputs plus any further sharpening that might occur at the cortical level. In the quantitative model, the responses of cortical neurons to single sources are taken as surrogates for the spatial sensitivity of the thalamocortical projection. As noted in Sect. 6.3.1, the spatial sensitivity of most cortical neurons was fairly broad, with the majority of neurons showing less than 50% modulation of their spike rates by varying sound-source location.
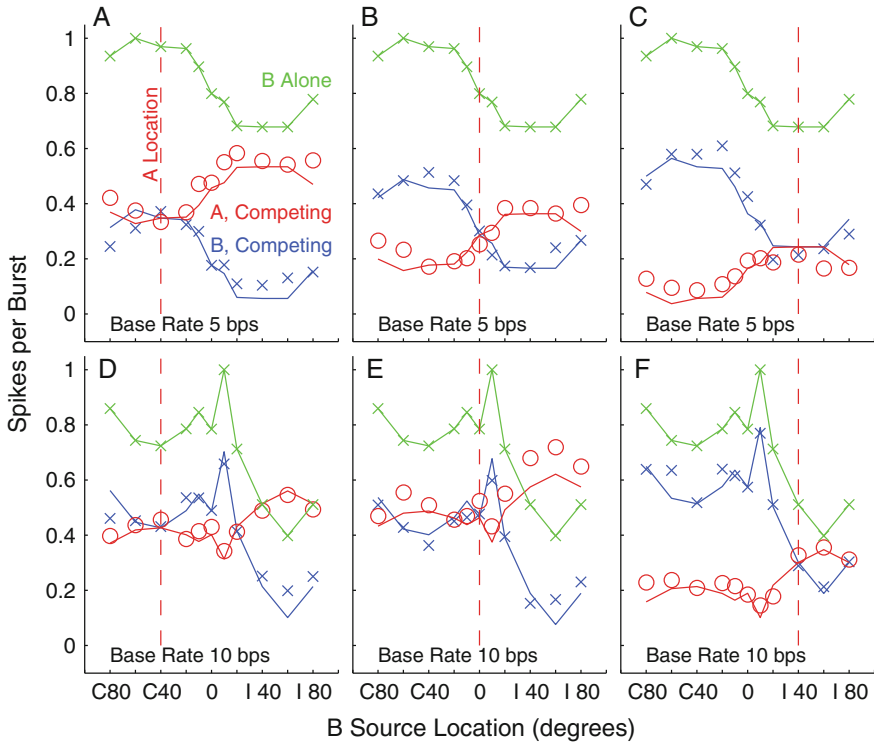
Fig. 6.10 Model predictions of neural stream segregation. The panels represent responses of one cortical neuron to stimuli at a base rate of 5/s (**A, B, C**) and of a different neuron to stimuli at a base rate of 10/s (**D, E, F**). Symbols represent mean spike bursts of the neurons, and curves represent model predictions based on responses to noise bursts from a single source (green) scaled by a forward suppression term. The neural spike count synchronized to source A (red) as a function of A and B locations θ_A and θ_B was given by $RA(\theta_A, \theta_B) = RSgl(\theta_A) - FS \times RSgl(\theta_B)$, and the response synchronized to source B (blue) was $RB(\theta_A, \theta_B) = RSgl(\theta_B) - FS \times RSgl(\theta_A)$, where $RSgl(\theta_A)$ and $RSgl(\theta_B)$ were the responses to sources A and B presented alone. The forward suppression term, FS, was estimated from the spike rates synchronized to trains of noise bursts at 5/s (top row) or 10/s (bottom row) divided by rates synchronized to trains at half those rates; ratios of spike rates were averaged across sources at contralateral 40°, 0°, and ipsilateral 40° (Middlebrooks and Bremen 2013)

Nevertheless, the spatial tuning for single sources tended to predict the spatial preference of the usually sharper tuning seen in the presence of a competing sound source. In the two examples in Fig. 6.10, the "B, competing" source plots (shown in blue) lie parallel to the single-source "B-alone" plots (green). The similarity between single- and competing-source spatial sensitivity was less obvious in the example in Fig. 6.7, but plots for the two conditions shared approximately the same peak locations and same signs of slopes. Across all the sampled units, the d' for discrimination of interleaved sound sequences from sources separated by 20° correlated highly ($r = 0.82$) with the d' for discrimination of single sources at the

equivalent locations, although d' in the competing-source condition tended to be about twice that in the single-source condition.

The response to a single source consistently was suppressed by addition of a competing sound. That is evident in Fig. 6.10 by the downward shift of the blue (competing-source) lines compared to the green (single-source) lines. The shifts tended to be a linear offset; that is, responses were attenuated by subtraction by a value that was constant within each panel, rather than by multiplication by a gain factor. The offset tended to be greater when the competing sound was at a location that would elicit a strong response to a single source (e.g., the A source fixed at 0°; Fig. 6.10E) than when the competing sound was located at a less favored location (e.g., the A source fixed at ipsilateral 40°; Fig. 6.10F). The response to sound A in the presence of competing sound B could be predicted by the response to a single source at the A location minus a forward suppression term times the response to a single source at the B location. That expression yielded the blue and red model lines in Fig. 6.10. The goodness of fit of the model (R^2) averaged 0.64 across 382 units tested with 5/s stimulus rates and averaged 0.46 across 295 units tested with 10/s rates.

The forward-suppression term in the Middlebrooks and Bremen model reflects the inability of cortical neurons to respond to rapid stimulus rates. Modulation transfer functions of primary auditory cortical neurons tend to peak around 10–30 Hz (Schreiner and Urbas 1988), which is consistent with the stimulus rates at which stream segregation is observed in the cortex (Fishman et al. 2001; Middlebrooks and Bremen 2013). Also, the relevant interstimulus times are on the same scale as forward masking that has been demonstrated in auditory cortex (Calford and Semple 1995; Brosch and Schreiner 1997). In contrast, neurons in the MGB respond well to stimulus rates in excess of 100 Hz (Creutzfeldt et al. 1980), considerably faster than the time scale of perceptual stream segregation and of the stream segregation demonstrated in the cortex. One possibility to consider for the failure of cortical neurons to follow rapid stimulus rates is that the minimum interspike time might be limited by the refractoriness of cortical neurons. That possibility was rejected in the case of stream segregation by Middlebrooks and Bremen (2013), who showed that the probability of a single neuron firing a spike to a particular noise burst was independent of whether or not that neuron had fired a spike to the immediately preceding noise burst. That indicates that rate-limiting step must be prior to the spiking activity of neurons in primary auditory cortex. The most likely alternative explanation is that forward suppression arises somewhere in the thalamocortical projection, possibly due to presynaptic inhibition or to synaptic depression in the thalamocortical synapses.

6.4 "Common" Versus "Dedicated" Spatial Representations for Localization and Spatial Stream Segregation

Basic bottom-up mechanisms for spatial stream segregation almost certainly are shared with pathways for sound localization per se; that is, for identification of the locations of sound sources. Those mechanisms include analysis of sound magnitude and phase spectra in the cochleae, interaural comparison of magnitude and phase in the superior olivary complex, identification of spectral cues for vertical locations, and some level of convergence leading to the spatial sensitivity of neurons in primary auditory cortex and in other cortical areas. It is less clear, however, whether the ultimate levels of stream segregation and localization occur within common cortical areas or whether there are particular cortical areas dedicated specifically to segregation and others to localization.

One might think of a hypothetical "common" cortical spatial representation that could be accessed for spatial segregation, localization, and possibly other spatial functions. In favor of such a common representation are the observations that spatial sensitivity of neurons in early levels of auditory cortex can both segregate interleaved sound sequences (Middlebrooks and Bremen 2013) and can identify sound-source locations; localization by cortical neurons with accuracy comparable to behavioral performance, however, can be accomplished only by integrated activity of multiple neurons in that cortical area (e.g., Mickey and Middlebrooks 2003; Miller and Recanzone 2009). These results suggest that, even if primary auditory cortex is not the ultimate locus of segregation and localization functions, it likely serves as a common pathway.

Seemingly a key prediction of a hypothesis of a common spatial representation is that the spatial acuity of all spatial functions (including localization and stream segregation) would vary in parallel as stimulus conditions are varied to favor spatial cues that differ in acuity (e.g., ITDs for azimuth compared to spectral shape cues for elevation). That prediction clearly was violated in the psychophysical results by Middlebrooks and Onsan (Middlebrooks and Onsan 2012, reproduced in Fig. 6.5 of the present chapter). That study used MAA as a measure of localization acuity. Distributions of MAA were largely constant across broadband, low-band, and high-band conditions in the horizontal dimension and across three pulse durations in the vertical dimension (Fig. 6.5B). In contrast, spatial acuity of stream segregation varied dramatically across stimulus pass-band and pulse-duration conditions (Fig. 6.5A). The difference between stream segregation compared to localization with respect to stimulus conditions suggests that spatial cues are utilized differently by dedicated pathways for spatial segregation and for localization, most likely in differing cortical areas.

Another indication that spatial segregation and localization involve differing brain structures, or at least differing mechanisms, comes from work by Edmonds and Culling (2005a, b). Speech targets were presented with maskers consisting of noise or competing speech. Speech reception was improved by introduction of

differences in ITD and/or ILD between target and masker. A summation of effects was observed for target and masker differing in both ITD and ILD (Edmonds and Culling 2005a) or in ITD in two distinct frequency bands (Edmonds and Culling 2005b). Surprisingly, that summation was observed whether the spatial direction was consistent across the interaural cues for a particular target or masker sound or whether the cues for a particular sound pointed in opposite directions. That is, in the opposite-direction case, spatial unmasking was possible even though the interaural cues for target and/or masker did not correspond to a plausible location in space. Localization was not requisite for spatial segregation.

At least some level of sensitivity to sound-source location is ubiquitous among auditory cortical neurons studied in animals. Although there are quantitative differences, qualitatively similar spatial sensitivity has been observed in every cortical area that has been studied in cats (Harrington et al. 2008), ferrets (Bizley et al. 2009), and nonhuman primates (Woods et al. 2006). One cannot yet say whether neurons in all those cortical areas also show stream segregation. Despite the widespread presence of cortical spatial sensitivity, behavioral studies show that cortical areas vary in their importance for localization and, presumably, other spatial tasks. For instance, Lomber and Malhotra (2008) compared the roles in behavior of two cortical areas in cat, the posterior auditory field (PAF) and anterior auditory field (AAF). Although differing in detail, neurons in both of those cortical areas are known to exhibit spatial sensitivity (Harrington et al. 2008). In the Lomber and Malhotra study, cats learned two tasks: they could identify the locations of sound sources and they could discriminate temporal patterns. Temporary inactivation of PAF disrupted performance of the localization task while preserving performance of the temporal pattern discrimination, whereas temporary inactivation of AAF disrupted performance of the temporal task and preserved localization. Those results suggest that the spatially sensitive neurons in PAF and AAF participate in dedicated networks that support localization in the case of PAF and temporal pattern analysis in the case of AAF. Hypothetically, the temporal pattern analysis by spatially sensitive neurons in AAF might participate in spatial stream segregation. Those dedicated networks might exist within PAF and AAF themselves and/or might reflect differential anatomical projections from those areas.

Human clinical results provide evidence for dedicated cortical substrates for particular auditory spatial functions. Thiran and Clarke (2003) and Duffour-Nikolov and colleagues (2012) evaluated 13 patients having unilateral cortical lesions varying in etiology. Of those patients, three showed pronounced deficits both in a lateralization task and in spatial release from masking, five showed lateralization deficits with preserved spatial release, and one showed intact lateralization with impaired spatial release. The dissociation of lateralization and spatial release from masking in 6 of 13 patients supports the view that these auditory spatial functions involve distinct cortical substrates.

6.5   Selection of Objects of Attention

Previous sections have demonstrated the importance of spatial cues for segregation of competing sounds (Sect. 6.2) and have demonstrated that interleaved sound sequences that human listeners would hear as segregated streams activate distinct neural populations in auditory cortex (Sect. 6.3). In real-world listening situations, however, humans or other animals must not only segregate multiple sounds but, from those segregated streams, must also select particular sound objects for attention and action. In Chap. 2, Shinn-Cunningham, Best, and Lee provide a broad overview of object selection. The present section considers task-dependent sharpening of spatial or spectral sensitivity of cortical neuron and presents other animal and human results that, although not specifically spatial, might be extrapolated to selection of objects specifically on the basis of location.

6.5.1   Task-Dependent Modulation of Stimulus Specificity in Behaving Animals

Two research groups have demonstrated that responses of cortical neurons can adapt on a rapid time scale to optimize performance when an animal is engaged in a sensory task. Neither of these groups has evaluated selection among simultaneous or interleaved sounds, but both have shown modulation of stimulus tuning during presentation of a reference sound that would enhance detection of a change from reference to a target sound.

Fritz and Shamma trained ferrets to detect the change from a broadband reference sound to a single- or multitone target (Fritz et al., 2003, 2007) or to discriminate between the directions of frequency shifts of two-tone sequences (Yin et al. 2014). The broadband reference probed the spectrotemporal receptive fields (STRFs) of neurons. Neurons in primary auditory cortex showed rapid changes in STRFs when the animal engaged in a task compared to during passive sound exposure. The STRF changes indicated changes in neuronal tuning that were adaptive in the sense that they would enhance discrimination of the target tone from the broadband reference or would enhance discrimination between two targets presented in separate time windows. The task-dependent modulation of stimulus tuning could be interpreted as a mechanism for enhancing the response to a particular object of attention, although the experiments tested only single targets, not simultaneous or interleaved target sequences.

Studies in trained cats have demonstrated task-dependent changes in selectivity for stimulus locations in space (Lee and Middlebrooks 2011, 2013). Cats pressed a pedal to initiate presentation of reference sounds consisting of a succession of broadband noise bursts from varying locations in the horizontal plane. The cat could release the pedal to receive a food reward when the sound changed to one of two targets. In "periodicity" trial blocks the target was a periodic click train,

whereas in "localization" blocks the target shifted in location to a higher elevation. Compared to the condition with unattended sound exposure, location specificity of neurons sharpened during performance of the periodicity task, which required attention to an auditory task. Further sharpening was observed during the performance of the localization task, which demanded evaluation of the location of each stimulus. The most common observed change in location specificity was that of increased suppression of responses to sounds from nonfavored locations, generally narrowing responses from omnidirectional sensitivity to responses restricted to locations contralateral to the recording site. Changes in location tuning were evident as soon as they could be evaluated, a few tens of seconds after the onset of task engagement. Again, that study did not test conditions of competing sounds, but the changes in neuronal spatial tuning such as those that accompanied engagement in that single-source task presumably would serve to enhance segregation and/or to selection of a sound sequence from one source among interleaved sequences from multiple sources.

In both the ferret and cat studies, the act of listening for a target resulted in changes in stimulus tuning to the reference sound, either to the broadband STRF probe or to the broadband noise bursts in the horizontal plane. A study in ferrets looked specifically at neural responses to the target sound. Atiani et al. (2009) evaluated the contrast between responses to reference and target sounds between on- and off-task conditions. Neurons developed robust, sustained responses to targets and suppressed their responses to reference sounds during task performance whereas they showed relatively little contrast between responses to target and reference during passive sound exposure. The task-dependent contrast in response to target and reference was moderate in the primary auditory area, stronger in a nonprimary area, and essentially binary in prefrontal cortex.

6.5.2 Object Selection in Human Neurophysiology

Recent studies of nonprimary auditory cortical areas in humans have demonstrated neurophysiological correlates of object selection by demonstrating enhanced synchrony of neural activity to the one of two competing speech streams that receives a listener's attention. We consider here work from two research groups (Ding and Simon 2012a; Mesgarani and Chang 2012) that attempted to reconstruct features of speech stimuli from patterns of activity in neural populations. Both groups utilized auditory stimuli consisting of speech utterances from two talkers, and both found that the reconstructions varied markedly depending on which of the talkers received the listener's attention. Neither group addressed specifically the influence of locations of sound sources, the principal topic of this chapter, although Ding and Simon (2012b) evaluated a condition of competing sounds delivered to the two ears. Nevertheless, one might take these studies as examples of selection of objects of attention, regardless of whether the objects are segregated by speech pitch and timbre or by location.

Ding and Simon (2012a) analyzed patterns of far-field neural activity recorded with MEG. Stimuli were 1-minute narratives uttered by two talkers mixed into a single audio signal. In successive blocks the listeners were instructed to attend to one or the other talker. The envelope of the attended talker was reconstructed from the neural activity by optimal integration across time and across MEG sensors. The reconstruction of attended speech generally correlated more closely with the envelope of the attended speech than with the envelope of the competing narrative. The unattended speech envelope also could be reconstructed, by optimization of different MEG sensors, although those correlations were lower than for the reconstruction of attended speech. That both attended and unattended signals could be reconstructed suggests that both signals are represented by neural populations, with the superior reconstruction of the attended compared to the unattended envelope reflecting relative facilitation of the representation of the attended signal. The STRFs reconstructed from MEG activity demonstrated that the modulation of neural responses by attention was largely limited to a magnetic component having approximately 100-ms latency and localized on the planum temporale. There was no significant attention effect on an earlier component having 50-ms latency and localized to Heschl's gyrus, presumably a primary auditory area.

Ding and Simon also tested a condition in which two speech narratives from the same talker were presented dichotically to the two ears (Ding and Simon 2012b). This was an MEG counterpart of the classic study by Cherry (1953), in which listeners could attend to the narrative at one or the other ear. Again, the envelope of the attended speech signal could be reconstructed from the MEG recordings. In this particular case, the selection of the object of attention was given by the ear of entry, which can be regarded as a coarse spatial cue, rather than by the identity of the talker. The results from these MEG studies indicate that selection of an auditory object of attention arises at a cortical level beyond the primary area, and that selection can be based on spectrotemporal cues (Ding and Simon 2012a) or on an approximation of a spatial cue (Ding and Simon 2012b).

A subsequent study by that group showed that the envelope of an attended speech signal presented in a background of speech-spectrum noise could be reconstructed from MEG recordings, but that both cortical synchrony and perceptual intelligibility were lost when the temporal fine structure of the speech was degraded with a four- or eight-channel vocoder (Ding et al. 2013). That result indicates that cortical neural populations in the planum temporale synchronize to attended auditory objects, defined in this case by the fine spectral and temporal characteristics needed to recognized speech sounds and likely analyzed at subcortical levels, rather than simply to the low-resolution temporal envelope of the stimulus.

Mesgarani and Chang (2012) recorded from arrays of cortical electrodes on the surface of the posterior superior temporal lobe in patients who were being evaluated for epilepsy surgery. Speech reconstruction filters were derived during passive listening conditions from responses to a corpus of sentences distinct from those used as test stimuli. Then, those filters were used to estimate spectrograms based on test utterances of single talkers and of mixtures of two talkers. When tested with the

mixed stimuli, the estimated spectrograms captured spectral and temporal features that correlated well with the spectrograms of the utterance of the attended talker, with substantially lower correlation with the unattended spectrogram. The listeners were asked to report words spoken by one of the two talkers. Reconstructions of the attended spectrogram were successful on trials in which the listeners answered correctly, and the reconstructions were degraded on trials in which the reports were incorrect. That result suggests a trial-by-trial correspondence between the patterns of cortical activity and task performance. Activity at single recording sites showed tuning for particular spectral features, but responses to those features were greater when contained in the speech stream from the attended talker than when the same features were present in the stream from the unattended talker. That observation supports the notion that neural responses in this nonprimary auditory cortex represent attended auditory objects rather than just particular acoustic features.

6.6   Summary, Synthesis, and Future Directions

It is clear from the available perceptual and physiological data that the locations of targets and competing sounds are key factors in parsing the auditory scene. Spatial separation of sources turns out to have remarkably little effect on the perception of multiple sound components that are otherwise bound by common onset time, fundamental frequency, and even visual cues, as shown in tests of obligatory streaming, of concurrent vowels, and of the ventriloquism effect (Stein and Meredith 1993). Spatial separations of signals and maskers, however, clearly are potent cues for voluntary stream segregation and object selection. Studies of primary auditory cortex in anesthetized cats demonstrate that distinct neural populations synchronize to sound sequences that presumably would be segregated perceptually on the basis of differences in target and interferer locations. That segregation as distinct neural populations is analogous to segregation that has been demonstrated previously on the basis of spectral differences. Studies in behaving ferrets and cats show that stimulus selectivity is modulated during task performance to enhance detection and discrimination of single targets. Neurophysiological studies in humans demonstrate enhanced synchrony of neurons in nonprimary cortical areas to attended speech streams.

   The results lead to a general hypothetical organization for the neural substrates for spatial stream segregation and object selection. Spatial stream segregation appears to be a largely bottom-up phenomenon beginning with basic brainstem analysis of spatial cues, including interaural time and level differences and spectral shape. Forward suppression at the level of the thalamocortical projection leads to the first appearance of spatial stream segregation on temporal and spatial scales similar to those in perception. Task-dependent sharpening of spatial tuning for single sources could contribute to sharpening of segregation of sounds from multiple sources. Distinct neural populations in primary and possibly higher-order auditory cortical areas appear to represent both attended and competing sounds.

The evidence in human neurophysiology for enhanced cortical synchrony to attended sounds suggests that a top-down executive mechanism in some way facilitates activity of neural populations that represent attended sounds and/or suppresses populations that represent competitors. That is, selection of objects of attention could correspond to selection among simultaneously active neural populations. This hypothetical neural substrate for object selection invites future studies designed to confirm or reject the notion of selection among distinct neural populations in low-level auditory cortical areas, to explore how neural populations distinguished by spatial sensitivity might integrate spectral and other cues for segregation, and to identify sources of the executive signal(s) that could accomplish such selection.

Compliance with Ethics Requirements

John Middlebrooks has no conflicts of interest.

References

Atiani, S., Elhilali, M., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron, 61,* 467–480.

Bizley, J. K., Walker, K. M., Silverman, B. W., King, A. J., & Schnupp, J. W. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *The Journal of Neuroscience, 29,* 2064–2075.

Boehnke, S. E., & Phillips, D. P. (2005). The relation between auditory temporal interval processing and sequential stream segregation examined with stimulus laterality differences. *Perception and Psychophysics, 67,* 1088–1101.

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

Briley, P. M., Kitterick, P. T., & Summerfield, A. Q. (2013). Evidence for opponent process analysis of sound source location in humans. *Journal of the Association for Research in Otolaryngology, 14,* 973–983.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *The Journal of the Acoustical Society of America, 29,* 708–710.

Brosch, M., & Schreiner, C. E. (1997). Time course of forward masking tuning curves in cat primary auditory cortex. *Journal of Neurophysiology, 77,* 923–943.

Calford, M. B., & Semple, M. N. (1995). Monaural inhibition in cat auditory cortex. *Journal of Neurophysiology, 73,* 1876–1891.

Carl, D., & Gutschalk, A. (2012). Role of pattern, regularity, and silent intervals in auditory stream segregation based on inter-aural time differences. *Experimental Brain Research, 224,* 557–570.

Cherry, C. E. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America, 25,* 975–979.

Creutzfeldt, O. D., Hellweg, F. C., & Schreiner, C. (1980). Thalamocortical transformations of responses to complex auditory stimuli. *Experimental Brain Research, 39,* 87–104.

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review, 2,* 114–140.

Ding, N., Chatterjee, M., & Simon, J. Z. (2013). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage, 88,* 41–46.

Ding, N., & Simon, J. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the USA, 109,* 11854–11859.

Ding, N., & Simon, J. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology, 107,* 78–89.

Dingle, R. N., Hall, S. E., & Phillips, D. P. (2010). A midline azimuthal channel in human spatial hearing. *Hearing Research, 268,* 67–74.

Duffour-Nikolov, C., Tardif, E., Maeder, P., Thiran, A. B., et al. (2012). Auditory spatial deficits following hemispheric lesions: Dissociation of explicit and implicit processing. *Neuropsychological Rehabilitation, 22,* 674–696.

Edmonds, B. A., & Culling, J. F. (2005a). The role of head-related time and level cues in the unmasking of speech in noise and competing speech. *Acta Acustica united with Acustica, 91,* 546–553.

Edmonds, B. A., & Culling, J. F. (2005b). The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay. *The Journal of the Acoustical Society of America, 117,* 3069–3078.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A., & Shamma, S. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron, 61,* 317–329.

Fishman, Y., Reser, D., Arezzo, J., & Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research, 151,* 167–187.

Fritz, J. B., Elhilali, M., & Shamma, S. A. (2007). Adaptive changes in cortical receptive fields induced by attention to complex sounds. *Journal of Neurophysiology, 98,* 2337–2346.

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. J. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience, 6,* 1216–1223.

Füllgrabe, C., & Moore, B. C. J. (2012). Objective and subjective measures of pure-tone stream segregation based on interaural time differences. *Hearing Research, 291,* 24–33.

Furukawa, S., Xu, L., & Middlebrooks, J. C. (2000). Coding of sound-source location by ensembles of cortical neurons. *The Journal of Neuroscience, 20,* 1216–1228.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Review of Neuroscience, 5,* 887–892.

Harrington, I. A., Stecker, G. C., Macpherson, E. A., & Middlebrooks, J. C. (2008). Spatial sensitivity of neurons in the anterior, posterior, and primary fields of cat auditory cortex. *Hearing Research, 240,* 22–41.

Hartmann, W. M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception, 9,* 155–184.

Hartmann, W. M., & Rakerd, B. (1993). Auditory spectral discrimination and the localization of clicks in the sagittal plane. *The Journal of the Acoustical Society of America, 94,* 2083–2092.

Hofman, P. M., & Van Opstal, J. A. (1998). Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America, 103,* 2634–2648.

Hukin, R. W., & Darwin, C. J. (1995). Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. *The Journal of the Acoustical Society of America, 98,* 1380–1387.

Ihlefeld, A., & Shinn-Cunningham, B. (2008a). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America, 123,* 4369–4379.

Ihlefeld, A., & Shinn-Cunningham, B. (2008b). Disentangling the effects of spatial cues on selection and formation of auditory objects. *The Journal of the Acoustical Society of America, 124,* 2224–2235.

Kidd, G., Jr., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America, 124,* 3793–3802.

King, A. J., & Middlebrooks, J. C. (2011). Cortical representation of auditory space. In J. Winer & C. Schreiner (Eds.), *The auditory cortex* (pp. 329–341). New York: Springer Science+Business Media.

Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America, 62,* 157–167.

Lee, C.-C., & Middlebrooks, J. (2011). Auditory cortex spatial sensitivity sharpens during task performance. *Nature Neuroscience, 14,* 108–114.

Lee, C.-C., & Middlebrooks, J. (2013). Specialization for sound localization in fields A1, DZ, and PAF of cat auditory cortex. *Journal of the Association for Research in Otolaryngology, 14,* 61–82.

Lomber, S., & Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. *Nature Neuroscience, 11,* 609–616.

Macpherson, E. A., & Middlebrooks, J. C. (2000). Localization of brief sounds: Effects of level and background noise. *The Journal of the Acoustical Society of America, 108,* 1834–1849.

Macpherson, E. A., & Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America, 111,* 2219–2236.

Magezi, D. A., & Krumbholz, K. (2010). Evidence of opponent-channel coding of interaural time differences in human auditory cortex. *Journal of Neurophysiology, 104,* 1997–2007.

May, B. J., & Huang, A. Y. (1996). Sound orientation behavior in cats. I. Localization of broadband noise. *The Journal of the Acoustical Society of America, 100,* 1059–1069.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485,* 233–236.

Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *Journal of the Association for Research in Otolaryngology, 11,* 709–724.

Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron, 48,* 139–148.

Mickey, B. J., & Middlebrooks, J. C. (2003). Representation of auditory space by cortical neurons in awake cats. *The Journal of Neuroscience, 23,* 8649–8663.

Middlebrooks, J. C., & Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. *The Journal of Neuroscience, 33,* 10986–11001.

Middlebrooks, J. C., Clock, A. E., Xu, L., & Green, D. M. (1994). A panoramic code for sound location by cortical neurons. *Science, 264,* 842–844.

Middlebrooks, J. C., & Green, D. M. (1990). Directional dependence of interaural envelope delays. *The Journal of the Acoustical Society of America, 87,* 2149–2162.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology, 42,* 135–159.

Middlebrooks, J. C., & Onsan, Z. A. (2012). Stream segregation with high spatial acuity. *The Journal of the Acoustical Society of America, 132,* 3896–3911.

Miller, L. M., & Recanzone, G. H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proceedings of the National Academy of Sciences of the USA, 106,* 5931–5935.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica, 88,* 320–332.

Moore, B. C. J., & Gockel, H. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367,* 919–931.

Phillips, D. P. (2008). A perceptual architecture for sound lateralization in man. *Hearing Research, 238,* 124–132.

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology, 18,* 1124–1128.

Sach, A. J., & Bailey, P. J. (2004). Some characteristics of auditory spatial attention revealed using rhythmic masking release. *Perception and Psychophysics, 66,* 1379–1387.

Salminen, N. H., May, P. J., Alku, P., & Tiitinen, H. (2009). A population rate code of auditory space in the human cortex. *PLoS ONE, 26,* e7600.

Saupe, K., Keoelsch, S., & Rubsamen, R. (2010). Spatial selective attention in a complex auditory environment such as polyphonic music. *The Journal of the Acoustical Society of America, 127,* 472–480.

Schadwinkel, S., & Gutschalk, A. (2010). Activity associated with stream segregation in human auditory cortex is similar for spatial and pitch cues. *Cerebral Cortex, 20,* 2863–2873.

Schreiner, C., & Urbas, J. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hearing Research, 32,* 49–63.

Shaw, E. A. G. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America, 56,* 1848–1861.

Snyder, J., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin, 133,* 780–799.

Stainsby, T. H., Fullgrabe, C., Flanagan, H. J., Waldman, S. K., & Moore, B. C. J. (2011). Sequential streaming due to manipulation of interaural time differences. *The Journal of the Acoustical Society of America, 130,* 904–914.

Stecker, G. C., Harrington, I. A., & Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biology, 3,* 520–528.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.*, Cognitive Neuroscience Series Cambridge, MA: MIT Press.

Takanen, M., Raitio, T., Santala, O., Alku, P., & Pulkki, V. (2013). Fusion of spatially separated vowel formant cues. *The Journal of the Acoustical Society of America, 134,* 4508–4517.

Thiran, A. B., & Clarke, S. (2003). Preserved use of spatial cues for sound segregation in a case of spatial deafness. *Neuropsychologia, 41,* 1254–1261.

Tollin, D. J., Populin, L. C., Moore, J. M., Ruhland, J. L., & Yin, T. C. (2005). Sound-localization performance in the cat: The effect of restraining the head. *Journal of Neurophysiology, 93,* 1223–1234.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* PhD dissertation, Eindhoven: University of Technology.

Vliegen, J., Moore, B. C., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *The Journal of the Acoustical Society of America, 106,* 938–945.

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America, 91,* 1648–1661.

Woods, T. M., Lopez, S. E., Long, J. H., Rahman, J. E., & Recanzone, G. H. (2006). Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. *Journal of Neurophysiology, 96,* 3323–3337.

Woods, W. S., & Colburn, H. S. (1992). Test of a model of auditory object formation using intensity and interaural time difference discrimination. *The Journal of the Acoustical Society of America, 91,* 2894–2902.

Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavior. *The Journal of Neuroscience, 34,* 4396–4408.