Chapter 50
Auditory Streaming at the Cocktail Party: Simultaneous Neural and Behavioral Studies of Auditory Attention

Mounya Elhilali, Juanjuan Xiang, Shihab A. Shamma, and Jonathan Z. Simon

Abstract We present a pair of simultaneous behavioral-neurophysiological studies in human subjects, in which we manipulate subjects' attention to different features of an auditory scene. In the first study, we embed a regular acoustic target in an irregular background; in the second study, we pair competing simultaneous regular acoustic streams. Our experimental results reveal that attention to the target, rather than to the background or unattended stream, correlates with a sustained increase in the neural target representation, as measured by magnetoencephalography (MEG), beyond auditory attention's well-known transient effects on onset responses. The enhancement originates in core auditory cortex and covaries with both behavioral states. Furthermore, for the slower streams, where the rhythmic rate is commensurate with that of speech prosody, the target's perceptual detectability improves over time, correlating strongly, within subjects, with the target representation's neural buildup.

Keywords Attention • Magnetoencephalography • MEG • Buildup • Auditory scene analysis

50.1 Introduction

Due to limited processing capacity of the auditory system, only a subset of the information available at the ear can be attended and processed in more detail at the high level of the auditory system. The cognitive process involved in this selection process is called as selective attention. Recent psychoacoustic studies on selective attention have demonstrated that human listeners can allocate attention not only to a particular location, but also to a particular feature, such as modulation rate

J.Z. Simon (✉)
Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
e-mail: jzsimon@umd.edu

E.A. Lopez-Poveda et al. (eds.), *The Neurophysiological Bases of Auditory Perception*,
DOI 10.1007/978-1-4419-5686-6_50, © Springer Science+Business Media, LLC 2010

(Grimault et al. 2002), pitch (Vliegen and Oxenham 1999) or timbre (Cusack and Roberts 2000; for review, see Moore and Gockel 2002).

The neural correlates of the auditory spatial attention have been extensively investigated using dichotic paradigms, where subjects attend to a series of tone pips in one ear and ignored concurrent tone pips in the other ear (Giraud et al. 2000; Hillyard et al. 1973; Woldorff and Hillyard 1991). On the other hand, there have been a limited number of studies (e.g., Bidet-Caulet et al. 2007; Gutschalk et al. 2008) that attempted to examine the neural correlates of feature-based auditory attention. Here, we take advantage of the auditory Steady-State Response (aSSR), an electrophysiological signature of modulated sounds. In this pair of studies, either one stream in a background of maskers, or two concurrent streams at different rhythmic rates, were diotically presented to human listeners. Each stream elicits an aSSR at the corresponding modulation rate. By requiring subjects to selectively attend to one stream, the modulatory effects of rate-based selective attention can be examined.

The first study employs stimuli consisting of a tone sequence repeated at 4 Hz in the midst of random maskers (Fig. 50.1a). The second employs stimuli consisting of a two parallel tone sequences with different rhythmic rates, 4 and 7 Hz (Fig. 50.1b). For each study, in separate tasks with identical stimulus ensembles, subjects are asked to detect deviants, either in a rhythmic stream or the background maskers while magnetoencephalography (MEG) responses are recorded. The evolution of neural representation of an attended stream is correlated with the behavioral performance. In addition, the neural source locations of the aSSR are assessed.

The stimuli, and hence the tasks, between studies are very different. In the first study, the rhythmic component's percept ranges from imperceptible to quiet-but-clear. In the second study, both streams are of similar loudness: each easily perceptible at all times.

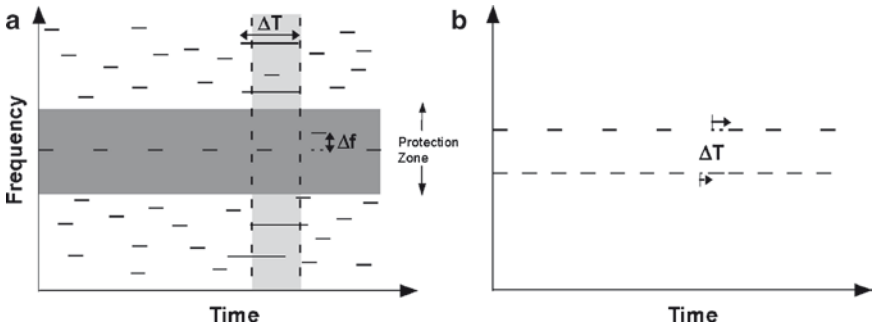

Fig. 50.1 Stimulus description. Cartoon spectrograms of typical stimuli. (**a**) The stimulus consists of a repeating target note embedded in random interferers. A 16 semitone spectral protection zone surrounds the target frequency (*gray band*). In the target task, subjects are instructed to detect a frequency shifted (Δf) deviant in the repeating target notes. In the masker task, subjects are instructed to detect a sudden temporal elongation (ΔT) of the masker notes. (**b**) The stimulus consists of two repeating target notes separated by 8 semitones. In each task, subjects are instructed to detect a temporally jittered (ΔT) deviant present in that stream

50.2 Methods

Subjects: 18 subjects participated in the first study; 28 in the second. Four subjects (respectively 2) were excluded from further analysis due to nonneural electrical artifacts or an inability to perform the tasks.

Stimuli: In the first study, sounds were 5.5 s in duration. Each stimulus contained a 75 ms target note, repeating at 4 Hz, with frequency randomly chosen in the range 250–500 Hz. The background consisted of random 75 ms tones uniformly distributed over time and log-frequency. The frequencies were randomly chosen from five octaves centered at 353 Hz, except for a 16-semitone protection zone. Fifteen exemplar stimuli were generated for each of the four condition types: no deviants; one target deviant per stimulus; one masker deviant per stimulus; and one target deviant and one masker deviant, at independent times, per stimulus. Each target deviant was the displacement (up or down) by 2 semitones. Each masker deviant was a single 500 ms time-window, in which all masker tones were elongated to 400 ms. The deviants were randomly distributed in time.

In the second study, the duration of sounds were randomly chosen from 5.25, 6.25, or 7.25 s. The spectral distance between the two streams was ±8 semitones, where the specific frequencies of each stream were randomly chosen in the range of 250–500 Hz. The loudness of each stream was approximately equal. Duration was 75 ms. Twelve exemplar stimuli were generated (and presented twice each) for each of the three condition types: no deviants; one deviant per 4 Hz stream; and one deviant per 7 Hz stream. Each deviant was the temporal displacement of a target tone by 70 ms (4 Hz stream) or 40 ms (7 Hz stream) from the regular interval. All subjects performed both tasks: T4 (4 Hz deviant detection) and T7 (7 Hz deviant detection).

In both studies, each subject performed both tasks (with order counterbalanced across subjects), and subjects were instructed to press a button held as soon as they heard the appropriate deviant.

MEG recording: Subjects were placed horizontally in a dimly lit magnetically shielded room. The signals were delivered with sound tubing attached to foam plugs inserted into the ear-canal, and presented at a comfortable loudness of approximately 70 dB SPL.

MEG recordings were made with a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan). Neural channels were denoised twice with an adaptive filter: using external reference channels (Ahmar and Simon 2005), and using the two channels with the strongest cardiac artifacts (Xiang et al. 2005).

Behavioral Analysis: The task performance was assessed by calculating a d-prime measure of performance (Kay 1993).

To investigate the buildup of the target object during a task, we divided the deviant trials according to temporal locations of the deviants. Because of the temporal uncertainty in the false alarm trials, we calculated an average false alarm rate, and combined it with the time-specific hit rate to derive a d-prime measure for each segment.

Neural Data Analysis: The concatenated neural responses were characterized by the magnitude and phase of the frequency component at the tone presentation rate and were used for localization. The remainder of the analysis was based on the normalized neural responses, defined to be the squared magnitude of the frequency component at the tone presentation rate divided by the average squared magnitude of the frequency components between 1 Hz above and below that rate, averaged over the 20 channels with the strongest normalized neural responses. The same analysis is also done at the two adjacent frequency bins, ±1/4.25 Hz or ±1/4 Hz.

To investigate the buildup of the target object in target task, the analysis epochs were divided into temporal segments, 750 ms duration for the first study, 1,000 ms for the second, extracted and concatenated. The first segment began at 1,250 ms post stimulus.

The behavioral curves for each subject were interpolated to match the sampling rate of the neural data. Subsequently, these two curves were then fitted by a line to derive the slope relating them. The slope was transformed into an angle, and combined across subjects using circular statistics to yield an angular mean (Fisher 1993). Confidence measures were then derived from the bootstrap statistics (Efron and Tibshirani 1993).

Neural Source Localization: Source localization for the neural response to the target was obtained by calculating the complex current-equivalent dipole best fitting the complex magnetic field configuration at 4 Hz peak, in each hemisphere (Simon and Wang 2005). Significance of the relative displacement between the (previously obtained) M100 and target dipole sources were determined by a two-tailed paired *t*-test in each of three dimensions.

50.3 Results

Depending on listeners' attentional focus, the neural representations of the streams mirror the percept of the scenes.

In the first study, during the performance of the target task (mean d-prime: 3.0), the rhythm of the stream emerges as a strong 4 Hz component in the neural signal of an individual subject (Fig. 50.2a, left panel). In contrast, during the performance of the masker task (mean d-prime: 3.0), the cortical response entrained at 4 Hz is relatively suppressed (Fig. 50.2a, right panel).

In the second study, during the T4 task (mean d-prime: 2.9), the rhythm of the slow stream emerges as a strong 4 Hz component for an individual subject (Fig. 50.2b, top left panel). In contrast, during the T7 task (mean d-prime: 1.8), the cortical response entrained at 4 Hz is relatively suppressed (Fig. 50.2a, top right panel). This effect is correspondingly reversed for the cortical representations of the fast stream: the neural response at 7 Hz is stronger in the T7 task than the T4 task (Fig. 50.2a, lower panels).

The neural response change between tasks was averaged across all subjects (Fig. 50.2c, d). In the first study, a significant positive (bootstrap across subjects, $p < 10^{-4}$) change at the 4 Hz aSSR is observed, reflecting enhanced phase-locked,

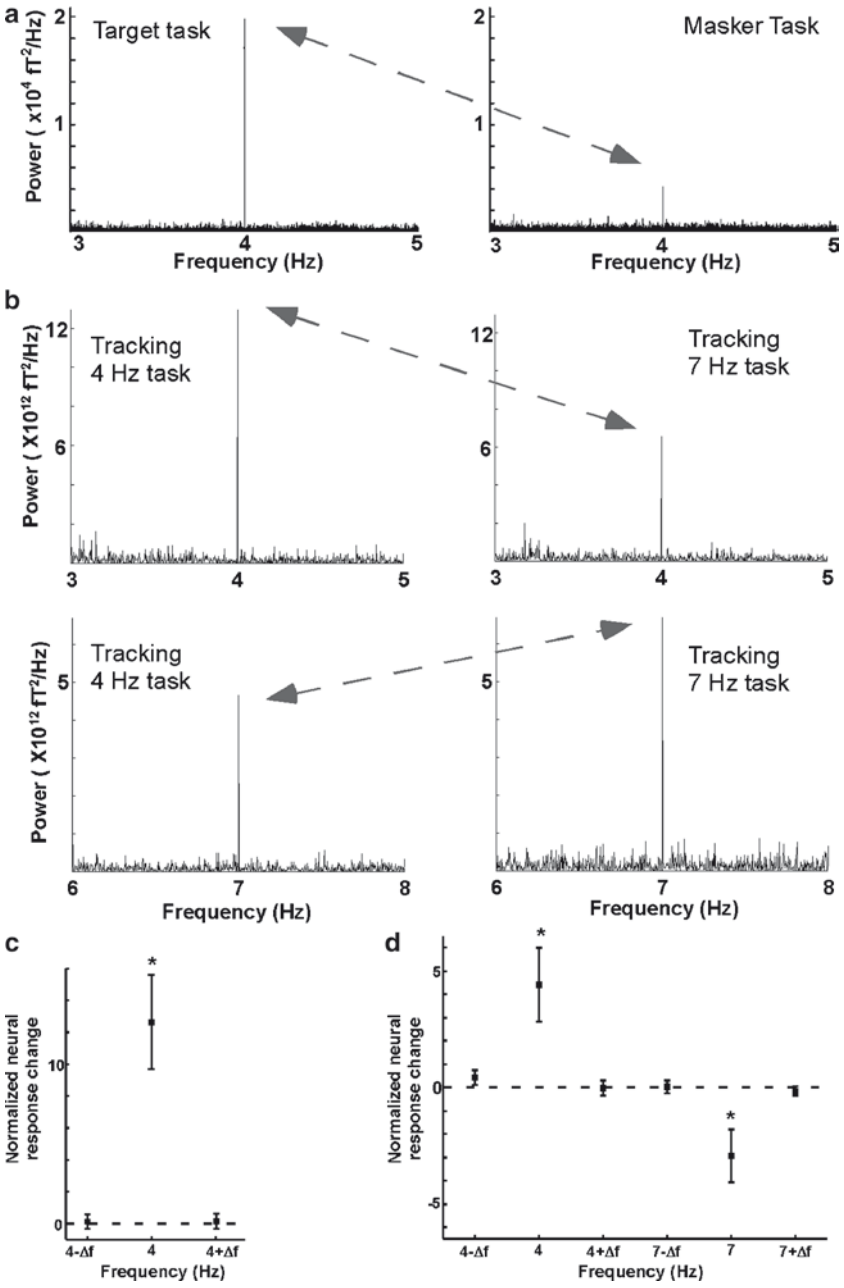

Fig. 50.2 Neural responses. (**a**) Power spectral density of MEG responses for a single subject in the first study in target (*left*) and masker (*right*) tasks, averaged over 20 channels. (**b**) Power spectral densities of MEG responses for a single subject in the second study in T4 (*left column*) and T7 (*right column*) tasks, averaged over 20 channels. (**c**, **d**) Normalized neural responses of one task relative to the other task ((**c**) target task minus masker task; (**d**) T4 task minus T7 task) shows enhancement exclusively at the frequency of the target rhythm. Each data points represents the average difference between normalized neural responses; *error bars* represent standard error

sustained activity when subjects' attention is directed toward the target stream. In the second study, a significant positive ($p < 10^{-3}$) change at the 4 Hz aSSR and a significant negative ($p < 0.002$) change at the 7 Hz aSSR are observed, reflecting an enhanced phase-locked, sustained activity when subjects' attention is directed toward the target stream. In contrast, there is no significant change in normalized neural response at adjacent frequencies, confirming that this feature-based selective attention precisely modulates the cortical representation of the specific feature, but not overall neural activities.

The neural sources of all the target rhythm response components originate in auditory cortex. In the first study, the neural source's mean displacement from the source of the auditory M100 response (Naatanen and Picton 1987) was significantly different (for the left auditory cortex only) (two-tailed *t*-test; $p = 0.017$) by 14 ± 5 mm in the anterior direction. In the second study, the total significant displacement (two-tailed *t*-test; $p = 0.016$) was 19 ± 6 mm in the anterior direction (for both hemispheres combined). Assuming an M100 origin of *planum temporale*, this is consistent with an origin for the neural response to the target rhythm in Heschl's gyrus, the site of core auditory cortex, and *not* consistent with the aSSR arising from a concatenation of periodic M100 responses.

For the 4 Hz targets in both studies, subjects' performance during the target task improves over several seconds as shown in Fig. 50.3a, b (solid lines). Moreover, the neural response to the target rhythm also displays a statistically significant buildup (Fig. 50.3a, b, dashed line) closely aligned with the behavioral curve, and, consequently, decoupled from the actual acoustics. The remarkable correspondence between these two measures strongly suggests that the enhanced perception of the target over time is mediated by an enhancement of the neural signal representation. In the second study, no such buildups are present in the 7 Hz task, which remains a puzzle (particularly so since the 4 Hz tasks across studies are so different).

We also note that the sub-segments over which the neural buildup is measured are required to span several rhythmic periods (at least three). There is no buildup using intervals with shorter durations, despite sufficient statistical power, implicating temporal phase coherence (in contrast to spatial phase coherence) as critical to the buildup of the neural target representation. The need for a longer time window cannot be attributed to increased power at 4 Hz; since this explanation would imply that using one rhythmic period would also show buildup. We have constructed a computational model (not shown) that further supports this interpretation.

50.4 Discussion

These studies build on previous work in stream segregation (Fishman et al. 2001; Gutschalk et al. 2005; Micheyl et al. 2005; Snyder et al. 2006) but keeping the physical parameters of the stimulus fixed while manipulating only the attentional state of the listeners. One major finding is that auditory attention strongly modulates the sustained neural representation of the target. This neural representation is

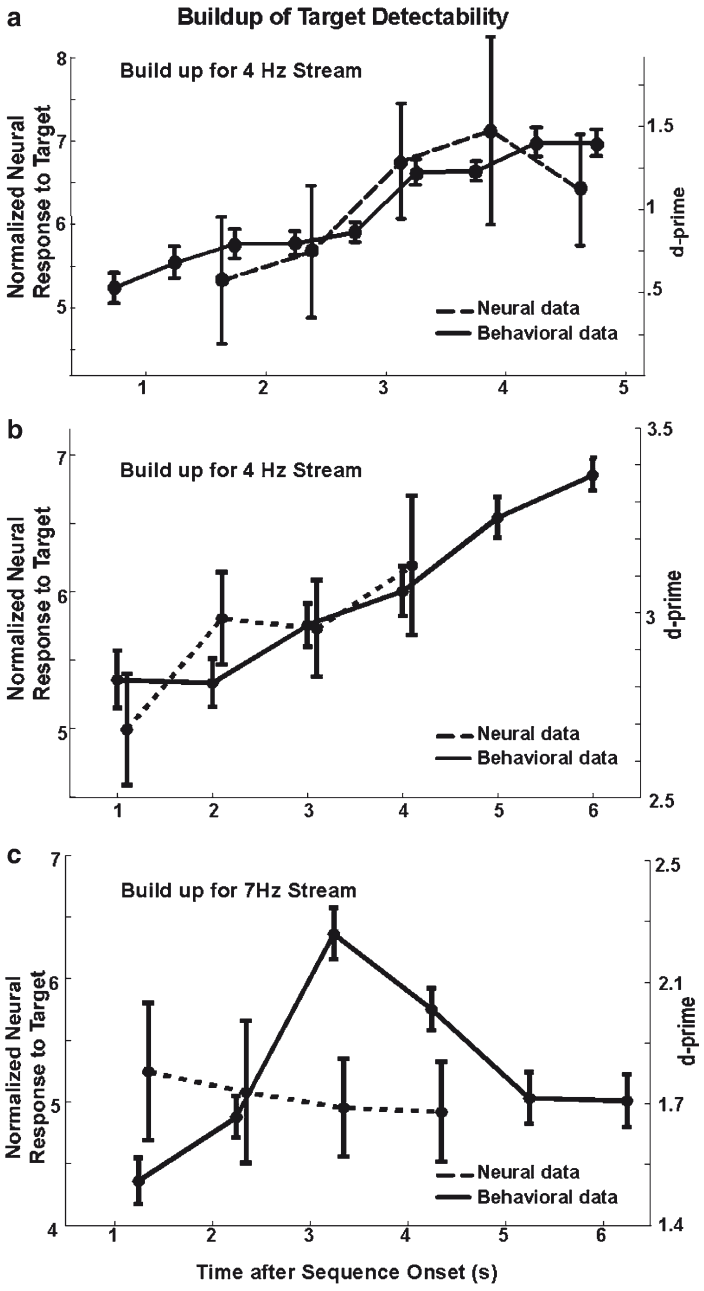

Fig. 50.3 Buildup over time of behavioral and neural responses in target task averaged over subjects (*error bars* represent standard error). (**a**) Normalized neural response to target rhythm, and behavioral performance, in the first study. (**b**, **c**) Normalized neural response to target rhythms, and behavioral performance, for the two tasks and their respective rates, in the second study

not merely a repeated M100 response, since its location is inconsistent with that interpretation, rather its location is consistent with core auditory cortex.

Finally, this study offers the first demonstration of the top-down mediated buildup over time of the neural representation of a target signal that also follows the same temporal profile of the buildup based on listeners' detectability performance in the same subject, *as long as the target is slow rather than fast* (4 Hz rather than 7 Hz). Using the current experimental paradigm, we are able to monitor the evolution in time of attentional processes as they interact with the sensory input. Many studies overlook the temporal dynamics of the neural correlates of attention, either by using cues that prime subjects to the object of attention (thereby stabilizing attention before the onset of the stimulus), or by explicitly averaging out the buildup of the neural signal in their data analysis (focusing instead on the overall contribution of attention in different situations, and not monitoring the dynamics by which the process builds up). Our findings reveal that even though the sensory target signal is unchanged, attention allows its neural representation to grow over time, closely following the time-course of the perceptual representation of the signal.

Acknowledgments Support has been provided by NIH grants R01DC008342, 1R01DC007657 and (via the CRCNS NSF/NIH joint mechanism) 1R01AG027573. We thank Jonathan Fritz and David Poeppel for comments and discussion. We are grateful to Jeff Walker for excellent technical support.

References

Ahmar N, Simon JZ (2005) MEG adaptive noise suppression using fast LMS. In: International IEEE EMBS conference on neural engineering 2005

Bidet-Caulet A, Fischer C, Besle J, Aguera PE, Giard MH, Bertrand O (2007) Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J Neurosci 27(35):9252

Cusack R, Roberts B (2000) Effects of differences in timbre on sequential grouping. Percept Psychophys 62:1112–1120

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall/CRC, New York

Fisher NI (1993) Statistical analysis of circular data. Cambridge University Press, New York

Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear Res 151:167–187

Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R et al (2000) Representation of the temporal envelope of sounds in the human brain. J Neurophysiol 84(3):1588–1598

Grimault N, Bacon SP, Micheyl C (2002) Auditory stream segregation on the basis of amplitude modulation rate. J Acoust Soc Am 111:1340–1348

Gutschalk A, Micheyl C, Melcher JR, Rupp A, Scherg M, Oxenham AJ (2005) Neuromagnetic correlates of streaming in human auditory cortex. J Neurosci 25:5382–5388

Gutschalk A, Micheyl C, Oxenham AJ (2008) Neural correlates of auditory perceptual awareness under informational masking. PLoS Biol 6(6):e138

Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical signs of selective attention in the human brain. Science 182(4108):177–180

Kay SM (1993) Fundamentals of statistical signal processing: estimation theory, Prentice-Hall, Inc. Upper Saddle River, NJ, USA

Micheyl C, Carlyon RP et al (2005) Performance measures of auditory organization. Auditory Signal Processing. Physiology, Psychoacoustics, and Models. D. Pressnitzer, A. de Cheveigne, S. McAdams and L. Collet. New York, NY, Springer: 203–11

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

Naatanen R, Picton T (1987) The N1 wave of the human electric and magnetic response to sound – a review and an analysis of the component structure. Psychophysiology 24:375–425

Simon JZ, Wang Y (2005) Fully complex magnetoencephalography. J Neurosci Methods 149(1):64–73

Snyder JS, Alain C, Picton TW (2006) Effects of attention on neuroelectric correlates of auditory stream segregation. J Cogn Neurosci 18:1–13

Vliegen J, Oxenham AJ (1999) Sequential stream segregation in the absence of spectral cues. J Acoust Soc Am 105:339–346

Woldorff MG, Hillyard SA (1991) Modulation of early auditory processing during selective listening to rapidly presented tones. Electroencephalogr Clin Neurophysiol 79(3):170–191

Xiang J, Wang Y, Simon JZ (2005) MEG responses to speech and stimuli with speechlike modulations. In: International IEEE EMBS conference on neural engineering 2005