HHS Public Access

Author manuscript

Int J Psychophysiol. Author manuscript; available in PMC 2016 February 01.

Published in final edited form as:

Int J Psychophysiol. 2015 February ; 95(2): 184–190. doi:10.1016/j.ijpsycho.2014.05.005.

The Encoding of Auditory Objects in Auditory Cortex: Insights from Magnetoencephalography

Jonathan Z. Simon[1,2,3]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, College Park, MD, 20742 USA

[2]Department of Biology, University of Maryland, College Park, College Park, MD, 20742 USA

[3]Institute for Systems Research, University of Maryland, College Park, College Park, MD, 20742 USA

Abstract

Auditory objects, like their visual counterparts, are perceptually defined constructs, but nevertheless must arise from underlying neural circuitry. Using magnetoencephalography (MEG) recordings of the neural responses of human subjects listening to complex auditory scenes, we review studies that demonstrate that auditory objects are indeed neurally represented in auditory cortex. The studies use neural responses obtained from different experiments in which subjects selectively listen to one of two competing auditory streams embedded in a variety of auditory scenes. The auditory streams overlap spatially and often spectrally. In particular, the studies demonstrate that selective attentional gain does not act globally on the entire auditory scene, but rather acts differentially on the separate auditory streams. This stream-based attentional gain is then used as a tool to individually analyze the different neural representations of the competing auditory streams.

The neural representation of the attended stream, located in posterior auditory cortex, dominates the neural responses. Critically, when the intensities of the attended and background streams are separately varied over a wide intensity range, the neural representation of the attended speech adapts only to the intensity of that speaker, irrespective of the intensity of the background speaker. This demonstrates object-level intensity gain control in addition to the above object-level selective attentional gain.

Overall, these results indicate that concurrently streaming auditory objects, even if spectrally overlapping and not resolvable at the auditory periphery, are individually neurally encoded in auditory cortex, as separate objects.

Contact information: Jonathan Z. Simon, jzsimon@umd.edu, Phone: 1-301- 405-3645, Fax: 1-301-314-9281, Mailing Address: Department of Electrical & Computer Engineering, University of Maryland, College Park MD 20742, USA.

Keywords

auditory object; MEG; neural representation; cortical representation; speech

1 Introduction

Sounds originating from multiple distinct physical elements mix linearly in the air before the mixture is neurally transduced at the auditory periphery. Despite the seemingly insurmountable obstacle that the "unmixing" of such linearly added multiple sources is a mathematically ill-posed problem, the brain routinely accomplishes this task and with little apparent effort. The method by which this problem is solved is known as *Auditory Scene Analysis* (Bregman, 1990), acknowledging the strong analogy between the perception of these auditory mixtures and of visual scenes.

Related to auditory scene analysis is the *Cocktail Party Effect* (Cherry, 1953; McDermott, 2009), in which a listener to a complex auditory scene can attend at will to a single element of the scene. The percept of the listener is that, under the influence of selective attention, the attended element ("foreground") stands out as acoustically distinct from the rest of the auditory scene, where as the rest of the auditory scene ("background") becomes correspondingly indistinct. Spatial separation of scene elements and binaural hearing can greatly benefit the listener, but is not necessary (Brungart et al., 2001; Hawley et al., 2004).

This review covers a series of results in auditory scene analysis, utilizing neural recordings made from human subjects using magnetoencephalography (MEG) in which listeners attend to a single auditory element in a complex auditory scene. The types of *attended* auditory elements employed range from repeating tone rhythms to speech. The *competing* elements in the auditory scenes include simultaneous interfering speech, spectrally matched noise, simultaneous interfering repeating tone rhythms (at a different rhythmic rate), and random spectrotemporal tone clouds (Figure 1).

The auditory scenes used in these studies are created by mixing the individual components in a single acoustic channel, which is then presented diotically (i.e., identically to each ear). This does not allow any spatial separation of the separate components to aid in their neural identification and segregation. Avoiding spatial separation removes some potential confounds when investigating the fundamental mechanisms underlying auditory scene analysis. For instance, hemispheric lateralization due to ipsilateral/contralateral processing (Ding and Simon, 2012b) might be confounded with additional processing lateralization hypothesized to be employed in neural auditory analysis (see, e.g., Poeppel, 2003). Other experimental approaches, however, can and do benefit from employing spatial separation, instead of acoustic differences, as the primary segregation cue (see, e.g., Lee et al., 2012).

1.1 Perceptual Auditory Objects

The most appropriate definition for what constitutes an auditory object is still an open question (Ahveninen et al., 2006; Alain and Arnott, 2000; Dyson, 2010; Kubovy and Van Valkenburg, 2001; Schnupp et al., 2013; Shinn-Cunningham, 2008), particularly compared to the case of vision (Cohen and Andersen, 2004; Dyson, 2010; Shamma et al., 2011). We

do not here distinguish between auditory objects and streams, following Bregman (1990), for whom auditory streams play the same role as visual objects (but this is also an open point). Auditory objects may be punctate or streaming, and they may compete serially or in parallel: only the case of parallel (simultaneous) streaming objects is addressed here. The formal definition of an auditory object employed here is that of Griffiths and Warren (2004).

From this definition, first the auditory object must correspond to something in the sensory world, e.g., the acoustic output of a single person speaking, conveyed via sound waves to the auditory periphery. Visual objects must satisfy an analogous (more intuitive) criterion: one example would be the light reflected off a person's body, arriving at the eye as a retinal image of the person. Secondly, it must be possible to separate the auditory object from the information arriving from the rest of the sensory world, e.g., from the speech of an interfering speaker or from background noise. A visual analog would be the ability to separate the arriving image of a person from competing visual elements, even in the presence of occlusion. Thirdly, the auditory object must be abstractable, or generalizable, across specific sensory experiences. The visual analog of this criterion is again more intuitive than the auditory: e.g., the percept of a visual object should not be affected by the specific two-dimensional projected image on the retina, and must be independent of the three-dimensional spatial location or spatial orientation (e.g., profile vs. frontal view). The auditory version of this property may be less intuitive, but a useful example is that speech as an auditory object should not be influenced by its absolute loudness, or by loudness relative to other objects in the scene.

In all these cases, the auditory object is a high-order perceptual construct, associated with, but distinct from, the perceived featural properties of a specific instance of the object. For instance the percept of speech *as an auditory object* is invariant under the addition of competing objects to the auditory scene and is invariant to changes in loudness and spatial location. But also associated with it is the distinct perception of the particular instance of that speech (e.g., whether it is loud or soft), which is *not* invariant to such changes. In this way we can readily perceive and report changes in the relative loudness of speech stream, even while identifying its persistence and identity as the same speech stream. The visual analogy is even more straightforward: we see a person's figure to be a unified and self-contained visual object while still being able to report changes in his or her location and orientation in space.

1.2 Neural Auditory Object Definition & Criteria

While auditory objects are defined only perceptually, the goal of these studies is to ascertain the underlying neural representations of the perceptual auditory objects (Alain and Winkler, 2012), i.e., the neural foundations of auditory objects: *neural* auditory objects. Such neural auditory objects must obey the same three criteria as their perceptual descendants.

First, there must be a neural representation of the auditory object. While the definition of "neural representation" is itself has potential issues, here we use a practical definition. We require that a neural representation allow for both the encoding of sounds into neural responses and the decoding of the neural responses into sounds. That is, if to some specified level of accuracy we can 1) predict the neural responses to a sound based on its physical

acoustics, and 2) predict the physical acoustics the sound based on the neural response to it, we will call that combination the neural representation of the sound.

Such a neural representation, in auditory cortex, of an acoustic speech stream has now been demonstrated using MEG, at least at the level of the acoustic envelope of the speech (Ding and Simon, 2012b), for modulation rates below 10 Hz. This representation assumes linear models for both the encoding and decoding (though a more detailed model might give higher accuracy): the MEG responses are linearly filtered versions of the speech envelope. Spectrotemporal generalizations (e.g., representations based on acoustic envelopes of different spectral bands) are also included.

For this case of a single speech stream, in which the entire acoustic scene is also a single auditory object, whether the observed neural representation of this stream is of the auditory object (*object-based neural representation*), or of the acoustic features of the (entire) sound impinging upon the auditory periphery (*feature-based neural representation*), or some combination, cannot be determined. If the neural representation is indeed object-based, however, it must satisfy the same criteria as must be satisfied by the perceptual auditory object under conditions where the auditory stream is only one element in a more complex auditory scene (c.f., Winkler et al., 2006). Using the three criteria above, the neural representation of this element 1) must represent something in the sensory world (e.g., the speech stream), 2) must represent the speech stream as segregated out from the remaining auditory scene and not encode other elements of the scene, and 3) must be invariant to, or at least robust against, large acoustic changes in the scene that do not change the auditory object itself (e.g., auditory analogs of visual displacements and rotations).

1.3 Speech-based Neural Auditory Objects

It is straightforward to show that a speech stream does act as a perceptual auditory object, using the three (perceptual) criteria above. The first criterion, corresponding to something in the sensory world, is already met just by having any perceptual representation of the speech stream. An example of satisfying the second criterion, being separable from the rest of the auditory information in the sensory world, is the ability to pick out one speaker among many in a multi-speaker environment, i.e., the Cocktail Party effect. An example of satisfying the third criterion, being abstractable across specific sensory experiences, is the experience of maintaining a stable auditory percept of a person’s speech from across a crowded room, which typically involves accommodating dynamic acoustic changes at the periphery, including in loudness, reverberation-induced spectrally colored reflections, interaural time difference, etc..

The object-based *neural* representation of a speech stream must satisfy analogous criteria. As discussed above, the first criterion, that there be a neural representation of the sensory-world-based speech stream (Figure 2), has been met (Ding and Simon, 2012b). The second criterion would be met if, using selective attention, it could be demonstrated that for two competing speakers there is a neural representation of the attended speech stream separate from that of the unattended speech stream (Figure 3). The third criterion would be met by showing that the same representation is invariant to changes of level (loudness) difference between the attended and unattended speech streams (Figure 4).

1.4 Cortical Hierarchy of Object Representations

Of course there are many auditory cortical areas, each of which (presumably) has its own representation(s) of the auditory scene and/or its components, suitable for extracting or processing specific kinds of information, which are then propagated to other cortical areas. In this light, one may expect to find some cortical areas for which feature-based representations dominate (as they do in the periphery), some areas for which the representations are intermediate, and other areas for which the representations are dominantly object-based. As will be shown below, the recorded MEG responses arise predominantly from two auditory areas (in each hemisphere), one of which is dominantly object-based (with post-stimulus latency ~100 ms), and the other of which is dominantly feature-based (with post-stimulus latency ~50 ms).

1.5 Object-based Neural Representations for Other Auditory Scenes

The ideas introduced above will be most fully applied to the paradigm of a subject listening to two simultaneous competing speakers, but other paradigms have been employed as well, whether using speech streams or tone streams as the auditory objects. As will be seen below, results from the paradigm of speech and interfering stationary noise exhibits strong parallels with the case of two simultaneous competing speakers. We also discuss two paradigms using tone-stream based objects, one in the presence of a competing tone-stream at a different rhythmic rate, and the other in the presence of a spectrotemporal cloud of random masking tones. In all these cases, when the two auditory objects are asymmetric (speech + noise or tone-stream + tone-cloud), we characterize only one neural representation (of the speech or of the tone-stream), but when the auditory objects are symmetric (competing speech streams or competing tone streams) we take advantage and then characterize two neural representations.

For all these paradigms, the first two of the neural auditory object criteria have now been demonstrated; the third has only been established in the case of two competing speech streams. This wide range of stimulus types demonstrates that the general phenomena are not restricted to only simple sounds, nor are they restricted to the special case of speech.

Several recent reviews detail a variety of alternate approaches (Lee et al., 2013; Snyder et al., 2012).

2 Results

2.1 Competing Speech Streams

As reported in Ding and Simon (2012a), for subjects listening to two speakers of opposite gender but attending only to one, the envelope of the attended speech stream can indeed be reconstructed approximately from the MEG responses. At the group level, the reconstructed envelope is significantly ($p < 0.001$) more correlated with the envelope of the attended speaker than of the unattended speaker. In fact even at the individual trial level, the reconstructed envelope is more strongly correlated with the envelope of the attended speaker than of the unattended speaker in 92% of trials. A separate experiment with subjects listening to two speakers of the same gender produced results that were both qualitatively

and quantitatively similar. This demonstrates that the second criteria necessary for the neural representation of the speech stream to also be a representation of an auditory objects has been met, by showing that the neural representation of the attended speech stream is separate from that of the unattended speech stream.

The third criterion is tested by the presence or absence of changes in the neural representation of the speech stream when one, or the other, of the speech streams is attenuated in level by up to 8 dB (16 dB total range). In this case the result is that decoding the attended speech is unaffected by this broad range of relative level changes, thereby satisfying this criterion.

Beyond satisfying the three object criteria, additional insight into the origin of this object-based neural representation of the attended speech stream can be gleaned from examining the neural *encoding* of the speech stream (in contradistinction to the neural *decoding* of the neural response thus far analyzed). A linear encoding from acoustic envelope to neural response waveform is accomplished by convolving the acoustic envelope with a mathematical kernel function, the Temporal Response Function (TRF), to produce a predicted neural response. The TRF is estimated by reverse-correlating the acoustic envelope with the actual (not predicted) neural response waveform (David et al., 2007). Separate TRFs are generated by reverse-correlating with different speech stream components (e.g., attended vs. unattended, or attenuated vs. unattenuated). Because the TRF can also be thought of as an Impulse Response from linear filter theory, its positive and negative peaks can be interpreted (and localized) much as the positive and negative peaks of a response to an impulsive stimulus (e.g., tone pip) can.

In this case the result is that there are two prominent peaks. The first has the same (positive) polarity and similar latency as the analogous M50 (or P1m) response, and the second, larger, peak has the same (negative) polarity and similar latency as the analogous M100 (or N1m) response (Chait et al., 2004). As in the case of the M50 and M100, these neural generators can be localized to sources consistent with, respectively, Heschl's Gyrus (HG) (which contains Primary Auditory Cortex), and more posteriorly, Planum Temporale (PT) (which contains higher order auditory areas). Critically, the early peak is not modulated by selective attention, whereas the later peak is strongly positively modulated by selective attention. Equally critically, this same later peak is not modulated by changing the relative loudness of the attended speech stream. In short, the later peak, whose source is in PT, by being strongly modulated by attention but not by relative loudness, satisfies all three criteria necessary for the representation to be that of an auditory object. In contrast, the earlier peak, which does not distinguish between the attended and unattended streams, fails the second criterion, and so it cannot be a representation of an auditory object, but rather must be more featurally based. In contrast to the case of monaurally and dichotically presented stimuli (Ding and Simon, 2012b), no specialization by hemisphere was noted.

2.2 Speech in Noise

Ding and Simon (2013) describe a study where subjects listen to speech corrupted by different levels of spectrally-matched stationary noise (from +6 dB to −9 dB SNR). The accuracy of the reconstructions of the underlying speech envelope from the MEG response

waveform remains nearly constant across all SNRs tested (except −9 dB where the speech is barely audible in the noise, which is itself more than twice as strong as the speech), suggesting a stable neural representation of speech maintained by contrast gain control. This again meets the second criterion for the speech stream being an auditory object, that the neural representation of the speech stream be separable from other auditory information, in this case from the noise (as long as the speech stream remains audible). It has also been demonstrated that the decoding accuracy of the slowest (delta band) neural modulations predicts how well individual subjects recognize speech in noise. This shows that the ability to extract information about the slowest acoustic modulations of speech from the noise likely represents a serious potential bottleneck in the speech recognition process.

TRF analysis again shows an earlier M50-like peak and a later M100-like peak. The early peak is strongly modulated by the presence of noise, indicating that its neural source is a feature-based representation of the auditory scene. In contrast, the later peak is unaffected by the noise, indicating an object-based representation of the speech stream (again until −9 dB, when the peak suddenly disappears for this case where the speech stream is no longer audible in the noise). Again, no hemispheric specialization was noted.

2.3 Tone Streams

Xiang et al. (2010) report on a study where subjects listen to competing tone streams at different rates (4 Hz vs. 7 Hz) while attending to only one. In this simplified (compared to speech) case, the neural representations of the different tone streams are simply the strength and phase of the neural response at the relevant frequency. The representation of each stream is again positively modulated by attention, meeting the second criterion by having the neural representations be separable from each other under selective attention. It is additionally shown that there is significant correlation across subjects between changes in the behavioral ability to attend to the tone stream (as measured by in-stream deviant detection rate) and changes in the attentional modulation of the neural representation, under several manipulations of task difficulty. Unlike in the speech case, significant hemispheric asymmetry is observed (even though the stimulus is diotic). The neural representation of the attended stream, regardless of rate, is significantly enhanced in the right hemisphere over the left hemisphere. The neural representation of the unattended stream is also enhanced in the right hemisphere over the left (though less than for the attended case), but significantly so only at 7 Hz (marginally at 4 Hz).

Elhilali et al. (2009) report a study where subjects listen to a tone stream at 4 Hz competing with a spectrotemporally random tone cloud. The attentional modulation of the neural representation of the tone stream is so robust that it is observable in every one of the 14 subjects, again demonstrating the second criterion, that the neural representation of the tone stream is separable from its background under selective attention. As in the competing tone stream experiment, there is also significant correlation across subjects between changes in the behavioral ability to attend to the tone stream (as measured by in-stream deviant detection rate) and changes in the attentional modulation of the neural representation, under several manipulations of task difficulty. Again a significant hemispheric asymmetry was observed, though in the opposite direction than in the competing tone-stream case, reflecting

the different nature of the attentional tasks in the two cases (searching for the tone stream hidden in the tone cloud, vs. choosing one of two strongly salient tone streams). In this case the representation of the unattended tone stream was, as above, enhanced in the right hemisphere over the left, but the asymmetry was reversed when the tone stream was attended.

3 Discussion & Conclusion

By recording MEG responses from human subjects listening to complex acoustic scenes, we have demonstrated that neural representations of perceptual auditory objects are readily found. The auditory objects can be as simple as rhythmic tone streams or as complex as speech. These observed object-based neural representations are separable from the remainder of the acoustic scene via selective auditory attention, under a variety of acoustic backgrounds ranging from stationary noise, to tone clouds and streams, to competing speech. The object-based representation appears to be fully formed by 100 ms post stimulus latency (in the sense of reverse-correlation time delay between an envelope peak and the corresponding neural response) in posterior auditory cortex, but not at 50 ms in medial auditory cortex.

One question unanswered by the studies reviewed above is whether the MEG response simply tracks the envelope of the auditory object or, alternatively, reflects the collective responses of different neural networks representing the decomposition of the object into fundamental acoustic features. If the latter turns out to be the case, then the MEG signal may provide true insight as to the neural mechanism by which the auditory scene is decomposed into its constituent auditory objects. One way of addressing this question would be to use a family of stimuli with the same acoustic envelopes but differing spectrotemporal carriers (this is partially addressed in Ding et al., 2014).

Finally, similar, parallel evidence for a hierarchy of auditory representations from feature-based to object-based has recently been reported using electrocorticography (ECoG) recordings from auditory cortex. The results of that study indicate that lower order auditory areas represent both attended and background speech streams, but higher order auditory areas preferentially represent only the attended speech stream (Zion Golumbic et al., 2013).

4 Materials and Methods

This section contains an abbreviated description of the methods employed in the studies whose results are described above.

4.1 Participants

All MEG experiment participants were right-handed (Oldfield, 1971), and reported normal hearing with no history of neurological disorder. The experimental procedures were approved by the appropriate university institutional review board. Written informed consent was obtained from each subject before an experiment.

4.2 MEG Recordings

In the MEG studies, subjects lay horizontally in a dimly lit magnetically shielded room (Yokogawa Electric Corporation). Stimuli were presented using Presentation software (Neurobehavioral Systems). The signals were delivered to the subjects' ears with 50 Ω sound tubing (E-A-RTONE 3A; Etymotic Research), attached to E-A-RLINK foam plugs inserted into the ear canal, and presented at a comfortable loudness of ~70 dB sound pressure level.

MEG recordings were conducted using a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan). Its detection coils are arranged in a uniform array on a helmet-shaped surface of the bottom of the dewar, with ~25 mm between the centers of two adjacent 15.5-mm-diameter coils. Sensors are configured as first-order axial gradiometers with a baseline of 50 mm; their field sensitivities are 5 fT/√Hz or better in the white noise region. Three of the 160 channels are magnetometers separated from the others and used as reference channels for noise-filtering methods. The magnetic signals were bandpassed between 1 and 200 Hz, notch filtered at 60 Hz, and sampled at the rate of f_s = 1000 Hz. All neural channels were denoised either using Block-LMS (Ahmar and Simon, 2005; Ahmar et al., 2005) or TSPCA (de Cheveigne and Simon, 2007) and SNS (de Cheveigne and Simon, 2008b).

Before each main experiment, a pre-experiment was run in which a 50 ms tone pip was presented ~100 or ~200 times, with interstimulus interval (ISI) randomized between 750 and 1550 ms, and subjects were instructed to count the tone pips. This was to record the M100 response (a prominent peak ~100 ms after pip onset, also called N1m) used for differential source localization since its neural source is easy to localize within auditory cortex (Lutkenhoner and Steinstrater, 1998).

4.3 MEG Experiments using Auditory Scenes with Speech

The speech stimuli were taken from narrations of *Alice's Adventures in Wonderland* by Lewis Carroll, and *A Child's History of England* by Charles Dickens, obtained from the LibriVox public domain library of audiobooks (http://librivox.org). The sound recordings were low-pass filtered below 4 kHz and divided into 50- or 60-second duration sections, after long speaker pauses (> 300 ms) were shortened to 300 ms (in order to keep the speech streams flowing continuously) and presented diotically. In the competing speaker experiments (Ding and Simon, 2012a), the stimulus was a mixture of two spoken narratives from different speakers (of opposite gender), either with equal RMS value or different RMS values (with one or the other speaker reduced by 0, 5, or 8 dB). In the speech-with-noise experiment (Ding and Simon, 2013), the speech was mixed with spectrally matched stationary noise at one of six SNRs, i.e., quiet (no noise added in), +6 dB, +2 dB, −3 dB, −6 dB, and −9 dB. In all experiments, subjects were asked factual questions about the story (to encourage attentiveness), and also to subjectively rate what percentage of words were correctly recognized after the first listening to each stimulus. Denoising source separation (DSS) was used for MEG response component analysis (de Cheveigne and Simon, 2008a).

4.4 MEG Experiments using Auditory Scenes with Tone Streams

The tone-stream stimuli consisted of two auditory elements presented diotically, the first of which was a single-frequency stream of tones (75 ms duration) repeating at 4 Hz, with its tonal frequency randomly drawn from the range 250–500 Hz in two semitone intervals. In the competing stream experiment (Xiang et al., 2010), the other auditory component was an additional single-frequency stream of tones, but repeating at 7 Hz, and with a frequency differing by ±8 semitones such that it was in the same 250–500 Hz range. The stimuli duration was randomly chosen from 5.25, 6.25, or 7.25 s uniformly. In the stream-plus-cloud experiment (Elhilali et al., 2009), the other auditory component was a “cloud” of random tones at a density of 50 tones/s, uniformly distributed over time and log-frequency (except for a spectral protection region). The frequencies of the random notes were randomly chosen from the five-octave range centered at 353 Hz, in two semitone intervals, with the constraint that no masker components were permitted within eight semitones around the target frequency (the spectral protection region half-width). Each trial was 5.5 s in duration. Deviants were randomly inserted in each auditory component in such a way as to be difficult to detect when the subject was attending to the other auditory component. The subject’s task was, for each of two blocks, to detect deviants in only one of the two components, with deviants in the other component detected in the other block (order counterbalanced across subjects). In this way subjects were given incentive to attend to only one element per block, and their ability to detect the appropriate deviants was used as a measure of their ability to attend to the instructed stream. Neural representations of the tone streams were analyzed in the Fourier domain, i.e., by their complex responses (both amplitude and phase) at the frequencies of interest, 4 Hz and 7 Hz (Simon and Wang, 2005).

Acknowledgments

This research was supported by the National Institute of Deafness and Other Communication Disorders Grant R01-DC-008342.

References

Ahmar, N.; Simon, JZ. MEG Adaptive Noise Suppression using Fast LMS. International IEEE EMBS Conference on Neural Engineering; 2005; 2005. p. 29-32.

Ahmar, N.; Wang, Y.; Simon, JZ. Significance Tests for MEG Response Detection. International IEEE EMBS Conference on Neural Engineering; 2005; 2005. p. 21-24.

Ahveninen J, Jääskeläinen IP, Raij T, Bonmassar G, Devore S, Hämäläinen M, Levänen S, Lin FH, Sams M, Shinn-Cunningham BG, Witzel T, Belliveau JW. Task-modulated “what” and “where” pathways in human auditory cortex. Proc Natl Acad Sci U S A. 2006; 103:14608–14613. [PubMed: 16983092]

Alain C, Arnott SR. Selectively attending to auditory objects. Front Biosci. 2000; 5:D202–D212. [PubMed: 10702369]

Alain, C.; Winkler, I. The Human Auditory Cortex. Springer; 2012. Recording event-related brain potentials: Application to study auditory perception; p. 69-96.

Bregman, AS. Auditory scene analysis: the perceptual organization of sound. The MIT Press; Cambridge: 1990.

Brungart DS, Simpson BD, Ericson MA, Scott KR. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am. 2001; 110:2527–2538. [PubMed: 11757942]

Chait M, Simon JZ, Poeppel D. Auditory M50 and M100 responses to broadband noise: functional implications. Neuroreport. 2004; 15:2455–2458. [PubMed: 15538173]

Cherry EC. Some Experiments on the Recognition of Speech, with One and with 2 Ears. Journal of the Acoustical Society of America. 1953; 25:975–979.

Cohen, YE.; Andersen, RA. Multimodal spatial representations in the primate parietal lobe. In: Spence, C.; Driver, J., editors. Crossmodal space and crossmodal attention. Oxford University Press; USA: 2004. p. 154-176.

David SV, Mesgarani N, Shamma SA. Estimating sparse spectro-temporal receptive fields with natural stimuli. Network. 2007; 18:191–212. [PubMed: 17852750]

de Cheveigne A, Simon JZ. Denoising based on time-shift PCA. J Neurosci Methods. 2007; 165:297–305. [PubMed: 17624443]

de Cheveigne A, Simon JZ. Denoising based on spatial filtering. J Neurosci Methods. 2008a; 171:331–339. [PubMed: 18471892]

de Cheveigne A, Simon JZ. Sensor noise suppression. J Neurosci Methods. 2008b; 168:195–202. [PubMed: 17963844]

Ding N, Chatterjee M, Simon JZ. Robust Cortical Entrainment to the Speech Envelope Relies on the Spectro-temporal Fine Structure. NeuroImage. 2014; 88:41–46.

Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A. 2012a; 109:11854–11859. [PubMed: 22753470]

Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol. 2012b; 107:78–89. [PubMed: 21975452]

Ding N, Simon JZ. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci. 2013; 33:5728–5735. [PubMed: 23536086]

Dyson BJ. Auditory organization. The Oxford Handbook of Auditory Science: Hearing. 2010; 3:177.

Elhilali M, Xiang J, Shamma SA, Simon JZ. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol. 2009; 7:e1000129. [PubMed: 19529760]

Griffiths TD, Warren JD. What is an auditory object? Nat Rev Neurosci. 2004; 5:887–892. [PubMed: 15496866]

Hawley ML, Litovsky RY, Culling JF. The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. J Acoust Soc Am. 2004; 115:833–843. [PubMed: 15000195]

Kubovy M, Van Valkenburg D. Auditory and visual objects. Cognition. 2001; 80:126.

Lee AK, Larson E, Maddox RK, Shinn-Cunningham BG. Using neuroimaging to understand the cortical mechanisms of auditory selective attention. Hear Res. 2013

Lee AK, Rajaram S, Xia J, Bharadwaj H, Larson E, Hamalainen MS, Shinn-Cunningham BG. Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. Front Neurosci. 2012; 6:190. [PubMed: 23335874]

Lutkenhoner B, Steinstrater O. High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiol Neurootol. 1998; 3:191–213. [PubMed: 9575385]

McDermott JH. The cocktail party problem. Curr Biol. 2009; 19:R1024–1027. [PubMed: 19948136]

Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia. 1971; 9:97–113. [PubMed: 5146491]

Poeppel D. The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. Speech Commun. 2003; 41:245–255.

Schnupp, JWH.; Honey, C.; Willmore, BDB. Neural Correlates of Auditory Object Perception. In: Cohen, YE.; Popper, AN.; Fay, RR., editors. Neural Correlates of Auditory Cognition. Springer; New York: 2013.

Shamma SA, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 2011; 34:114–123. [PubMed: 21196054]

Shinn-Cunningham BG. Object-based auditory and visual attention. Trends Cogn Sci. 2008; 12:182–186. [PubMed: 18396091]

Simon JZ, Wang Y. Fully complex magnetoencephalography. J Neurosci Methods. 2005; 149:64–73. [PubMed: 16026851]

Snyder JS, Gregg MK, Weintraub DM, Alain C. Attention, awareness, and the perception of auditory scenes. Front Psychol. 2012; 3:15. [PubMed: 22347201]

Winkler I, van Zuijen TL, Sussman E, Horvath J, Naatanen R. Object representation in the human auditory system. Eur J Neurosci. 2006; 24:625–634. [PubMed: 16836636]

Xiang J, Simon J, Elhilali M. Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. J Neurosci. 2010; 30:12084–12093. [PubMed: 20826671]

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron. 2013; 77:980–991. [PubMed: 23473326]

Highlights

- Selective attentional gain operates on individual components of an auditory scene
- Intensity gain control operates on individual components of an auditory scene
- Necessary criteria for a neural representation of a perceptual auditory object are met
- Auditory scene components investigated range from repeating tone rhythms to speech

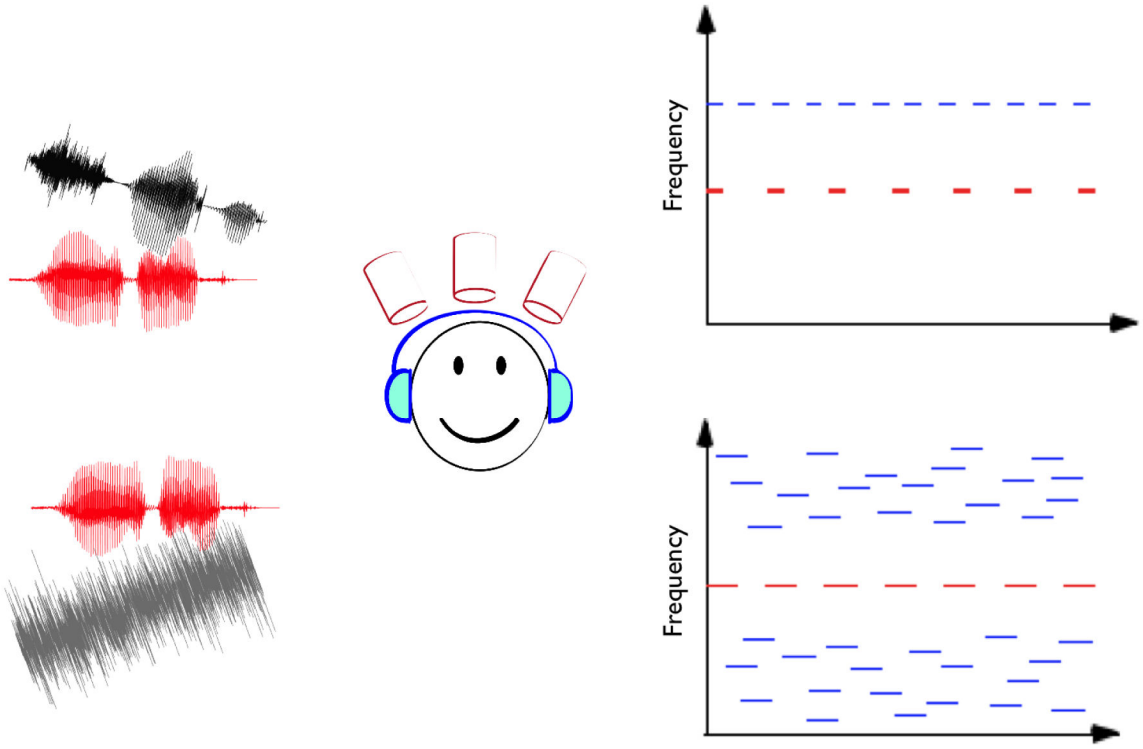

Figure 1.

Schematic of stimuli used in described experiments for listener during MEG recording. Left: Speech with competing speech, or speech with competing stationary noise. Right: Tone Stream with competing tone stream, or tone stream with competing tone cloud.

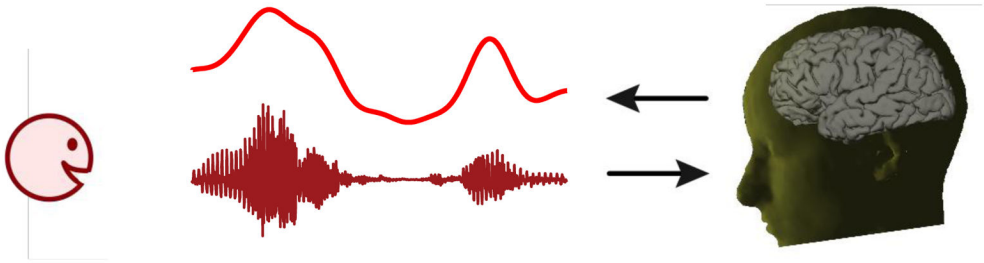

Figure 2.
Schematic example of satisfying the first criterion for a neural auditory object, a neural representation of something that exists in the sensory world. In this case it is (a filtered version of) the neural response phase-locked to the envelope of a speech stream.

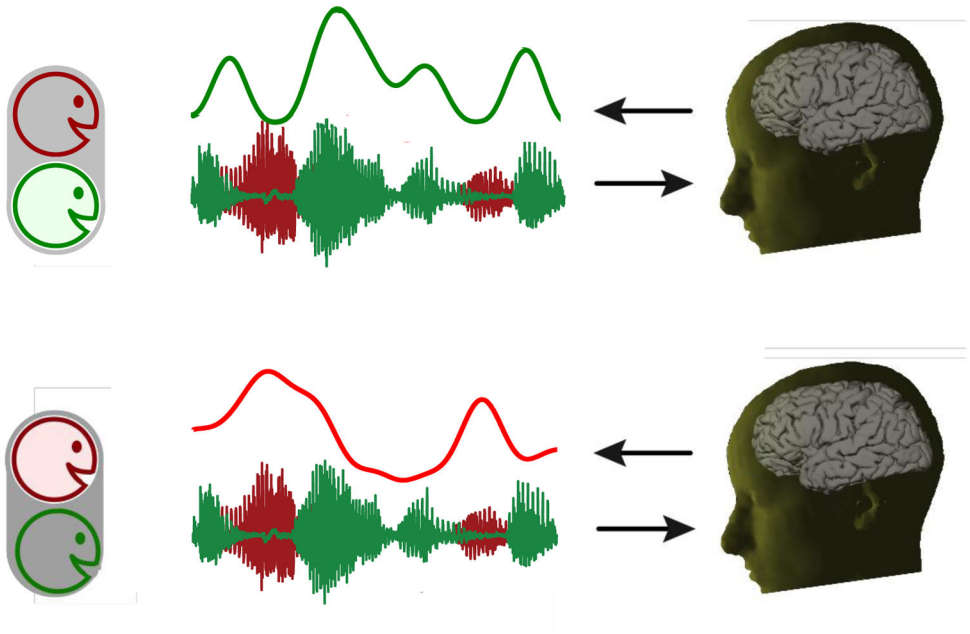

Figure 3.

Schematic example of satisfying the second criterion for a neural auditory object, that the neural representation must represent the object segregated out from the remaining auditory scene, not encoding the other elements of the scene. In this case the (filtered version of the) neural response is phase-locked to the envelope of the attended speech stream, not the unattended. It must be the same filter that applies to both attentional states.

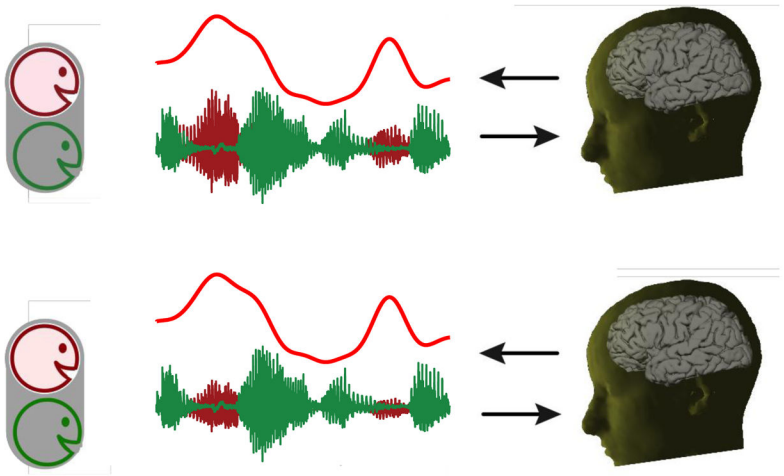

Figure 4.

Schematic example of satisfying the third criterion for a neural auditory object, that the neural representation must be invariant to, or at least robust against, large acoustic changes in the scene that do not change the auditory object itself. In this case the (filtered version of the) neural response phase-locked to the envelope of the attended speech stream is unchanged by attenuating the speech stream being attended to (as long as the attenuation is not so strong as to render the speech impossible to attend to).