Chimaeric sounds reveal dichotomies in auditory perception

Zachary M. Smith*†‡, Bertrand Delgutte*†‡ & Andrew J. Oxenham†‡

* *Eaton-Peabody Laboratory, Massachusetts Eye & Ear Infirmary, Boston, Massachusetts 02114, USA*
† *Research Laboratory of Electronics; and ‡ Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

By Fourier's theorem[1], signals can be decomposed into a sum of sinusoids of different frequencies. This is especially relevant for hearing, because the inner ear performs a form of mechanical Fourier transform by mapping frequencies along the length of the cochlear partition. An alternative signal decomposition, originated by Hilbert[2], is to factor a signal into the product of a slowly varying envelope and a rapidly varying fine time structure. Neurons in the auditory brainstem[3–6] sensitive to these features have been found in mammalian physiological studies. To investigate the relative perceptual importance of envelope and fine structure, we synthesized stimuli that we call 'auditory chimaeras', which have the envelope of one sound and the fine structure of another. Here we show that the envelope is most important for speech reception, and the fine structure is most important for pitch perception and sound localization. When the two features are in conflict, the sound of speech is heard at a location determined by the fine structure, but the words are identified according to the envelope. This finding reveals a possible acoustic basis for the hypothesized 'what' and 'where' pathways in the auditory cortex[7–10].

Combinations of features from different sounds have been used in the past to produce new, hybrid sounds for use in electronic music[11,12]. Our aim in combining features from different sounds was to study the perceptual relevance of the envelope and fine structure in different acoustic situations. To synthesize auditory chimaeras, two sound waveforms are used as inputs. A bank of band-pass filters is used to split each sound into 1 to 64 complementary frequency bands spanning the range 80–8,820 Hz. Such splitting into frequency bands resembles the Fourier analysis performed by the cochlea and by processors for cochlear implants. The output of each filter is factored into its envelope and fine structure using the Hilbert transform (see Methods). The envelope of each filter output from the first sound is then multiplied by the fine structure of the corresponding filter output from the second sound. These products are finally summed over all frequency bands to produce an auditory chimaera that is made up of the envelope of the first sound and the fine structure of the second sound in each band. The primary variable in this study is the number of frequency bands, which is inversely related to the width of each band. A block diagram of chimaera synthesis is shown in Fig. 1 with example waveforms for a single frequency band. For an audio demonstration of auditory chimaeras, see ref. 13.

Speech is a robust signal that can be perturbed in many different ways while remaining intelligible[14,15]. Speech chimaeras were created by combining either a speech sentence and noise or by combining two separate speech sentences. The speech material comprised sentences from the Hearing in Noise Test (HINT)[16]. Speech–noise chimaeras were synthesized from individual HINT sentences and spectrally matched noise. These chimaeras contain speech information in either their envelope or their fine structure. Speech–speech chimaeras were synthesized from two different HINT sentences of similar duration. The envelope of each speech–speech chimaera contains information about one utterance; its fine structure contains information about another.

Listening tests with speech–noise chimaeras showed that speech reception is highly dependent on the number of frequency bands used for synthesis (Fig. 2). When speech information is contained solely in the envelope, speech reception is poor with one or two frequency bands and improves as the number of bands increases. Good performance (>85% word recognition) is achieved with as few as four frequency bands, consistent with previous findings that bands of noise modulated by speech envelope can produce good speech reception with very limited spectral information[17]. In contrast, when speech information is only contained in the fine structure, speech reception is generally better with fewer frequency bands. The best performance is achieved with two bands; performance then deteriorates as the number of bands increases until, with eight or more bands, there is essentially no speech reception. Good performance with one and two frequency bands of fine structure is consistent with previous findings that peak-clipping (which flattens out the envelope) does not severely degrade speech reception[14]. Poorer performance with increasing numbers of bands is consistent with the auditory system's insensitivity to the fine structure of critical-band signals at high frequencies[18].

Speech–speech chimaeras measure the relative salience of the speech information transmitted through the envelope and fine structure when the two types of information are conflicting. Even though speech–speech chimaeras are constructed with two distinct

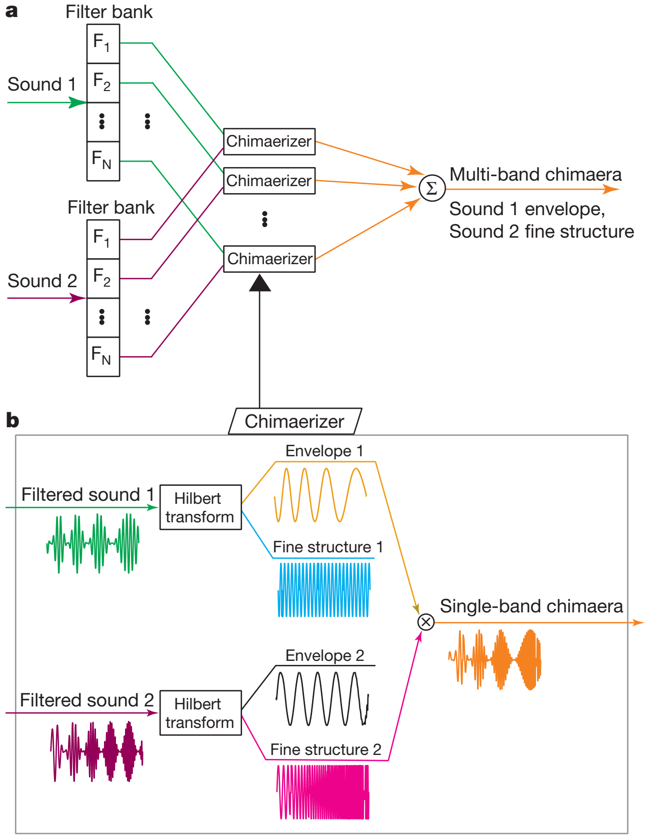

Figure 1 Auditory chimaera synthesis. **a**, Two sounds are used as input. Each sound is split into *N* complementary frequency bands with a perfect-reconstruction filter bank. Filtered signals from matching frequency bands are processed through a chimaerizer, which exchanges the envelope and the fine time structure of the two signals, producing a single-band chimaera. Partial chimaeras are summed over all frequency bands to produce a multi-band chimaera. **b**, Example waveforms within a chimaerizer, where band-limited input signals are factored into their envelope and fine structure using the Hilbert transform. A single-band auditory chimaera is made from the product of envelope 1 and fine structure 2.

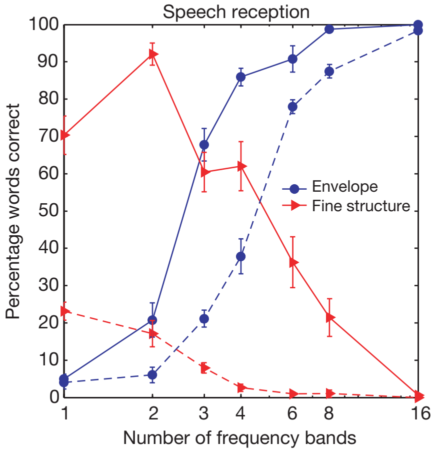

Figure 2 Speech reception of sentences in the envelope and fine structure of auditory chimaeras. Speech–noise chimaeras (solid lines) only contain speech information in either the envelope or the fine structure. Speech–speech chimaeras (dashed lines) have conflicting speech information in the envelope and the fine structure.

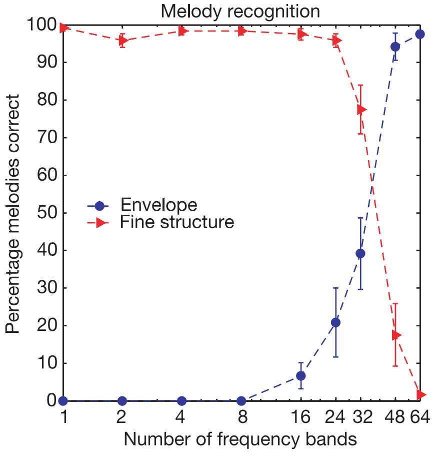

Figure 3 Recognition of melodies in the envelope and in the fine structure of auditory chimaeras. Melody–melody chimaeras contain conflicting melodies in the envelope and fine structure.

utterances, listeners almost invariably heard words from only one of the two sentences. In general, the speech information contained in the envelope dominated the information contained in the fine structure. Speech reception based on the fine structure was much poorer with speech–speech chimaeras than with speech–noise chimaeras, and was only above chance with one and two frequency bands. Speech reception based on envelope information was also degraded, but not as severely, and performance still exceeded 80% with eight or more frequency bands. Thus, envelope information in speech is more resistant to conflicting fine-structure information from another sentence than vice versa.

We used a melody recognition task to assess pitch perception of complex harmonic sounds. For this purpose, we synthesized chimaeras based on two different melodies, one in the envelope and the other in the fine structure. Melody–melody chimaeras show a reversal in the relative importance of envelope and fine structure when compared to speech–speech chimaeras (Fig. 3). Listeners always heard the melody based on the fine structure with up to 32 frequency bands. With 48 and 64 frequency bands, however, they identified the envelope-based melody more often than they did the melody based on fine structure. In responses to these melody–melody chimaeras, subjects sometimes reported hearing two melodies and often picked the two melodies represented in the envelope and the fine structure, respectively. This can be seen in the data with 16 to 48 frequency bands where scores based on envelope and fine structure add up to more than 100% correct.

The crossover point, where the envelope begins to dominate over the fine structure, occurs for a much higher number of frequency bands (about 40) for melody–melody chimaeras than it does for speech–speech and speech–noise chimaeras, further suggesting that speech reception depends primarily on envelope information in broad frequency bands. For melody recognition, this crossover occurs approximately when the bandwidths of the bandpass filters become narrower than the critical bandwidths. For such narrow bandwidths, precise information about the frequency of each spectral component, and hence the overall pitch, is available in the spectral distribution of the envelopes.

Sound localization in the horizontal plane is based on interaural differences in time and level. Interaural time differences (ITD) are the dominant cue for low-frequency sounds such as speech[19]. A delay of 700 μs was introduced into either the right or left channel of each HINT sentence to create ITDs that would produce completely lateralized sound images. To synthesize dichotic chimaeras, a sentence with an ITD pointing to the right was combined with a sentence having an ITD pointing to the left to produce a chimaera with its envelope information pointing to one side and its fine structure pointing to the other side. Two types of dichotic chimaeras were constructed, one using the same sentence for both the envelope and fine structure, and the other using different sentences. Lateralization of dichotic chimaeras was always based on the ITD of the fine structure (Fig. 4), consistent with results using non-speech stimuli[20,21]. Chimaeras synthesized with a small number of frequency bands were difficult to lateralize, but lateralization improved with increasing number of bands. Dichotic chimaeras based on the same sentence in the envelope and fine structure were more easily lateralized than those based on two different sentences.

When dichotic chimaeras based on different sentences were presented, listeners were asked to pick which of the two sentences they heard in addition to reporting the lateral position of the sound image. Consistent with our results for speech–speech chimaeras, subjects most often heard the sentence based on the fine structure with one and two frequency bands, whereas they heard the sentence based on the envelope for four or more bands. With eight or more frequency bands, subjects clearly identified the sentence based on the envelope but lateralized the speech to the side to which the fine structure was pointing. Thus the fine structure determines ‘where’ the sound is heard, whereas the envelope determines ‘what’ sentence is heard. In this respect, auditory chimaeras are consistent with evidence for separate ‘where’ and ‘what’ pathways in the auditory cortex[7–10].

The Hilbert transform provides a mathematically rigorous definition of envelope and fine structure free of arbitrary parameters. The squared Hilbert envelope contains frequencies up to the bandwidth of the original signal. Thus, for low numbers of frequency bands (less than about six), and hence large filter bandwidths, the fluctuations in the envelope can become rather rapid, making the functional distinction between fine structure and envelope based on fluctuation rate difficult. One of our most important findings is that the perceptual importance of the envelope increases with the number of frequency bands, while that of the fine structure diminishes (Figs 2 and 3). Had we used a different technique (such as rectification followed by low-pass filtering) to extract a smoother envelope, the smooth envelope would have contained even less information for small numbers of frequency bands, and therefore its perceptual importance would probably have been even smaller. Furthermore, a previous study[17] has indicated that eliminating all rapid envelope fluctuations from bands of noise modulated by speech envelope has little or no effect on speech

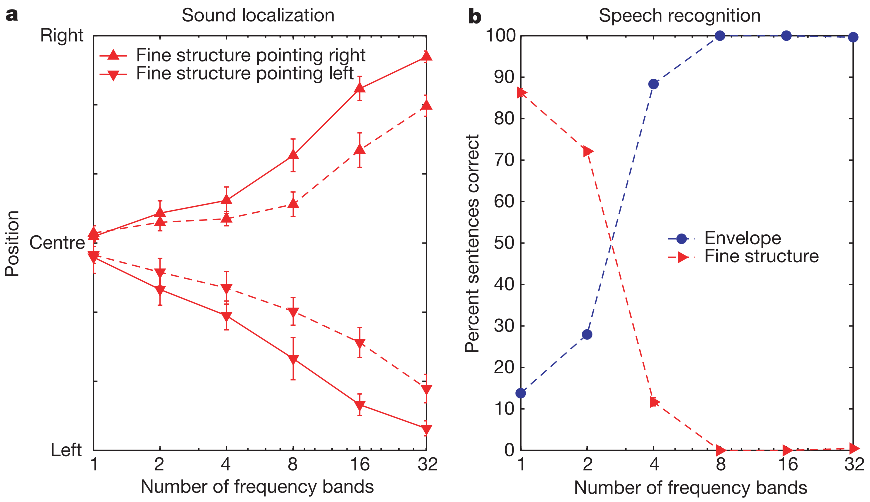

Figure 4 'What' and 'where' for dichotic chimaeras are determined by different cues. **a**, Dichotic chimaeras constructed with the same sentence are shown with solid lines and those made from different sentence are shown with dashed lines. **b**, Subjects heard the sentence in the envelope more than the sentence in the fine structure when four or more bands were used.

reception. Thus, for our purposes, it is valid to consider the envelope as slowly varying relative to the fine structure.

Cochlear implants are prosthetic devices that seek to restore hearing in the profoundly deaf by stimulating the auditory nerve via electrodes inserted into the cochlea. Current processors for cochlear implants discard the fine time structure, and present only about six to eight bands of envelope information[22]. Our results suggest that modifying cochlear implant processors to deliver fine-structure information may improve patients' pitch perception and sensitivity to ITD. Better pitch perception should benefit music appreciation. It should also help convey prosody cues in speech and may enhance speech reception among speakers of tonal languages, such as Mandarin Chinese, where pitch is used to distinguish different words. Better ITD sensitivity may help the increasing number of patients with bilateral cochlear implants in taking advantage of binaural cues that normal-hearing listeners use to distinguish speech among competing sound sources. Strategies for improving the representation of fine structure in cochlear implants have been proposed[23] and supported by single-unit data[24]. □

Methods

Stimulus synthesis

The perfect-reconstruction digital filter banks used for chimaera synthesis spanned the range 80–8,820 Hz, spaced in equal steps along the cochlear frequency map[25] (nearly logarithmic frequency spacing). For example, with six bands, the cutoff frequencies were 80, 260, 600, 1,240, 2,420, 4,650 and 8,820 Hz. The transition over which adjacent filters overlap significantly was 25% of the bandwidth of the narrowest filter in the bank (the lowest in frequency). Thus, for the six band case, each filter transition was 45 Hz wide.

To compute the envelope and fine structure in each band, we used the analytic signal[26] $s(t) = s_r(t) + is_i(t)$, where $s_r(t)$ is the filter output in one band, $s_i(t)$ the Hilbert transform of $s_r(t)$, and $i = \sqrt{-1}$. The Hilbert envelope is the magnitude of the analytic signal, $a(t) = \sqrt{s_r^2(t) + s_i^2(t)}$. The fine structure is $\cos\phi(t)$, where $\phi(t) = \arctan(s_i(t)/s_r(t))$ is the phase of the analytic signal. The original signal can be reconstructed as $s_r(t) = a(t)\cos\phi(t)$. In practice, the Hilbert transform was combined with the band-pass filtering operation using complex filters whose real and imaginary (subscripts r, i) parts are in quadrature[27].

Subjects and procedure

Six native speakers of American English with normal hearing thresholds participated in each part of the study. Five of these subjects participated in the entire series of tests. Speech reception, melody recognition, and lateralization tests were conducted separately. Within each individual experiment, the order of all conditions was randomized. Stimuli were presented in a soundproof booth through headphones at a root-mean-square sound pressure level of 67 dB.

In the speech reception experiment, subjects listened to the processed sentences and were instructed to type the words they heard into a computer. Each subject listened to a total of 273 speech chimaeras with an additional seven for training. Speech reception was measured as percentage words correct. 'The', 'a' and 'an' were not scored. When speech–speech chimaeras were used, each word in a subject's response could count for either a sentence in the envelope or in the fine structure, but this condition rarely occurred in practice.

Before the melody recognition experiment, each subject selected ten melodies that he/she was familiar with. Each melody was taken from a set of 34 simple melodies with all rhythmic information removed[28], consisting of 16 equal-duration notes and synthesized with MIDI software that used samples of a grand piano. During the experiment, subjects selected from their own list of ten melodies which one(s) they heard on each trial. Melodies were scored as percentage correct even when subjects reported multiple melodies in a single trial without penalty for incorrect responses.

In the lateralization experiment, subjects used a seven-point scale to rate the lateral position of the sound image inside the head. This scale ranges from -3 to +3, with -3 corresponding to the left ear and +3 to the right ear. Lateralization scores were averaged for each condition. In addition, subjects had to select which of two possible sentences they heard, one choice corresponding to the envelope and the other one to the fine structure.

Received 15 October; accepted 12 December 2001.

1. Fourier, J. B. J. La théorie analytique de la chaleur. *Mém. Acad. R. Sci.* **8,** 581–622 (1829).
2. Hilbert, D. *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen* (Teubner, Leipzig, 1912).
3. Rhode, W. S., Oertel, D. & Smith, P. H. Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *J. Comp. Neurol.* **213,** 448–463 (1983).
4. Joris, P. X. & Yin, T. C. Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences. *J. Neurophysiol.* **73,** 1043–1062 (1995).
5. Yin, T. C. & Chan, J. C. Interaural time sensitivity in medial superior olive of cat. *J. Neurophysiol.* **65,** 465–488 (1990).
6. Langner, G. & Schreiner, C. E. Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophysiol.* **60,** 1799–1822 (1988).
7. Rauschecker, J. P. & Tian, B. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl Acad. Sci. USA* **97,** 11800–11806 (2000).
8. Recanzone, G. H. Spatial processing in the auditory cortex of the macaque monkey. *Proc. Natl Acad. Sci. USA* **97,** 11829–11835 (2000).
9. Tian, B., Reser, D., Durham, A., Kustov, A. & Rauschecker, J. P. Functional specialization in rhesus monkey auditory cortex. *Science* **292,** 290–293 (2001).
10. Maeder, P. P. *et al.* Distinct pathways involved in sound recognition and localization: a human fMRI study. *NeuroImage* **14,** 802–816 (2001).
11. Dodge, C. & Jerse, T. A. *Computer Music: Synthesis, Composition, and Performance* (Schirmer Books, New York, 1997).
12. Depalle, P., Garcia, G. & Rodet, X. A virtual castrato (!?). *Proc. Int. Computer Music Conf., Aarhus, Denmark* 357–360 (1994).
13. Shen, C., Smith, Z. M., Oxenham, A. J. & Delgutte, B. Auditory Chimera Demo; available at ⟨http://epl.meei.harvard.edu/~bard/chimera⟩ (2001).
14. Licklider, J. Effect of amplitude distortion upon the intelligibility of speech.*J. Acoust. Soc. Am.* **18,** 429–434 (1946).
15. Saberi, K. & Perrot, D. R. Cognitive restoration of reversed speech. *Nature* **398,** 760 (1999).
16. Nilsson, M., Soli, S. D. & Sullivan, J. A. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* **95,** 1085–1099 (1994).
17. Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. Speech recognition with primarily temporal cues. *Science* **270,** 303–304 (1995).
18. Ghitza, O. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *J. Acoust. Soc. Am.* **110,** 1628–1640 (2001).
19. Wightman, F. L. & Kistler, D. J. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.* **91,** 1648–1661 (1992).
20. Henning, G. B. & Ashton, J. The effect of carrier and modulation frequency on lateralization based on interaural phase and interaural group delay. *Hear. Res.* **4,** 184–194 (1981).

21. Bernstein, L. R. & Trahiotis, C. Lateralization of low-frequency complex waveforms: The use of envelope-based temporal disparities. *J. Acoust. Soc. Am.* **77,** 1868–1880 (1985).
22. Wilson, B. S. *et al.* Better speech recognition with cochlear implants. *Nature* **352,** 236–238 (1991).
23. Rubinstein, J. T., Wilson, B. S., Finley, C. C. & Abbas, P. J. Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation. *Hear. Res.* **127,** 108–118 (1999).
24. Litvak, L., Delgutte, B. & Eddington, D. Auditory nerve fiber responses to electrical stimulation: modulated and unmodulated pulse trains. *J. Acoust. Soc. Am.* **110,** 368–379 (2001).
25. Greenwood, D. D. A cochlear frequency-position function for several species—29 years later. *J. Acoust. Soc. Am.* **87,** 2592–2604 (1990).
26. Ville, J. Théorie et applications de la notion de signal analytique. *Cables Transmission* **2,** 61–74 (1948).
27. Troullinos, G., Ehlig, P., ,Chirayil, R., Bradley, J. & Garcia, D. in *Digital Signal Processing Applications with the TMS320 Family* (ed. Papamichalis, P.) 221–330 (Texas Instruments, Dallas, 1990).
28. Hartmann, W. M. & Johnson, D. Stream segregation and peripheral channeling. *Music Percept.* **9,** 155–184 (1991).

Acknowledgements

We thank C. Shen for assistance with data analysis and J. R. Melcher and L. D. Braida for comments on an earlier version of the manuscript. A.J.O. is currently a fellow at the Hanse Institute for Advanced Study in Delmenhorst, Germany. This work was supported by grants from the National Institutes of Health (NIDCD).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.D. (e-mail: bard@epl.meei.harvard.edu).

Long-term plasticity in hippocampal place-cell representation of environmental geometry

Colin Lever*, Tom Wills*, Francesca Cacucci*, Neil Burgess*† & John O'Keefe*†

* *Department of Anatomy and Developmental Biology; and † Institute of Cognitive Neuroscience, University College London, WC1E 6BT, UK*

The hippocampus is widely believed to be involved in the storage or consolidation of long-term memories[1–4]. Several reports have shown short-term changes in single hippocampal unit activity during memory and plasticity experiments[5–12], but there has been no experimental demonstration of long-term persistent changes in neuronal activity in any region except primary cortical areas[13–16]. Here we report that, in rats repeatedly exposed to two differently shaped environments, the hippocampal-place-cell representations of those environments gradually and incrementally diverge; this divergence is specific to environmental shape, occurs independently of explicit reward, persists for periods of at least one month, and transfers to new enclosures of the same shape. These results indicate that place cells may be a neural substrate for long-term incidental learning, and demonstrate the long-term stability of an experience-dependent firing pattern in the hippocampal formation.

In rats, hippocampal lesions cause deficits in spatial behaviour[2,17–20]. One of the major behavioural correlates of the firing of hippocampal pyramidal cells is the animal's location. Previous experimental and theoretical work suggests that the major determinant of the location and shape of place-cell firing fields is the distance from two or more walls in particular directions[21,22]. This theory predicts that these cells will have related patterns of firing in enclosures of different shape. We tested this prediction (see Methods 'preliminary experiment' and Supplementary Information) and found, in each of seven rats, that place fields were very similar on initial exposure to square and circular boxes (Fig. 1a). 73% of the cells (48/66) had 'homotopic' fields in both shapes (that is, in the same location, see Methods for definition of 'homotopic'). Other environmental manipulations, such as translation (Fig. 1b), removal (Fig. 1c), and reconfiguration of the box into shapes other than squares and circles (not shown), showed that firing patterns relate to the box walls and not to other cues in the testing arena. This finding of similarity across shapes conflicts with earlier experiments[7,23], performed on animals that had had considerable experience of the testing enclosures. We asked whether experience was the critical factor in producing neuronal discrimination between different shapes.

We recorded hippocampal CA1 cells from a new group of three animals during repeated exposures to different shapes of enclosure (see Methods 'main experiment'). Recording was performed during the animals' entire experience (up to three weeks) of these enclosures: unlike previous studies (for example, refs 5,6,7,8,9,10,11,12, 21, 23,24,25,26,27,28,29,30), there was no unrecorded training phase. Four boxes were used: two identical circular-walled, and two identical square-walled boxes (hereafter simply 'circle' and 'square'), all made of the same materials so that discrimination on the basis of geometry could be separated from discrimination on the basis of other differences between boxes.

We recorded on successive days monitoring the activity of the same group of neurons where possible, in some cases following individual cells for over a week. In other cases we obtained different samples from the same overall population. Either way, on the first day the firing patterns in the two shapes were similar, replicating our previous work; on later trials, the patterns diverged while those in

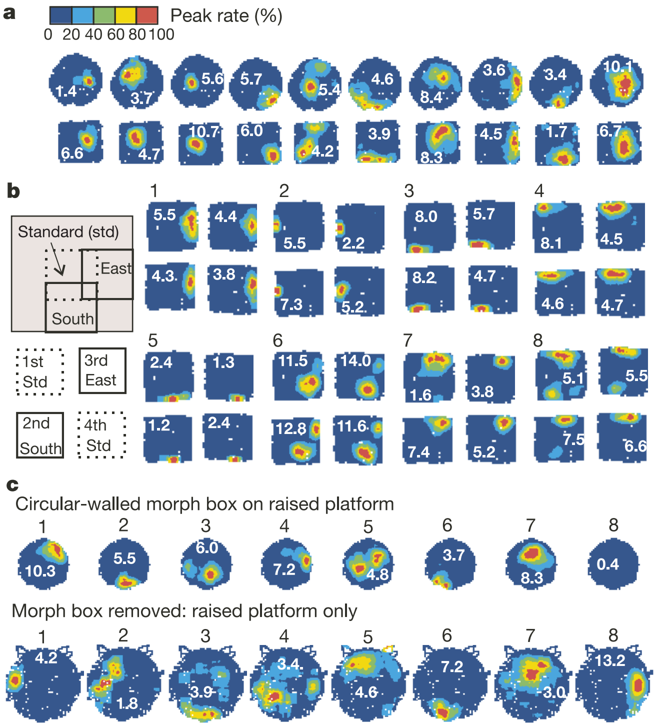

Figure 1 Similarity of spatial firing during early exposure to circular- and square-walled enclosures. **a**, Similar place fields of 10 representative simultaneously recorded CA1 neurons in circle and square. Probe trials on two different groups of cells (**b**, **c**) show that this similarity is determined by box walls rather than similar sets of background cues in the testing arena in both circle and square conditions. Box-wall translation by 40 cm (**b**) eastwards (upper right) or southwards (lower left) does not affect firing fields relative to the box frame, while box-wall removal (**c**) induces remapping. Fields with less than a 1.0-Hz peak rate are not shown.