An ear for statistics

Israel Nelken & Alain de Cheveigné

A study finds that sound textures are stored in auditory memory as summary statistics representing the sound over long time scales; specific events are superimposed, forming a 'skeleton of events on a bed of texture'.

What do we retain of the incessant streams of sound that reach our ears? Potentially, there is a lot to remember: the popular MP3 format, designed on the basis of human perceptual limits, stores anywhere from 32 to 320 kbits for each second[1]. Remembering sensory input is useful, for example to spot recurring patterns that might be of survival value, such as the soft footsteps of a predator. And yet we cannot retain it all because our memory would fill up, and patterns would be hard to find. Besides, most of the detail of our sensory experience is of little use, as argued eloquently by Jorge Luis Borges in the short story "Funes, the Memorious." So what should we keep and what should we discard? In this issue, McDermott *et al.*[2] make a major contribution to this question, showing that for a certain class of sounds that they call "textures," all that we remember is a small set of summary statistics.

Psychophysicists like to annoy their subjects by presenting them over and over (and over) with pairs of stimuli to be compared. ("Which sound had the higher pitch?" "Which picture contained a faint change in contrast?") A constant observation in such studies is that the longer the stimulus is, the better perceptual judgments are[3]. McDermott *et al.*[2] found that the opposite is true for textures, sounds that are characterized by a random, stationary pattern of fluctuations, such as those produced by the superposition of many individual events. Examples are the crackle of fire, the patter of rain or the buzz of a swarm (**Fig. 1**). When comparing different sounds with a similar texture, longer sounds are harder to discriminate than shorter sounds. This counterintuitive finding provides a fascinating glimpse into the nature of auditory working memory.

Key to the result of McDermott *et al.*[2] is a technique grounded in their previous work on the controlled synthesis of textures[4]. Starting from a natural sound such as rain, 'rain-like' exemplars can be produced by warping white noise 'seeds' so that their statistics, such as averages and correlations of energy in narrow frequency bands, match those of rain. In many cases, such artificial textures sound remarkably natural. An implication of that work is that our perceptual representation of an exemplar of a texture is limited to such statistics.

When we compare two sounds that are presented sequentially, a representation of the first sound must be maintained long enough that it can be compared with the second sound. The longer the sound, the more information is available for making the discrimination, and hence the common finding that performance increases with sound duration. Indeed, McDermott *et al.*[2] observed such a relationship between performance and stimulus duration with speech sounds (one speaker talking at a time) or a sequence of drum beats (if not too dense), or when subjects were asked whether two sounds belonged to the same texture (admittedly, a different task from telling whether or not two sounds are identical). In contrast, sounds that shared the same statistics became harder to discriminate the longer they lasted. In other words, the countless acoustic differences between two exemplars of the same sound texture are perceptually salient if the sounds are short, but they become inaccessible if the sounds are long, as if the perceptual representation was reduced to summary statistics over time.

The authors carefully excluded several trivial explanations for their result. For example, the observed improvement in discrimination of tokens from distinct texture classes with the duration of their presentations implies that the auditory system is using the longer sounds for accumulating information, ruling out the hypothesis of a lack of access to acoustic details for longer sounds. In addition to their artificial textures, McDermott *et al.*[2] extended their findings to textures created as dense mixtures of many voices ('cocktail party' sounds) or of multiple streams of drum beats, showing the relevance of their observations to real-world sounds.

The finding that sounds that share a small number of statistical descriptors are difficult to discriminate is an unusual observation: many studies (for example, ref. 5) suggest that perceptual decisions are remarkably sensitive to the details of the spectro-temporal representations of the incoming sounds that are computed by the early auditory system. Using only statistics, rather than the full peripheral representation, implies a huge loss of information. Acoustic detail seems to be discarded somewhere along the auditory processing stream and replaced by much simpler summary statistics gathered over the duration of the sound. Potentially useful information is lost in the process: what might justify the loss?

One driving factor may be the cost of maintaining a higher-resolution representation for ongoing, spectro-temporally complex sounds. The flux of sensory information must somehow be reduced, and statistics are one way of summarizing the data. Another driving factor may be the benefit of abstracting away ecologically irrelevant detail. In the real world, the differences between different chunks of the same texture are meaningless: it is important that the grass is rustling, not the exact sound it makes when rustling. It is reasonable to argue that storing all the details of such sounds is both unnecessarily expensive and potentially counterproductive, loading memory with information that is unlikely to be of use. Thus, this summarizing may reflect the operation of a scheme in which auditory representations are reduced in such a way that they keep the relevant information about the environment, while weeding out the irrelevant detail (see ref. 6 for attempts to formalize similar tradeoffs).

There are several examples of perceptual invariance in audition. For example, Ohm's acoustic law states that sounds that only differ in the relative phases of their sinusoidal components sound alike. Ohm's law can be traced to a large extent to the function of the peripheral auditory system. It may be useful in reducing sensitivity to phase distortion due to propagation and reverberation in the environment. Another, more complex perceptual invariance is categorical perception of speech sounds, according to which sounds that fall in the same phonemic category (for example, "e") are harder to distinguish than sounds that fall on either side of a boundary between phonemes (for example, "e" and "o"). Different tokens of the same phoneme differ acoustically and at the level of the early auditory system and yet are perceptually very

Israel Nelken is in the Department of Neuroscience, the Alexander Silberman Institute of Life Sciences, and the Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem, Israel, and Alain de Cheveigné is at the CNRS, the Université Paris Descartes and the Ecole normale supérieure, Paris, France, and University College London, London, UK.
e-mail: israel@cc.huji.ac.il

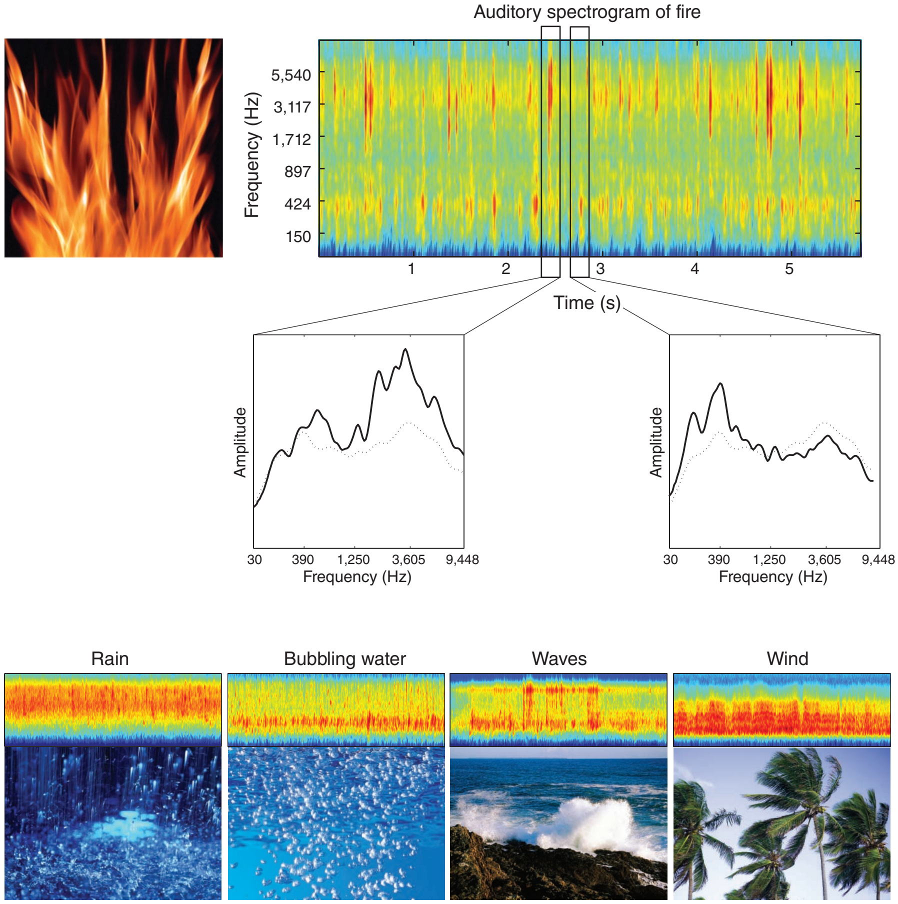

Figure 1 Auditory spectrograms of different textures. Auditory spectrograms approximate the representation of sounds at the level of the cochlea and the auditory nerve. The textures used here are examples of textures studied by McDermott and Simoncelli[4] and used in this work[2] (source: http://www.cns.nyu.edu/~jhm/texture_examples/). The top example sounds like crackling fire. The auditory spectrogram shows the presence of short, wide-band events with a typical spectral shape. Such events are captured with statistics by specifying the average value of the spectrum at each moment in time, as well as the correlations between spectral energy in different frequency bands. The two insets below the fire spectrogram represent the individual spectra of two short segments of the sound (solid lines) compared to the mean spectrum of this sound texture (dotted lines). The four examples at the bottom show other textures. Their different statistics are manifest as a different visual texture in each respective spectrogram.

similar. As in the case of textures, discrimination efficiency of speech in noise may be reduced when the acoustic-to-phonetic transformation of the sound signals is engaged[7].

Nevertheless, the auditory system can store, under the right conditions, more than just statistics. As an example, a faint bird song over the texture-like sound of a stream is somehow perceived separately from the texture, although it may barely disturb the summary statistics. This implies that extra storage space is available. Why is it not used to store details of other exemplars of textures, possibly improving their discrimination? One would think that using the full memory capacity should be ecologically advantageous. A possible explanation, sketched by McDermott *et al.*[2], is that low-level features are retained but are overwritten as new sounds arrive, so that comparison is not possible. Alternatively, the auditory system may store statistics only when it somehow determines that the sound is a texture. Obviously this decision cannot be made on the basis of the summary statistics alone, as exemplars with and without faint extraneous sounds share very similar statistics. Furthermore, sounds that belong to the same family of textures can be 'individualized' by using tricks, such as repetition. For example, individual chunks of white noise can be stored precisely in memory and retrieved weeks later[8]. Thus, the picture that emerges is that of a flexible sensory memory mechanism that stores not only the slowly varying statistics of ongoing textures but also the detailed patterns of temporally localized events. Applied to textures, statistics serve both to reduce storage costs and to generalize over irrelevant detail.

These findings raise many questions. Statistics that describe longer-term patterns would be expected to be remembered longer than the more detailed information that characterizes shorter events. We could conjecture, therefore, the existence of yet higher level statistics applicable to even longer sounds, describing, for example, the pattern of variation over time of the statistics proposed by McDermott *et al.*[2] ('meta-textures'?). Next, are all the necessary statistics produced in parallel, directly from incoming sensory information? Or are higher-order statistics elaborated from lower-level representations according to a hierarchical compaction process? What triggers their calculation? Are they extracted only for textures, or are they extracted generally but are the only information that is stored in the case of textures? How does the auditory system decide that a texture is a texture?

This study opens new perspectives on sound and hearing. In the past, countless studies have surveyed our ability to detect subtle differences between sounds, such as the presence of a weak target, a minute temporal gap, or a small increment in frequency or amplitude, leading to the notion that every detail of a sound counts. We now discover that there exists a class of sounds, textures, for which radically different waveforms evoke the same percept if they share a minimum set of statistical features. This finding has implications both for the processing taking place in the auditory system and for the general design principles of sensory systems. It suggests that statistics may be applicable to elements of the scene and that the auditory system may maintain a multi-stream representation involving events and statistics of multiple sources: a 'skeleton of events on a bed of texture'.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Grill, B. & Quackenbush, S. MPEG-1 audio. <http://mpeg.chiariglione.org/standards/mpeg-1/audio> (2005; accessed 5 February 2013).
2. McDermott, J.H., Schemitsch, M. & Simoncelli, E.P. *Nat. Neurosci.* **16**, 493–498 (2013).
3. Moore, B.C.J. *J. Acoust. Soc. Am.* **54**, 610–619 (1973).
4. McDermott, J.H. & Simoncelli, E.P. *Neuron* **71**, 926–940 (2011).
5. Green, D.M. in *Human Psychophysics* (eds. Yost, W.A., Popper, A.N. & Fay, R.R.) 13–55 (Springer, New York, 1993).
6. Tishby, N. & Polani, D. in *Perception-Reason-Action Cycle: Models, Algorithms and Systems* (eds. Vassilis, C., Hussain, A. & Taylor, J.G.) 601–636 (Springer, New York, 2011).
7. Nahum, M., Nelken, I. & Ahissar, M. *PLoS Biol.* **6**, e126 (2008).
8. Agus, T.R., Thorpe, S.J. & Pressnitzer, D. *Neuron* **66**, 610–618 (2010).