Auditory Perception: Hearing the Texture of Sounds

A recent study provides intriguing insights into how we recognize the sound of everyday objects from the statistical properties of the textures they produce.

Neil C. Rabinowitz and Andrew J. King

Sound and time are inextricably linked. We are used to time freezing in other sensory modalities: that geeky mugshot of you as a teenager isn't ever going to change, while the smell of frying onions, once wafted, lingers. But sounds — the bleating of your goat or the jabbering of your spouse — usually start and stop in quick succession.

With these examples in mind, we tend to think of our hearing as being geared towards processing rapidly changing events. Indeed, the temporal precision of the auditory system far exceeds that of other sensory modalities [1,2], a property that allows sounds to 'capture' the perceived timing of accompanying visual stimuli [3,4]. A new study by McDermott and Simoncelli [5], however, challenges this perspective of sounds as exclusively transient events. They step back and consider sounds on the time scale of seconds, rather than milliseconds, and describe a class of sound 'textures' that do not change with time on this scale, and which can be recognized purely on the basis of their statistics.

From Statistics to Sound Recognition

McDermott and Simoncelli [5] began by constructing a simple model of the auditory pathway. In this model, an incoming signal is broken down into separate frequency channels, mimicking the filtering provided by the cochlea. Next, the envelopes of energy modulations within each channel are extracted, and these envelopes are then decomposed through an additional modulation filterbank. The authors then fed a number of real sounds — from chirping insects to falling rain — into the model and looked at the averages over time of the model's various outputs, such as how much (and how sparsely) the cochlear channels were activated, how much the modulation channels were activated, and how correlated each of these channels were with each other.

The next step involved synthesizing brand new sounds: the authors started with white noise trains, and perturbed them until their statistics according to the model matched those of the original sounds. Surprisingly, a large number of the synthesized sounds were indistinguishable to listeners from the original sounds (Figure 1). But omitting any one of the statistics from the model often impaired the ability of subjects to recognize them. For many auditory textures, this fairly simple set of statistics thus appears to capture the same information that our auditory system does.

Such sound textures are therefore the static images of the auditory system. They are perceptually stable: despite variations in its short-term acoustic structure, the sound of a waterfall continues to sound like a waterfall. The longer-term texture statistics are all that seem to matter for us to be able to make out people applauding, a raging fire, or a swarm of bees from the signals reaching our ears.

Representing Sound Texture in the Brain

These findings suggest that, for the purposes of stimulus recognition, the auditory system 'abstracts away' the short-term, contingent features of the sound texture that are encoded in its low-level representations. In doing so, it produces a compact, high-level representation of these sounds whose content is some combination of the statistics of lower-level activity. If this is the case, then sound textures, perhaps paradoxically, may offer a novel opportunity for understanding how information travels backwards through the auditory pathway.

A powerful theoretical idea is that of predictive coding: given that sensory systems are organized in both a hierarchical and recurrent fashion, high-level sensory stations are able to learn the statistical regularities present in the activity of low-level stations, and feed this information back to the lower level at an appropriate time. This could be used to 'explain away' the pattern of low-level activities. Under this hypothesis, the predictable pattern of low-level activities could be suppressed by top-down feedback.

Such a process has been described as 'analysis by synthesis', whereby the brain constructs high-level hypotheses, based on low-level processing, about the structure or content of the scene. These hypotheses are not just summaries: they are generative models for how the data came about. The models manufacture synthetic data, which can then be compared with the low-level activity [6,7]. The residual activity that remains is the unexplained data, and therefore constitutes a prediction error signal [8,9].

Evidence that the auditory cortex might engage in predictive coding has come from studies of the effects of stimulus repetition on neural responses. Thus, we know that auditory cortical neurons are acutely sensitive to the patterns in a sound sequence, and respond more strongly to rare events than to common ones [10]. Even task-irrelevant background sounds elicit different electro-encephalogram (EEG) traces

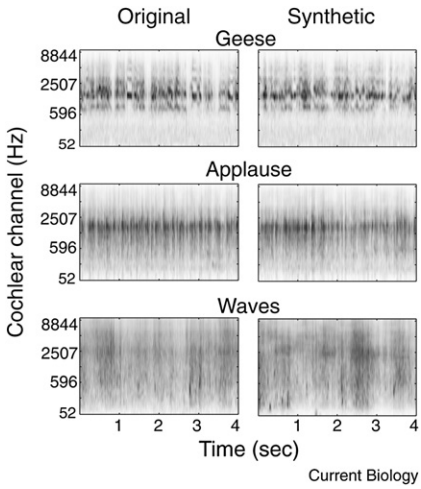

Figure 1. Example results from sound synthesis.

Spectrograms of original sound textures (left), and synthesized versions (right). Listeners generally correctly identified the synthetic sounds. From McDermott and Simoncelli [5].

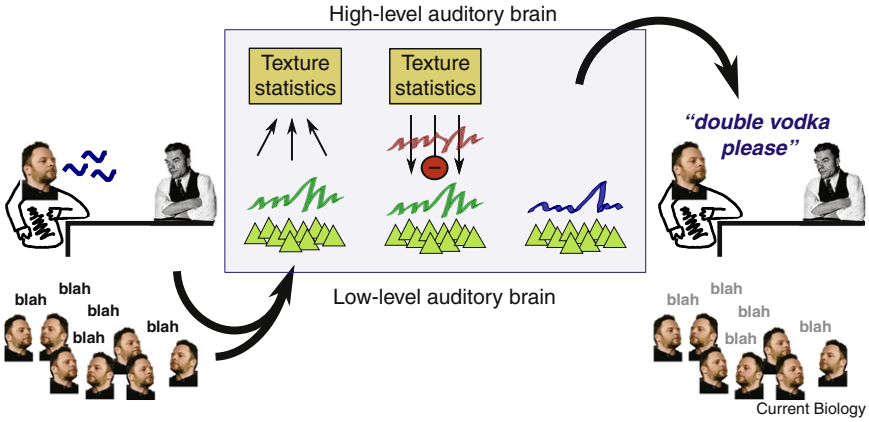

Figure 2. The classification of sound textures by their statistics may assist in auditory scene analysis.

Sound from a complex scene enters the brain, and is represented by the activity of a population of low-level neurons (triangles). The higher-level brain captures the statistics of the low-level activity, hypothesizing that it consists of a stationary background sound texture, whose statistics are collected, and a foreground non-stationary sound. The background could then be subtracted to improve the perception of the foreground sound.

depending on whether they are regular or surprising [11].

It has been suggested that repetition suppression whereby the neural activity evoked by a sound is reduced following repeated presentations of the sound has a top-down origin, such that cortical activity encodes prediction error [12]. Recently, evidence supporting this has come from Todorovic and colleagues [13], who showed that repetition suppression depends on expectation. When human subjects were presented with repeated tones, the magneto-encephalogram (MEG) activity measured from their auditory cortex was reduced in response to the second tone. In those blocks where the repetition was expected, however, the reduction in activity was greater than when the repeated sound was unexpected. The predictive coding framework has also been used to link the amount of sensory noise in auditory cortex to errors made in an auditory detection task [14].

What does sound texture processing have to do with this? There are several pieces of evidence that suggest that the 'explaining away' of predictive coding might also be the basis for 'ignoring' unimportant sounds. First, responses to unattended, neutral sounds have been shown in many experiments to be suppressed below those to attended sounds, an effect that is even stronger for actively ignored stimuli [15]. Second, the kinds of stimuli that we have difficulty ignoring, which can be described as having bottom-up saliency, are often those that stand out as statistical abnormalities from their spatially or temporally local environment [16–18]. Finally, many of the sound textures that were captured well by the model of McDermott and Simoncelli [5] were what we would class as background sounds — insects chirping, rain falling, air conditioners blowing, and so on. These are, coincidentally, just the kinds of sounds that we are good at ignoring.

This leads us to the tantalizing hypothesis that, if our auditory system can characterize a particular sound texture by a small set of statistics, then it may also be able to feed this back to the lower level to cancel out the impact the texture might otherwise have on perception. This could be described as 'catalysis by synthesis', and is potentially useful for auditory scene analysis, the process by which the sound elements belonging to a particular object or event are grouped together and segregated from those belonging to other objects or events [19] (Figure 2).

Thus, at that cocktail party that seems to go on all night, the barman has to drown out your colleagues' endless, superimposed banter to hear your request for something stronger. In doing so, he must capture the statistics of the complex background signal, collapse it into a texture, and subtract the resultant model's predictions from the ongoing babble. Barmen are pretty good at that.

References

1. Tyler, C.W., and Hamer, R.D. (1990). Analysis of visual modulation sensitivity. IV. Validity of the Ferry-Porter law. J. Opt. Soc. Am. A *7*, 743–758.
2. Viemeister, N.F., and Plack, C.J. (1993). Time analysis. In Human Psychophysics, Springer Handbook of Auditory Research, W.A. Yost, A.N. Popper, and R.R. Fay, eds. (New York: Springer), pp. 116–154.
3. Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions. What you see is what you hear. Nature *408*, 788.
4. Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. Atten. Percept. Psychophys. *72*, 871–884.
5. McDermott, J.H., and Simoncelli, E.P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron *71*, 926–940.
6. Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. *10*, 301–308.
7. Rao, R.P.N., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.
8. Friston, K. (2005). A theory of cortical responses. Phil. Trans. Roy. Soc. B. Biol. Sci. *360*, 815–836.
9. Lee, T.S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. J. Opt. Soc. Am. A *20*, 1434–1448.
10. Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. Nat. Neurosci. *6*, 391–398.
11. Winkler, I., Teder-Sälejärvi, W.A., Horváth, J., Näätänen, R., and Sussman, E. (2003). Human auditory cortex tracks task-irrelevant sound sources. Neuroreport *14*, 2053–2056.
12. Baldeweg, T. (2006). Repetition effects to sounds: evidence for predictive coding in the auditory system. Trends Cogn. Sci. *10*, 93–94.
13. Todorovic, A., van Ede, F., Maris, E., and de Lange, F.P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: An MEG Study. J. Neurosci. *31*, 9118–9123.
14. Hesselmann, G., Sadaghiani, S., Friston, K.J., and Kleinschmidt, A. (2010). Predictive coding or evidence accumulation? false inference and neuronal fluctuations. PLoS One *5*, e9926.
15. Chait, M., de Cheveigné, A., Poeppel, D., and Simon, J.Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. Neuropsychologia *48*, 3262–3271.
16. Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Analysis and Machine Intelligence *20*, 1254–1259.
17. Kayser, C., Petkov, C.I., Lippert, M., and Logothetis, N.K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. Curr. Biol. *15*, 1943–1947.
18. Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. Vision Res. *49*, 1295–306.
19. Schnupp, J., Nelken, I., and King, A.J. (2010). Auditory Neuroscience: Making Sense of Sound (Cambridge MA: MIT Press).

Department of Physiology, Aantomy and Genetics, University of Oxford, Parks Road, Oxford OX1 3PT, UK.
E-mail: neil.rabinowitz@merton.ox.ac.uk, andrew.king@dpag.ox.ac.uk

DOI: 10.1016/j.cub.2011.10.027