Listening in complex acoustic scenes

Andrew J King and Kerry MM Walker

Being able to pick out particular sounds, such as speech, against a background of other sounds represents one of the key tasks performed by the auditory system. Understanding how this happens is important because speech recognition in noise is particularly challenging for older listeners and for people with hearing impairments. Central to this ability is the capacity of neurons to adapt to the statistics of sounds reaching the ears, which helps to generate noise-tolerant representations of sounds in the brain. In more complex auditory scenes, such as a cocktail party — where the background noise comprises other voices, sound features associated with each source have to be grouped together and segregated from those belonging to other sources. This depends on precise temporal coding and modulation of cortical response properties when attending to a particular speaker in a multi-talker environment. Furthermore, the neural processing underlying auditory scene analysis is shaped by experience over multiple timescales.

Address
Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3PT, UK

Corresponding author: King, Andrew J (andrew.king@dpag.ox.ac.uk)

Introduction

Most research on the auditory system focuses on the way single sound sources are processed and perceived. In real life, however, the sounds reaching our ears usually comprise a mixture of signals arising from multiple sources. A major challenge faced by the auditory system is therefore to group together the sound attributes associated with a particular source and segregate them from those belonging to other sources. This is auditory scene analysis [1•]. In order to follow a conversation in a busy restaurant, for example, the brain has to be able to separate the voice of the person speaking to you from the babble of other, superimposed voices that overlap in time. This 'cocktail party problem' [2,3] represents a particular challenge for

the auditory system since each of the voices will likely resemble one another both acoustically and perceptually.

There are two reasons why the presence of other voices may make it difficult to pick out the speech signal of interest [4,5]. First, due to their overlapping spectra, the voices compete to activate the same frequency channels in the auditory system, which is known as energetic masking. This reduces the audibility of the target voice by weakening its representation in both the cochlea and the central auditory pathway. Second, competing voices may result in informational masking, drawing the listener's attention away from the target voice. The unpredictability of the background voices and shared higher-level statistical features with the target speaker impede our attempts to ignore them, introducing perceptual uncertainty as to what that speaker was saying (Figure 1).

The brain's solution to this problem is theorized to be composed of at least two processes, which are not entirely independent [1•,3]. First, the features of the various sounds in the environment that reach the ear as a mixture need to be extracted and segregated into those that correspond to different sound sources. In this manner, the features of the target sound are bound together into a single perceived object, such as a person's voice. Second, there must exist a mechanism for attending to the bound features of the target sound source of interest, moving other auditory information into the perceptual background.

While cocktail parties represent a particularly challenging situation for the auditory system, most behaviorally important sounds are heard against a background of everyday noise — from street and workplace sounds to music — which therefore represents a simpler example of auditory scene analysis. Understanding the neural basis for listening in noise has considerable clinical and societal relevance since many people experience difficulties with this vital task. This is particularly the case with increasing age [6] and in individuals with hearing impairments, even when amplification is provided to compensate for raised thresholds [7]. But irrespective of age, people with audiograms within the normal range can show marked differences in speech-in-noise performance [8]. This has been attributed to physiological differences in the way individuals process the temporal structure of sounds [8,9], and in their ability to attend to specific speech streams [9,10] or group sound elements as belonging to foreground or background sounds [11]. Although cochlear abnormalities that do not show up in the audiogram are likely to

Figure 1
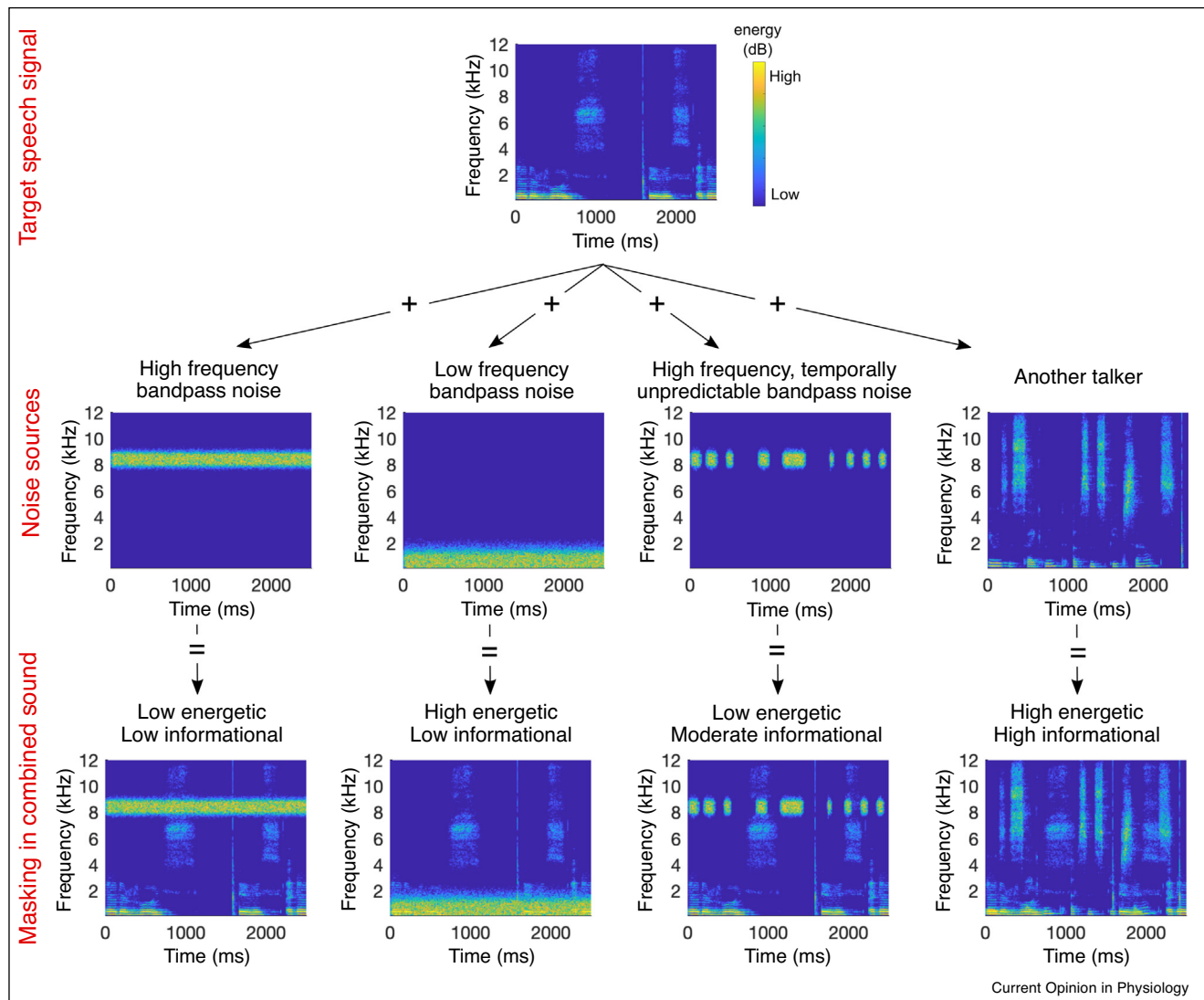
Energetic and informational masking of speech. The spectrogram of a speech stream of interest is shown in the top panel. The spectrograms of four example noise sources are shown in the middle row. Spectrograms of the mixtures of the target speech stream and each noise are shown in the bottom row. A steady-state, high-frequency bandpass noise provides less energetic masking of the speech (column 1) than a similar noise with a lower frequency band (column 2), as most of the energy in the speech is low frequency. Reducing the temporal predictability of the noise increases the informational masking (column 3), as the listener's attention is captured by the noise and the auditory system cannot as easily adapt to it. The unpredictable noise will therefore be more disruptive to speech intelligibility even though it provides less energetic masking than the noise sources in the two left columns. Finally, the speech of a second person talking provides high amounts of energetic and informational masking (column 4), making it more difficult to segregate and attend to the target speech.

contribute too [12], these findings indicate that impairments of listening in noise at least partially reflect a deficit in central auditory processing.

Adaptation and the background noise problem

An effective solution has evolved for coping with the challenge of hearing in noisy environments, at least when the statistics of the foreground and background sounds differ. A key feature of sensory neurons is that they can adapt their responses to match these statistics. If the frequency content and level of the background noise vary relatively slowly compared to that of the human voice or other sounds of interest, neuronal adaptation can serve to reduce neuronal responses to the noise and therefore improve the audibility of the target sound [13].

Efferent control of the cochlea

Evidence for adaptation to background noise has been obtained in studies of human hearing, since speech-in-noise recognition improves if the masking noise starts before the speech sounds rather than at the same time [14,15]. Because they predominantly innervate outer hair cells in the contralateral cochlea, medial olivocochlear neurons in the superior olivary complex have long been implicated in enhancing sound detection in background noise [16,17]. However, recent studies in which noise adaptation was reduced by introducing fluctuations into the noise [18] and of noise adaptation during word recognition in cochlear implant users [19], in whom the medial olivocochlear reflex is not thought to operate, has cast doubt on this. Furthermore, measurements of otoacoustic emissions indicate that post-adaptation improvements in sensitivity to amplitude modulation for tones presented in noise are unlikely to be due to an efferent-dependent reduction in cochlear responses [20]. Nevertheless, through their influence on outer hair cells, medial olivo-cochlear efferents can regulate cochlear gain and therefore the responses of auditory nerve fibers, and their role in listening in noise remains a controversial area [21,22].

Neuronal adaptation to sound level and contrast

In the last few years, investigation of the neurophysiological basis for noise adaptation has focused primarily on central auditory processing. Adaptation to sound level statistics is found throughout the auditory system from the auditory nerve to the cortex [23–27]. The dynamic range — the range of stimulus values encoded by a neuron by a change in its firing rate — can shift to compensate for changes in mean stimulus level, thereby maintaining maximum sensitivity over the most commonly encountered values. In addition to changing the mean overall sound level, the presence of background noise will alter the contrast, that is the variance in the sound level distribution [13]. Again, the brain can adapt to this by scaling neuronal response gain to compensate for changes in stimulus contrast. Contrast gain control is also a common property of neurons along the auditory pathway [28,29•,30,31•].

The functional consequences of adaptation to sound statistics are still relatively unexplored. In monkeys, adaptation of inferior colliculus (IC) neurons to mean sound level does not appear to affect their neurometric thresholds or the animals' psychometric thresholds for tones presented in noise [32]. On the other hand, dynamic range adaptation is associated with perceptual adjustments in human spatial hearing [33,34]. Moreover, sound level discrimination thresholds in human listeners vary with stimulus contrast and the strength of this perceptual adaptation can be predicted from the contrast gain control exhibited by neurons at both subcortical and cortical levels in the mouse [31•] (Figure 2).

Noise-tolerant coding of sounds in the auditory cortex

Several studies have shown that adaptive coding gradually builds up along the auditory pathway [25,29•,31•]. As a consequence, by the level of the primary auditory cortex (A1), adaptation to mean level and contrast enables speech to be represented in a way that is relatively robust to the presence of stationary background noise [29•]. Other studies have also reported a role for adaptation in generating noise-tolerant cortical representations of speech [35,36,37••]. For example, electrocorticography (ECoG) data obtained from neurosurgical patients listening to speech in the presence of abruptly changing background noise have shown that auditory cortical neurons rapidly adapt to the noise, resulting in enhanced neural coding and perception of the phonetic features of speech [37••] (Figure 3). Furthermore, fMRI responses to natural sounds presented in isolation and in real-world noise are more noise invariant in non-primary auditory cortex than in primary areas [38].

How stimulus processing changes in the presence of noise is actually not straightforward. The effects of noise on frequency selectivity in rat A1 and on word recognition performance in humans depend not only on the signal-to-noise ratio, but also on the absolute levels of the foreground tones and background noise [39]. Moreover, cortical neurons differ in how accurately they encode target stimuli in the presence of noise [40]. Surprisingly, the presence of continuous broadband noise has been found to improve tone discrimination for small frequency differences in mice and this behavioral improvement could be replicated by optogenetically activating parvalbumin-positive interneurons in order to make A1 tuning curves resemble those recorded in the presence of noise [41•]. Furthermore, A1 neuronal sensitivity to tones presented in noise is enhanced if coherently modulated sidebands are added to the noise masker, a condition that improves signal detection thresholds in humans [42]. Like the release from masking found for speech-in-noise recognition noise by human listeners [14,15], prior adaptation to the noise is required to produce substantial comodulation masking release in A1 and this was reduced by inhibiting cortical activity during the priming period [42].

Given the importance of the cortex in listening in noise, attention is turning to the cellular circuits and synaptic mechanisms responsible for the adaptive processing of auditory information [43,44]. Nevertheless, we should not ignore what happens at subcortical levels. Although recent work in guinea pigs has confirmed that cortical neurons are more tolerant to changes in noise level, populations of neurons in the medial geniculate body (MGB) and particularly the IC actually show greater discrimination performance for communication calls presented in noise than those recorded in the cortex [45]. It is possible that descending corticofugal projections

Figure 2
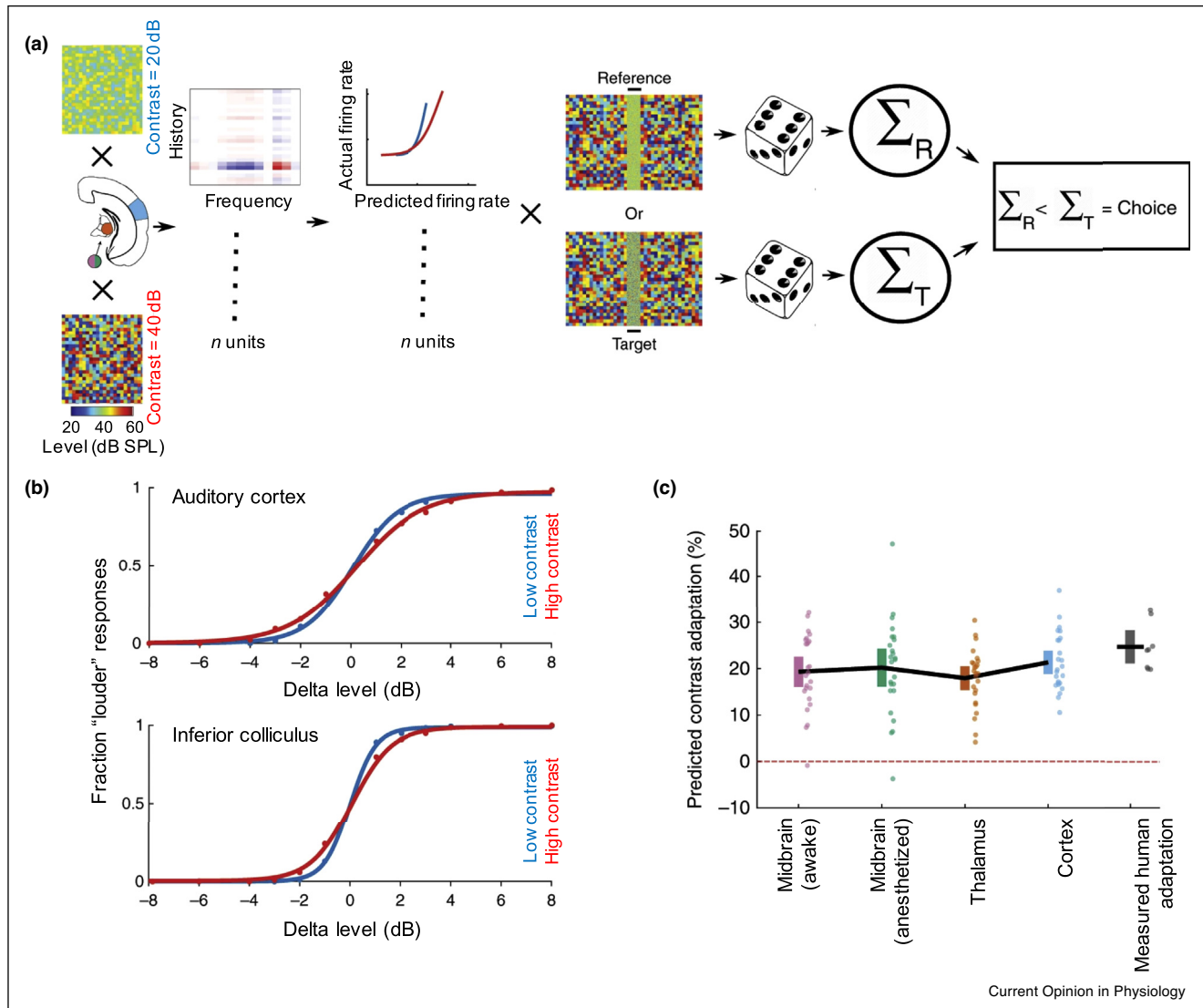
The strength of perceptual contrast adaptation can be predicted from the contrast adaptation exhibited by auditory neurons. **(a)** Schematic of a computational model that uses neuronal responses recorded from neurons in the mouse auditory midbrain, thalamus or cortex to predict performance on a 2-alternative, forced-choice sound level discrimination task for pairs of broadband noise bursts presented in different contrast environments. Simulated responses to noise stimuli of different levels (reference, R: 70 dB SPL, target, T: 62–78 dB SPL), embedded in low- or high-contrast dynamic random chords, were derived from the contrast-dependent linear/non-linear model estimated from the individual neuronal responses recorded at each processing level. Psychometric functions were determined by asking which noise stimulus elicited most spikes across all recorded units in the simulated trial. If the reference noise elicited fewer spikes than the target noise stimulus a "louder" response was predicted. See Lohse et al. [31•] for more details. **(b)** Psychometric functions produced by the model based on responses recorded in the primary auditory cortex (top) or in the midbrain from the central nucleus of the inferior colliculus (bottom) in low-contrast (20 dB, blue) and high-contrast (40 dB, red) conditions. Level discrimination improved when the contrast of the flanking sounds was low, as indicated by the steeper psychometric functions. **(c)** Predicted strength of contrast adaptation from neuronal responses recorded in the auditory midbrain of awake mice or in the midbrain, thalamus or cortex under anesthesia, compared with measured perceptual contrast adaptation in human listeners. Note the similarity in each case. Solid black lines connect mean values after 25 runs of the model (or across the eight participants in the measured human adaptation). Colored error bars denote 95% confidence intervals around the mean (for clarity, individual data points are displayed next to the corresponding error bars). Adapted from Lohse *et al*. [31•].

Figure 3
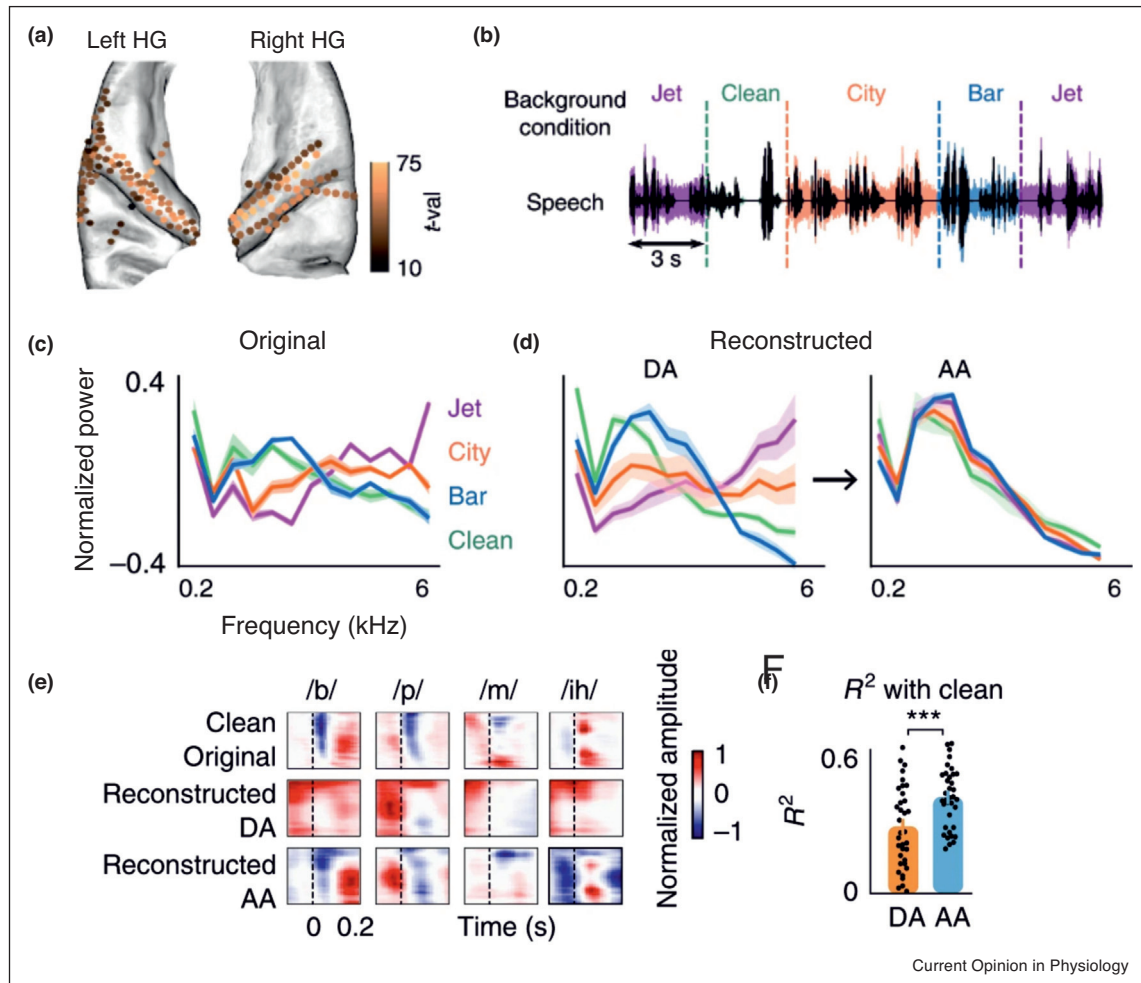
Adaptation of the human auditory cortex to changing background noise enables robust representation of the phonetic features of speech. **(a)** Recordings in human auditory cortex showing electrode locations where significant responses to speech (t-val > 10, t-test speech versus silence) were found. HG, Heschl's gyrus. **(b)** Waveforms of the sounds used: speech (shown in black) was presented alone (Clean) or with different types of background noise (shown by the colors for Bar, City, Jet backgrounds), which changed randomly every 3 or 6 s. **(c)** Average frequency power from the spectrograms for each stimulus type. **(d)** A reconstruction model was trained on the responses to clean speech and used to reconstruct spectrograms from the neural responses to speech with added background noise. Left panel shows the average reconstructed frequency profiles during adaptation (DA), which resemble the frequency profiles for each noise type. Right panel shows that after adaptation (AA) of cortical responses, the average reconstructed frequency profile in each case closely resembles the frequency profile of clean speech. **(e)** Original and reconstructed spectrograms of four example phonemes. The spectrotemporal features that distinguish each of these phonemes in the spectrograms reconstructed from cortical activity are initially distorted during adaptation to the noise (0–0.4 s after the change in stimulus), but are evident after adaptation (2.0–2.4 s after the transition). For example, the phoneme /b/ is characterized by an onset gap followed by low-frequency spectral power. Both the gap and the low-frequency feature are masked during adaptation, but are subsequently restored after adaptation. **(f)** Correlation between the reconstructed phoneme spectrograms during and after adaptation with the clean phoneme spectrograms. Adapted from Khalighinejad *et al.* [37••].

contribute to context-dependent processing at subcortical levels. Neither adaptation to mean level by IC neurons [26] nor to contrast by IC or MGB neurons [31•] depends on corticofugal inputs, but cortical inactivation does slow down sound level adaptation by IC neurons when the same sound statistics are re-encountered [26]. This suggests that descending projections might play a role in

adaptation to stimulus statistics in rapidly changing acoustic environments.

Environmental statistics and scene analysis

The brain is faced with the challenge of not only identifying different sources from the mixture of sounds reaching the ears, but also of separating those sources from

environmental information that may be present too. Within rooms and other enclosed spaces, sound arrives directly from its source, accompanied by multiple delayed versions due to reflections off the walls and other surfaces. Reverberation is useful because it helps the listener to estimate room size [46] and source distance [47], but it also distorts the sound arriving from the source. However, listeners can perceptually separate sound sources and the accompanying reverberation, though this ability is impaired if the environmental statistics deviate markedly from natural values measured in a range of inner-city and rural locations [48•]. This study therefore suggests that prior experience of these statistics is important for perception.

Sound source segregation and selective attention

Listening to speech in the presence of a constant noise source, like an airplane passing overhead, is challenging, but is aided by the way neurons adapt to low-level statistics, like sound level and contrast. However, it is considerably more difficult to filter out less predictable noise sources, such as background speech, because they provide more informational masking [49].

Several 'primitive' perceptual features, which can be derived from the bottom-up stimulus statistics of most sounds, have been shown to contribute to source segregation. These include differences in the level, spatial location, timbre (e.g. the spectral envelopes), temporal modulation and harmonicity of the sounds produced by separate sources [1•,3]. Each of these features is extracted through specialized mechanisms throughout the ascending auditory system, which are beyond the scope of this review. Individual neurons in ferret A1 can represent multiple features by multiplexing several neural codes [50], and recent evidence suggests that this strategy may also be employed in human auditory cortex [51,52]. When multiplexing cortical neurons are co-activated by several features of a single sound source, they may bind these features through their simultaneous activity within the wider cortical network, consistent with predictions of the temporal coherence theory of auditory scene analysis [53].

Neural synchronization may also help to bind coincident features across sensory systems. The addition of temporally coherent visual cues can enhance the representation of a target speech stream in auditory cortex [54], and this has been shown to be mediated by visual cortex [55••]. These neural processes may account for why face reading helps human listeners to selectively attend to a single speaker in multi-talker listening situations [56].

The process of grouping sound features belonging to a single source depends critically upon encoding their onsets and offsets [53,57]. The precise representation of the timing of acoustical events in the auditory system is well suited to provide this essential information. Accurate neural representations of sound offsets are generated through specialized post-inhibitory rebound mechanisms in the dorsal cochlear nucleus [58] and superior paraolivary nucleus [59], and remain temporally precise in regions of the MGB [60]. These thalamic neurons in turn form offset-encoding synapses onto auditory cortex neurons that appear to be distinct from those representing sound onsets [61]. The auditory cortex also makes use of local inhibitory inputs to sharpen temporal spiking responses [62,63]. Furthermore, Fishbach et al. [64] have shown that A1 neurons produce enhanced responses to feature onsets, which may highlight the beginning of an auditory event in complex scenes.

Many natural sounds, including voices, are composed of frequencies that are harmonics of a common fundamental frequency. Such sounds are perceived as a single auditory object with a pitch at the fundamental frequency. Therefore, harmonicity is a useful cue for grouping sound sources in busy auditory scenes, including multi-talker environments [65••]. Harmonic grouping cues are often examined experimentally by mistuning one tone of a harmonic tone complex, which results in perception of the mistuned tone as a separate auditory object (e.g. Ref. [66]). Human [67] and macaque [68] auditory cortex produce enhanced responses to tone complexes that contain a mistuned harmonic, which correlate with subjects' perception of a second auditory source [67]. Descending feedback from cortex may play a key role in the process of harmonic scene segregation. Deactivating the auditory cortex disrupts the neural representation of the relative levels of two concurrent harmonic sounds in the IC [69], while lesioning the connections from A1 to the MGB impairs ferrets' performance on a mistuning detection task [70•].

Sensory experience and scene analysis

Listeners can learn to group known feature combinations, or use new statistical features, in order to better segregate sound sources within their individual auditory environments. In a process described by Bregman as 'schema-based integration' [1•], a listener can rely on a particular combination of sound features to segment a familiar sound source, be it the sound of their spouse's voice [71], a familiar language [72], or a statistical regularity encountered during an experiment [73]. The latter study demonstrated that schema learning can be rapid and implicit, and is derived from the statistics of the current acoustic environment [73]. Młynarski and McDermott [74] further showed that grouping is not limited to the well-studied primitive grouping features (harmonicity, common onsets and offsets, etc), but is also based on previously unexplored spectrotemporal features that commonly co-occur in natural sounds, such as speech and music. Thus, the features used for auditory scene analysis are more numerous and varied than previously appreciated.

Prior exposure to the target speech stream facilitates speech segregation in a multi-talker listening task, and magnetoencephalographic (MEG) recordings have shown that this principally involves reduced tracking in the auditory cortex of the non-primed speech stream [75]. Thus, auditory scene analysis is influenced by the statistical properties of the input, as well by linguistic information held in working memory. However, the neural basis by which we flexibly derive and use grouping features to segregate sounds is largely unexplored, particularly at the cellular level.

As mentioned in the introductory section of this review, the ability of listeners to understand speech in the presence of other sounds varies between individuals, even when factors like age and hearing status are taken into account [8,10]. Musical experience is likely to play a role here [6,8], and various forms of training have been shown to be effective in improving speech-in-noise perception [76,77•]. Furthermore, raising rats in the presence of noise with various spectrotemporal statistics leads to enhanced behavioral performance and A1 encoding of vocalizations in noise [78••]. Collectively, these studies highlight the importance of experience in shaping the capacity of the brain to segregate sound sources.

Cortical correlates of attending to a single talker

Once features from different sound sources are segmented, we can guide our attention to a single source of interest. Studies throughout the past decade have substantially improved our understanding of how selectively listening to a single speaker in a multi-talker environment alters cortical representations of sounds. Studies using MEG [79], EEG [80], ECoG arrays [81,82••] and depth electrodes [82••] have described an enhanced representation of the attended speech in human auditory cortex during these selective attention tasks (Figure 4). In fact, the attended speech stream dominates the cortical response to the extent that it can be accurately reconstructed from neural responses to the sound mixture as if it were presented alone [81,82••]. Several studies suggest that this form of attentional enhancement is observed in secondary, but not primary, auditory cortex [79,82••,83]. This is indicated by the late (>100 ms) timing of attentional effects that have

Figure 4
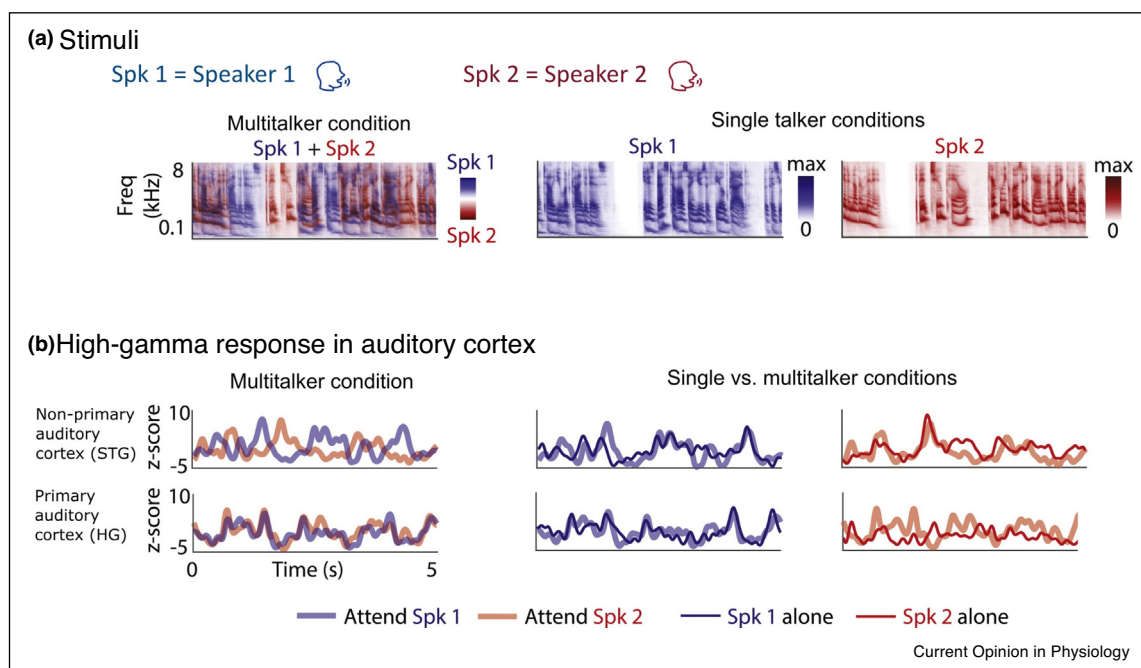
Representations of attended and unattended speech in human auditory cortex. **(a)** Each plot shows a spectrogram of speech presented to listeners, produced by speaker 1 (Spk 1; male; shown in blue) or speaker 2 (Spk 2; female; shown in red). The left panel is a spectrogram of speech from both speakers when presented together. The spectrotemporal content of the two speech streams largely overlaps, as in many real-world environments. The same speech streams presented in isolation are shown in the spectrograms on the right. **(b)** Example high-gamma responses from two depth electrodes recorded in one subject: one in the non-primary auditory cortex (superior temporal gyrus; STG) and the other in primary auditory cortex (Heschl's gyrus; HG). The response in the non-primary auditory cortex changes depending on which speaker is being attended, and resembles the response to that speaker in isolation. Conversely, the response in primary auditory cortex is similar when attending to either speaker, and is dominated by the spectrotemporal content of speaker 1. Therefore, responses in the primary auditory cortex are determined by the spectrotemporal content of the speech, irrespective of attention, whereas responses in non-primary auditory cortex better represent the attended speaker. Adapted from O'Sullivan *et al*. [82••].

been reported in MEG [79], EEG [80] and ECoG [81,82••] studies, even if the two speech streams are presented to different ears, providing a low-level binaural cue [84]. Studies using depth electrodes [82••] and fMRI [85] further support the functional localization of these attentional effects to the superior temporal gyrus.

The cellular mechanisms giving rise to selective listening effects in the presence of high energetic and informational masking remain largely unexplored in animal models. Because two competing speech streams often overlap substantially in their frequency content and simple stimulus statistics, the neural processes are likely to act on higher-order perceptual features (such as voice timbre and pitch [86]), which simple frequency gain filters alone are insufficient to explain.

Applications beyond sensory neuroscience

Our growing understanding of the neurophysiology of auditory scene analysis has important applications in artificial intelligence and clinical practice. The budding field of computational auditory scene analysis draws inspiration from neural algorithms to improve automatic speech recognition [87]. In another promising area of research, Han *et al.* [88] have demonstrated that the acoustics of an attended auditory source can be recovered online by measuring and decoding its enhanced representation in auditory cortex — even without knowledge of how the voices sound in isolation. The hope is that this information can be used to amplify target-relevant acoustical features in a listener's hearing prosthetic within a crowded auditory scene. This essentially involves reading the answer to the cocktail party problem from the brain itself, and feeding it back into the listener's hearing aid. As our knowledge of the biological solutions to the cocktail party problem improve, so will our ability to support hearing and communication throughout people's lifespans.

Author statement

Andrew J. King and Kerry M. M. Walker wrote the manuscript.

Conflict of interest statement

Nothing declared.

Acknowledgement

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Bregman AS: *Auditory Scene Analysis*. MIT Press; 1990
Albert Bregman's book provided the first detailed conceptualization of auditory scene analysis, and continues to guide current research in this field.

2. Cherry C: **Some experiments on the recognition of speech, with one ear and two ears**. *J Acoust Soc Am* 1953, **25**:975-979.

3. Middlebrooks JC, Simon JZ, Popper AN, Fay RR: *The Auditory System at the Cocktail Party*. Springer Handbook of Auditory Research; 2017.

4. 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.

5. Rennies J, Best V, Roverud E, Kidd G: **Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort**. *Trends Hear* 2019, **23**:1-21.

6. Alain C, Zendel BR, Hutka S, Bidelman GM: **Turning down the noise: the benefit of musical training on the aging auditory brain**. *Hear Res* 2014, **308**:162-173.

7. Johannesen PT, Pérez-González P, Kalluri S, Blanco JL, Lopez-Poveda EA: **The influence of cochlear mechanical dysfunction, temporal processing deficits, and age on the intelligibility of audible speech in noise for hearing-impaired listeners**. *Trends Hear* 2016, **20**:233121651664105.

8. Coffey EBJ, Chepesiuk AMP, Herholz SC, Baillet S, Zatorre RJ: **Neural correlates of early sound encoding and their relationship to speech-in-noise perception**. *Front Neurosci* 2017, **11**:479.

9. Ruggles D, Bharadwaj H, Shinn-Cunningham BG: **Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication**. *Proc Natl Acad Sci U S A* 2011, **108**:15516-15521.

10. Shinn-Cunningham B: **Cortical and sensory causes of individual differences in selective attention ability among listeners with normal hearing thresholds**. *J Speech Lang Hear Res* 2017, **60**:2976-2988.

11. Holmes E, Griffiths TD: **'Normal' hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception**. *Sci Rep* 2019, **9**:1-11.

12. Grant KJ, Mepani AM, Wu P, Hancock KE, de Gruttola V, Liberman MC, Maison SF: **Electrophysiological markers of cochlear function correlate with hearing-in-noise performance among audiometrically normal subjects**. *J Neurophysiol* 2020, **124**:418-431.

13. Willmore BDB, Cooke JE, King AJ: **Hearing in noisy environments: noise invariance and contrast gain control**. *J Physiol* 2014, **592**:3371-3381.

14. Cervera T, Ainsworth WA: **Effects of preceding noise on the perception of voiced plosives**. *Acta Acust United Acust* 2005, **91**:132-144.

15. Ben-David BM, Avivi-Reich M, Schneider BA: **Does the degree of linguistic experience (native versus nonnative) modulate the degree to which listeners can benefit from a delay between the onset of the maskers and the onset of the target speech?** *Hear Res* 2016, **341**:9-18.

16. Winslow RL, Sachs MB: **Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise**. *J Neurophysiol* 1987, **57**:1002-1021.

17. Kawase T, Ogura M, Sato T, Kobayashi T, Suzuki Y: **Effects of contralateral noise on the measurement of auditory threshold**. *Tohoku J Exp Med* 2003, **200**:129-135.

18. Marrufo-Pérez MI, Sturla Carreto D del P, Eustaquio-Martín A, Lopez-Poveda EA: **Adaptation to noise in human speech recognition depends on noise-level statistics and fast dynamic-range compression**. *J Neurosci* 2020, **40**:6613-6623.

19. Marrufo-Pérez MI, Eustaquio-Martín A, Lopez-Poveda EA: **Adaptation to noise in human speech recognition unrelated to the medial olivocochlear reflex**. *J Neurosci* 2018, **38**:4138-4145.

20. Wojtczak M, Klang AM, Torunsky NT: **Exploring the role of medial olivocochlear efferents on the detection of amplitude modulation for tones presented in noise**. *J Assoc Res Otolaryngol* 2019, **20**:395-413.

21. Mishra SK: **The role of efferents in human auditory development: efferent inhibition predicts frequency discrimination in noise for children**. *J Neurophysiol* 2020, **123**:2437-2448.

22. Rao A, Koerner TK, Madsen B, Zhang Y: **Investigating influences of medial olivocochlear efferent system on central auditory processing and listening in noise: a behavioral and event-related potential study**. *Brain Sci* 2020, **10**:428.

23. Dean I, Harper NS, McAlpine D: **Neural population coding of sound level adapts to stimulus statistics**. *Nat Neurosci* 2005, **8**:1684-1689.

24. Watkins PV, Barbour DL: **Specialized neuronal adaptation for preserving input sensitivity**. *Nat Neurosci* 2008, **11**:1259-1261.

25. Wen B, Wang GI, Dean I, Delgutte B: **Dynamic range adaptation to sound level statistics in the auditory nerve**. *J Neurosci* 2009, **29**:13797-13808.

26. Robinson BL, Harper NS, McAlpine D: **Meta-adaptation in the auditory midbrain under cortical influence**. *Nat Commun* 2016, **7**:1-8.

27. Herrmann B, Maess B, Johnsrude IS: **Aging affects adaptation to sound-level statistics in human auditory cortex**. *J Neurosci* 2018, **38**:1989-1999.

28. Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ: **Contrast gain control in auditory cortex**. *Neuron* 2011, **70**:1178-1191.

29. Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH:
• **Constructing noise-invariant representations of sound in the auditory pathway**. *PLoS Biol* 2013, **11**:e1001710
This study shows that neuronal adaptation to the mean and contrast of sounds increases along the auditory pathway in ferrets, and that this can account for the gradual emergence of noise-tolerant sound representations at the level of the primary auditory cortex.

30. Cooke JE, King AJ, Willmore BDB, Schnupp JWH: **Contrast gain control in mouse auditory cortex**. *J Neurophysiol* 2018, **120**:1872-1884.

31. Lohse M, Bajo VM, King AJ, Willmore BDB: **Neural circuits**
• **underlying auditory contrast gain control and their perceptual implications**. *Nat Commun* 2020, **11**:1-13
This study shows that neurons in the auditory midbrain, thalamus and cortex of mice exhibit comparable levels of contrast gain control, although the time constant of adaptation increases along the auditory pathway. Subcortical adaptation is not affected by optogenetic cortical silencing, and these physiological properties can account for adaptive changes in human perceptual thresholds.

32. Rocchi F, Ramachandran R: **Neuronal adaptation to sound statistics in the inferior colliculus of behaving macaques does not reduce the effectiveness of the masking noise**. *J Neurophysiol* 2018, **120**:2819-2833.

33. Dahmen JC, Keating P, Nodal FR, Schulz AL, King AJ: **Adaptation to stimulus statistics in the perception and neural representation of auditory space**. *Neuron* 2010, **66**:937-948.

34. Gleiss H, Encke J, Lingner A, Jennings TR, Brosel S, Kunz L, Grothe B, Pecka M: **Cooperative population coding facilitates efficient sound-source separability by adaptation to input statistics**. *PLoS Biol* 2019, **17**:e3000150.

35. Ding N, Simon JZ: **Adaptive temporal encoding leads to a background-insensitive cortical representation of speech**. *J Neurosci* 2013, **33**:5728-5735.

36. Mesgarani N, David SV, Fritz JB, Shamma SA: **Mechanisms of noise robust representation of speech in primary auditory cortex**. *Proc Natl Acad Sci U S A* 2014, **111**:6792-6797.

37. Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N: **Adaptation**
•• **of the human auditory cortex to changing background noise**. *Nat Commun* 2019, **10**:1-11
Recordings from the auditory cortex in human patients show that neuronal adaptation to the spectrotemporal statistics of background noise enhances the encoding of the phonetic structure of speech.

38. Kell AJE, McDermott JH: **Invariance to background noise as a signature of non-primary auditory cortex**. *Nat Commun* 2019, **10**:1-11.

39. Teschner MJ, Seybold BA, Malone BJ, Hüning J, Schreiner CE: **Effects of signal-to-noise ratio on auditory cortical frequency processing**. *J Neurosci* 2016, **36**:2743-2756.

40. Malone BJ, Heiser MA, Beitel RE, Schreiner CE: **Background noise exerts diverse effects on the cortical encoding of foreground sounds**. *J Neurophysiol* 2017, **118**:1034-1054.

41. Christensen RK, Lindén H, Nakamura M, Barkat TR: **White noise**
• **background improves tone discrimination by suppressing cortical tuning curves**. *Cell Rep* 2019, **29**:2041-2053.e4
The authors found that the presence of continuous broadband noise improved the discriminability of nearby tone frequencies by A1 neurons in mice, with comparable results observed behaviorally. They also showed that manipulating the frequency selectivity of A1 neurons optogenetically reproduced the enhancement in tone discrimination behavior, implying that changes in A1 responses are directly linked to these effects on tone-in-noise perception.

42. Sollini J, Chadderton P: **Comodulation enhances signal detection via priming of auditory cortical circuits**. *J Neurosci* 2016, **36**:12299-12311.

43. Natan RG, Rao W, Geffen MN: **Cortical interneurons differentially shape frequency tuning following adaptation**. *Cell Rep* 2017, **21**:878-890.

44. Cooke JE, Kahn MC, Mann EO, King AJ, Schnupp JWH, Willmore BDB: **Contrast gain control occurs independently of both parvalbumin-positive interneuron activity and shunting inhibition in auditory cortex**. *J Neurophysiol* 2020, **123**:1536-1551.

45. Souffi S, Lorenzi C, Varnet L, Huetz C, Edeline JM: **Noise-sensitive but more precise subcortical representations coexist with robust cortical encoding of natural vocalizations**. *J Neurosci* 2020, **40**:5228-5246.

46. Kolarik AJ, Pardhan S, Cirstea S, Moore BCJ: **Using acoustic information to perceive room size: effects of blindness, room reverberation time, and stimulus**. *Perception* 2013, **42**:985-990.

47. Bronkhorst AW, Houtgast T: **Auditory distance perception in rooms**. *Nature* 1999, **397**:517-520.

48. Traer J, McDermott JH: **Statistics of natural reverberation**
• **enable perceptual separation of sound and space**. *Proc Natl Acad Sci U S A* 2016, **113**:E7856-E7865
By analyzing the acoustical properties of a large number of natural scenes, the authors show that reverberation exhibits consistent statistical properties that human listeners utilize to separate sound sources from their environments. They argue that hearing in rooms and other reverberant situations can therefore be regarded as a form of auditory scene analysis.

49. Stone MA, Canavan S: **The near non-existence of "pure" energetic masking release for speech: extension to spectro-temporal modulation and glimpsing**. *J Acoust Soc Am* 2016, **140**:832-842.

50. Walker KMM, Bizley JK, King AJ, Schnupp JWH: **Multiplexed and robust representations of sound features in auditory cortex**. *J Neurosci* 2011, **31**:14565-14576.

51. Allen EJ, Burton PC, Olman CA, Oxenham AJ: **Representations of pitch and timbre variation in human auditory cortex**. *J Neurosci* 2017, **37**:1284-1293.

52. Patel P, Long LK, Herrero JL, Mehta AD, Mesgarani N: **Joint representation of spatial and phonetic features in the human core auditory cortex**. *Cell Rep* 2018, **24**:2051-2062.e2.

53. Shamma SA, Elhilali M, Micheyl C: **Temporal coherence and attention in auditory scene analysis**. *Trends Neurosci* 2011, **34**:114-123.

54. Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D: **Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party"**. *J Neurosci* 2013, **33**:1417-1426.

55. Atilgan H, Town SM, Wood KC, Jones GP, Maddox RK, Lee AKC,
•• Bizley JK: **Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding**. *Neuron* 2018, **97**:640-655.e4

The authors show that a visual cue enhances the representation of a temporally coherent sound in auditory cortex. This potential signature of cross-sensory binding is dependent upon visual cortex activity.

56. Sumby WH, Pollack I: **Visual contribution to speech intelligibility in noise**. *J Acoust Soc Am* 1954, **26**:212-215.

57. Darwin CJ: **Perceptual grouping of speech components differing in fundamental frequency and onset-time**. *Q J Exp Psychol Sect A* 1981, **33**:185-207.

58. Ding J, Benson TE, Voigt HF: **Acoustic and current-pulse responses of identified neurons in the dorsal cochlear nucleus of unanesthetized, decerebrate gerbils**. *J Neurophysiol* 1999, **82**:3434-3457.

59. Kopp-Scheinpflug C, Tozer AJB, Robinson SW, Tempel BL, Hennig MH, Forsythe ID: **The sound of silence: ionic mechanisms encoding sound termination**. *Neuron* 2011, **71**:911-925.

60. Anderson LA, Linden JF: **Mind the gap: two dissociable mechanisms of temporal processing in the auditory system**. *J Neurosci* 2016, **36**:1977-1995.

61. Scholl B, Gao X, Wehr M: **Nonoverlapping sets of synapses drive on responses and off responses in auditory cortex**. *Neuron* 2010, **65**:412-421.

62. Wehr M, Zador AM: **Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex**. *Nature* 2003, **426**:442-446.

63. Zhang LI, Tan AYY, Schreiner CE, Merzenich MM: **Topography and synaptic shaping of direction selectivity in primary auditory cortex**. *Nature* 2003, **424**:201-205.

64. Fishbach A, Yeshurun Y, Nelken I: **Neural model for physiological responses to frequency and amplitude transitions uncovers topographical order in the auditory cortex**. *J Neurophysiol* 2003, **90**:3663-3678.

65. Popham S, Boebinger D, Ellis DPW, Kawahara H, McDermott JH:
•• **Inharmonic speech reveals the role of harmonicity in the cocktail party problem**. *Nat Commun* 2018, **9**:2122
By manipulating the harmonic content of spoken sentences on multi-talker listening tasks, the authors demonstrate that harmonicity plays a key role in speech segmentation.

66. Moore BCJ, Glasberg BR, Peters RW: **Thresholds for hearing mistuned partials as separate tones in harmonic complexes**. *J Acoust Soc Am* 1986, **80**:479-483.

67. Dyson BJ, Alain C: **Representation of concurrent acoustic objects in primary auditory cortex**. *J Acoust Soc Am* 2004, **115**:280-288.

68. Fishman YI, Steinschneider M, Micheyl C, Micheyl C: **Neural representation of concurrent harmonic sounds in monkey primary auditory cortex: implications for models of auditory scene analysis**. *J Neurosci* 2014, **34**:12425-12443.

69. Nakamoto KT, Shackleton TM, Palmer AR: **Responses in the inferior colliculus of the guinea pig to concurrent harmonic series and the effect of inactivation of descending controls**. *J Neurophysiol* 2010, **103**:2050-2061.

70. Homma NY, Happel MFK, Nodal FR, Ohl FW, King AJ, Bajo VM: **A**
• **role for auditory corticothalamic feedback in the perception of complex sounds**. *J Neurosci* 2017, **37**:6149-6161.
This study shows that auditory corticothalamic feedback plays a critical role in the ability of ferrets to perform a mistuning detection task, indicating a possible role for descending projections from A1 in auditory scene analysis.

71. Johnsrude IS, Mackey A, Hakyemez H, Alexander E, Trang HP, Carlyon RP: **Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice**. *Psychol Sci* 2013, **24**:1995-2004.

72. Cooke M, Garcia Lecumberri ML, Barker J: **The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception**. *J Acoust Soc Am* 2008, **123**:414-427.

73. Woods KJP, McDermott JH: **Schema learning for the cocktail party problem**. *Proc Natl Acad Sci U S A* 2018, **115**:E3313-E3322.

74. Młynarski W, McDermott JH: **Ecological origins of perceptual grouping principles in the auditory system**. *Proc Natl Acad Sci U S A* 2019, **116**:25355-25364.

75. Wang Y, Zhang J, Zou J, Luo H, Ding N: **Prior knowledge guides speech segregation in human auditory cortex**. *Cereb Cortex* 2019, **29**:1561-1571.

76. Song JH, Skoe E, Banai K, Kraus N: **Training to improve hearing speech in noise: biological mechanisms**. *Cereb Cortex* 2012, **22**:1180-1190.

77. Whitton JP, Hancock KE, Shannon JM, Polley DB: **Audiomotor**
• **perceptual training enhances speech intelligibility in background noise**. *Curr Biol* 2017, **27**:3237-3247.e6
This is one of several studies demonstrating that training in adulthood can improve speech-in-noise perception by human listeners. In this case, training elderly hearing-impaired subjects on a computerized audio game for several weeks led to improvements in speech perception in levels of background noise comparable to those found in a crowded restaurant.

78. Homma NY, Hullett PW, Atencio CA, Schreiner CE: **Auditory**
•• **cortical plasticity dependent on environmental noise statistics**. *Cell Rep* 2020, **30**:4445-4458.e5
The authors show that exposing rat pups to different spectrotemporal noise statistics during a critical period of development results in enhanced vocalization-in-noise detection in adulthood, which was associated with improved encoding of vocalizations in the presence of noise in A1. This work therefore highlights the importance of early sensory experience in shaping signal-in-noise processing.

79. Ding N, Simon JZ: **Emergence of neural encoding of auditory objects while listening to competing speakers**. *Proc Natl Acad Sci U S A* 2012, **109**:11854-11859.

80. Hambrook DA, Tata MS: **Theta-band phase tracking in the two-talker problem**. *Brain Lang* 2014, **135**:52-56.

81. Mesgarani N, Chang EF: **Selective cortical representation of attended speaker in multi-talker speech perception**. *Nature* 2012, **485**:233-236.

82. O'Sullivan J, Herrero J, Smith E, Schevon C, McKhann GM,
•• Sheth SA, Mehta AD, Mesgarani N: **Hierarchical encoding of attended auditory objects in multi-talker speech perception**. *Neuron* 2019, **104**:1195-1209.e3
The authors found that the human primary auditory cortex represents both attended and unattended speech streams during a multi-talker listening task, while non-primary auditory cortex provided an enhanced representation of the attended speech.

83. Hausfeld L, Riecke L, Valente G, Formisano E: **Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes**. *Neuroimage* 2018, **181**:617-626.

84. Ding N, Simon JZ: **Neural coding of continuous speech in auditory cortex during monaural and dichotic listening**. *J Neurophys* 2012, **107**:78-89.

85. Scott SK, Rosen S, Beaman CP, Davis JP, Wise RJS: **The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes**. *J Acoust Soc Am* 2009, **125**:1737-1743.

86. Maddox RK, Shinn-Cunningham BG: **Influence of task-relevent and task-irrelevant feature continuity on selective auditory attention**. *J Assoc Res Otolaryngol* 2012, **13**:119-129.

87. Zeremdini J, Ben Messaoud MA, Bouzid A: **A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation**. *Brain Inform* 2015, **2**:155-166.

88. Han C, O'Sullivan J, Luo Y, Herrero J, Mehta AD, Mesgarani N: **Speaker-independent auditory attention decoding without access to clean speech sources**. *Sci Adv* 2019, **5**:eaav6134.