Trends in Neurosciences

CellPress REVIEWS

Review

Efficient Neural Coding in Auditory and Speech Perception

Judit Gervain[1,2] and Maria N. Geffen[3,*]

Speech has long been recognized as 'special'. Here, we suggest that one of the reasons for speech being special is that our auditory system has evolved to encode it in an efficient, optimal way. The theory of efficient neural coding argues that our perceptual systems have evolved to encode environmental stimuli in the most efficient way. Mathematically, this can be achieved if the optimally efficient codes match the statistics of the signals they represent. Experimental evidence suggests that the auditory code is optimal in this mathematical sense: statistical properties of speech closely match response properties of the cochlea, the auditory nerve, and the auditory cortex. Even more interestingly, these results may be linked to phenomena in auditory and speech perception.

Highlights

Efficient neural coding may support the selectivity for speech in the auditory pathway.

The auditory neuronal code matches the statistics of natural and behaviorally relevant sounds.

Speech perception may rely on the same auditory coding mechanisms that facilitate efficient coding of other natural sound statistics.

The Relevance of Efficient Neural Coding for Speech Perception

Speech has long been recognized as 'special' [1–6]. We prefer it over other sounds from birth onwards [6], and we are able to make fine-grained discriminations that allow us to convey an infinite amount of messages. The special status of speech has been studied from a variety of perspectives. Researchers of social cognition approach it as our species-specific communicative signal, and as the basis of learning and cultural transmission [7,8]. Others have claimed that speech is special because it is the only auditory signal that we 'feel', in the sense of perceiving the movement of our articulators, when producing it [1]. Here, we review experimental evidence for the hypothesis that speech is special for another reason, because our auditory system has evolved to encode it in an efficient way.

Organisms need to process the environmental signals they encounter, and the efficiency with which they do so may considerably impact their survival. The ability to process environmental signals efficiently is, therefore, assumed to be an important principle shaping the evolution of the sensory systems. Specifically, the theory of efficient neural coding [9,10] argues that our perceptual systems have evolved to encode environmental stimuli in the most efficient way. Information theory provides a mathematically precise and empirically testable framework to evaluate this theory. It defines efficient or optimal coding as one that transmits the highest fidelity information at the lowest cost (i.e., if the encoding maximally reduces the redundancy in the signal). Mathematically, this can be achieved if the optimally efficient codes match the statistics of the signals they represent [11].

In the past decades, the hypothesis that the neural code used by the perceptual systems is optimal in this mathematical sense has gained considerable empirical and theoretical support in vision [12]. More recently, experimental evidence has suggested that the auditory code may also be optimal [13–17]. Here, we link these findings to auditory perception, with special attention to speech perception. Conceiving of speech as an auditory signal that is particularly well-suited to match the encoding capabilities of the auditory system may contribute to a better

[1]Laboratoire Psychologie de la Perception, Université Paris Descartes, Paris, France
[2]Laboratoire Psychologie de la Perception, CNRS, Paris, France
[3]Departments of Otorhinolaryngology, Neuroscience and Neurology, University of Pennsylvania, Philadelphia, PA, USA

*Correspondence: mgeffen@pennmedicine.upenn.edu (M.N. Geffen).

understanding of speech perception phenomena and the 'special' nature of speech. This hypothesis is now gaining momentum [13,14,18–20], motivating the current review.

The Statistical Structure of Sounds

To test whether the mammalian auditory system codes sound in a mathematically optimal way, it is first necessary to describe the statistical structure of sounds. The space (in the mathematical sense) of all potential sounds is vast (Box 1). Within this space, natural sounds, including speech, comprise a compact yet multi-dimensional subspace. Analyses of statistical regularities in natural sounds have identified several prominent features. The temporal structure of

Box 1. The Efficient Auditory Coding Hypothesis

The potential space (in the abstract, mathematical sense) of all possible sounds is vast, but environmental sounds, animal vocalizations, and speech occupy specific subspaces. These subspaces are determined by the spectro-temporal statistics of the acoustic properties of sounds from different groups (Figure I A). It is, however, difficult to identify the relevant dimensions in the sound space. Research in exploring the space of environmental, vocalization, and speech sounds has focused on complementary approaches (Figure IB): (i) by shaping random noise according to some statistical constraints to generate sounds from different groups, or (ii) by using recorded sounds, and applying directed perturbations to these sounds along specific dimensions within this complex space to produce distorted sounds. The first approach allows researchers to test whether a particular statistical constraint is sufficient to define a sound category. The second approach tests whether the particular statistical constraint is required to define a sound category. We propose that throughout evolutionary history and during human development, transformations inconstraints over the perceived sounds produce an auditory code that encodes speech in an efficient fashion (Figure IC).

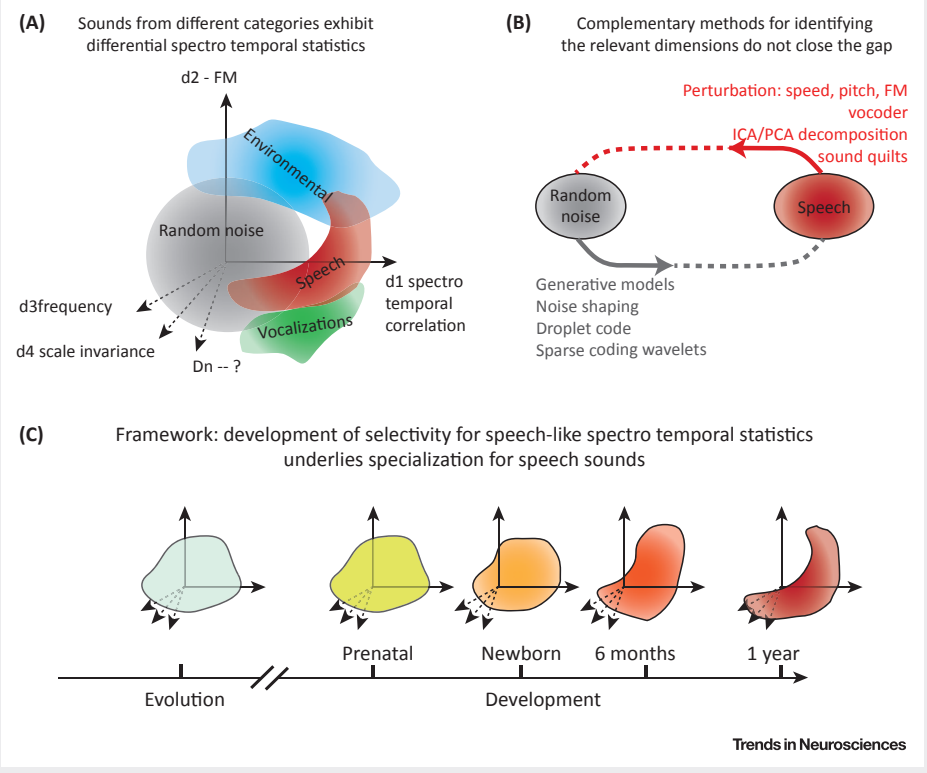

Figure I. Framework for Understanding the Development of Speech Selectivity. (A) Diagram of the spectro-temporal statistical space of different types of sounds projected on a subset of dimensions, including d1, spectro-temporal correlation; d2, frequency modulation (FM); d3, frequency; d4, scale-invariant coefficient; dn, other components to be identified. (B) Diagram of complementary methods to identify the relevant dimension. (C) Speech statistics are shapes throughout evolution and development.

many natural environmental sounds has a self-similar property: its power spectrum scales as 1/f [21], which means that the signals exhibit correlations across multiple time-scales. These spectral correlations translate into statistical dependencies across frequency and time, which can be captured with a histogram of the statistical features of sounds in the spectrotemporal domain [22] (Figure 1A). These dependencies can be encoded by a neuronal population that processes the inputs at multiple time-scales with varying degrees of resolution across scales [23,24]. Scale-invariant dependency occurs not just within the amplitude spectrum of sounds, but also across spectral bands; if we consider the spectrogram of a natural sound, we observe that the temporal fluctuations occur on a faster timescale in higher frequency bands than in lower frequency bands. As result, the temporal correlations in the spectrogram are shorter at high than at low frequencies [25,26].

Interestingly, the relationship between frequency and temporal correlations drives a differential perception of sounds that are generated under this statistical relationship. Varying the value of a single statistical parameter that controls the correlation within the temporal structure can yield a range of sound percepts to which both adults and infants exhibit sensitivity [26]. More generally, controlling a small number of statistical parameters for first- and second-order distributions of the means and variance of spectrotemporal channel components of sounds can reproduce 'sound textures', yielding percepts that range from a chorus of insects to helicopter sounds [15,27]. In contrast to environmental sounds, mammalian vocalizations, which often have a strongly harmonic structure, show peaks over their 1/f spectra, corresponding to the fundamental frequency and its harmonics [16].

The speech signal shows properties of both environmental sounds and harmonic vocalizations. Globally, speech also has a 1/f spectrum [21]. More locally, vowels have a harmonic structure, while different consonant classes show acoustic transience to different degrees, resulting in

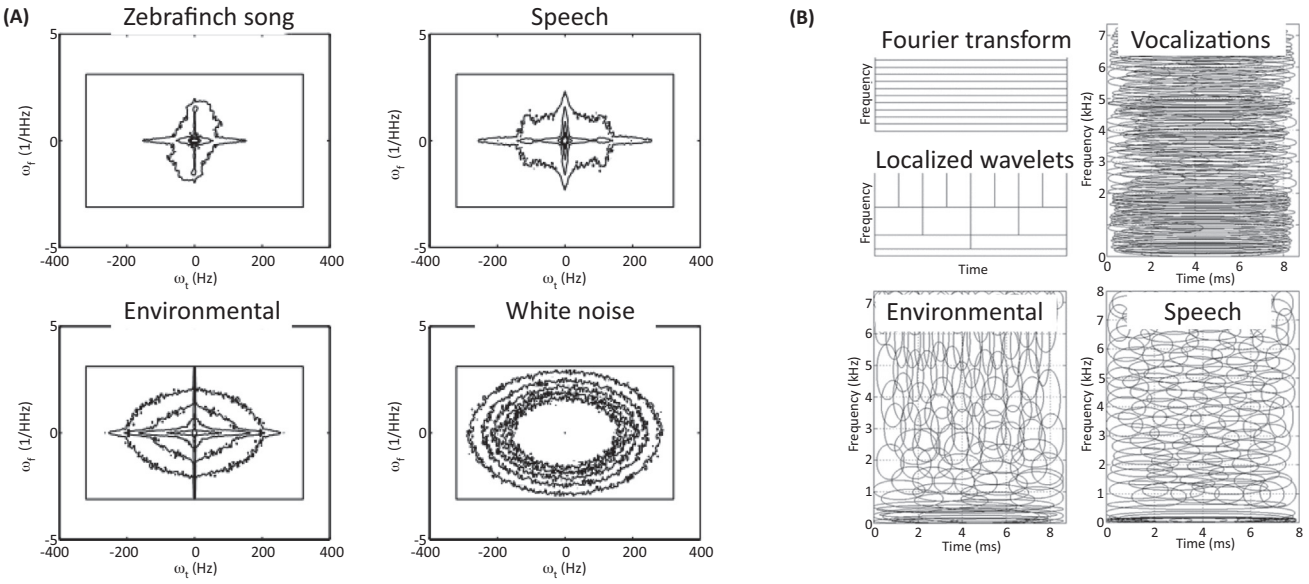

Figure 1. Spectrotemporal Characteristics of Different Types of Sounds. (A) The time-frequency histogram of a speech signal shows that speech shares characteristics with both animal vocalizations and environmental sounds (adapted from [22]). (B) The spectrotemporal characteristics of mathematically computed optimal filters also suggest that speech resembles both vocalizations and environmental sounds (adapted from [13]).

systematic variations in the local statistical structure of the speech signal. This led to conceptualizing speech as a modulated carrier signal [28,29]. The carrier signal produced by the vocal folds is modulated slowly in amplitude and frequency as a result of the dynamic changes of the vocal tract during phonation. The amplitude modulation corresponds to the envelope of the speech signal, while the frequency modulation corresponds to its temporal fine structure. The amplitude and frequency modulations can be obtained using a Hilbert transform applied to the speech signal. Based on the modulated carrier signal view of speech, powerful analysis and synthesis algorithms, called vocoders, have been developed, and can selectively manipulate the acoustic components of speech [30]. The amplitude and frequency modulation spectra of the speech signal have recently received considerable attention. It has been shown that the amplitude modulation spectrum, believed to correspond to the percept of speech rhythm, has a peak between 4–5 Hz (Figure 2). This temporal modulation is found across a wide range of different languages [31,32], with slight variations corresponding to well-established rhythmic and other prosodic differences between them [32].

Nonredundant, Optimal Mathematical Models of Sounds

According to the efficient coding hypothesis, the brain has evolved to efficiently process and respond to stimuli that occur in nature, reducing redundancy in their neural representations [9]. This principle posits that the statistical properties of neuronal responses should match the statistical structure of natural stimuli, and should maximize the efficiency in representation [10,33]. This is best achieved if neuronal responses constitute a sparse, nonredundant code, meaning that the code should be as parsimonious as possible yet capture the full range of variability in the signal structure along the relevant dimensions [34].

Following these principles, recent studies have derived sparse codes for different categories of sounds and compared them to the response properties of components of the auditory pathway. So far, to our knowledge, two mathematical approaches have been used. The first [13,18,19,35] uses independent component analysis (ICA). Imposing a sparsity constraint improves the decomposition of sounds into independent components [35], which is a statistical analysis that identifies the most informative dimensions in the spectrotemporal space of sounds. Optimal filter populations derived using ICA [13] for three categories of sounds

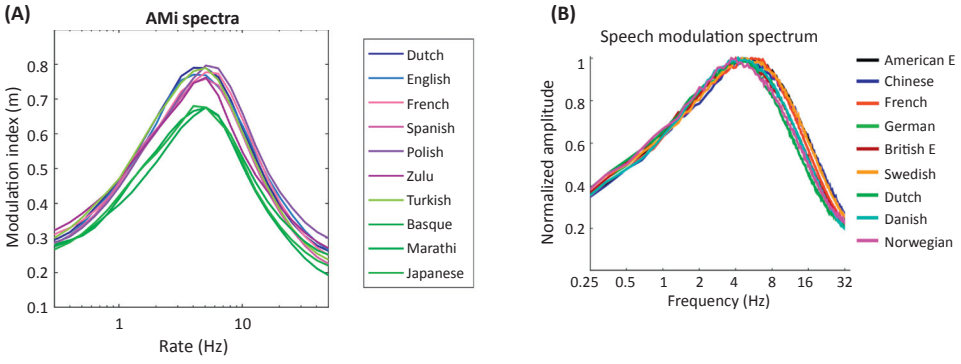

Figure 2. The Amplitude Modulation Spectra of Speech in Different Languages. (A) The amplitude modulation spectra of speech in 10 different languages from a database of clearly articulated, well-controlled recordings show both a strong similarity with a modulation peak at around 4 Hz as well as slight differences across languages in the strength and rate of modulation (adapted from [32]). (B) The amplitude modulation spectra of speech in nine different languages from large corpora with a wide range of different speech styles and registers show the same ~4 Hz modulation peak, but not the smaller cross-linguistic differences (adapted from [31]).

(i.e., environmental sounds, animal vocalizations, and speech) differ in their spectrotemporal properties, reflecting the statistics of the sound classes (Figure 1B). Thus, the optimal filters for animal sounds resemble a Fourier decomposition in conformity with the harmonic structure of these sounds; the filters for environmental sounds approximate wavelets, reflecting the fast transients in these sounds; whereas the filters for speech are in between these two representations. Importantly, the filters derived for speech as well as for a mix of environmental sounds and animal vocalizations, but not for environmental sounds alone or vocalizations alone, very closely match auditory nerve fiber tuning properties. At the more local level, filter populations for vowels match the global properties of speech, whereas those of different consonant classes vary from Fourier-like to wavelet-like representations. Crucially, these filter populations are well aligned with the response properties of cochlear nuclei [18]. Interestingly, the filter populations for speech in different languages well match the acoustic correlates of the percept of speech rhythm [19]. Furthermore, the basis for binaural sounds reproduced sound localization networks with a small number of components, suggesting that sound localization can be carried out with reduced representation that is optimized to the distribution of binaural dependencies in the natural world [36].

As a second mathematical approach, sparse coding models have been proposed. Identifying a sparse and efficient representation of sounds in terms of spikes and imposing a sparse binary code constraint on sound encoding replicates encoding features observed in the mammalian auditory system [15]. Specifically, the spike code representation of speech approximates time-domain cochlear filter estimates, and the frequency-bandwidth dependence of auditory nerve fibers.

These nonredundant codes stand in contrast to more traditional representations of sound in terms of a waveform over different spectral bands, such as a spectrogram or cochleogram. These traditional representations require a large number of parameters to fit the sound waveform. The assumption of sparsity in acoustic signals reduces the number of parameters required to represent a sound waveform.

Sounds with Naturalistic Statistics are Special for the Mammalian Auditory System

According to the efficient coding hypothesis, identifying the statistical dependencies in the structure of sounds yields insight into the structure of the neuronal code. This was tested by constructing artificial codes that were optimized according to some set of constraints to best represent natural sounds, and then compared to experimental measurements of responses of neurons in the auditory pathway. Such advanced mathematical models were, for instance, used to better understand the structure of receptive fields of auditory neurons. The assumption of sparsity responses in conjunction with analysis of a library of sounds yields spectrotemporal filters with different spectrotemporal relations that capture the diversity observed in the auditory pathway [24]. Using independent component analysis on a library of natural and speech sounds furthermore yielded a correlation between the bandwidth and center frequency of tuning, and predicted overrepresentation of the frequency of an overexposed tone. Such a relationship was identified experimentally for primary auditory cortical neurons [37]. Imposing a sparse coding constraint on natural sounds yielded single and multi-peaked frequency response units, such as found in the primate A1 [38]. Enforcing sparseness and a specific form of scaling of inputs, termed divisive normalization, in a network of neurons reproduced the set of auditory features within the auditory processing pathway [39]. Extending this code by adding a layer of neuronal connectivity captured the nonuniform distribution of spatial tuning that was observed experimentally in mammals [40].

Natural sounds, or sounds that exhibit naturalistic statistics, evoke enhanced responses throughout the auditory system. Neurons throughout the auditory pathway represent complex sounds as a population that integrates across neurons tuned to different spectrotemporal features of sounds. These can be described through the spectrotemporal receptive fields of neurons (STRF) [41,42]. The STRF of neurons indicates the range of frequencies and transformations in time in the stimulus amplitude that evoke a strong response and can predict the ability of neurons to represent and discriminate between complex sounds [43,44]. STRFs are typically determined from a set of randomized sounds that obey certain statistical constraints. Modifying those statistical constraints to capture statistics of natural sounds, such as conspecific vocalizations, amplifies cortical responses [43,45–47] (Figure 3). Furthermore, modifying the stimulus to exhibit a 1/f frequency spectrum yields tuned responses [48–50]. Indeed, such dependence is consistent with the 1/f structure of responses within the cochlea [51], and the enhanced information transmission at the level of auditory primary afferents [17] for naturalistic stimuli.

Particularly relevant to auditory communication are conspecific vocalizations. In many species, auditory cortical neurons exhibit enhanced tuning for natural vocalizations [52–56]. Vocalizations are encoded at a higher information rate when their statistics are unperturbed [57]. Furthermore, predictive models for auditory processing were able to predict activity more accurately in the primary auditory cortex when the stimulus was comprised of sounds with the statistical structure of conspecific vocalizations [46]. Interestingly, the responses of cortical neurons in mammals to speech can approach estimates for perceptual speech discrimination [58], and neuronal responses to phonetic features of speech sounds can be related to their spectrotemporal tuning properties [59,60].

Can Efficient Coding Explain Perception?

Few studies to date have directly addressed whether efficient coding principles can account for auditory percepts. Among these, one series of studies [25,26,61] tested how human adults, infants, and newborns perceive water sounds generated by a mathematical model (Figure 4) that consisted of a population of randomly spaced gamma tone chirps from a wide range of frequencies [25]. This model generated scale-invariant sounds when the temporal structure of the chirps scaled relative to their center frequency, and variable-scale sounds when chirps in different spectral bands varied in their temporal structure relative to their center frequency. Adults rated the scale-invariant sounds generated by the model as natural, and qualitatively described them as water sounds (e.g., rain, shower, ocean, etc.), whereas they rated the variable-scale sounds as unnatural and qualitatively described them as noise or machine-like sounds [25], suggesting that scale-invariance is indeed a statistical property underlying the percept of naturalness in sounds. Similarly, when habituated with the same scale-invariant water sounds, 5-month-old infants readily dishabituated to the variable-scale sounds, suggesting that they formed a perceptual category for the scale-invariant sounds during habituation. If, however, they were habituated to the variable-scale sounds, they did not dishabituate when hearing the scale-invariant ones, indicating that they did not perceive the variable-scale sounds as constituting a well-formed perceptual category [26]. The perceptual advantage of scale-invariant sounds appears to be present even earlier in human development. The newborn brain also discriminates between the scale-invariant and variable-scale water sounds [61].

Speech perception may also obey efficient coding principles. When speech is degraded by preserving only six frequency bands in a noise vocoder, adult listeners' speech recognition performance is better if the vocoder uses mathematically derived efficient filters rather than linear or cochleotopic filters [14]. Speech perception also shows scale-invariance in time: adult

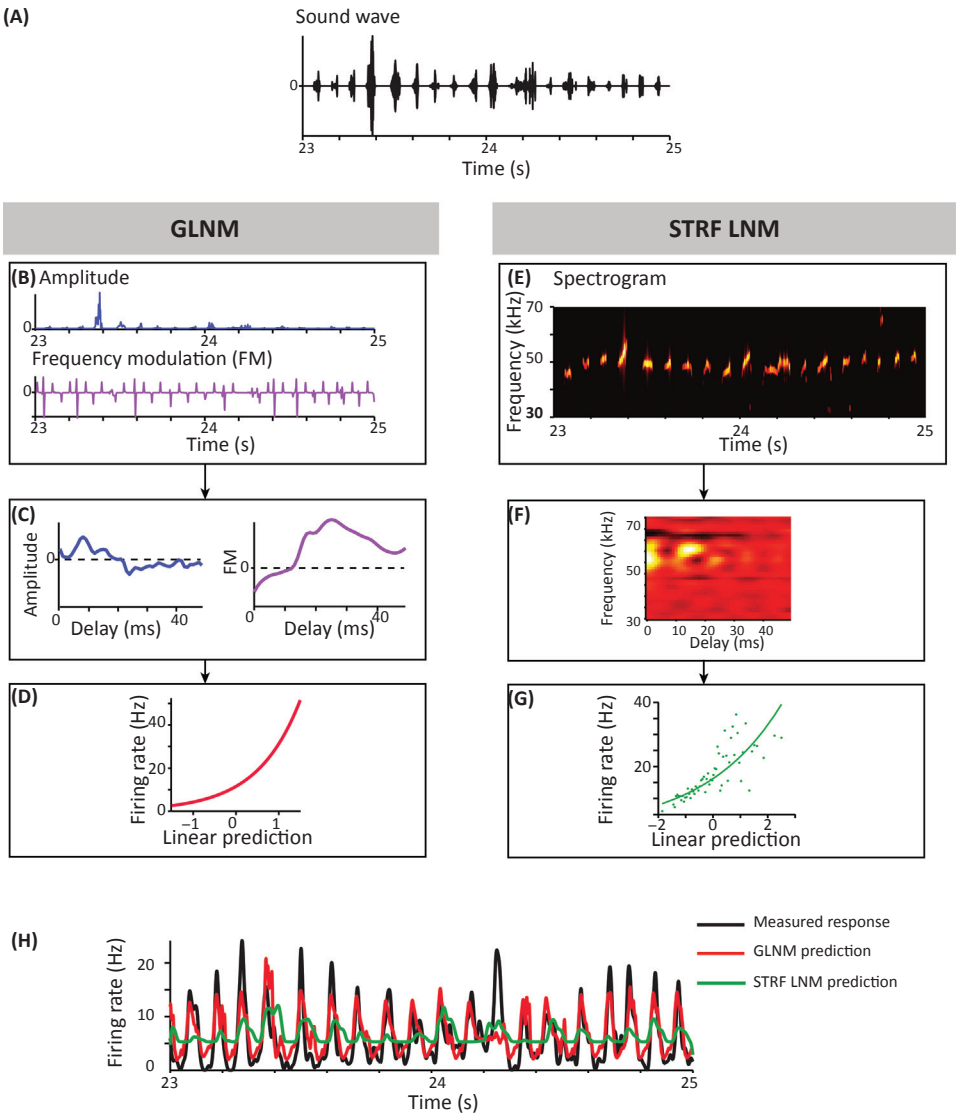

Figure 3. Predictive Model for Neuronal Responses to a Sequence of Conspecific Vocalizations in Awake Rat A1 Is Improved through Low-Dimensional Parametrization of the Stimulus (adapted from [46]). (A) stimulus waveform, which consisted of rat ultra-sonic vocalizations, concatenated at the naturalistic rate of production of 10 Hz. (B,E) Representation of the stimulus in (B) a two-dimensional space of frequency modulation and amplitude or (E) as a spectrogram. (C,F) Linear filters for responses of the neuron for the two models. (D,G) Instantaneous non-linearities used for the two models. (H) Firing rate (black), and model prediction based on two-parameter generalized linear model (GLNM, red) and on the full spectrogram linear-non-linear model (STRF LNM, green).

[62–66], child [67,68], and even newborn [20] listeners readily adapt to time-compressed speech in their native language as well as in rhythmically similar non-native languages. This indicates that adaptation happens at the auditory, rather than at the abstract linguistic level, confirming the auditory system's ability to encode scale-invariance in time. It needs to be noted, however, that listeners can only adapt to speech compressed maximally to about 30% of its original duration. Beyond that, adaptation breaks down. Some researchers thus interpret

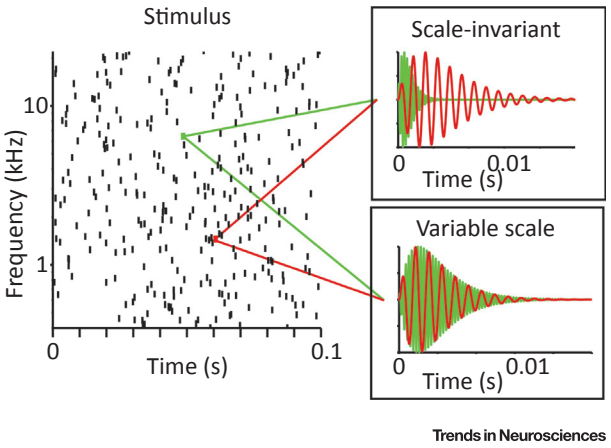

Figure 4. The Mathematical Model of Water Sounds (adapted from [25]). A scale-invariant or variable-scale random sound was constructed as superposition of gammatone chirps with the same time and frequency. Left: the onset timing and center frequency of each gammatone used for both scale-invariant and variable-scale sounds. Right insets: two representative gammatones for either sound. Scale-invariant gammatones had a constant cycle constant of decay, whereas variable-scale sounds had a constant time constant of delay; as a result, scale-invariant sounds differed in duration across frequencies, whereas variable-scale sounds did not. In perceptual judgement experiments, adult subjects rated scale-invariant, but not variable-scale sounds as natural for a wide range of parameters.

adaptation to time-compressed speech not as a manifestation of scale-invariant processing [69]. Rather, it is taken as evidence in favor of the multiple time-scale model of speech perception [70–72], which posits that speech is simultaneously processed at a few privileged time-scales, roughly corresponding to the linguistic units of (sub)phonemes, syllables, and phrases and sustained in the brain by a hierarchy of embedded neural oscillations in the low gamma (25–35 Hz), theta (4–8 Hz), and delta (1–2 Hz) bands. This model predicts that speech perception is not fully scale-invariant in time. Rather, adaptation to compression is only possible if the rhythm of the signal remains within these privileged frequency ranges, and the lower limit on listeners' ability to adapt to compression is seen as an indication that the theta rhythm is no longer maintained, making speech perception impossible.

Concluding Remarks and Future Perspectives

The research findings discussed in this review suggest that auditory perception may obey the principles of efficient neural coding, relying on the informational, theoretical notion of optimality. The existing studies demonstrate that the approaches for understanding the mathematical structure of sounds can yield predictions about neuronal encoding throughout the auditory pathway. The correspondence between neuronal responses and model predictions, conversely, is consistent with the notion that the neuronal representation of sounds is optimized for the statistical features of sounds found in nature. The auditory neural code appears to be particularly well-matched to the statistical properties of speech.

The efficient neural coding approach opens up an interesting perspective on auditory and speech perception. Nevertheless, a number of issues remain unresolved (see Outstanding Questions). First, efficient coding assumes that neural representations are optimal. This stands in apparent contrast to the redundancy that is well attested in biological systems. Since both signals and the computing units (neurons) are noisy, and can be damaged, a certain amount of redundancy is necessary and even desirable to make auditory representations robust and resilient. Future models of efficient auditory coding will need to take into account the need for resilience in the face of noise or damage.

Outstanding Questions

How can redundancy and optimality in the neuronal code be reconciled?

Does the efficient coding principle operate across levels, both peripheral and central, of the auditory pathway?

How are statistical regularities in speech across different temporal scales, such as phonemes and words, combined in their representation?

How does efficient neural coding relate to other theories of auditory and speech perception?

Second, the general mathematical principle of coding efficiency does not specify the aspects of the signal that need to be encoded, nor the neural structures that are involved. The theory leaves underspecified whether efficient coding principles should operate at the level of individual neurons, neuronal assemblies, or even larger structures. For a better understanding of these issues, efficient coding models need to be integrated with anatomical and neurophysiological as well as acoustic and linguistic accounts.

Third, it remains open how the efficient neural coding account relates to other theories of auditory perception. As discussed above, the temporal scale-invariance prediction of the efficient coding of speech stand in (apparent) contradiction with the multiple time-scale model of speech perception [70]. Whether these models may be integrated, and if so how, remains an important question for future research.

Acknowledgements

This work was supported by Human Frontier in Science Foundation Young Investigator Award to M.N.G. and J.G.; National Institutes of Health (Grant numbers NIH R01DC014700, NIH R01DC015527), and the Pennsylvania Lions Club Hearing Research Fellowship to M.G.N.; an ERC Consolidator Grant 773202 ERC-2017-COG 'BabyRhythm', the LABEX EFL (ANR-10-LABX-0083) and the ANR grant ANR-15-CE37-0009-01 awarded to J.G. M.N.G. is the recipient of the Burroughs Wellcome Award at the Scientific Interface. We thank Janet Werker and Marcelo Magnasco for helpful discussions. This research was supported in part by NSF Grant No. PHY-1748958, NIH Grant No. R25GM067110, and the Gordon and Betty Moore Foundation Grant No. 2919.01.

References

1. Liberman, A.M. *et al.* (1967) Perception of the speech code. *Psychol. Rev.* 5, 552–563
2. Liberman, A.M. (1984) On finding that speech is special. In *Handbook of Cognitive Neuroscience* (Gazzaniga, M.S., ed.), pp. 169–197, Springer Verlag
3. Marler, P. and Peters, S. (1981) Birdsong and speech: evidence for special processing. In *Perspectives on the Study of Speech* (Eimas, P. D. and Miller, J.L., eds), pp. 75–112, Lawrence Erlbaum Associates
4. Pinker, S. and Jackendoff, R. (2005) The faculty of language: what's special about it? *Cognition* 95, 201–236
5. Vatakis, A. *et al.* (2008) Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *J. Vis.* 8, 14. 1–11
6. Vouloumanos, A. and Werker, J.F. (2004) Tuned to the signal: the privileged status of speech for young infants. *Dev. Sci.* 7, 270–276
7. Csibra, G. and Gergely, G. (2009) Natural pedagogy. *Trends Cogn. Sci.* 13, 148–153
8. Tomasello, M. and Farrar, M.J. (1986) Joint attention and early language. *Child Dev.* 57, 1454–1463
9. Attneave, F. (1954) Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193
10. Barlow, H.B. (1961) Possible principles underlying the transformation of sensory messages. In *Sensory Communication* (Rosenblith, W., ed.), pp. 217–234, MIT Press
11. Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423, 623–656
12. Simoncelli, E.P. and Olshausen, B.A. (2001) Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216
13. Lewicki, M.S. (2002) Efficient coding of natural sounds. *Nat. Neurosci.* 5, 356–363
14. Ming, V.L. and Holt, L.L. (2009) Efficient coding in human auditory perception. *J. Acoust. Soc. Am.* 126, 1312–1320
15. Smith, E.C. and Lewicki, M.S. (2006) Efficient auditory coding. *Nature* 439, 978–982
16. Attias, H. and Schreiner, C. (1997) Temporal low-order statistics of natural sounds. *Adv. Neural Inf. Process. Syst.* 9, 27–33
17. Rieke, F. *et al.* (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. Biol. Sci.* 262, 259–265
18. Stilp, C.E. *et al.* (2013) Speech perception in simulated electric hearing exploits information-bearing acoustic change. *J. Acoust. Soc. Am.* 133, EL136–EL141
19. Guevara Erra, R. and Gervain, J. (2016) The efficient coding of speech: cross-linguistic differences. *PLoS One* 11, e0148861
20. Issard, C. and Gervain, J. (2017) Adult-like processing of time-compressed speech by newborns: a NIRS study. *Dev. Cogn. Neurosci.* 25, 176–184
21. Voss, R.F. and Clarke, J. (1975) '1/f noise' in music and speech. *Nature* 258, 317–318
22. Singh, N. and Theunissen, F. (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* 114, 3394–3411
23. Chi, T. *et al.* (2005) Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887
24. Fritz, J. *et al.* (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223
25. Geffen, M.N. *et al.* (2011) Auditory perception of self-similarity in water sounds. *Front. Integr. Neurosci.* 5, 15
26. Gervain, J. *et al.* (2014) Category-specific processing of scale-invariant sounds in infancy. *PLoS One* 9, e96278
27. 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
28. Plomp, R. (1964) The ear as a frequency analyzer. *J. Acoust. Soc. Am.* 36, 1628–1636
29. Houtgast, T. and Steeneken, H.J.M. (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.* 77, 1069–1077
30. Drullman, R. (1995) Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am.* 97, 585–592
31. Ding, N. *et al.* (2017) Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* 81, 181–187
32. Varnet, L. *et al.* (2017) A cross-linguistic study of speech modulation spectra. *J. Acoust. Soc. Am.* 142, 1976
33. MacKay, D.M. (1956) Towards an information-flow model of human behaviour. *Br. J. Psychol.* 47, 30–43

34. Olshausen, B.A. and Field, D.J. (2004) Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487
35. Hyvärinen, A. *et al.* (2002) *Independent Component Analysis,* Wiley
36. Mlynarski, W. (2014) Efficient coding of spectrotemporal binaural sounds leads to emergence of the auditory space representation. *Front. Comput. Neurosci.* 8, 26
37. Saxe, A.M. *et al.* (2011) Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. *NIPS* 24, 1971–1979
38. Kadia, S.C. and Wang, X. (2003) Spectral integration in A1 of awake primates: neurons with single- and multipeaked tuning characteristics. *J. Neurophysiol.* 89, 1603–1622
39. Kozlov, A.S. and Gentner, T.Q. (2016) Central auditory neurons have composite receptive fields. *Proc. Natl. Acad. Sci. U. S. A.* 113, 1441–1446
40. Mlynarski, W. (2015) The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS Comput. Biol.* 11, e1004294
41. Depireux, D.A. *et al.* (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J. Neurophysiol.* 85, 1220–1234
42. Theunissen, F.E. *et al.* (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.* 20, 2315–2331
43. Woolley, S. *et al.* (2005) Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat. Neurosci.* 8, 1371–1379
44. Elie, J.E. and Theunissen, F.E. (2015) Meaning in the avian auditory cortex: neural representation of communication calls. *Eur. J. Neurosci.* 41, 546–567
45. Nelken, I. *et al.* (1999) Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* 397, 154–157
46. Carruthers, I.M. *et al.* (2013) Encoding of ultrasonic vocalizations in the auditory cortex. *J. Neurophysiol.* 109, 1912–1927
47. Carruthers, I.M. *et al.* (2015) Emergence of invariant representation of vocalizations in the auditory cortex. *J. Neurophysiol.* 114, 2726–2740
48. Escabi, M.A. *et al.* (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J. Neurosci.* 23, 11489–11504
49. Garcia-Lazaro, J. *et al.* (2006) Tuning to natural stimulus dynamics in primary auditory cortex. *Curr. Biol.* 16, 264–271
50. Blackwell, J.M. *et al.* (2016) Stable encoding of sounds over a broad range of statistical parameters in the auditory cortex. *Eur. J. Neurosci.* 43, 751–764
51. Robles, L. and Ruggero, M.A. (2001) Mechanics of the mammalian cochlea. *Physiol. Rev.* 81, 1305–1352
52. Gehr, D.D. *et al.* (2000) Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. *Hear. Res.* 150, 27–42
53. Huetz, C. *et al.* (2009) A spike-timing code for discriminating conspecific vocalizations in the thalamocortical system of anesthetized and awake guinea pigs. *J. Neurosci.* 29, 334–350
54. Wang, X. *et al.* (1995) Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J. Neurophysiol.* 74, 2685–2706
55. Galindo-Leon, E.E. *et al.* (2009) Inhibitory plasticity in a lateral band improves cortical detection of natural vocalizations. *Neuron* 62, 705–716
56. Liu, R.C. and Schreiner, C.E. (2007) Auditory cortical detection and discrimination correlates with communicative significance. *PLoS Biol.* 5, e173
57. Holmstrom, L.A. *et al.* (2010) Efficient encoding of vocalizations in the auditory midbrain. *J. Neurosci.* 30, 802–819
58. Mesgarani, N. *et al.* (2008) Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.* 123, 899–909
59. Ahissar, E. *et al.* (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13367–13372
60. Mesgarani, N. *et al.* (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010
61. Gervain, J. *et al.* (2016) The neural correlates of processing scale-invariant environmental sounds at birth. *Neuroimage* 133, 144–150
62. Banai, K. and Lavner, Y. (2012) Perceptual learning of time-compressed speech: more than rapid adaptation. *PLoS One* 7, e47099
63. Nourski, K.V. *et al.* (2009) Temporal envelope of time-compressed speech represented in the human auditory cortex. *J. Neurosci.* 29, 15564–15574
64. Sebastian-Galles, N. *et al.* (2000) Adaptation to time-compressed speech: phonological determinants. *Percept. Psychophys.* 62, 834–842
65. Pallier, C. *et al.* (1998) Perceptual adjustment to time-compressed speech: a cross-linguistic study. *Mem. Cogn.* 26, 844–851
66. Dupoux, E. and Green, K. (1997) Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 914–927
67. Orchik, D.J. and Oelschlaeger, M.L. (1977) Time-compressed speech discrimination in children and its relationship to articulation. *J. Am. Audiol. Soc.* 3, 37–41
68. Guiraud, H. *et al.* (2013) Adaptation to natural fast speech and time-compressed speech in children. *INTERSPEECH*, 1370–1374
69. Ghitza, O. (2011) Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front. Psychol.* 2, 130
70. Giraud, A.L. and Poeppel, D. (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517
71. Poeppel, D. (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as asymmetric sampling in time. *Speech Commun.* 41, 245–255
72. Ghitza, O. *et al.* (2012) Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Front. Hum. Neurosci.* 6, 340