A THEORY OF MAXIMIZING SENSORY INFORMATION

Hateren, J.H. van

Biol. Cybern. 68, 23–29 (1992)

Biological Cybernetics

A theory of maximizing sensory information

J. H. van Hateren

Department of Biophysics, University of Groningen, Westersingel 34, 9718 CM Groningen, The Netherlands

Received April 2, 1992/Accepted in revised form May 22, 1992

Abstract. A theory is developed on the assumption that early sensory processing aims at maximizing the information rate in the channels connecting the sensory system to more central parts of the brain, where it is assumed that these channels are noisy and have a limited dynamic range. Given a stimulus power spectrum, the theory enables the computation of filters accomplishing this maximizing of information. Resulting filters are band-pass or high-pass at high signal-to-noise ratios, and low-pass at low signal-to-noise ratios. In spatial vision this corresponds to lateral inhibition and pooling, respectively. The filters comply with Weber's law over a considerable range of signal-to-noise ratios.

1 Introduction

What principles govern the first stages of sensory processing? An answer due to Attneave (1954) and Barlow (1961) is that the first stages of sensory processing are designed to remove redundancy from the stimulus. Redundancy in the stimulus is associated with those features and characteristics that are repetitive and predictable. For instance, Kretzmer (1952; see also Srinivasan et al. 1982) showed that elements (pixels) in an image are on average not very different from their immediate neighbours. By removing this predictability, e.g. by predictive coding (Kretzmer 1952), the stimuli can be more efficiently transmitted through the available channels connecting the early sensory system to more central parts of the brain.

As already noted by Kretzmer (1954), reducing redundancy has its drawbacks. In particular, when the stimulus is very noisy, redundancy is beneficial because it increases the reliability of transmission. In these circumstances it may even be better to increase redundancy, thus using a repetitive code with reasonable success, rather than using a compact code that results in many errors.

In this article I discuss a more general strategy for early sensory processing. It results in a reduced redundancy of the stimulus at high signal-to-noise ratios, and in an increased redundancy at low signal-to-noise ratios. The general assumption it is based upon is that early sensory processing maximizes the amount of information transmitted through a noisy channel of limited dynamic range. I will derive the general theory, including several properties and approximations, and give a numerical example showing that the theory is consistent with Weber's law for sensory processing. The purpose of this article is to present the theory in a way as general as possible. Elsewhere I present specific applications of the theory to spatiotemporal processing in the early visual system (van Hateren 1992a, b).

2 Theory

Below I will sketch the theory (Sect. 2.1), derive the main results (Sect. 2.2), give a numerical example (Sect. 2.3), derive several properties of the theory (Sect. 2.4), and show that the theory leads to a close approximation of Weber's law (Sect. 2.5).

2.1 Theoretical scheme

The scheme underlying the theory is shown in Fig. 1; the equations on the right are explained in Sect. 2.2. A stimulus is acquired by a transducer which also acts as a prefilter. The prefilter limits the frequency-bandwidth of the stimulus, in order to restrict the amount of further processing required. The resulting prefiltered stimulus is contaminated by noise, either due to the stimulus (e.g., photon noise) or due to the transducer. For simplicity, we will lump both of these noise sources to a source of additive noise introduced after the prefiltering.

The noisy, prefiltered stimulus is eventually transferred to a channel (e.g., a neuron) that has a limited dynamic range (i.e., it can only support a certain range of response values), and that produces some noise of its

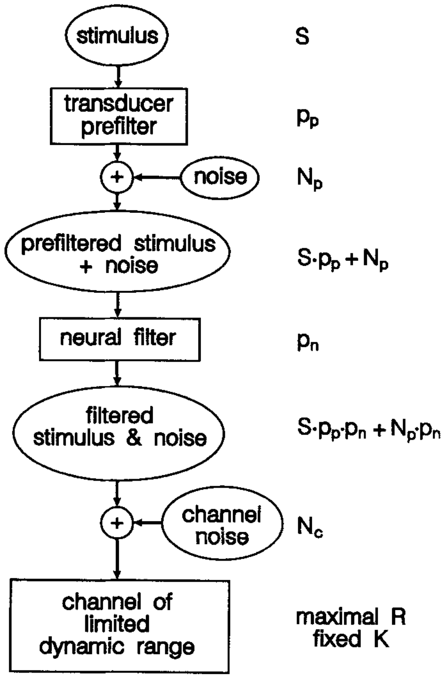

Fig. 1. Scheme of early sensory processing. A stimulus is transduced and low-pass filtered by a prefilter, and to the result noise is added. A neural filter subsequently filters the prefiltered stimulus in such a way, that as much as possible of the information it contains is transmitted through a channel of limited dynamic range that produces some noise of its own

own. Thus the channel has a limited information capacity (see e.g. Goldman 1953). It is the purpose of the neural filter in Fig. 1 to condition the prefiltered stimulus in such a way that the available channel capacity is utilized as much as possible. Thus the neural filter must be chosen such that the signal present in the channel gives as much information on the stimulus as possible, with the constraint that the signal values stay within the channel's dynamic range. In the next section I show how the unique neural filter with this property can be computed.

2.2 Derivation of the theory

The theory will be developed in the frequency domain. The frequency f then denotes a component of a suitable Fourier transform of the stimulus. For example, when applying the theory to spatial vision, f could represent a spatial frequency. Although f is treated below as a one-dimensional variable, generalization to more dimensions is straightforward (e.g., three dimensions for images changing in time, with two dimensions for spatial frequencies, and one for the temporal frequency).

The power spectrum of the stimulus (see e.g. van der Ziel 1970) is denoted by $S(f)$. Obviously, if $S(f)$ is completely different for different stimuli, the resulting neural filter (see Fig. 1) will be different as well. Then each stimulus would require its own filter. Fortunately, the power spectrum of stimuli is often much more constant than one might expect: it is mainly the phase of the spectrum, rather than its amplitude (or power), that determines differences between individual stimuli (Oppenheim and Lim 1981). For example, the spatial power spectrum of natural images is usually proportional to $1/f^2$ in good approximation (Field 1987; see also van Hateren 1992a). If $S(f)$ has fixed characteristics, with individual stimuli only varying slightly around it, we can find a fixed neural filter optimized for this average power spectrum.

If the prefilter has a power transfer function $p_p(f)$, the resulting prefiltered stimulus will have a power spectrum $S_p(f)$, with

$$S_p(f) = S(f)p_p(f)\,. \tag{1}$$

To this power spectrum, noise with a power spectrum $N_p(f)$ is added, and the result is filtered by a neural filter with a power transfer function $p_n(f)$, resulting in

$$S_p(f)p_n(f) + N_p(f)p_n(f)\,. \tag{2}$$

Finally, the channel adds noise with a power spectrum $N_c(f)$, thus yielding the total signal and noise power in the channel, $P_c(f)$

$$P_c(f) = S_p(f)p_n(f) + N_p(f)p_n(f) + N_c(f)\,. \tag{3}$$

$P_c(f)$ consists of a part due to the signal $(S_p(f)p_n(f))$ and a part due to the noise $(N_p(f)p_n(f) + N_c(f))$. Thus the signal-to-noise ratio $\sigma(f)$ (signal amplitude divided by noise amplitude) in the channel is

$$\sigma(f) = \left[\frac{S_p(f)p_n(f)}{N_p(f)p_n(f) + N_c(f)}\right]^{1/2}. \tag{4}$$

From $\sigma(f)$ we can find the information rate R (see e.g. Goldman 1953)

$$\begin{aligned} R &= \int_{\Delta f} \log_e(1 + \sigma^2(f))\,\mathrm{d}f \\ &= \int_{\Delta f} \log_e\left(1 + \frac{S_p(f)p_n(f)}{N_p(f)p_n(f) + N_c(f)}\right)\mathrm{d}f\,, \end{aligned} \tag{5}$$

where Δf is the frequency-bandwidth selected by the prefilter. I have chosen $\log_e$ rather than $\log_2$ in (5) for mathematical convenience. Thus R must be divided by $\log_e 2$ if the information is wanted in terms of bits.

How much of the dynamic range of the channel will be occupied by the total power in the channel, as given by (3)? We will approximate the dynamic range needed for signal and noise by the mean square value K^2 of the response in the channel, which follows from (see e.g. van der Ziel 1970)

$$\begin{aligned} K^2 = \int_{\Delta f} P_c(f)\,\mathrm{d}f = \int_{\Delta f} (S_p(f)p_n(f) &+ N_p(f)p_n(f) \\ &+ N_c(f))\,\mathrm{d}f\,. \end{aligned} \tag{6}$$

K is a measure of the range of response values occurring in the channel (due to both signal and noise). For a given channel with a limited dynamic range, K has to be fixed to a certain maximum.

The final step in this derivation is the execution of the main strategy of the theory: the information rate R, as given in (5), has to be maximized (through varying $p_n(f)$), while keeping the response within the channel's dynamic range (i.e., $K = \text{constant}$, see (6)). A similar problem was solved by Goldman (1953, p. 159) using the method of Lagrange multipliers, and we will follow that approach here. For simplicity of notation, I will drop from here on the arguments (f) of spectra and transfer functions. With λ a Lagrange multiplier, we require that

$$\frac{\partial}{\partial p_n} \log_e\left(1 + \frac{S_p p_n}{N_p p_n + N_c}\right) + \lambda \frac{\partial}{\partial p_n}(S_p p_n + N_p p_n + N_c) = 0\,. \quad (7)$$

This leads to

$$p_n = \frac{-N_c(2N_p + S_p) + \left(N_c^2 S_p^2 - 4N_p S_p N_c \frac{1}{\lambda}\right)^{1/2}}{2N_p(N_p + S_p)}\,, \quad (8)$$

where the negative root of p_n was dropped, because a power transfer function has to be nonnegative. Now we can find p_n by choosing λ such that (6) is satisfied with K fixed. This has to be done numerically by varying λ, thus varying p_n (8) and thereby (3), and (6).

2.3 Example

Figure 2 shows a numerical example of the application of the theory for an average signal-to-noise ratio ($\overline{\text{SNR}}$, see Sect. 2.5) of 100. In Fig. 2a the amplitude spectrum (square root of the power spectrum S) of a stimulus is shown. The power spectrum of the stimulus here depends as $1/f^2$ on the frequency f. The stimulus is subsequently filtered by a prefilter with an amplitude transfer function (square root of the power transfer function p_p) shown in Fig. 2b. The result is shown as the solid line in Fig. 2c. The dashed line shows additive noise with a power spectrum N_p. The neural filter of Fig. 2d is such that it maximizes the information rate in the channel (see below). The solid line in Fig. 2e shows the stimulus after pre- and neural filtering, the line with the short dashes the noise of Fig. 2c filtered by the neural filter, and the line with the long dashes additive channel noise (with a power spectrum N_c). The amplitude spectra of Fig. 2e are such, that if they are squared (yielding power spectra), added, and integrated over f, as in (6), the result is a mean square response K^2 that exactly fits the channel's capabilities (i.e., its dynamic range). Fig. 2f shows the signal-to-noise ratio σ, as in (4), derived from the signal and noise shown in Fig. 2e. The information (Fig. 2g) then follows from the integrand of (5), and the total information rate R from the integral over the curve in Fig. 2g. The neural filter in Fig. 2d is the filter that maximizes R, under the constraint that the response does not exceed the channel's dynamic range. Finally, Fig. 2h shows the combined action of the prefilter and the neural filter.

Note that the neural filter (Fig. 2d) has a maximum for the frequency where the signal and noise in Fig. 2c are equal (see Sect. 2.4). Both lower and higher frequencies are reduced. Lower frequencies are reduced because they are so strongly present in the stimulus (Fig. 2a), that they threaten to occupy too much of the channel's dynamic range. Although reducing them costs some information, this is more than regained by the boosting of other frequencies, protecting those against the detrimental effects of channel noise. The reason of this surplus of information is the logarithmic relationship between signal-to-noise ratio and information, as in the integrand of (5). Thus it is better to have many frequencies of moderate signal-to-noise ratio, than to have a mixture of very low and very high ones.

Higher frequencies are also reduced by the neural filter, because their original signal-to-noise ratio (see Fig. 2c) is so small that they carry very little information. Therefore, it is better to reduce them and prevent the associated noise (N_p) at these frequencies from wasting part of the channel's dynamic range.

2.4 Properties of the theory

As suggested by Fig. 2, the neural filter has a maximum at a frequency where the signal-to-noise ratio right after the prefilter equals 1. Below I will show that this is actually the case, under certain conditions. Furthermore, I will derive simple approximations of p_n for very large and very small signal-to-noise ratios.

2.4.1 Maximum of p_n. For convenience of notation I will substitute μ for $-1/\lambda$, and rewrite (8) to

$$p_n = \frac{-N_c(2N_p + S_p) + (N_c^2 S_p^2 + 4N_p S_p N_c \mu)^{1/2}}{2N_p(N_p + S_p)}\,. \quad (9)$$

For the derivation of the maximum of p_n, relationships between μ and N_c, and between μ and N_p are needed, which I will derive first (15 and 16). I assume that for $f \gg 0$

$$S_p \ll N_p \quad (10)$$

and

$$S_p \ll N_c\,. \quad (11)$$

This assumption holds for the example of Fig. 2 (see Figs. 2c and 2e), and holds in general if the characteristic stimulus power spectrum and the prefilter transfer function decrease much more strongly than the noise power spectra for large f. Now the second term of the square root of (9) must be much larger than the first, because otherwise the square root would be of the order of $N_c S_p$. This would lead to a contradiction, because then the numerator (and thus p_n) would be negative (using 10), whereas a power transfer function must be nonnegative. Thus with (10), (9) becomes

$$p_n \approx \frac{-2N_c N_p + 2N_p^{1/2} S_p^{1/2} N_c^{1/2} \mu^{1/2}}{2N_p^2} \quad \text{for} \quad f \gg 0\,. \quad (12)$$

Requiring $p_n \geq 0$ for (12) leads to

$$2N_p^{1/2} S_p^{1/2} N_c^{1/2} \mu^{1/2} \geq 2N_c N_p\,, \quad (13)$$

or

$$\mu \geq N_c N_p / S_p\,. \quad (14)$$

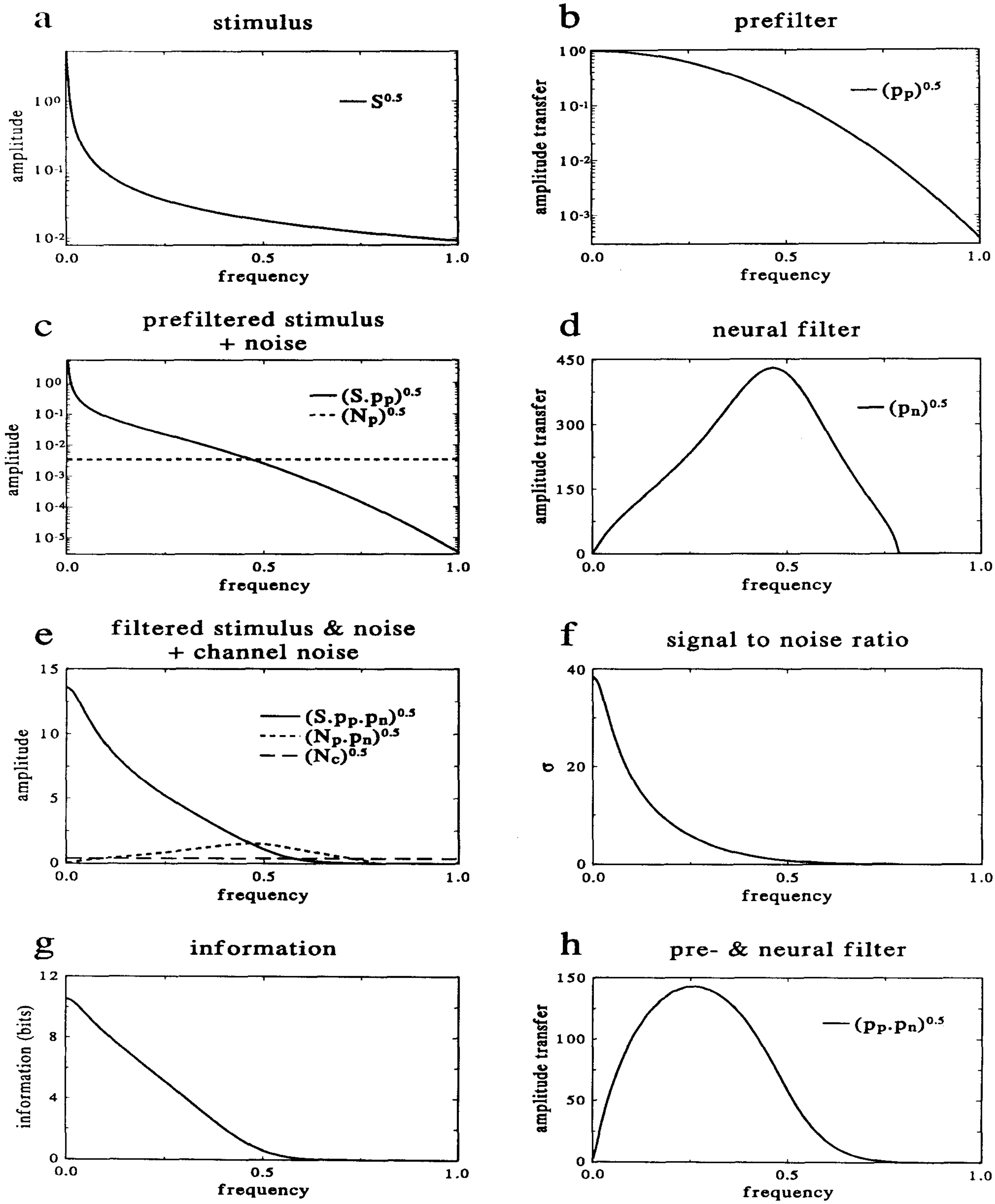

Fig. 2a–h. Example of the theory for $\overline{SNR} = 100$. **a** Average amplitude spectrum of the stimulus; the power spectrum S is proportional to $1/f^2$, with the total power at $f \neq 0$ equal to 40% of the power at $f = 0$. The frequency axis is in arbitrary units. **b** Gain of the gaussian prefilter. **c** Continuous line: prefiltered stimulus; dashed line: additive noise. **d** Neural filter optimizing information flow. **e** Continuous line: stimulus after pre- and neural filtering; short dashes: noise filtered by neural filter; long dashes: additive noise due to the channel itself. **f** Signal-to-noise ratio resulting from **e**, **g**. Information following from **f**, **h**. Combination of pre- and neural filter. All amplitudes are in arbitrary units

Combining (14) with (10) gives

$$\mu \gg N_c \,, \tag{15}$$

and (14) with (11) yields

$$\mu \gg N_p \,. \tag{16}$$

Although (15) and (16) were derived for $f \gg 0$, they are valid for any f. This is because μ is a constant, and because we assume that N_c and N_p are constants as well (as in Fig. 2), at least in good approximation.

Now we can proceed to show that p_n has a maximum for $S_p/N_p = 1$. We will for that purpose study the behaviour of p_n for f in a neighbourhood where

$$S_p \approx N_p \,. \tag{17}$$

Combining (17) with (15) yields

$$4N_pS_pN_c\mu \gg N_c^2S_p^2 \,. \tag{18}$$

Then we find for p_n from (9) and (18)

$$p_n \approx \frac{-N_c(2N_p + S_p) + 2N_p^{1/2}S_p^{1/2}N_c^{1/2}\mu^{1/2}}{2N_p(N_p + S_p)} \,. \tag{19}$$

With $N_p^{1/2}S_p^{1/2}$ of the order of $(2N_p + S_p)$, which follows from (17), and $\mu^{1/2} \gg N_c^{1/2}$, which follows from (15), we find

$$2N_p^{1/2}S_p^{1/2}N_c^{1/2}\mu^{1/2} \gg N_c(2N_p + S_p) \,, \tag{20}$$

and thus

$$p_n = \frac{N_p^{1/2}S_p^{1/2}N_c^{1/2}\mu^{1/2}}{N_p(N_p + S_p)} = \frac{S_p^{1/2}N_c^{1/2}\mu^{1/2}}{N_p^{1/2}(N_p + S_p)} \,. \tag{21}$$

Requiring that the derivative of p_n to f equals zero, and assuming that $\mathrm{d}N_p/\mathrm{d}f$ and $\mathrm{d}N_c/\mathrm{d}f$ are zero in good approximation, we finally get

$$S_p/N_p = 1 \,, \tag{22}$$

thus the signal-to-noise ratio right after the prefilter ($S_p^{1/2}/N_p^{1/2}$) also equals 1.

2.4.2 p_n for large signal-to-noise ratios. Assuming

$$N_p \to 0 \,, \tag{23}$$

we find from (9)

$$p_n \approx \frac{-N_cS_p + N_cS_p\left(1 + \frac{4N_p\mu}{N_cS_p}\right)^{1/2}}{2N_pS_p} \approx \frac{-N_cS_p + N_cS_p\left(1 + \frac{2N_p\mu}{N_cS_p}\right)}{2N_pS_p} = \frac{\mu}{S_p} \,, \tag{24}$$

where the approximation of the square root is valid if N_p is sufficiently small to have $N_p/S_p \ll N_c/\mu$. Then, using (1),

$$p_n \propto (Sp_p)^{-1} \,, \tag{25}$$

i.e., for very large signal-to-noise ratios the power transfer function of the neural filter is proportional to the inverse of the prefiltered stimulus power spectrum. As S, and in particular p_p, have in general low-pass characteristics, p_n will thus be high-pass. This leads to phenomena in sensory physiology like lateral inhibition and self-inhibition.

2.4.3 p_n for small signal-to-noise ratios. Assuming

$$N_p \gg S_p \,, \tag{26}$$

we derived already, (12),

$$p_n \approx \frac{-2N_cN_p + 2N_p^{1/2}S_p^{1/2}N_c^{1/2}\mu^{1/2}}{2N_p^2} \,. \tag{27}$$

Assuming that S_p is not very much smaller than N_c (which is a reasonable assumption for a certain range of low frequencies f), we known that $S_p^{1/2}N_c^{1/2}$ is of the order of magnitude, or larger than N_c. Furthermore, $N_p^{1/2}\mu^{1/2} \gg N_p$ because of (16). This yields

$$(N_p^{1/2}\mu^{1/2})(S_p^{1/2}N_c^{1/2}) \gg N_pN_c \,, \tag{28}$$

and with (27)

$$p_n \approx \frac{N_p^{1/2}S_p^{1/2}N_c^{1/2}\mu^{1/2}}{N_p^2} \,. \tag{29}$$

As μ is a constant as a function of frequency, and assuming again that N_p and N_c are also constants in good approximation, we get, using (1)

$$p_n \propto (Sp_p)^{1/2} \,, \tag{30}$$

i.e., for very small signal-to-noise ratios the power transfer function of the neural filter is proportional to the square root of the prefiltered stimulus power spectrum. With S and p_p of a low-pass nature, p_n will be low-pass as well. This leads to phenomena in sensory physiology like (spatial) pooling and temporal summation.

2.5 Weber's law

In Sect. 2.3 I have presented an example of a neural filter for a certain level of N_p. As can be seen in Fig. 2c, the signal-to-noise ratio at the prefilter (i.e., $(S_p/N_p)^{0.5}$) is a function of frequency. It would be convenient, however, to have a scalar variable indicating the signal quality, and for this purpose I define the average signal-to-noise ratio, $\overline{SNR}$, as

$$\overline{SNR} = \left[\frac{\int_{\Delta f} S_p(f)\,\mathrm{d}f}{\int_{\Delta f} N_p(f)\,\mathrm{d}f}\right]^{1/2} \,. \tag{31}$$

How will the filter change as a function of $\overline{SNR}$? Figure 3 shows the total filter (combined prefilter and neural filter) as a function of frequency for various $\overline{SNR}$s (0.1, 1, 10, and 1000; for $\overline{SNR} = 100$ see Fig. 2h). As expected from the theoretical results in the previous sections, the filter gradually changes from low-pass to band-pass for increasing $\overline{SNR}$. Note that the gain is defined relative to a normalized stimulus power spectrum of fixed total power (Fig. 2a): the $\overline{SNR}$ is varied here by shifting the level of N_p (see Fig. 2c).

Figure 4 shows the total filter as a function of $\overline{SNR}$ for several frequencies, again for the numerical example of Fig. 2. In fact, cross sections through Fig. 4 were already shown in Fig. 3. At low $\overline{SNR}$ the filter's gain is

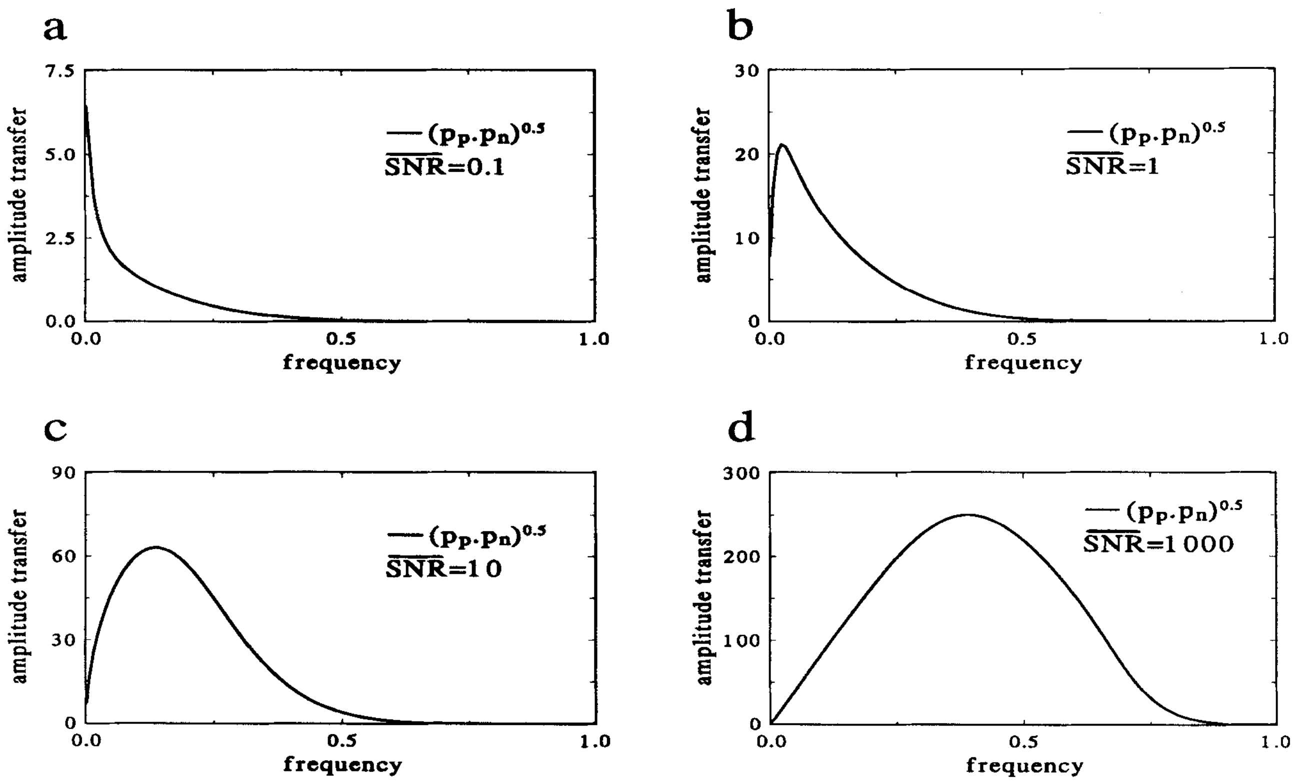

Fig. 3a–d. Transfer function of the total filter (pre- and neural filter) for $\overline{SNR} = 0.1$ (**a**), $\overline{SNR} = 1$ (**b**), $\overline{SNR} = 10$ (**c**), $\overline{SNR} = 1000$ (**d**). The corresponding case of $\overline{SNR} = 100$ is shown in Fig. 2h

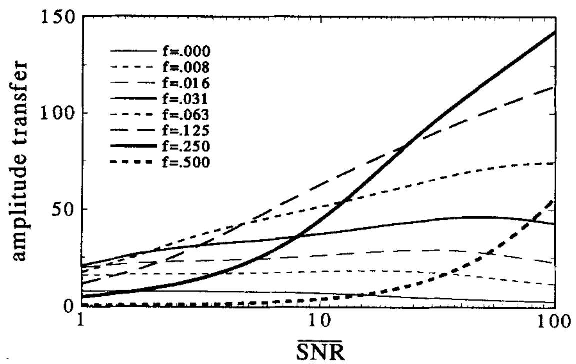

Fig. 4. Total filter as a function of $\overline{SNR}$, for various frequencies f. Curves comply with Weber's law in regions where they are approximately constant as a function of $\overline{SNR}$

largest for low frequencies, whereas this optimal frequency shifts to higher values at higher $\overline{SNR}$. Interestingly, for an appreciable range of frequencies, the gain is approximately constant over a substantial range of $\overline{SNR}$s. This is related to Weber's law (the just-noticeable difference of a stimulus is proportional to the background stimulus strength). Figure 4 was calculated for a normalized stimulus power spectrum (i.e., normalized background stimulus strength), thus the just-noticeable difference will be inversely proportional (at least in good approximation) to the filter's gain. Therefore, the flat curves in Fig. 4 shows that the Weber fraction is approximately constant as a function of $\overline{SNR}$ (with the exception of low $\overline{SNR}$ and high f). This phenomenon will be further enhanced by the fact that usually the dependence of $\overline{SNR}$ on the average stimulus level is not linear, but saturating (for an example of this behaviour for photoreceptors see e.g. Howard and Snyder 1983).

3 Conclusion

The theory I derived in the previous section is, at least qualitatively, consistent with the experimental finding that the senses generally 'sharpen' the stimulus (e.g. show lateral inhibition) at sufficiently high signal-to-noise ratios. It is also consistent with the finding that, at low signal-to-noise ratios, the signals from many sensory receptors are usually pooled, apparently to obtain a more reliable signal. Finally, it is consistent with the general finding that Weber's law holds in a certain range of average stimulus strengths.

In those cases where it is possible to find a general (statistical) description of the natural stimuli for a certain sensory modality, the theory allows quantitative predictions of how the sensory data are to be processed by the early sensory system. Indeed, the theory was recently successfully applied to spatiotemporal vision in both the blowfly visual system and the human visual system (van Hateren 1992a, b). Here the statistical structure of the stimuli is determined by the spatial power spectrum of natural images (Field 1987) combined with the distribution of velocities perceived by the visual system when it is moving around. The resulting spatiotemporal filters are very similar to those that can be measured in second order neurons in the fly visual system (van Hateren 1992a). Furthermore, the theory generates spatiotemporal contrast sensitivities close to those measured psychophysically, e.g. by Kelly (1979), in the human visual system (van Hateren 1992b). It seem likely that the theory can also be successfully applied to other aspects of the visual sense, and to other senses as well.

Acknowledgements. I wish to thank Nomdo Jansonius and Doekele Stavenga for useful comments. This research was supported by the Netherlands Organization for Scientific Research (NWO).

References

Attneave F (1954) Some informational aspects of visual perception. Psychol Rev 61:183–193

Barlow HB (1961) Possible principles underlying the transformations of sensory messages. In Rosenblith WA (ed) Sensory communication. MIT Press, Cambridge, pp 217–234

Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A 4:2379–2394

Goldman S (1953) Information theory. Dover Publications, New York

Hateren JH van (1992a) Theoretical predictions of spatiotemporal receptive fields of fly CMCs, and experimental validation. J Comp Physiol A 171:157–170.

Hateren JH van (1992b) Spatiotemporal contrast sensitivity of early vision. Vision Res, in press

Howard J, Snyder AW (1983) Transduction as a limitation on compound eye function and design. Proc R Soc London Ser B 217:287–307

Kelly DH (1979) Motion and vision. II. Stabilized spatio-temporal threshold surface. J Opt Soc Am 69:1340–1349

Kretzmer ER (1952) Statistics of television signals. Bell Syst Tech J 31:751–763

Kretzmer ER (1954) Redundancy in television. Bell Lab Record, November 1954, pp 401–404

Oppenheim AV, Lim JS (1981) The importance of phase in signals. Proc IEEE 69:529–541

Srinivasan MV, Laughlin SB, Dubs A (1982) Predictive coding: a fresh view of inhibition in the retina. Proc R Soc London Ser B 216:427–549

Ziel A van der (1970) Noise. Sources, characterization, measurement. Prentice-Hall, Englewood Cliffs