Learning Temporally Stable Representations from Natural Sounds: Temporal Stability as a General Objective Underlying Sensory Processing

Armin Duff[1,2], Reto Wyss[3], and Paul F.M.J. Verschure[2,4]

[1] Institute of Neuroinformatics, UNI - ETH Zürich, Winterthurerstrasse 190,
CH-8057 Zürich, Switzerland
duffar@ini.phys.ethz.ch

[2] SPECS, IUA, Technology Department, Universitat Pompeu Fabra, Ocata 1,
E-08003 Barcelona, Spain

[3] CSEM Centre Suisse d'Electronique et de Microtechnique SA,
Untere Gründlistrasse 1, CH-6055 Alpnach-Dorf, Switzerland

[4] ICREA Institució Catalana de Recerca i Estudis Avançats,
Passeig Lluís Companys 23, E-08010 Barcelona, Spain

Abstract. In order to understand the general principles along which sensory processing is organized, several recent studies optimized particular coding objectives on natural inputs for different modalities. The homogeneity of neocortex indicates that a sensitive objective should be able to explain response properties of different sensory modalities. The temporal stability objective was successfully applied to somatosensory and visual processing. We investigate if this objective can also be applied to auditory processing and serves as a general optimization objective for sensory processing. In case of audition, this translates to a set of non-linear complex filters optimized for temporal stability on natural sounds. We show that following this approach we can develop filters that are localized in frequency and time and extract the frequency content of the sound wave. A subset of these filters respond invariant to the phase of the sound. A comparison of the tuning of these filters to the tuning of cat auditory nerves shows a close match. This suggests that temporal stability can be seen as a general objective describing somatosensory, visual and auditory processing.

1 Introduction

The human neocortex is to a great extent homogeneous throughout all its areas [1, 2]. This suggests that it should be possible to describe its structure and dynamics with general concepts and models. This line of thinking appears most prominent in experimental and theoretical work that proposes a "canonical microcircuit" as a basic computational unit [3]. Such a general view does imply that any relevant computational neuronal model should explain response properties of neurons of different areas of the neocortex. Thus, one would expect that

J. Marques de Sá et al. (Eds.): ICANN 2007, Part II, LNCS 4669, pp. 129–138, 2007.

also in sensory processing, the same model can be applied to replicate response properties of different modalities.

It has long been assumed, that sensory systems adapt to the statistical properties of their input leading to the general approach of explaining receptive field properties of sensory systems by optimizing a particular coding objective for natural stimuli [4]. In a series of theoretical studies, several general coding principles have been exploited for different sensory modalities. In visual processing, learning sparse codes from natural images leads to simple-cell like receptive fields as found in primary visual cortex [5]. An extension of this approach to a multi-layer network enabled to learn contour coding in natural images [6]. Similarly, optimizing for temporal stability can replicate properties of V1 simple cells [7], but also invariant representations similar to V1 complex cells [8, 9, 10, 11, 12], color selective cells [13] and viewpoint invariant object representations [14, 15]. Recently it has been shown that the optimization of a multi layered network for temporal stability combined with local memory can account for a complete visual hierarchy, including place fields, by processing a continuous natural input stream generated by a mobile robot [16]. In a different approach, a hierarchical model was optimized based on a MAX-like operation for object recognition [17]. Another objective function, predictability, was proposed to yield self-emergent symbols in an optimization process [18]. In auditory processing, it was shown that optimizing a set of filters for efficiency can explain the formation of adequate auditory filters [19, 20, 21]. Further, optimizing a spectrographic representation of speech for sparseness leads to similar spectro-temporal receptive fields (STRF) as observed in primary auditory cortex [22]. Temporal stability optimization is not restricted to a visual input stream but was also successfully applied to preprocess data for somatosensory discrimination of texture [23]. This diversity of approaches to explain sensory processing contradicts several theoretical and anatomical studies suggesting that the same computational strategy is likely to be involved in processing information from different sensory modalities [1, 2, 24, 25, 26, 27].

In this study we investigate whether temporal stability may be an appropriate objective for the auditory domain and serve as a general objective to replicate neural responses in somatosensory, visual and auditory sensory processing. Auditory processing begins when the cochlea transforms sound energy into electrical signals and passes them to the auditory nerves. Ignoring the nonlinearities and amplification features of the cochlea and the primary auditory system, the response properties of auditory nerves can be described by a set of filters, with different frequency tunings, forming a spectro-temporal representation of the sound [28]. It is not clear how this representation is tuned to the statistics of the sound environment. In order to get a filter set adapted to the input statistics we optimize a set of complex filters to show a maximally stable response across time for natural sounds. Following this approach we do not describe the details of how cochlea and auditory nerve realize the spectro-temporal separation but reveal the general principle around which sensory processing is organized. While different types of filters emerge, we find that the tuning of the filters is in accordance with experimental data from cat physiological data. This suggests that the response properties of auditory nerve fibers can be explained in terms of temporal stability optimization. Thus the somatosensory, visual and auditory sensory processing can be explained within the same framework of temporal stability.

2 Methods

2.1 Input

The natural sounds used to learn the stable representations consist of different words in 6 distinct languages, spoken by three male and three female native speakers. The sound samples are re-sampled to 22050 Hz from the language illustrations accompanying the handbook of the International Phonetic Association (IPA) [29]. In addition to the raw sound wave we investigate band-passed sound waves. We consider four different band-pass filters with characteristic frequency bands of 1 Hz - 324 Hz, 324 Hz - 1050 Hz, 1050 Hz - 3402 Hz and 3402 Hz - 11025 Hz. According to the logarithmic organization of the cochlea we increased the bandwidth of the different filters logarithmically.

2.2 Filters

We optimized a set of complex filters with an analysis window covering 5.8 ms of the raw sound wave which corresponds to 128 data points. At each iteration of the optimization the sound wave is shifted through this window by the time interval τ. Thus, by changing τ we can change the overlap of subsequent analysis windows.

The activity (A_i) of the filter i is given by the absolute value of the scalar product between the input $\boldsymbol{I}$ and the *complex* filter function $\boldsymbol{h_i} \in \mathbb{C}^{128}$. Real and imaginary values can change independently and thus each filter is defined by 256 parameters.

$$A_i = abs(\boldsymbol{h}_i \cdot \boldsymbol{I}) \tag{1}$$

The absolute value function implies that the filters are not linear. Mathematically they are equivalent to the energy model proposed by Adelson and Bergen [30] and applied in different studies [9, 16, 23].

2.3 Optimization

In the present study we optimized a goal function which contains two terms such that the total objective Ψ is given by:

$$\Psi = (1 - \gamma)\Psi_{stab} + \gamma\Psi_{decor} \tag{2}$$

where γ is used to balance between the relative contribution of the two objectives. The first part is the temporal stability objective Ψ_{stab} given by:

$$\Psi_{stab} = -\sum_i \frac{\langle \dot{A}_i^2 \rangle_t}{var_t(A_i)} \tag{3}$$

A_i is the activity of the filter and the sum is over all filters i. $\dot{A}$ denotes the discrete temporal differentiation given by:

$$\dot{A}_i = A_i(t) - A_i(t - \tau) \tag{4}$$

where τ is the time interval by witch the analysis window is shifted each iteration. The floating average at time t, $\langle\ \ .\ \ \rangle_t$ is calculated iteratively with a time constant $\zeta = 500\ ms$ and defined by:

$$\langle A_i \rangle_t = (1 - \frac{1}{\zeta})\langle A_i \rangle_{t-1} + \frac{1}{\zeta} A_i(t) \tag{5}$$

The temporal derivative in equation 3 is divided by the variance in order to avoid the trivial solution where all the parameters of the filter are zero. The variance is computed using:

$$var_t(A_i) = \langle A_i^2 \rangle_t - \langle A_i \rangle_t^2 \tag{6}$$

As the filters should have different receptive fields, collectively representing the statistics of the input, each of them must encode different information. Therefore, the second term, i.e the decorrelation objective Ψ_{decor} (7), is introduced to augment the statistical independence of the filters.

$$\Psi_{decor} = -\sum_{j \neq i} (\rho_{ij}(A_i, A_j))^2 \tag{7}$$

$$\rho_{ij}(A_i, A_j) = \frac{\langle A_i A_j \rangle_t - \langle A_i \rangle_t \langle A_j \rangle_t}{\sqrt{var_t(A_i) var_t(A_j)}} \qquad (correlation) \tag{8}$$

The filter functions h_i change following an on-line learning algorithm along the gradient in order to maximize the total objective function Ψ.

2.4 Time and Frequency Analysis

To characterize each of the filters we extracted the center frequency (CF), the spectral bandwidth (BW), the quality factor (Q), the temporal extent (TE) and the relative shift ϕ as key characteristics. The CF of the filter corresponds to the maximum of the power spectrum of the filter function. This is the frequency for which the filter is most sensitive. The BW is the width of the frequency response measured at 10 dB down from the peak at the CF. Q corresponds to the sharpness of the filter and is defined by the CF divided by the BW. The subscript in Q_{10dB} indicates the level at which the BW is measured. The TE of a filter is defined as the width that is used to cover 90 % of the filter power. ϕ is given by the relative phase shift of the real and the imaginary part of the filter and is calculated with respect to the CF of the filter.

3 Results

We optimize 64 filters on the raw sound wave using an update interval τ between 0.1 - 2.8 ms in steps of $\approx$ 0.1 ms for different simulations. Each filter is defined by its 256 parameters $h_i \in \mathbb{C}^{128}$. Most of the resulting filters are sinusoidal,

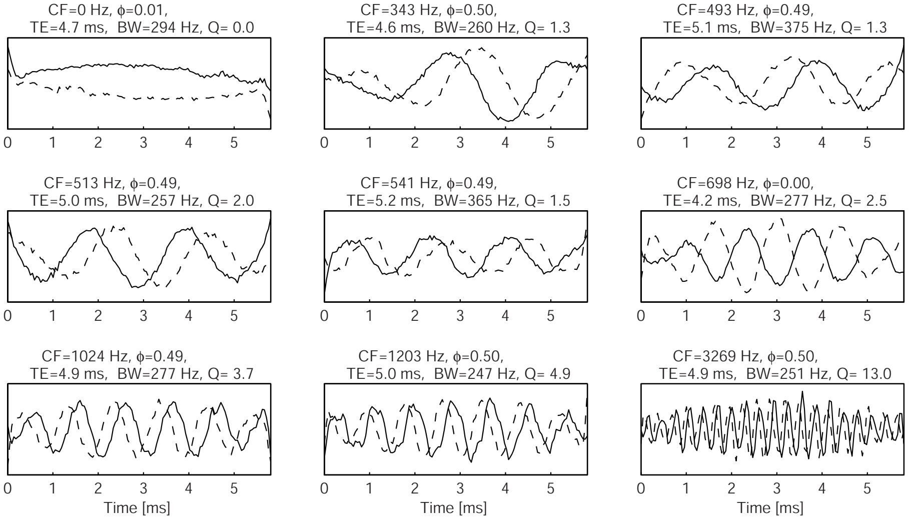

Fig. 1. Auditory filters after temporal stability optimization using an update interval τ of 0.7 ms. The plot shows a representative subset of the total population of 64 filters. The filters are plotted in the time domain where the solid line corresponds to the real part of the filter and the dashed line to the imaginary part. On top of each waveform the key characteristics of the filter are indicated: center frequency (CF), relative shift (ϕ), temporal extent (TE), spectral bandwidth (BW) and quality factor (Q).

amplitude modulated, and cover the whole window width (Figure 1). The filters have a single peak frequency tuning such that their CF is well defined.

The investigation of the distribution of the CFs, shows that it matches the Power Spectrum Density (PSD) of the sound (Figure 2). The higher the PSD in a frequency interval, the higher the number of filters with a CF in this interval.

A part of the nonlinear sinusoidal filters ($\approx$ 60%) exhibit a relative shift of $\approx \pi/2$ between their real and imaginary components that conforms to a filter selective for a particular frequency while being invariant to the phase of the sound wave. Other filters have no relative shift and therefore do depend on the phase.

As pointed out above, the distribution of the CFs is correlated with the PSD such that we only obtained filters within the lower part of the frequency spectrum. To enable the system to form filters with CFs in other frequency ranges we band-passed the signal of the speech ensemble before optimizing. For each band-passed signal we optimized 16 filters. The filter set that emerged covers a frequency range of 120 - 5500 Hz. Frequencies higher than 5500 Hz are not covered as the main energy of the band-passed signal with the highest pass range, lies between 3400 Hz and 5500 Hz. Preliminary experiments showed that higher CFs can be obtained by applying higher band-pass (data not shown).

To get an impression of the distribution and coverage of the filter sets in time and frequency, we considered the extent of the filters on the time-frequency plane. (Figure 3 A). The filter set possesses both, temporally localized and non-

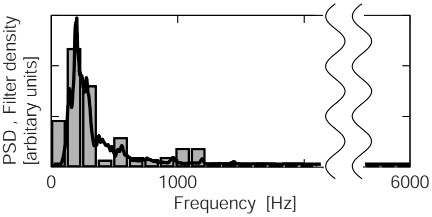

Fig. 2. Distribution of the CFs (bars) and the PSD of the signal (line). The bars indicate the relative density of filters for the corresponding frequency range. The PSD is superimposed where the energy density is given in an arbitrary scale.

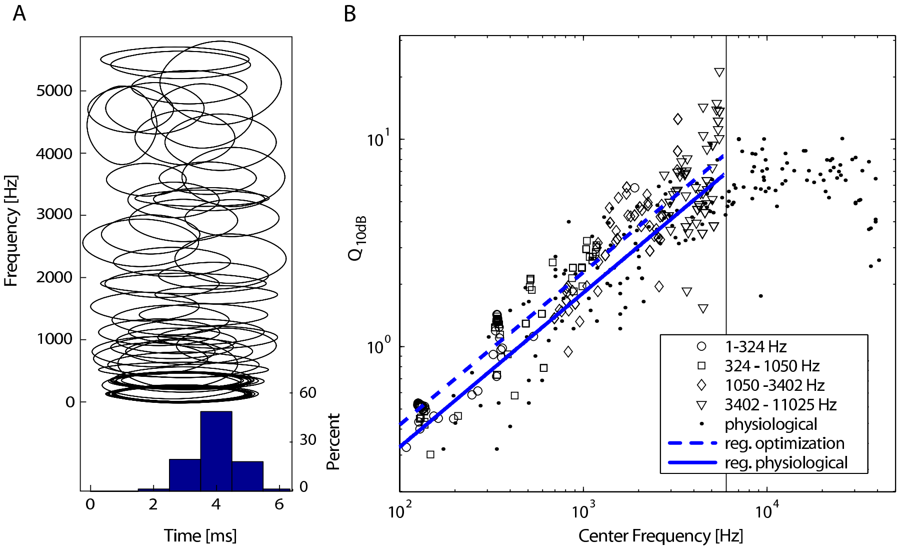

Fig. 3. Characteristics of the filter sets for the band-passed signals. **A** Tiling the time frequency plane. The extent of each filter is represented by an ellipse. The height of the ellipse corresponds to the frequency bandwidth and the width to the temporal extent of the filter. The histogram at the bottom indicates the distribution of the temporal extent for each filter set. **B** Q_{10dB} of the optimized filters compared to the Q_{10dB} measured from cat auditory nerve fibers. For comparison the regression line for the optimized filters and the physiological data are superimposed. Notice that the regression for the physiological data only includes the data points with a frequency lower than 5500 Hz. The physiological data is replotted from [28].

localized filters. Corresponding to the time frequency uncertainty relation, filters with narrower temporal extent (TE) have a broader frequency extent (BW). Further, one can observe that the TE decreases for higher frequencies.

In order to validate the emerged filters against physiological data we compare the tuning of the filters to cat auditory nerve fibers [28]. The sharpness (Q) of the optimized filters is consistent with the sharpness measured for cat auditory nerve fibers (Figure 3 B). Linear regression on the physiological data in a range of $1-5500$ Hz i.e. the range covered by the filters that emerged from temporal stability optimization, yields a stiffness of 0.744 $[0.675, 0.831]$ and an offset of -4.542 $[-5.176, -3.908]$. For the optimized filter set linear regression yields a stiffness of 0.736 $[0.703, 0.769]$ and an offset of -4.259 $[-4.487, -4.031]$. The numbers in brackets correspond to the 95% confidence interval. Thus, both the stiffness and the offset of the two curves are very close. The deviation for the stiffness is 1% and for the offset 6%. The confidence interval for the optimized filters lie within the confidence intervals of the physiological data. This suggests that some of the response properties of the auditory nerve fibers can be well explained in terms of temporal stability optimization.

4 Discussion

In this study we investigated if temporal stability is a general objective underlying sensory processing and can be applied in the auditory domain. We have shown that temporal stability optimization in the auditory domain leads to filters that extract the frequency content of speech. Due to a relative shift of $\pi/2$ between the real and imaginary parts of the complex filters, some filters have an invariant response to the phase of the sound wave. Further we found that the distribution of the CFs is related to the PSD such that frequency bands with high energy have a high filter density. Our approach rendered filters which show a quantitative match to the filter properties of the cat auditory nerve. This suggests that the physiological properties of this part of the auditory pathway can be explained in terms of temporal stability optimization.

The design of filters involves an inevitable trade-off between time resolution and frequency resolution [31]. To get precise information about the frequency content one has to integrate over a characteristic time length of the sound signal, leading to a decrease in temporal precision. In other words, it is not possible to design a filter that captures both the frequency and the timing of a sound with arbitrary accuracy. However, in order to be able to process natural sounds it is often important to have information about both frequency and timing. Which spectro-temporal representation is optimal depends on a number of factors such as the importance of the information available, the biological or computational constraints and the statistics of the sound. A common mis-characterization of the peripheral auditory system is that it performs a short time Fourier transform (STFT) or a kind of wavelet decomposition. For the STFT, the bandwidth of the signals is approximately constant whereas for a wavelet representation the sharpness of the filters remains constant for all frequencies. Neither of the two properties are observed experimentally [28]. Instead, similar to the filters found in our approach, the sharpness of the auditory nerve fibers follows a sub-linear power law (Figure 3 B). Therefore, an optimal set of filters for the analysis of speech must exhibit both, Fourier and wavelet characteristics.

This study is complementary to the work of Lewicki and Smith [19, 20, 21] who optimized a set of real-valued linear filters, for sparseness on different sound ensembles including speech. While all the filter responses resulting from Lewicki's approach must vary with the phase of the incoming signal due to linearity, we have found that a part of the nonlinear complex filters are phase invariant. Similarly, some of the auditory nerve fibers at the cochlea do code the phase of the signal, which is important to determine sound location in the early auditory pathway. At higher levels of processing, this phase information is lost, suggesting that phase invariance constitutes a first step towards sound classification [32]. A further difference to the work of Lewicki can be found in the distribution of the CF of the filters. Optimizing for sparseness leads to a distribution of CFs covering the whole frequency range up to the Nyquist frequency. Optimizing for temporal stability, however, results in a distribution which is correlated to the PSD and therefore confined to reagions with high energy. The question that arises is to what extent the features that carry the main energy also carry the relevant information. For speech, the slowly varying features are the vowels whereas the consonants vary much faster and have higher frequencies. Therefore, temporal stability optimization tends to extract the information that is contained in the vowels but mostly ignores the consonants whereas sparseness mainly extracts features with higher frequencies. However, for word discrimination both vowels and consonants are important. We have shown that we can account for the whole range of frequencies by band-passing the signal before optimizing for temporal stability. Given the acoustic properties of the cochlea, such band-passing is likely to happen already at a mechanical level at the basilar membrane [32]. The subsequent levels of auditory information processing would therefore already be supplied with an appropriately band-pass filtered signal.

In order to validate our results against experimental data, we compared the learned filters with cat physiological data. Primary auditory cortex (A1) neurons are characterized by Spectro-Temporal Receptive Fields (STRF) [33, 34, 35]. As our filters do not include a temporal component their activity is only dependent on the spectral content of the sound and thus a direct comparison is not possible. We could however compare the characteristics of the optimized filters with the characteristics of the cat auditory nerve fibers. For this comparison it is not clear to what extent human speech is an adequate auditory stimulus for cats or whether animal vocalization would be more appropriate. Speech contains a mixture of various auditory features which are present in different classes of natural sounds [19]. Thus, given that any animal is exposed to a mixture of environmental sounds and vocalization, speech constitutes a good compromise. We have found, that the sharpness of the filters for the band-passed speech ensemble matched the sharpness measured for cat auditory nerve fibers. Hence some of the response properties of auditory nerve fibers can be explained in terms of optimizing a set of complex filters for temporal stability. Therefore, temporal stability can be considered a general objective underlying sensory processing.

References

[1] Sur, M., Rubenstein, J.L.R.: Patterning and plasticity of the cerebral cortex. Science 310(5749), 805–810 (2005)

[2] Horng, S.H., Sur, M.: Visual activity and cortical rewiring: activity-dependent plasticity of cortical networks. Prog. Brain Res. 157, 3–11 (2006)

[3] Douglas, R., Martin, K.: Neocortex. In: Shepherd, G. (ed.) The Synaptic Organization of the Brain, pp. 459–509. Oxford University Press, New York (1998)

[4] Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annu. Rev. Neurosci. 24, 1193–1216 (2001)

[5] Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. NatureEmergence of simple-cell receptive field properties by learning a sparse code for natural images 381(6583), 607–609 (1996)

[6] Hoyer, P.O., Hyvärinen, A.: A multi-layer sparse coding network learns contour coding from natural images. Vision. Res. 42(12), 1593–1605 (2002)

[7] Hurri, J., Hyvärinen, A.: Simple-cell-like receptive fields maximize temporal coherence in natural video. Neural Comput 15(3), 663–691 (2003)

[8] Berkes, P., Wiskott, L.: Slow feature analysis yields a rich repertoire of complex cell properties. J Vis. 5(6), 579–602 (2005)

[9] Körding, K.P., Kayser, C., Einhäuser, W., König, P.: How are complex cell properties adapted to the statistics of natural stimuli? J. Neurophysiol. 91(1), 206–212 (2004)

[10] Hashimoto, W.: Quadratic forms in natural images. Network 14(4), 765–788 (2003)

[11] Einhäuser, W., Kayser, C., König, P., Körding, K.P.: Learning the invariance properties of complex cells from their responses to natural stimuli. Eur. J. Neurosci. 15(3), 475–486 (2002)

[12] Kayser, C., Einhäuser, W., Dümmer., O., König, P., Körding, K.P.: Extracting slow subspaces from natural videos leads to complex cells. In: Dorffner, G., Bischof, H., Hornik, K. (eds.) ICANN 2001. LNCS, vol. 2130, pp. 1075–1080. Springer, Heidelberg (2001)

[13] Einhäuser, W., Kayser, C., Körding, K., König, P.: Learning distinct and complementary feature-selectivities from natural colour videos. Journal of Neuroscience-Learning distinct and complementary feature-selectivities from natural colour videos 21, 43–52 (2003)

[14] Stringer, S.M., Rolls, E.T.: Invariant object recognition in the visual system with novel views of 3D objects. Neural Comput. 14(11), 2585–2596 (2002)

[15] Einhäuser, W., Hipp, J., Eggert, J., Körner, E., König, P.: Learning viewpoint invariant object representations using a temporal coherence principle. Biol. Cybern. 93(1), 79–90 (2005)

[16] Wyss, R., König, P., Verschure, P.F.M.J.: A Model of the Ventral Visual System Based on Temporal Stability and Local Memory. PLoS Biol. 4(5), e120 (2006)

[17] Riesenhuber, R., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019–1025 (1999)

[18] König, P., Krüger, N.: Symbols as Self-emergent Entities in an Optimization Process of Feature Extraction and Predictions. Biol. Cybern. 94(4), 325–334 (2006)

[19] Lewicki, M.S.: Efficient coding of natural sounds. Nat. Neurosci. 5(4), 356–363 (2002)

[20] Smith, E., Lewicki, M.S.: Efficient coding of time-relative structure using spikes. Neural. Comput. 17(1), 19–45 (2005)

[21] Smith, E.C., Lewicki, M.S.: Efficient auditory coding. Nature 439(7079), 978–982 (2006)

[22] Klein, D., König, P., Körding, K.: Sparse spectrotemporal coding of sounds. Eurasip JaspSparse spectrotemporal coding of sounds 3, 659–667 (2003)

[23] Hipp, J., Einhäuser, W., Conradt, J., König, P.: Learning of somatosensory representations for texture discrimination using a temporal coherence principle. Network 16(2-3), 223–238 (2005)

[24] Shamma, S.: On the role of space and time in auditory processing. Trends Cogn. Sci. 5(8), 340–348 (2001)

[25] Sur, M., Leamey, C.A.: Development and plasticity of cortical areas and networks. Nature Reviews Neuroscience 2, 251–262 (2001)

[26] Dennis, D., O'Leary, M.: Do cortical areas emerge from a protocortex? Trends in Neuroscience 12(10), 400–406 (1989)

[27] Sur, M., Garraghty, P., Roe, A.: Experimentally induced visual projections into auditory thalamus and cortex. Science 242, 1437–1441 (1988)

[28] Rhode, W.S., Smith, P.H.: Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. Hearing Research 18, 159–168 (1985)

[29] International Phonetic Association, ed.: Handbook of the International Phonetic Association. Cambridge University Press, Cambridge, UK, (1999) Available at: `http://web.uvic.ca/ling/resources/ipa/handbook.htm`

[30] Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. J Opt. Soc. Am. A 2(2), 284–299 (1985)

[31] Rioul., O., Vetterli, M.: Wavelets and signal processing. IEEE Signal Processing Magazine 8, 14–38 (1991)

[32] Hudspeth, A.J.: Hearing. In: Kandel, E.R., Schwartz, J.H., Jessell, T.M. (eds.) Principles of Neural Science, 4th edn., pp. 590–613. McGraw-Hill, New York (2000)

[33] Depireux, D.A., Simon, J.Z., Klein, D.J., Shamma, S.A.: Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85(3), 1220–1234 (2001)

[34] Klein, D.J., Simon, J.Z., Depireux, D.A., Shamma, S.A.: Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex. J Comput. Neurosci. 20(2), 111–136 (2006)

[35] Theunissen, F.E., Woolley, S.M.N., Hsu, A., Fremouw, T.: Methods for the analysis of auditory processing in the brain. Ann. N Y Acad. Sci. 1016, 187–207 (2004)