# Learning of sparse auditory receptive fields

Konrad P. Körding, Peter König, and David J. Klein
Institute of Neuroinformatics, ETH/UNI Zürich
Winterthurerstr. 190, 8057 Zürich, Switzerland

**Abstract - It is largely unknown how the properties of the auditory system relate to the properties of natural sounds. Here, we analyze representations of simulated neurons that have optimally sparse activity in response to spectrotemporal speech data. These representations share important properties with auditory neurons as determined in electrophysiological experiments.**

## I. Introduction

The properties of any sensory system should be matched to the statistics of the natural stimuli it is typically operating on [1]. Thus, it is interesting to compare the properties of sensory systems with the statistics of natural stimuli; and to analyze to what extend the neural properties can be understood in terms of optimally handling those stimuli.

Significant parts of the visual system can be understood in terms of leading to optimally sparse neural responses in response to pictures of natural scenes. Searching for sparse representations on such pictures allows one, for example, to reproduce the properties of neurons in the lateral geniculate nucleus (LGN)[2] and of simple cells in the primary visual cortex [3-5]. Here a sparse representation has two distinct albeit related meanings. (1) The different neurons of the population should have significantly different properties to avoid redundancy. (2) At the same time the neurons should have sparse activities over time implying that they often have an activity close to zero and then sometimes have very high activitiy. Searching for such responses is also at the heart of independent component analysis (ICA)[6].

Simple cells are visual neurons that are specific to position, orientation and spatial frequency of bars or edges. They can approximately be understood in terms of computing a localized linear function over their LGN inputs. It has recently been shown that neurons in the central auditory system share similar properties [7-9]. In particular, neurons in the primary auditory cortex AI can also be understood as linear filters acting upon an input that is local in time and local in tonotopic space (the space of the sounds frequency). These neurons are also often specific to orientation, that is, to changes of the underlying frequency.

Here, we analyze if these linear neurons in the auditory system can also be understood in terms of leading to sparse activity in response to natural input, which in this case is speech data.

## II. Methods

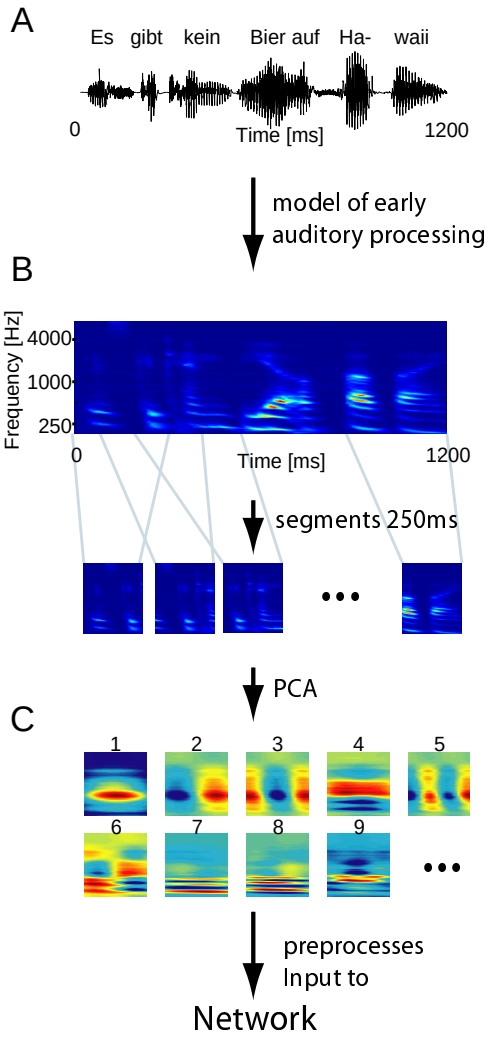


**Figure 1: Methods**
A) Text from various German and English sources is read and the raw waveforms are recorded. This data is input to a model of the auditory system's early stages resulting in a spectrogram B), where the strength of the activity is color-coded. These spectrograms are subsequently cut into overlapping pieces of length 250 ms each. They are whitened using the first nPCA components only. The basis vectors of this PCA are shown in C), color-coded in a scale where blue represents small values and red represents large values.

We obtain speech data from one voluntary human subject (KPK) using a standard microphone (Escom) and Cool Edit Pro software (Syntrillium

software, Phoenix, USA) recording mono at 44kbit, 16 bits precision (Fig 1A). This data is pre-processed to mimic the properties of the early auditory system [10, 11] using the "NSL Tools" MATLAB package (courtesy of the Neural Systems Laboratory, University of Maryland, College Park, downloadable from http://www.isr.umd.edu/CAAR/pubs.html).

The resulting data can be viewed as spectrograms, which represent the time-dependent spectral energy of sound. An example is shown in the Figures 1A and 1B, which show, respectively, the input and output of the model to the utterance "Es gibt kein Bier auf Hawaii" (the title of a German folksong).

Spectrograms are produced with two different sets of model parameters, one "high resolution" and one "low resolution". The high-resolution spectrograms have 64 points along the tonotopic axis, covering a frequency range from 185 to 7246 Hz; over this range the low-resolution spectrograms only have 16 points. Temporally, the data is arranged into overlapping blocks of 25 points, covering 250 milliseconds, and is subjected to a principal component analysis. The first nPCA components (200 for the high-resolution set and 100 for the low-resolution set) are all set to have unit variance (corresponding to whitening) and are subsequently used as input I(t) to the optimization algorithm. 50,000 subsequent samples are used as input.

100 neurons are simulated, each of which has a weight vector W of length nPCA. The activities of the neurons are defined as:

$$A_i(t) = \mathbf{I}(t)\,\mathbf{W}_i$$

The parameters of the simulated neurons are optimized by scaled gradient descent [12] to maximize the following objective function, where <*> denotes the average over all stimuli:

$$\Psi_{\text{skewness}} = \sum_i \frac{\langle A_i^3 \rangle}{\langle A_i^2 \rangle^{3/2}}$$

The skewness is high for asymmetrical, skewed distributions, for which high positive values are sometimes reached, but not many significantly negative values are reached. There is an important difference between the visual system and the auditory system: In the visual systems bright edges are often seen on dark background and dark edges are also often seen on bright background. In the auditory system however the inputs from the cochlea that largely represent the presence of energy are bound to be positive. This makes skewness an appropriate objective function for such data.

The optimal neuronal receptive fields are characterized by a weighted sum of the principle components. For each neuron, the relative weight of each principle component is determined by the optimized set of weights. The resulting functions, referred to as *spectrotemporal receptive fields* (STRFs), are then characterized to facilitate comparison with physiological data.

Characteristics include the *best frequency* (BF), which is the spectral location of the maximum weight, and the *excitatory-tuning bandwidth* (Qn value), defined as the spectral width of the portion of the STRF within 1/sqrt(n) of the peak value, divided by the BF.
Furthermore, we assess the separability and quadrant-separability of some STRFs. A separable STRF is one that is fully described by the product of a spectral function with a temporal function. A quadrant-separable STRF is not separable; however, its two-dimensional Fourier transform has separable quadrants. This pervasive feature of cortical neurons has recently been described in detail [7].

## III. Results

The first several principle components of our high-resolution speech data are shown in Figure 1C. A number of features of these components can be observed. Those components that represent much of the variance change slowly in spectrum and in time. Thus, using only the first nPCA principal components for learning effectively low-passes the stimuli; however, these components alone account for more than 90 % of the total variance of the data. Interestingly, we have found that the form of the lowest components is extremely robust to changes to the spectrotemporal resolution of the peripheral auditory model. We perform two simulations, one with high tonotopic resolution, the other one with low tonotopic resolution to analyze the effects of the chosen representation

## IV. High Resolution

In the high-resolution case the number of PCA components used (nPCA) is set to 200 and the number of neurons is 100. The resulting STRFs of several of the optimized neurons is

shown in Figure 2A. Note the horizontal-striped appearance of many of the functions; this is a common feature of the high-resolution results, and arises from the dominant presence of narrow harmonics in the spectrum of voiced speech. In fact, out of the 100 neurons, 37 are selective for voice pitch, 16 are selective for changing pitch, and 29 of them are primarily selective for a single frequency band, presumably due to a single harmonic (numbers obtained by manual inspection).

Time-frequency localization is a very common characteristic of these simulated neurons; in this respect they have much in common with nature. Of the 100 neurons 98 are localized in time and all 100 are localized in spectrum (as estimated by visual inspection).

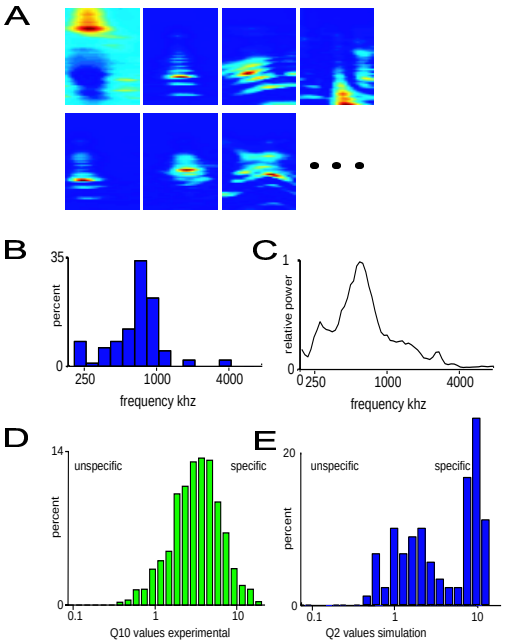


**Figure 2: High Resolution Results**

A) The color-coded spectrotemporal receptive fields of 7 out of the 100 neurons are shown. B) The histogram of BF for our optimized neuronal population is shown. C) The average energy is shown as a function of frequency. D) The histogram of frequency-specificity, quantified by Q10, measured from cat auditory cortex[13]. E) The Q2 value for our simulated neurons is shown. Theoretically the Q2 values should be a factor of log(10)/log(2) (about 3.33) higher then the Q10 values if we may assume Gaussian behavior of the response around the maximum.

To further quantify these effects, we calculated various measures for the STRFs of our neurons (see Methods). As a first measure we calculate *BF*, the frequency that results in the strongest activity of a neuron. In Figure 2B, the results accumulated from all neurons are shown in histogram form. It can be seen however that the lower frequencies, over which voice pitch is predominantly manifested, are represented far stronger than the higher frequencies. This clearly deviates from the more uniform distributions commonly obtained in animal-physiology experiments (e.g. [14]), and rather seems to reflect the frequency distribution of our particular speech dataset, shown in Figure 1C.

The distribution of Q10, the excitatory-tuning bandwidth obtained from electrophysiological investigation of cat auditory cortex [13] is shown in Figure 2D. The distribution of Q2 of our neuronal population is shown in Figure 2E. This Q measure quantifies the notion that the neuronal representations are spectrally localized.

It should be noted that, as in [13], the STRFs are divided into two groups: A group with one peak, and a group with multiple spectrally separated peaks. This division is reflected in the multimodal distribution of Q2; the peak at low Q is mostly due to multi-peaked STRFs, while the higher-Q peak represents single-peaked STRFs. However, the physiological data shown in Figure 2D only represents neurons with single-peaked frequency tuning curves.

These simulations thus show that without preprocessing, many of the neurons represent pitch, a behavior that interestingly is not commonly observed in vivo. At the same time, much of the qualitative properties of cortical neurons can be reproduced. The neurons are specific to time and spectral frequency and partially also for frequency sweeps, just as observed in physiology [8, 14].

## V. Low Resolution

In the low resolution case the spectrograms are smoothed and resampled to only have 16 points along the tonotopic axis, nPCA is set to 100 and the number of neurons is 60. The resampling alleviated the computation. It also suppressed the strong tendency of the network to represent pitch. Furthermore, spectrally smoothed spectrograms are expected to better approximate the input to the central auditory system, since neurons there do not respond differentially to fine spectral features [15]. The optimal STRFs of several neurons is shown in Figure 3A. Note that the

horizontal stripes, visible in Figure 2A, has vanished due to the smoothing.

Figure 3B and 3C show jointly that again the preferred frequency histogram tends to follow the average energy distribution. In Figure 3E we find that the rightmost peak of the Q distribution has shifted to lower values, relative to the high-resolution case (Figure 2E). This is expected since the downsampling lowpasses the data along the spectral axis. In the upper panel of Figure 3F, an STRF of a neuron measured from ferret auditory cortex [16] is shown; this can be compared with the STRF from a simulated neuron, shown in the lower panel. They share a certain degree of localization and smoothness, although the feature size of the simulated STRF is somewhat larger.

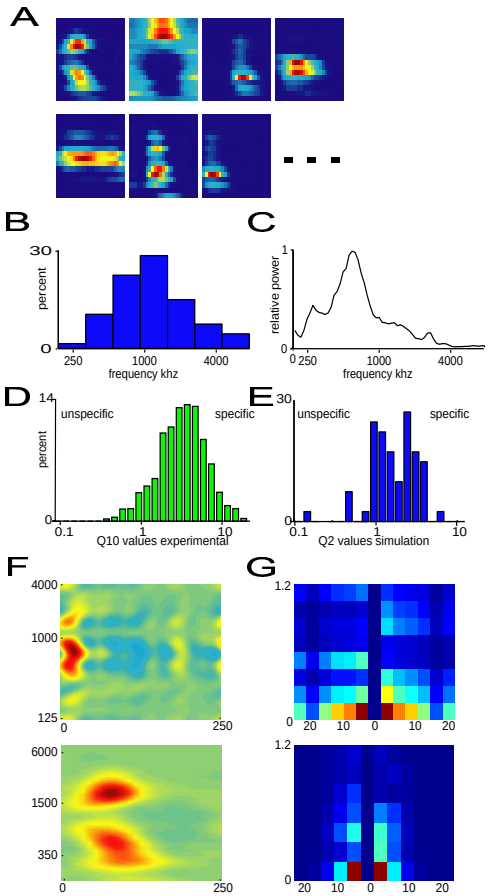


**Figure 3: Low Resolution Results**
A) The color-coded spectrotemporal receptive fields of 7 out of the 100 neurons are shown. B) The histogram of BF for our optimized neuronal population is shown. C) The average energy is shown as a function of frequency. D) The histogram of Q10 measured from cat auditory cortex [13] is shown. E) The histogram of Q2 for our simulated neurons is shown. F) The STRF of a neuron measured from ferret auditory cortex [16] is shown (upper panel), along with the STRF of a simulated neuron. G) The upper two quadrants of the Fourier transforms for each of the STRFs are displayed. Both are well described by quadrant-separable functions (see text).

A particularly intriguing result arises from assessing the separability of these two functions. Neither STRF is well described as being separable into a product of a spectral function with a temporal function (For the simulated STRF, the relative root-mean-squared (RMS) error is 31.1 %. Analysis of the real STRF is more involved, and follows from [7]). However, the quadrants of the Fourier transforms of both STRFs, shown in Figure 3G, are well described as being separable (Relative RMS error is 11.1 % for the simulated STRF). This non-trivial property, called *quadrant separability*, is commonly observed in auditory-cortical neurons [7], although its specific function is as yet unclear. Therefore, one promising point of value in this continuing investigation is the possibility to describe how statistics of natural auditory scenes might give rise to ill-understood functional cortical properties, such as quadrant separability.

## VI. Discussion

The type of representation chosen as input to a learning system is very important and strongly influences the resulting representations. When trying to understand the computational properties of neural systems, it is therefore helpful to have a thorough understanding of the appropriate input. Fortunately, there are a number of studies on the early stages of the auditory system that allow us to at least approximately choose the right input representation [10, 11].

The selection of the input representation might also explain the differences to other theoretical studies of auditory learning. Anthony Bell [17] studied the learning of auditory temporal filters using ICA (searching for super-Gaussian projections on whitened data). He used raw waveforms as input, obtaining sets of filters that represent both the frequency envelope and the phase information. Our filters however only represent the envelope information. Only for impulsive sounds, such as tooth tapping, does the ICA algorithm lead to responses that are localized in time. Michael Lewicki [18] also studies learning from raw waveforms on auditory systems and shows that using overcomplete representations can significantly improve the

model quality. Our study shows the expected effect, that choosing the type of input representation is important for the results that are obtained with such an ICA method. Using spectrograms instead of waveforms better represents the cortical situation and thus leads to responses that can better be compared to cortical responses.

Many properties of the mammalian auditory system largely diverge from the values we found in this study. The preferred modulation frequencies both in time and in spectrum were clearly too low frequency. Furthermore, they did not feature the bands of spectral and temporal inhibition that are commonly revealed by physiological experiments. We are currently only able to demonstrate qualitative behavior. It thus remains an important problem for further research to devise a compact learning mechanism that is able to reproduce much of the neuronal properties found in the auditory cortex, lending itself to a thorough comparison.

It is interesting to compare learning of auditory receptive fields with the learning of visual receptive fields. Spectrotemporal receptive fields in the auditory system are best compared to spatiotemporal receptive fields in the visual system. In both cases some of the neurons are separable in space and time respectively spectrum and time. Other neurons are inseparable. In the visual system motion selective neurons are among those inseparable neurons. They can nevertheless be learnt maximizing the sparseness of the ensemble[19]. In the auditory system a large number of neurons are not separable in spectrum and time but instead are quadrant separable in Fourier space. Whether or not this property can be generalized to visual neurons remains an important question for further research.

Interestingly the properties of the visual system are understood far better, and analyzed far more vigorously, than those of the auditory system. We consider it important to also analyze the auditory system since the statistics of their inputs, such as principal and independent components, at least superficially seem very different. If parts of the cortical algorithm are conserved over both auditory and visual system, studies of learning in the auditory system might lead to insights that can be helpful in understanding the visual system as well.

All in all it is an interesting endeavor to describe representations of simulated neurons that are efficient with respect to some criterion for natural stimuli[1-3, 5, 12, 17-22]. This leads to very compact models that are able to at least qualitatively describe a number of features of the biological system [21]. In addition to that it might lead to insights why sensory systems have the properties they have and link computation to anatomical and electrophysiological detail.

## VII. Acknowledgements

We want to thank ETH Zürich, UNI Zürich, SPP, Boehringer Ingelheim Fonds (KPK), Collegium Helveticum (KPK) and the Swiss National Fonds (PK – grant no: 3100-51059.97) for financial support. We would like to acknowledge the ADA – “the intelligent room” project that is part of EXPO 2002, that in part inspired this project. We would furthermore like to thank Bruno Olshausen, Tony Bell, and Heather Read for inspiring discussions and technical assistance. We thank Jörg Conradt and Christoph Kayser for reading previous versions of this manuscript.

## Bibliography

[1] H. B. Barlow, "Possible principles underlying the transformation of sensory messages.," in *Sensory Communication*, W. Rosenblith, Ed.: M.I.T. Press, Cambridge MA, 1961, pp. 217.

[2] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network: Computation in Neural Systems*, vol. 3, pp. 213-251, 1992.

[3] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607-9., 1996.

[4] J. H. Van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex.," *Proc R Soc Lond B Biol Sci*, vol. 265, 1998.

[5] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Res*, vol. 37, pp. 3327-3338, 1997.

[6] A. Hyvärinen, "Survey on Independent Component Analysis," *Neural Computing Surveys*, vol. 2, pp. 94-128, 1999.

[7] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol*, vol. 85, pp. 1220-1234, 2001.

[8] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neruons," *Science*, vol. 280, pp. 1439-1444, 1998.

[9] S. A. Shamma, " On the role of space and time in auditory processing," *TRENDS Cog Sci*, vol. 5, pp. 340-348, 2001.

[10] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans Inf Theory Special Issue on Wavelet Transforms and Multiresolution Signal Analysis*, vol. 38, pp. 824-839, 1992.

[11] K. Wang and S. A. Shamma, "Representation of spectral profiles in primary auditory cortex," *IEEE Trans Speech Audio Process*, vol. 3, pp. 382-395, 1995.

[12] A. Hyvärinen and P. Hoyer, "Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural Comput*, vol. 12, pp. 1705-20., 2000.

[13] C. E. Schreiner, H. L. Read, and M. L. Sutter, "Modular Organization of Frequency Integration in Primary Auditory Cortex," *Annu. Rev. Neurosci*, vol. 23, pp. 501-529, 2000.

[14] P. Heil, R. Rajan, and D. R. F. Irvine, "Sensitivity of neurons in cat primary auditory cortex to tones and frequency-modulated stimuli. I: Effects of variation of stimulus parameters," *Hear. Res.*, vol. 63, pp. 108-134, 1992.

[15] M. A. Escabí, C. E. Schreiner, and L. M. Miller, "Dynamic time-frequency processing in the cat auditory idbrain, thalamus, and auditory cortex: spectrotemporal receptive fields obtained using dynamic ripple spectra," *Soc. Neurosci. Abstr.*, vol. 24, pp. 1879, 1998.

[16] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robost spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design," *J. Comp. Neurosci.*, vol. 9, pp. 85-111, 2000.

[17] A. J. Bell, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, vol. 7, pp. 261-266, 1996.

[18] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput*, vol. 12, pp. 337-365, 2000.

[19] B. A. Olshausen, "Sparse codes and spikes. ," in *Probabilistic Models of the Brain: Perception and Neural Function.* , R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, Eds.: MIT Press, 2001.

[20] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Res*, vol. 37, pp. 3311-25., 1997.

[21] E. P. Simoncelli and B. A. Olshausen, "Natural Image Statistics and Neural Representation,," *Annu. Rev. Neurosci*, vol. 24, pp. 1193-1216, 2001.

[22] C. Kayser, W. Einhäuser, O. Dümmer, P. König, and K. P. Körding, "Extracting slow subspaces from natural videos leads to complex cells.," *International conference on artificial neural networks*, vol. 9, 2001.