Efficient Coding of Natural Sounds

David Ferrone

The University of Connecticut

September 10, 2011

Overview

"In fact, our ear uses a wavelet transform when analyzing sound, at least in the very first stage." - *Ingrid Daubechies*

"The ear perceives the bare signal f and processes f into a representation that provides the simultaneous information about time and frequency that we call music." - *Karlheinz Grochenig*

- Discrete Fourier Transform (DFT) vs. Discrete Wavelet Transform (DWT)
- Efficient Coding and Sensory Coding
- Filter Theory and Independent Component Analysis
- Auditory System
- Efficient Coding of Natural Sounds (Lewicki, 2002)

Discrete Fourier Transform

For a signal of length N ($\{x_i\}_0^{N-1}$) the Discrete Fourier Transform (DFT) is defined

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi i}{N} kn}, \quad k = 0, 1, \ldots, (N-1),$$

with the inverse defined as

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n e^{\frac{2\pi i}{N} kn}.$$

Discrete Fourier Matrix

Another way to explicitly define the Discrete Fourier Transform is with a Vandermonde matrix. Then $X = Fx$, where $\omega = e^{\frac{-2\pi i}{N}}$, and

$$F = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & \omega^{1\cdot 1} & \omega^{1\cdot 2} & \dots & \omega^{1\cdot(N-2)} & \omega^{1\cdot(N-1)} \\ 1 & \omega^{2\cdot 1} & \omega^{2\cdot 2} & \dots & \omega^{2\cdot(N-2)} & \omega^{2\cdot(N-1)} \\ 1 & \omega^{3\cdot 1} & \omega^{3\cdot 2} & \dots & \omega^{3\cdot(N-2)} & \omega^{3\cdot(N-1)} \\ & & & \vdots & & \\ 1 & \omega^{(N-1)\cdot 1} & \omega^{(N-1)\cdot 2} & \dots & \omega^{(N-1)\cdot(N-2)} & \omega^{(N-1)\cdot(N-1)} \end{pmatrix}$$

Wavelets

A *scaling function* is a function $\phi : \mathbb{R} \to \mathbb{R}$ which satisfies the following equation:

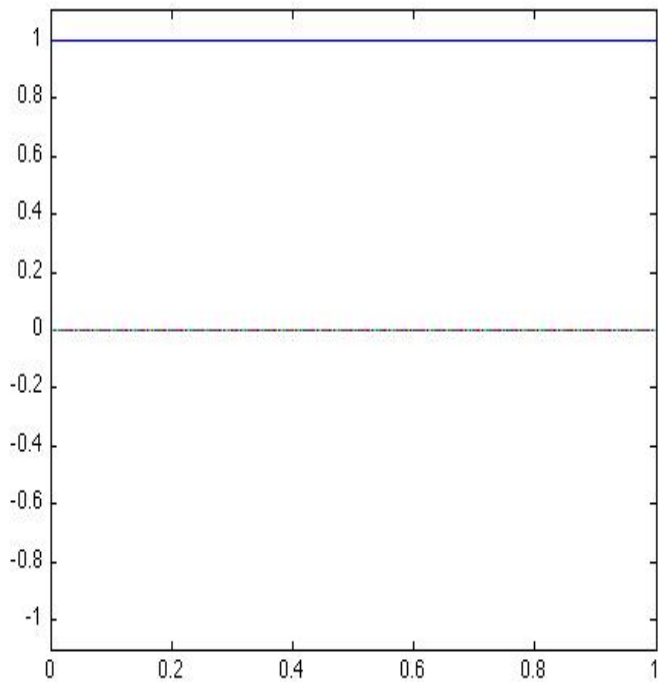
$$\phi(x) = \sum_n s_n \phi(2x - n)$$

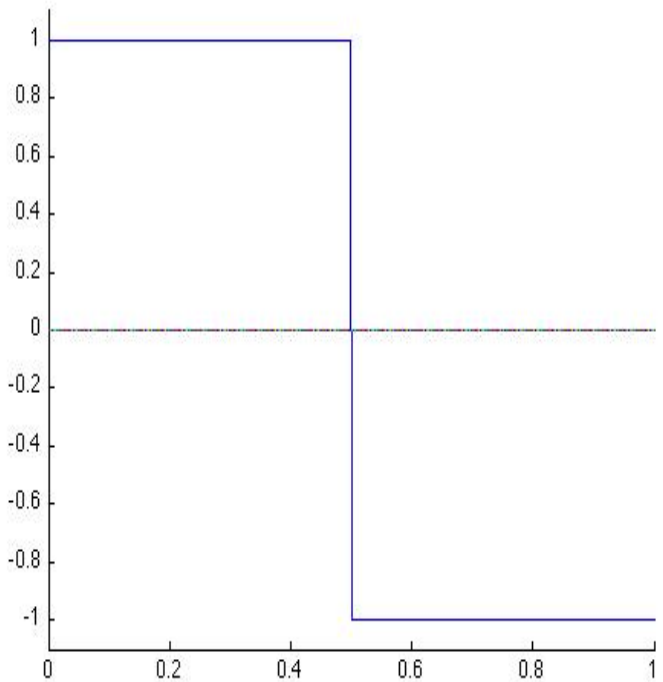
The real numbers $\{s_n\}$ are called the *scaling coefficients*. The *wavelet coefficients* are given by $w_n = (-1)^n s_{1-n}$, and these are used to define the corresponding wavelet function

$$\psi(x) = \sum_n w_n \phi(2x - n)$$

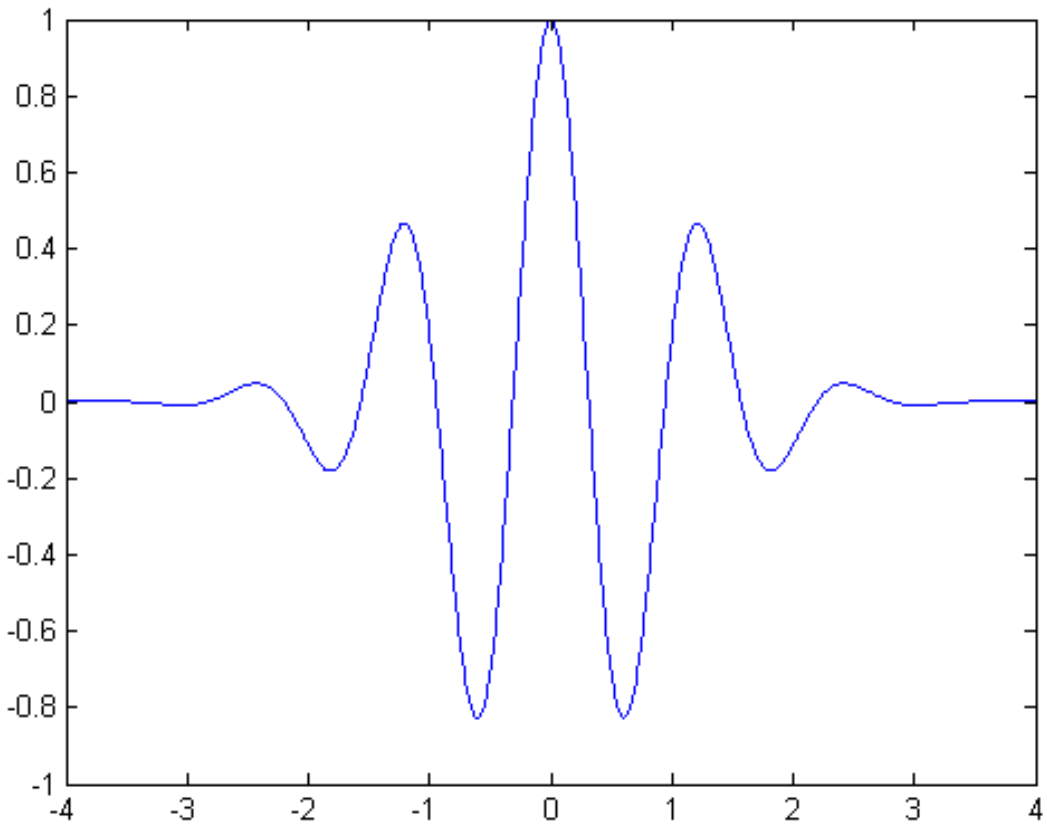
Haar Wavelet

The Haar function is the characteristic (indicator) function over $[0, 1]$ with scaling coefficients $\{s_0, s_1\} = \{1, 1\}$ and wavelet coefficients $[1, -1]$.





Morlet Wavelet



Discrete Wavelet Transform

$$T = \begin{pmatrix}
s_0 & s_1 & s_2 & s_3 & \dots & s_{2n+1} & 0 & \dots & 0 & 0 \\
w_0 & w_1 & w_2 & w_3 & \dots & w_{2n+1} & 0 & \dots & 0 & 0 \\
0 & 0 & s_0 & s_1 & s_2 & s_3 & \dots & s_{2n+1} & 0 & \vdots \\
0 & 0 & w_0 & w_1 & w_2 & w_3 & \dots & w_{2n+1} & 0 & \vdots \\
\vdots & \vdots & & & \ddots & \ddots & & & & \vdots \\
0 & 0 & 0 & \dots & s_0 & s_1 & s_2 & \dots & s_{2n} & s_{2n+1} \\
0 & 0 & 0 & \dots & w_0 & w_1 & w_2 & \dots & w_{2n} & w_{2n+1} \\
s_{2n} & s_{2n+1} & 0 & 0 & \dots & \ddots & \ddots & & s_{2n-2} & s_{2n-1} \\
w_{2n} & 2_{2n+1} & 0 & 0 & \dots & & \ddots & \ddots & w_{2n-2} & w_{2n-1} \\
\vdots & & \ddots & & & & & \ddots & \vdots & \vdots \\
s_2 & s_3 & \dots & s_{2n+1} & 0 & \dots & \dots & 0 & s_0 & s_1 \\
w_2 & w_3 & \dots & w_{2n+1} & 0 & \dots & \dots & 0 & w_0 & w_1
\end{pmatrix}$$

Comparison

Here is a matrix transformation for a signal of length 4 using the Haar wavelet.

$$T = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

Here is the Fourier matrix:

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}$$

Efficient Coding

Neurons communicate by firing nerve impulses (also called action potentials or spikes). It is believed that all of the information that a sensory neuron encodes is captured by these spikes, and one of the goals of sensory neuroscience is to understand this sensory code.

The efficient coding hypothesis (suggested by Horace Barlow, 1961) is the belief that the brain will minimize the number of spikes used to transmit information. An efficient code should minimize redundancy.

Another belief is that neurons in our visual or auditory systems should be optimized for coding images or sounds that we find in nature.

Filters

Given a signal x, applying a linear filter is the same as taking the convolution with a function.

$$a_i(t) = \sum_k x(k) h_i(t - k)$$

Filters

Given a signal x, applying a linear filter is the same as taking the convolution with a function.

$$a_i(t) = \sum_k x(k) h_i(t-k)$$

$$\mathbf{a}_t = \mathbf{H}_t x$$

Independent Component Analysis

A method for decomposing a mixed signal into its constituent parts, assuming the originals signals were independent. (e.g. The Cocktail Party Problem)

The method used in Lewicki's article to find filters which encode natural sounds into independent channels is based on a method called *infomax*.

Filter Estimation

Let

$$\mathbf{x} = \sum_i a_i \phi_i$$

$$\mathbf{x} = \mathbf{a}\Phi$$

Then

$$\mathbf{a} = \Phi^{-1}\mathbf{x}$$

The finite impulse response filters h_i are given by the rows of Φ^{-1}.
The data likelihood is now $p(\mathbf{x}|\Phi) = p(\mathbf{a})/|\det \Phi|$.
(Although technically the true density $p(\mathbf{x})$ is unknown and is also being estimated.)

Maximizing Likelihood

The quantity $p(\mathbf{x}|\Phi) = \frac{p(\mathbf{a})}{|\det \Phi|}$ is maximized using a method called stochastic gradient descent, under the model

$$\nabla\Phi \propto \Phi\Phi^T \frac{\partial}{\partial\Phi} \log p(\mathbf{x}|\Phi)$$

The idea is that an objective function, $M(\omega)$, with dependence on a parameter (ω) can be minimized by iterating the following (with step-size α):

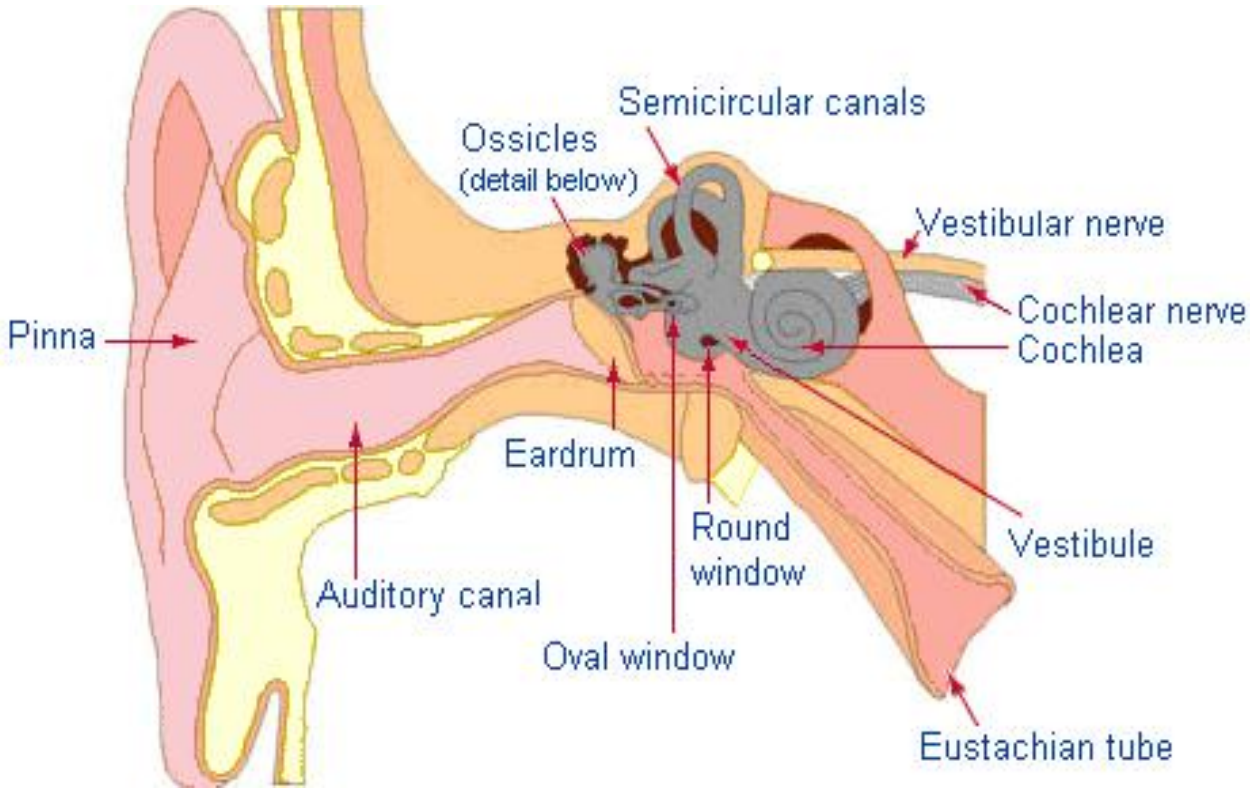
$$\omega = \omega - \alpha\nabla M(\omega)$$

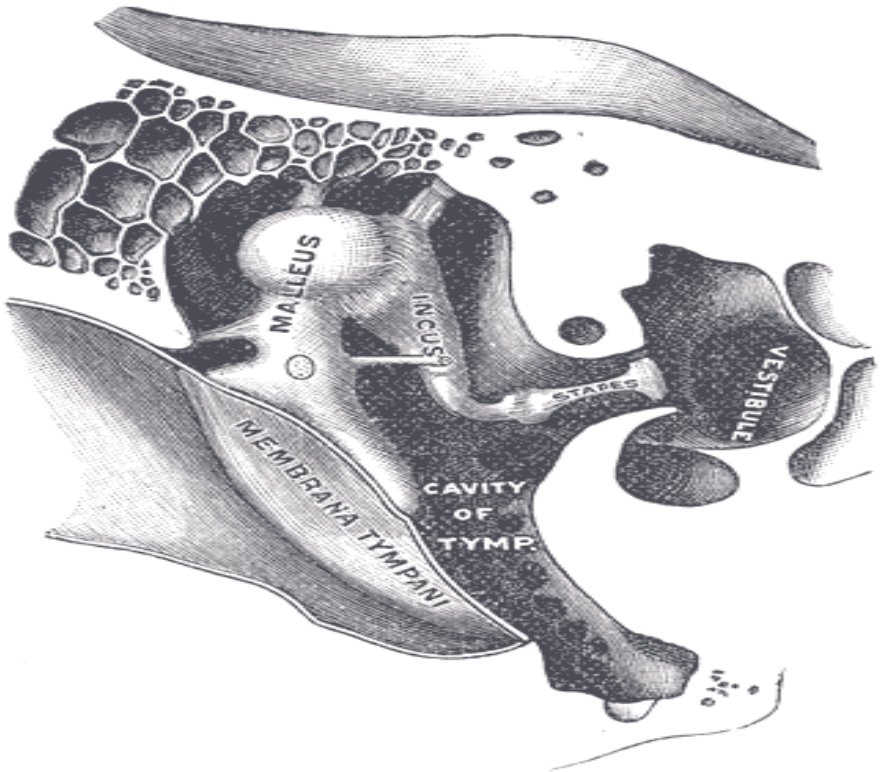
While this estimation is occurring, the probability density for $\mathbf{a}$ is also being estimated. The coefficients $\mathbf{a}$ are assumed to be generalized Gaussians $(\mathbf{a} \sim e^{-|t|^\beta})$ with the largest exponent possible given the data.

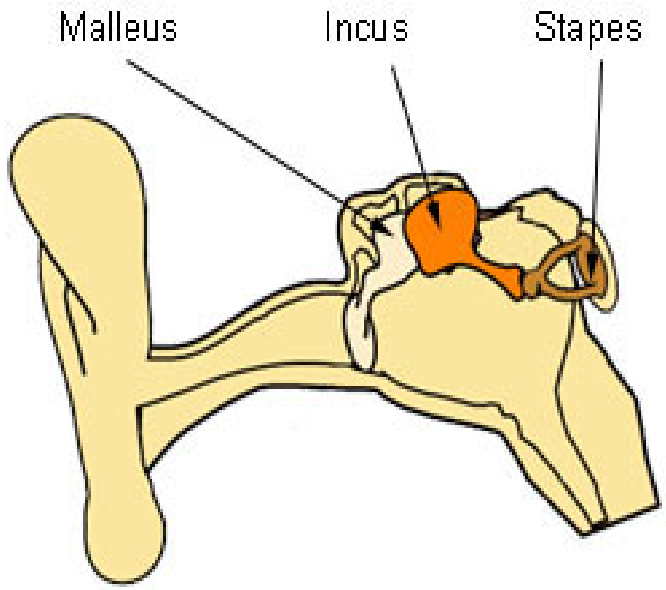
Reverse Correlation

The filters that actually reflect what your auditory system are using have also been estimated. The auditory nerve filters were estimated using a method called reverse correlation. This is more generally called spike-triggered averaging (STA).

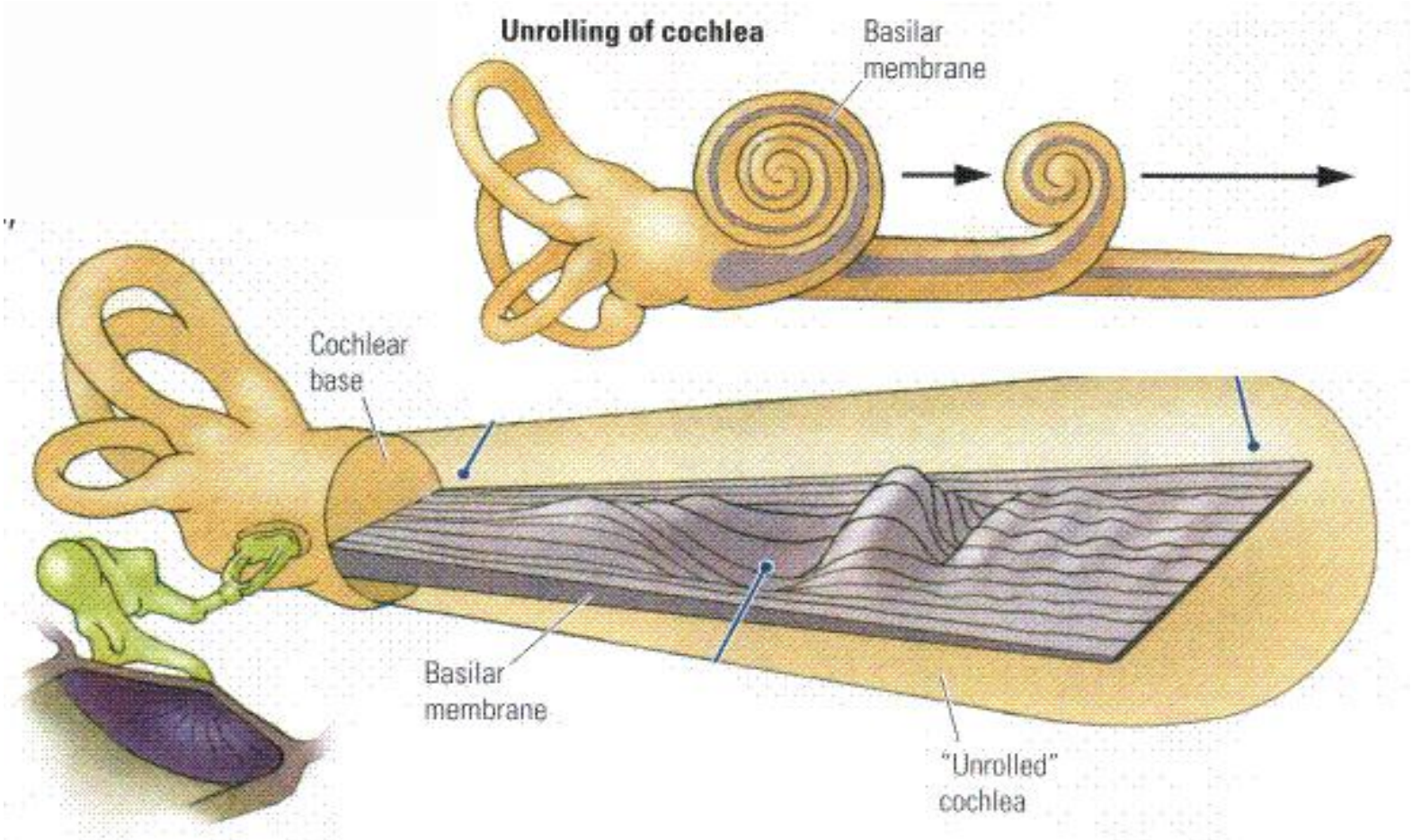
STA is the average stimulus preceding a spike. A known signal is sent through the auditory system, and a series of spikes from a neuron are recorded. A small section of the signal just before each spike occurred is used to construct an average and the result is called a 'revcor' filter.

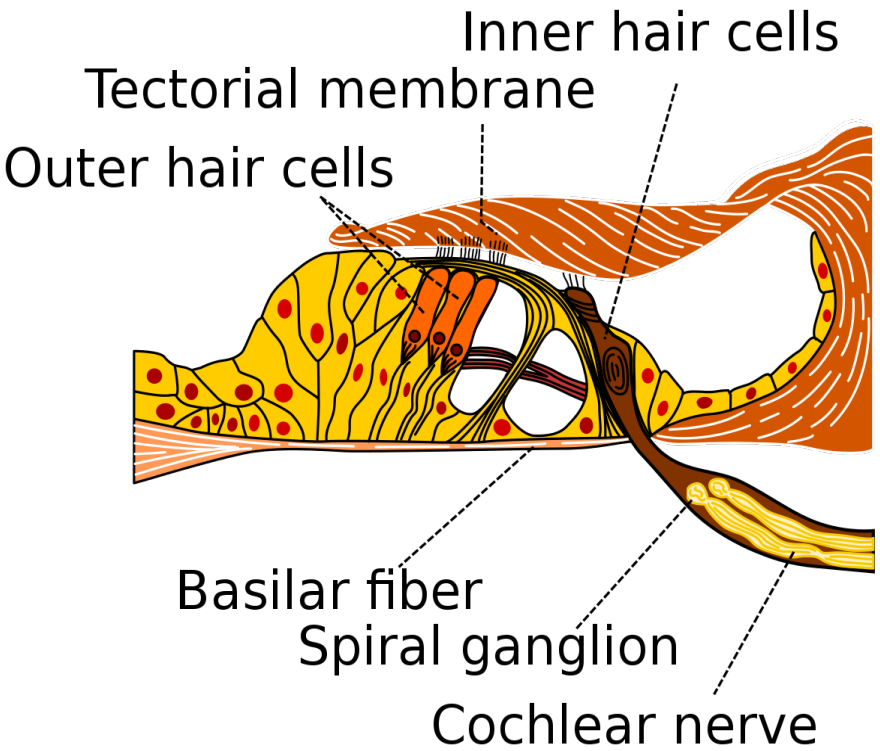

Semicircular canals
Ossicles
(detail below)
Vestibular nerve
Cochlear nerve
Pinna
Cochlea
Eardrum
Round
window
Vestibule
Auditory canal
Oval window
Eustachian tube

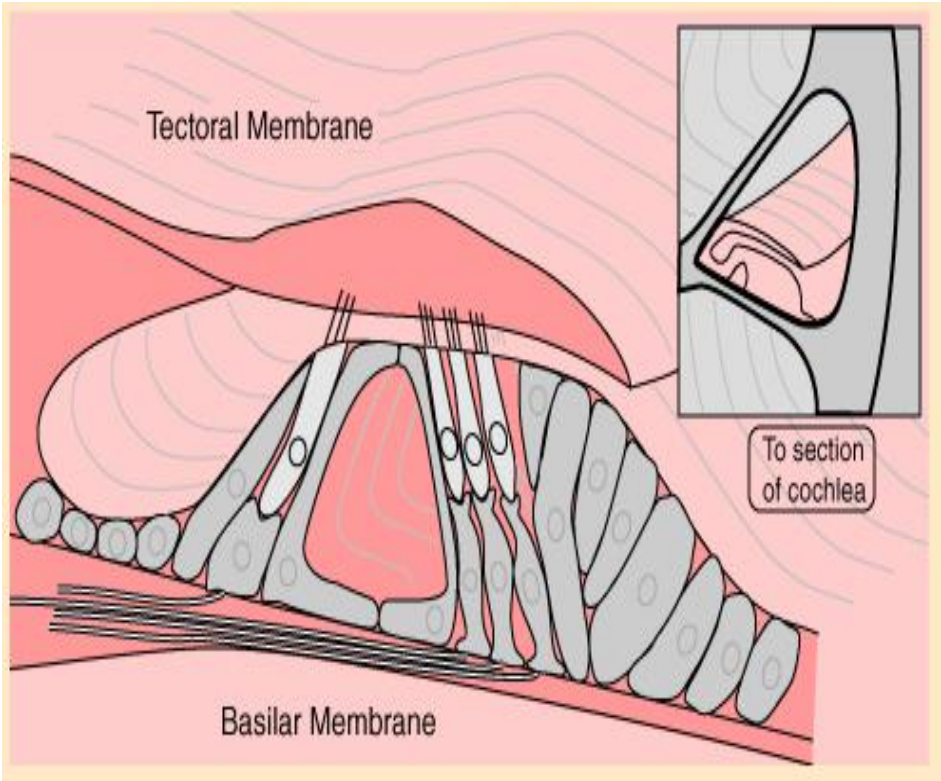




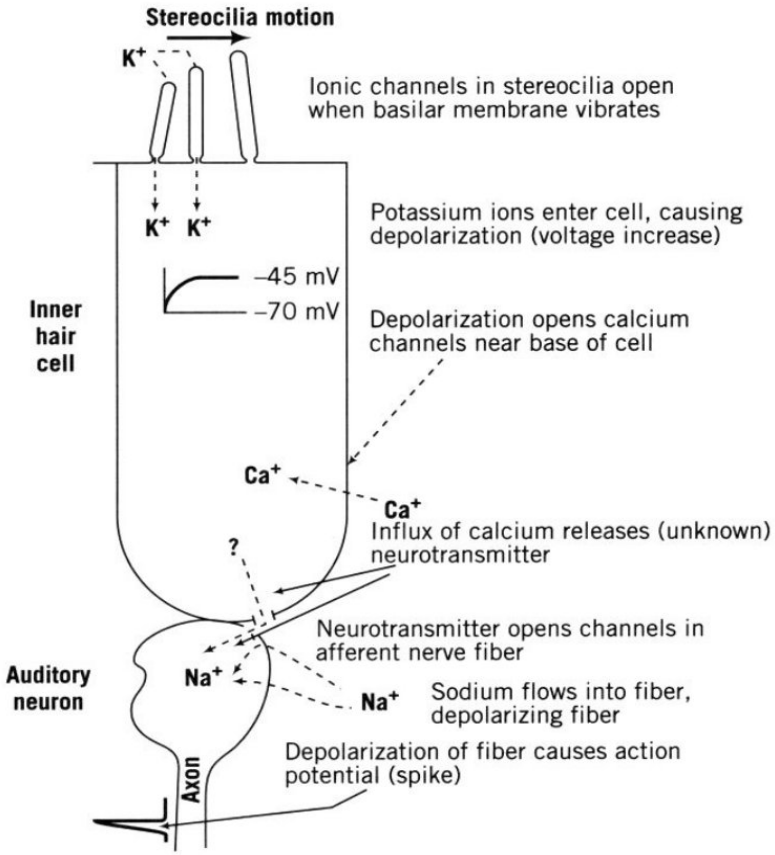
Cranial Bones


Unrolling of cochlea
Basilar membrane
Cochlear base
Basilar membrane
"Unrolled" cochlea

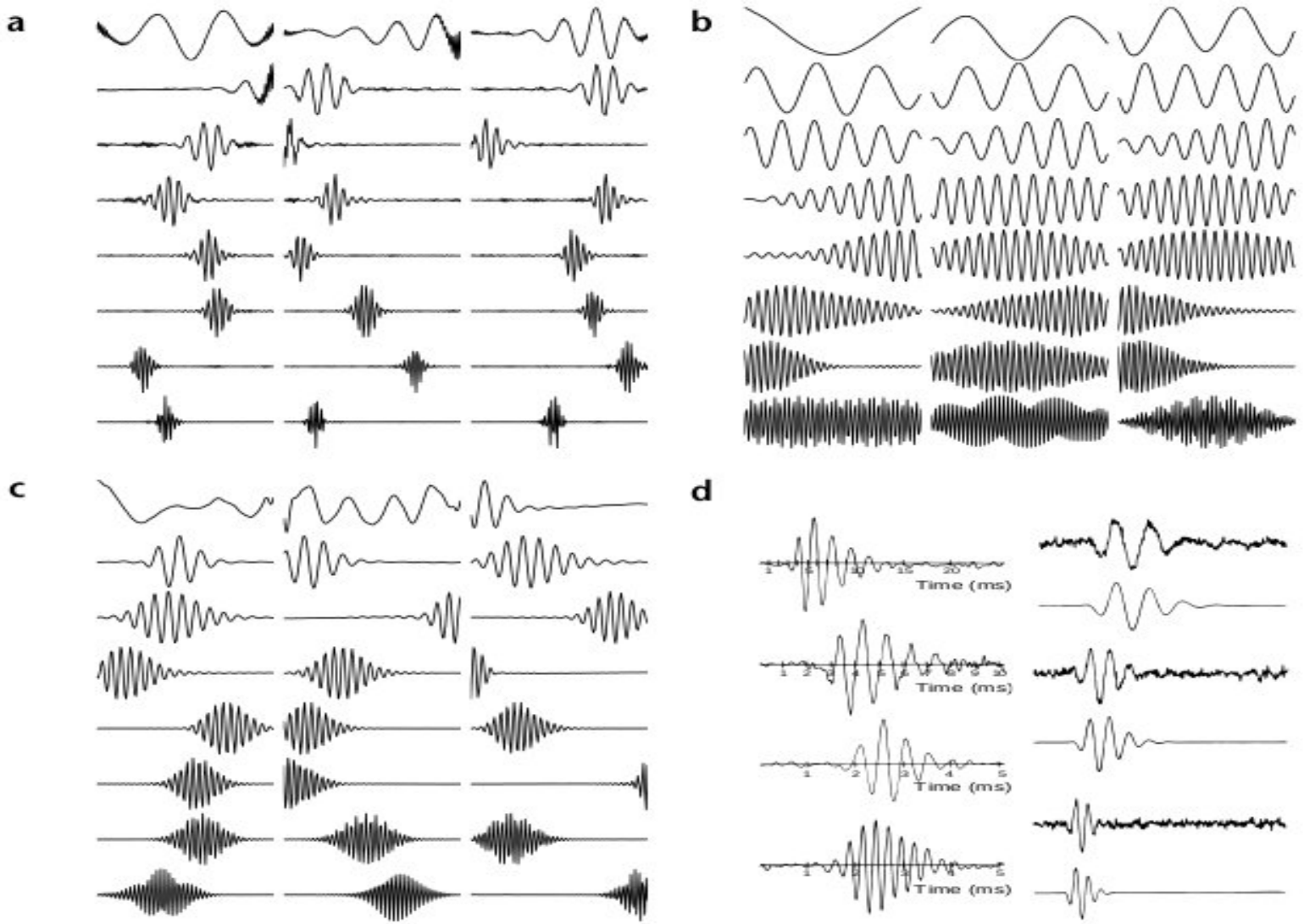

Inner hair cells
Tectorial membrane
Outer hair cells
Basilar fiber
Spiral ganglion
Cochlear nerve

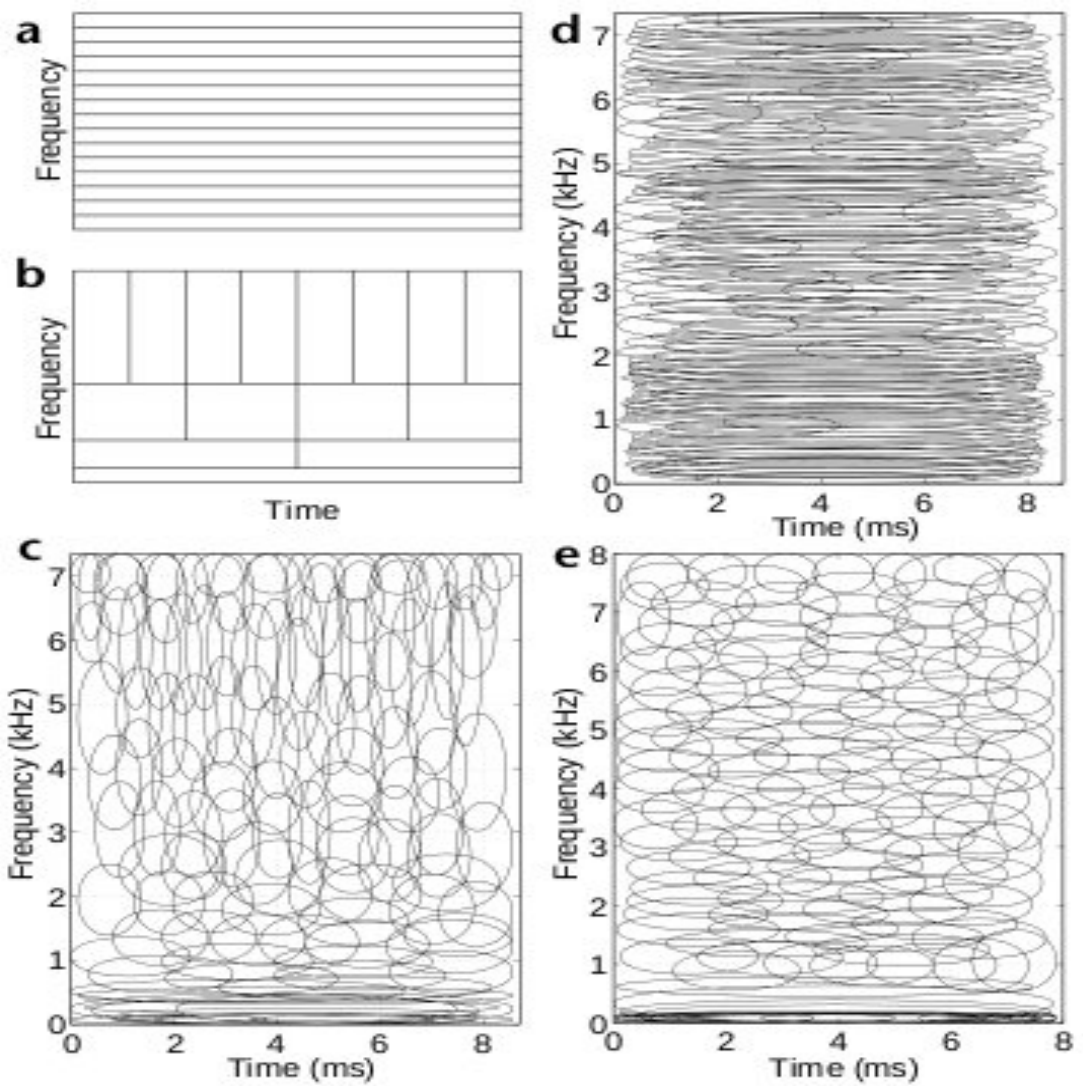

Tectoral Membrane
To section of cochlea
Basilar Membrane

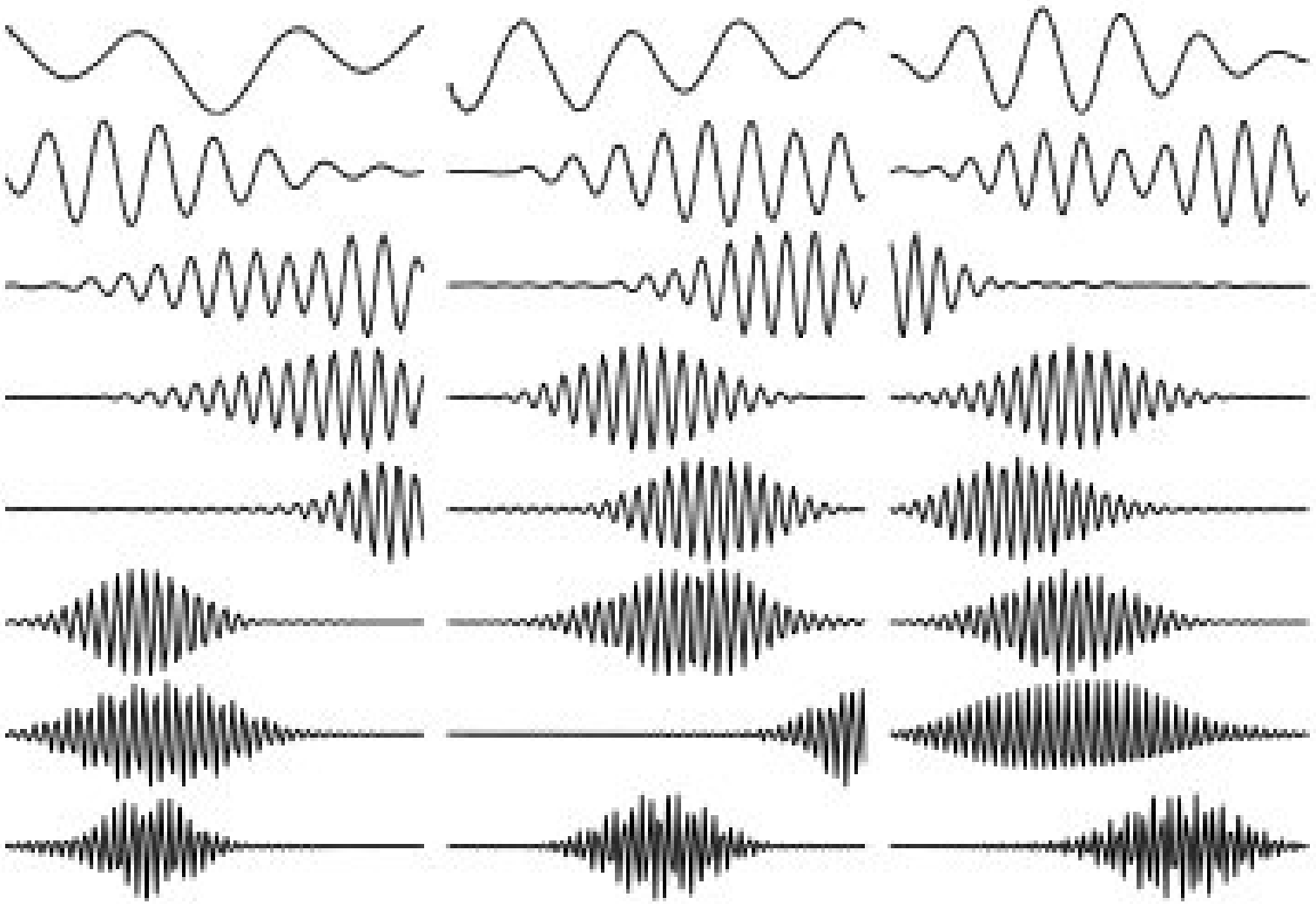
Hair-cell transduction

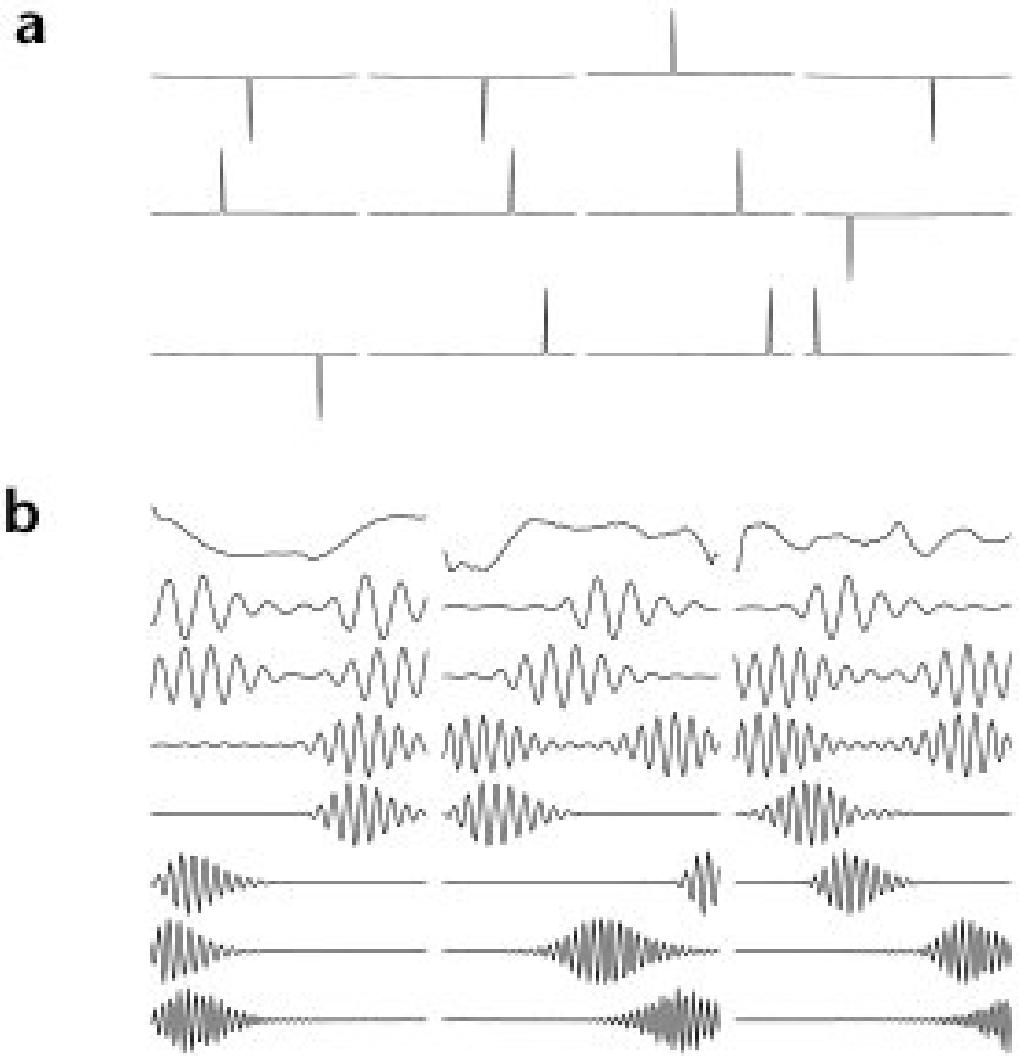


(Gold and Morgan, Speech and Audio Signal Processing, 2000)


a
b
c
d
Time (ms)
Time (ms)
Time (ms)
Time (ms)


a
Frequency
b
Frequency
Time
d
Frequency (kHz)
Time (ms)
c
Frequency (kHz)
Time (ms)
e
Frequency (kHz)
Time (ms)




a
b

References

Thank You!

- Auditoryneuroscience.com (https://mustelid.physiol.ox.ac.uk/drupal/) accessed September 10, 2011.
- Daubechies, I. (1992) *Ten Lectures On Wavelets*. Philadelphia: SIAM.
- Gröchenig, K. (2001) *Foundations of Time-Frequency Analysis*. New York: Birkhäuser.
- Lewicki, M. (2002). Efficient Coding of Natural Sounds. *Nature Neuroscience*, 5(4), 356-363.
- Port, R. (2007, September 1). *Audition for Linguists*. Retrieved August 16, 2011, from http://www.cs.indiana.edu/ port/teach/641/audition.for.linguists.Sept1.html
- Meddis, R., & Lopez-Peveda, E.A. (2010) *Computational Models of the Auditory System*. New York: Springer Science.