The efficient coding hypothesis for the peripheral auditory system

François Deloche (CAMS / EHESS / PSL University)
3-rd year PhD Student

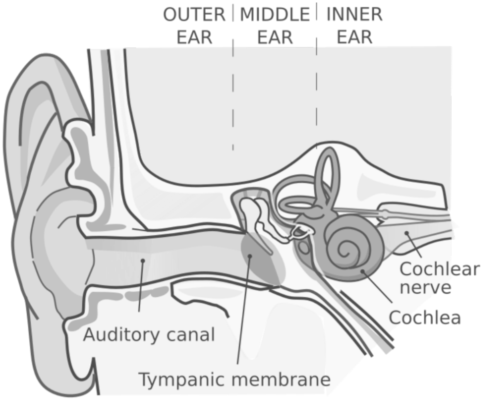
Séminaire Neuromathématiques, Collège de France
2019, April

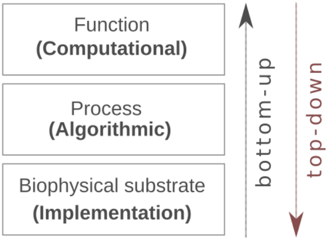
Project *SpeechCode*

Project *SpeechCode*: *Cracking the Speech Code: The Neural and Perceptual Encoding of the Speech Signal*

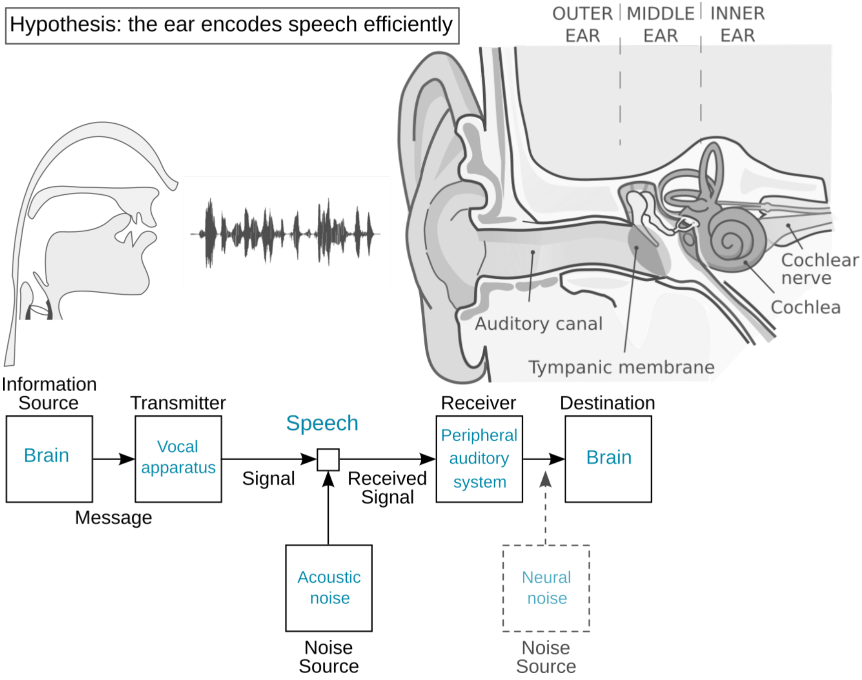
- **Laboratoire Psychologie de la Perception (LPP - Paris-Descartes)**
  Judit Gervain, Ramon Guevarra Erra
- **Laboratoire des systèmes perceptifs (LSP - ENS)**
  Christian Lorenzi, Léo Varnet
- **Centre d'analyse et de mathématique sociales (CAMS)**
  Jean-Pierre Nadal, François Deloche, Laurent Bonnasse-Gahot

The peripheral auditory system

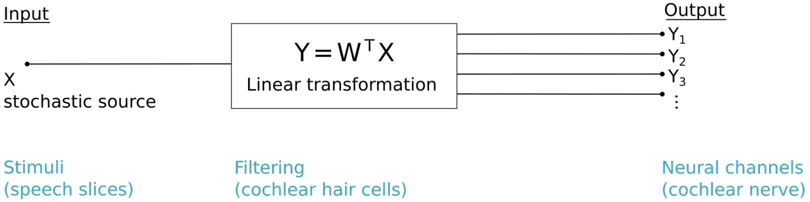




D. Marr's levels of analysis


Hypothesis: the ear encodes speech efficiently
OUTER EAR
MIDDLE EAR
INNER EAR
Cochlear nerve
Cochlea
Auditory canal
Tympanic membrane
Information Source
Brain
Message
Transmitter
Vocal apparatus
Signal
Speech
Received Signal
Acoustic noise
Noise Source
Receiver
Peripheral auditory system
Neural noise
Noise Source
Destination
Brain

Context



- Y: time-frequency decomposition of the signal X
- $W = (W_1, \ldots, W_m)$: filter bank

The efficient coding hypothesis

The efficient coding hypothesis:
sensory sytems encode natural stimuli efficiently.

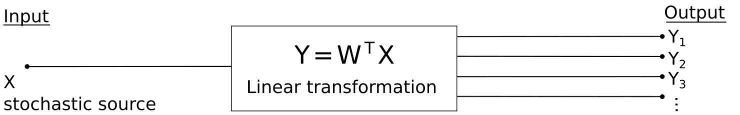
Efficiency ? Several criteria:

- Redundancy reduction [Barlow, 1961]
- Information-maximization [Linsker, 1988]
- Minimum entropy code [Barlow, 1989]
  - Independent feature coding
    $\rightarrow$ Independent Component Analysis (ICA) [Jutten and Herault, 1988]
  - Sparse coding [Olshausen and Field, 2004]

The efficient coding hypothesis

- **Evidence**
  - Empirical: Measures of information transfer in single nerve fibers. Example: higher rates for naturalistic sounds compared to white noise, in auditory nerves [Rieke et al., 1995], in midbrain and auditory cortex [Hsu et al., 2004]
  - Predictive power: Prediction of characteristics of sensory systems based on statistics of natural stimuli.
    Example: Prediction of visual Receptive Profiles (V1) based on statistics of natural images [Olshausen and Field, 1996].
- **Limitations**
  - Higher level processing, information bottleneck.
  - The neural code **is** redundant.
    - NB: not that many auditory hair cells ($\sim$ 1-10k IHCs)
    - Still the 'redundancy reduction' criterion has many benefits [Barlow, 2001] (e.g. general strategy to find good features of data)

Minimum entropy codes



Stimuli
(speech slices)

Filtering
(cochlear hair cells)

Neural channels
(cochlear nerve)

$$\boxed{\min_{W} h(W) = \min_{W} \sum_i H(Y_i) \; - H(Y)}$$

- $H(Y_i) = -\mathbb{E}(\log p(y_i))$: marginal entropy terms
- $H(Y)$: joint entropy
- entropy $\leftrightarrow$ quantity of information $\leftrightarrow$ coding/neuronal resources

$-H(Y)$ behaves as a **penalty term**. It prevents the collapse of filters W_i during learning (controls size and correlation).

$\rightarrow$ W square matrix: $-H(Y) = -H(X) - \log|\det W|$

Overcompleteness

Case where W is a rectangular matrix $n \times m$ with $m > n$.
What happens to the penalty term ?

- No natural expression
- Every overcomplete dictionaries have highly correlated components.
- Minimum entropy of outputs gain importance from decorrelation of filters.
- Still, we want the dictionaries to represent all directions of the space (**diversity** of filters)

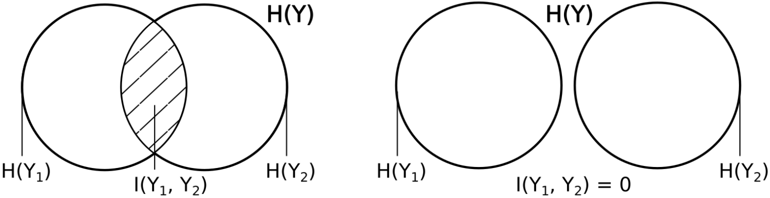
Overcomplete dictionaries of filters uniformly distributed in time/frequency/phase.
$\rightarrow$ just forget the penalty term.

$$h = \sum_i H(Y_i)$$

Independent Component Analysis (ICA)

Mutual information

$$I(Y_1, \cdots, Y_m) = \sum_i H(Y_i) \; - H(Y)$$



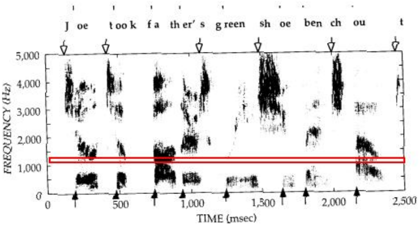
Bivariate case : $I(Y_1, Y_2) = H(Y_1) + H(Y_2) - H(Y_1, Y_2)$

- Type of redundancy
- Intuition: we want the output channels to code for independent features (factorial code).

ICA $\rightarrow$ minimization of $I(Y_1, \cdots, Y_m)$.

Statistical structure

minimum entropy code = structure



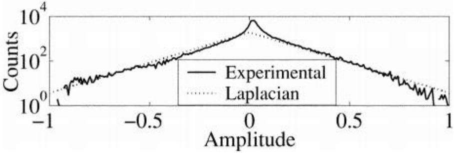
Typical spectogram of speech

In this special case:

structure = sparse activations (peaked distributions around 0)

Probalistic model

How can we estimate the entropy terms (probalistic prior) ?



Posterior distribution of amplitude for typical speech samples (from [Gazor and Zhang, 2003])

Laplace prior : $\log[p(y)] = \log \gamma/2 - \gamma|y|$

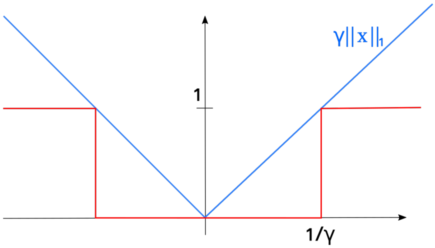
Sparse coding

With this prior, the objective to minimize is:

$$h = \sum_i \gamma_i \mathbb{E}(|Y_i|) = \gamma \mathbb{E}(||Y||_1)$$

(in reality the γ_i are different and depends on the power spectrum.)

Another way to derive the L_1 norm:
Sparse coding : reduce activation or number of neuron spikes (save energy).

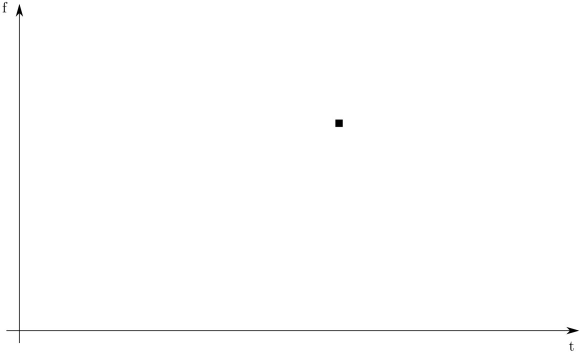


Quadratic time-frequency distributions

- f input signal, g analysis function (real non-negative functions)
- Cross Wigner-Ville distributions:

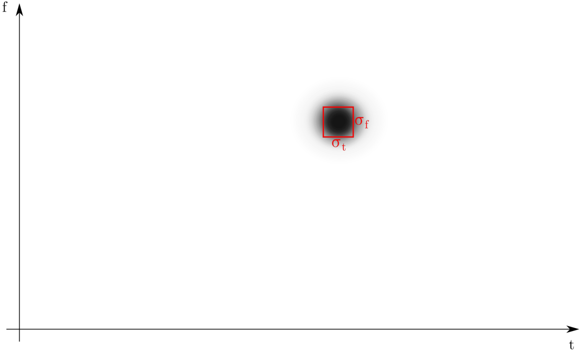
$$W_{f,g}(t,\omega) = \frac{1}{2\pi}\int_{\tau} f(t+\tau/2)\overline{g(t-\tau/2)}e^{-i\omega\tau}\,d\tau$$

Extra-sparse code: is it possible ?



$\sim$ *grandmother cell* for (t_0, f_0)

Heisenberg's uncertainty principle



Extra-sparse (or factorial) code impossible.
Best time-frequency resolution achieved by Gabor filters.

$$\sigma_t \sigma_f = \frac{1}{4\pi}$$

Lieb's uncertainty principle

$$h(f, g) = ||W_{f,g}||_1$$

$$\min_{f,g} h(f, g)$$

$$\text{s.t } ||f||_2 = ||g||_2 = 1 \text{ ?}$$

Lieb's uncertainty principle [Lieb, 1990]

$$||W_{f,g}||_1 \geq ||f||_2 ||g||_2$$

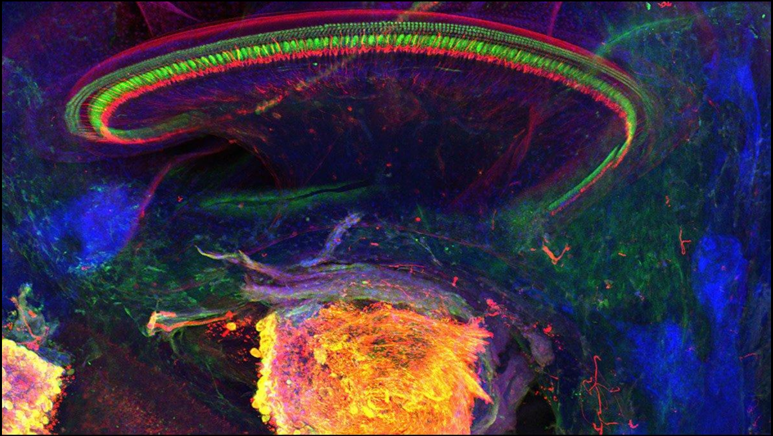
Case of equality: f is Gaussian and $f = g$

Lieb's uncertainty principle

Note:

$$||W_{f,g}||_1 \geq ||f||_2||g||_2 \geq < f, g >= \int_t \int_\omega W_{f,g}(t,\omega)dtd\omega$$

Case of equality: $f = g$ and $W_{f,f}$ is non-negative.
$\rightarrow$ Hudson's theorem : $W_{f,f}$ is non-negative iff f is a Gaussian.

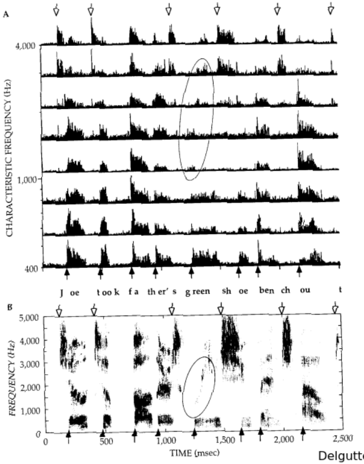


Credits: Glen MacDonald & Ed Rubel

Picture: 3D image (confocal miscroscopy) of a mouse cochlea.

- Sensory hair cells:
  3.5k inner hair cells (IHC) + 12k outer hair cells (OHC)
- Nerve fibers: afferent connections mostly on IHCs
- Tonotopy : place $\leftrightarrow$ frequency

Cochlea = frequency analyzer

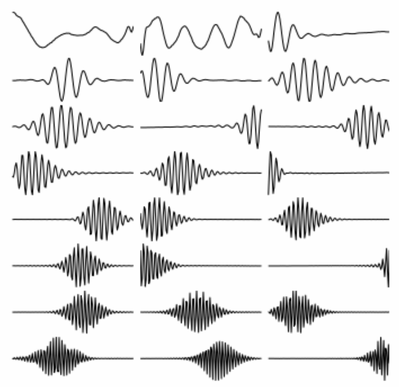


Delgutte 1997

Time histogram of neuron spikes of auditory nerve fibers (cat) in response to an utterance (Delgutte, 1999).

ICA applied to speech

ICA applied to speech produces a bank of filters similar to both Gabor wavelets and auditory filters [Lewicki, 2002] :

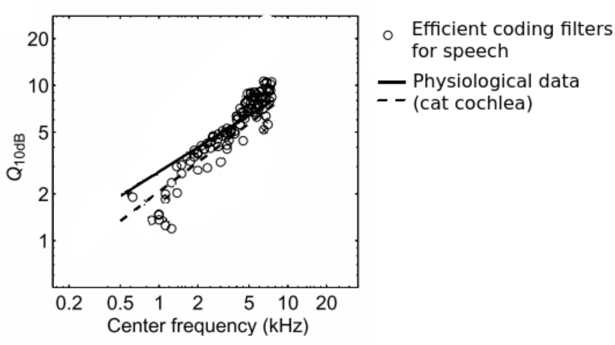


Input X : 128 samples/8ms slices of speech

ICA applied to speech

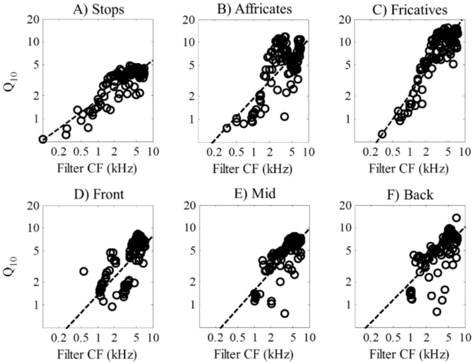
Frequency selectivity is expressed by the **quality factor**:

$$\boxed{Q_{10} = \frac{f_c}{\Delta f_{10dB}}}$$



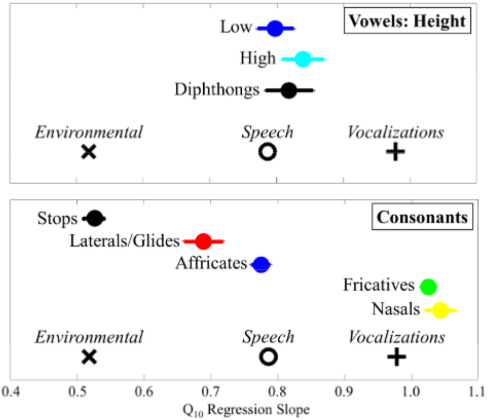
The quality factor Q_{10} is characterized by the same power law for learned filters and auditory filters.

Stilp and Lewicki carried out ICA on different phonetic categories (TIMIT Database: American English).[Stilp and Lewicki, 2013]



β **parameter : slope** Q_{10} **on** f_c (*log-log scale*)

Further analysis



High	iy, uw, ux, ih, ix, uh, er, axr, eh
Low	ah, ax, ax-h, ao, ae, aa
Front	iy, ih, ix, eh, ae
Mid	er, axr, ah, ax, ax-h, aa
Back	uw, ux, uh, ao
Diphthong	ey, ay, oy, aw, ow

Closure	bcl, dcl, gcl, pcl, tcl, kcl, q
Stops	b, d, dx, g, p, t, k
Fricatives	s, z, f, v, sh, zh, th, dh
Affricates	ch, jh
Laterals/glides	r, l, el, w, y, hh, hv
Nasals	m, em, n, en, nx, ng, eng

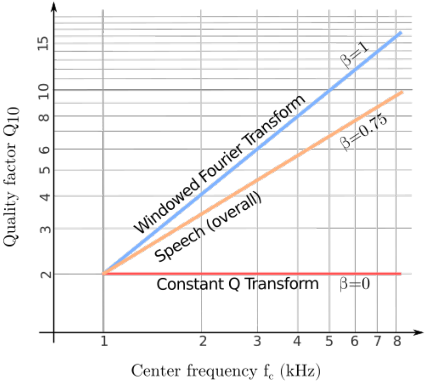
The β parameter depends on the phonetic class (from [Stilp and Lewicki, 2013]).

What is the meaning of the β parameter ?

1. Controls the time-frequency trade-off in the high frequency range

$$Q_{10}(f) = Q_0 \left(\frac{f}{f_0}\right)^{\beta}, \; f_0 = 1.0kHz, \; Q_0 = 2.0$$

2. Separates unique resolution from multi-resolution decompositions



Questions

- Why do we obtain different values for β ?
- What is the meaningful division of speech for stat. structure ?
- What are the **signal/acoustic features** relevant to β ?
- Are there some regularities at a finer level that can be exploited by efficient coding schemes ?

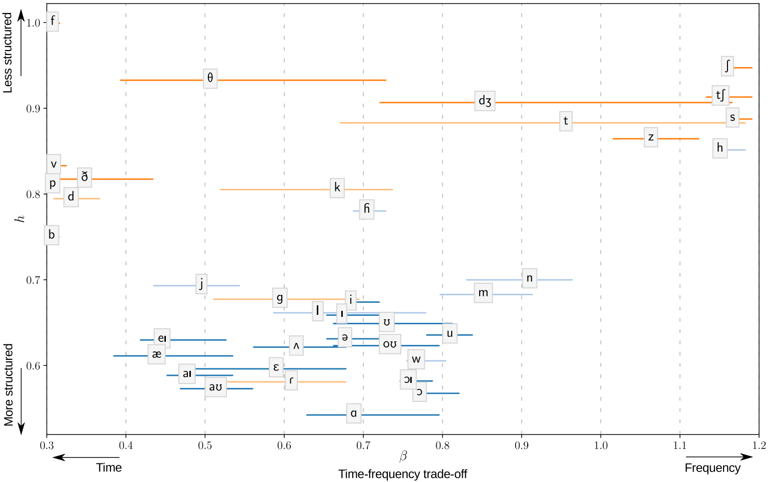
Parametric approach

To go further in the description of the statistical structure of speech, I propose a **parametric approach** instead of ICA.
Method:

1. Create a set of 30 overcomplete dictionaries W_β of Gabor wavelets from $\beta = 0.3$ to $\beta = 1.2$
2. Compute the scores $h(\beta) = \mathrm{E}\left(||W(\beta)^T X||_1\right)$
3. Select $\beta^\star = \arg\min_\beta h(\beta)$.

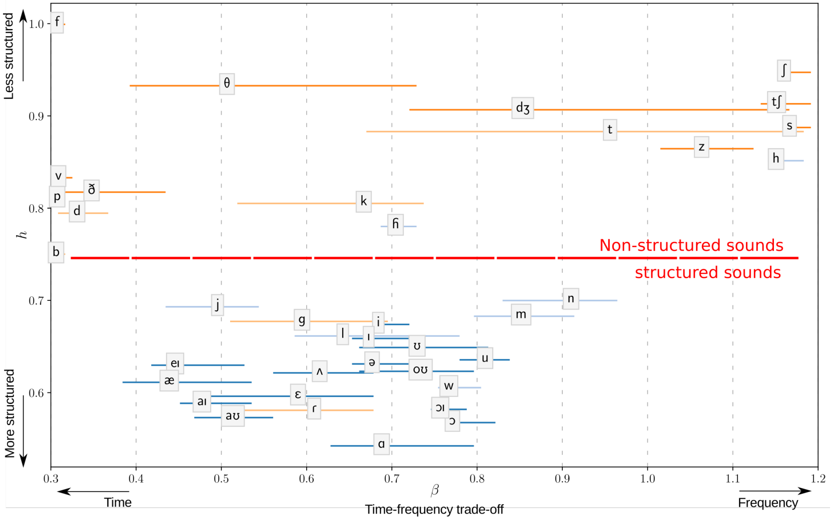
Done for 400 or 800 (normalized) 16ms-slices of speech
Confidence intervals are computed with a bootstrap procedure.

Results

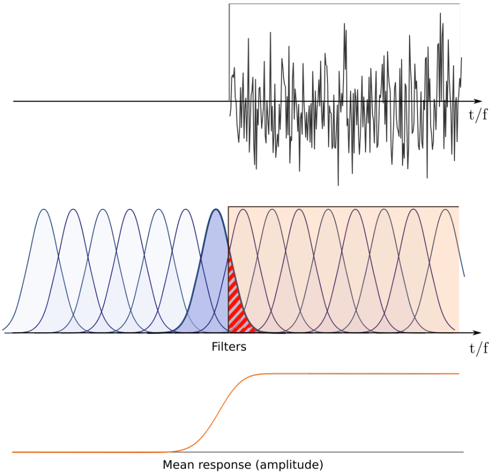


Deloche, 2018

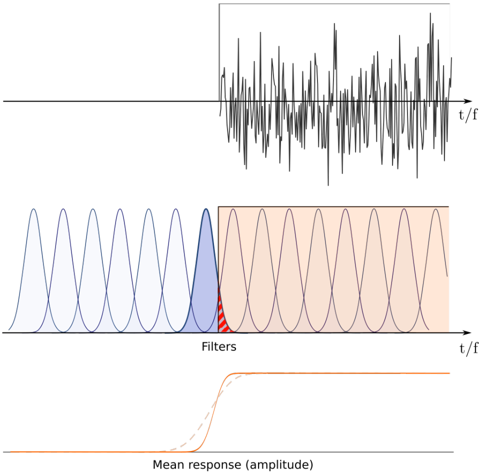
Non-structured sounds and structured sounds

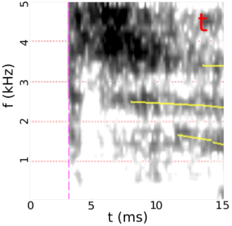
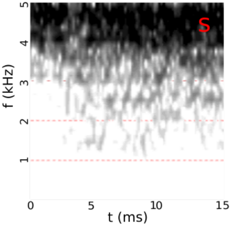


Non-structured sounds

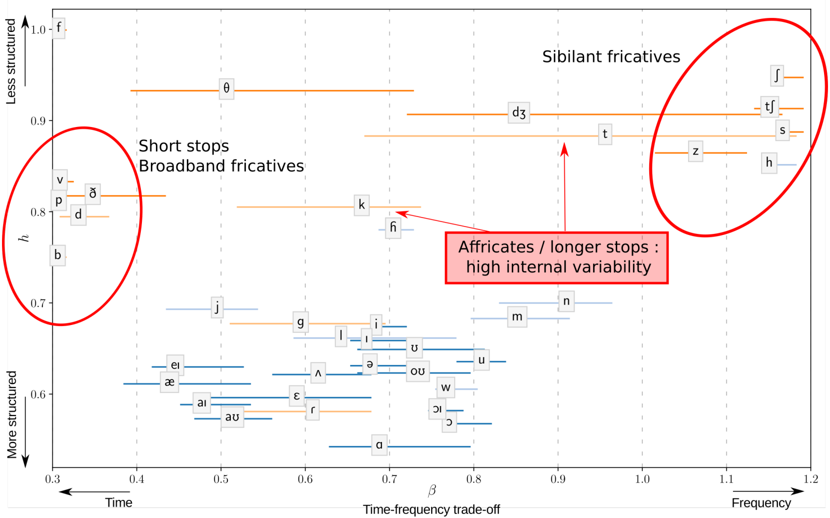


Non-structured sounds

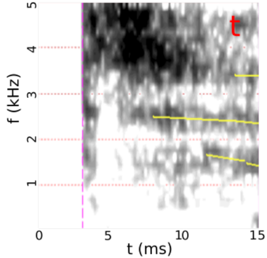


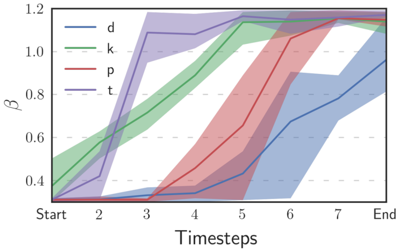
<table>
<tr><td>Stop bursts
(b, p, d, t...)</td><td>Time structure
(low β)</td><td>
</td></tr>
<tr><td>Sibilant fricatives
(s, z, ʃ...)</td><td>Freq. structure
(high β)</td><td>
</td></tr>
</table>

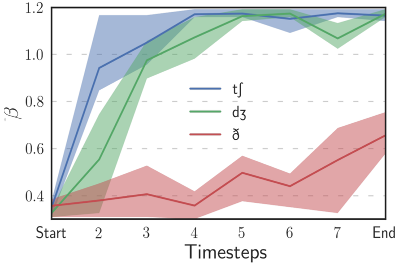
Non-structured sounds



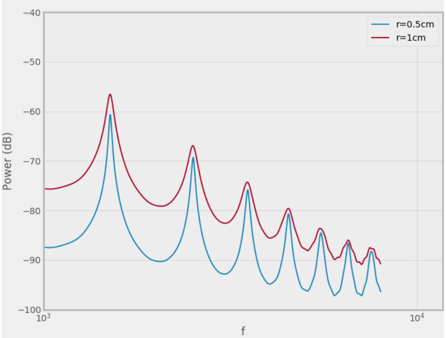
Plosives and affricates are *biphasic*







Vowels

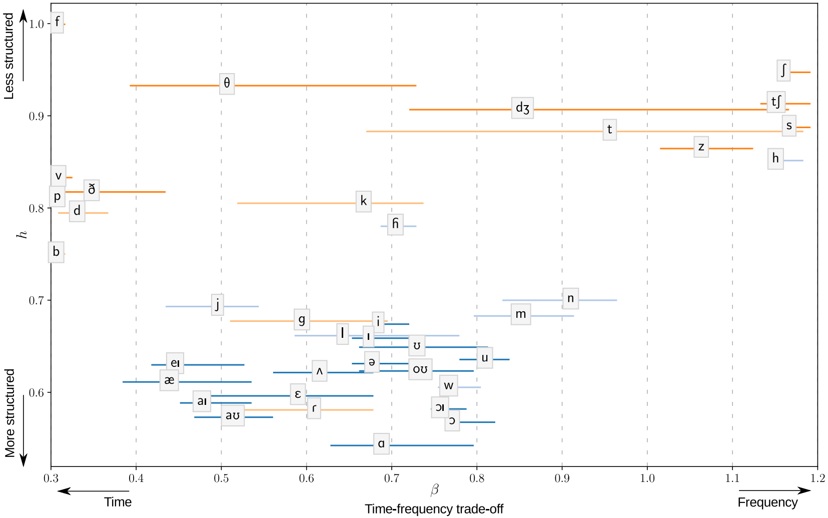


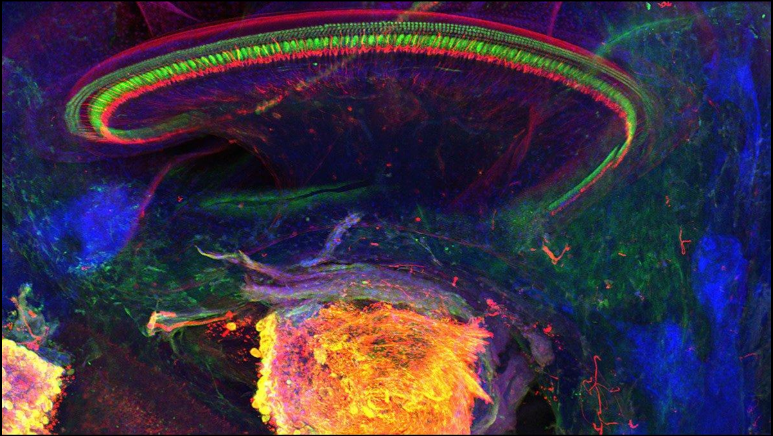
Power spectrum of generated sound at the output of a cylindrical waveguide (for 2 different radii). Greater aperture (=greater loss) results in larger bandwidths.

Two concurrent effects of greater aperture:

1. Larger bandwidths
2. Higher sound intensity level

Structured sounds: Vowels





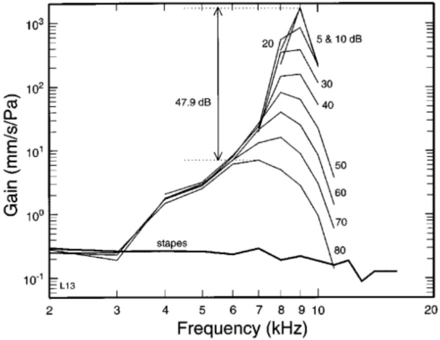
Credits: Glen MacDonald & Ed Rubel

Sensory hair cells: 3.5k inner hair cells (IHC) + 12k outer hair cells (OHC)
Role of outer hair cells ?

amplify signal + increase frequency selectivity

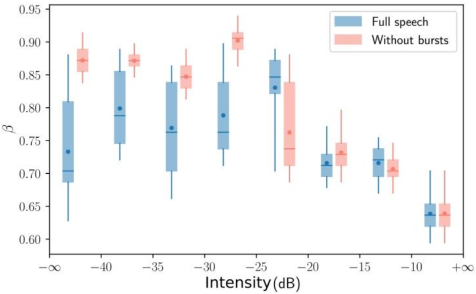
Level dependence

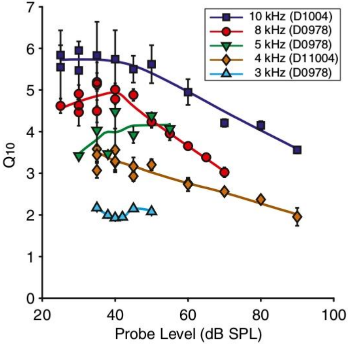
OHC have a **non-linear** behavior.



Effect of cochlear compression: cochlear filter bandwidths increase with sound intensity level [Ruggero et al., 1997].

Comparison





Left: Theoretical behavior of β with respect to intensity level in dB (ref:max) by intervals of 5dB.
Right: Physiological measures (cat cochlea) [Verschooten et al., 2012]

Further directions/Conclusions

Conclusions

- Gabor filters achieve the most sparse patterns for Wigner-Ville distributions.
- ICA applied to speech produces filters similar to cochlear filters.
- Several acoustic features explain the fine-grained statistical structure of speech (but they are different for consonants and vowels).
- Level-dependent auditory filters may be part of an advanced efficient coding scheme.

Further directions

- Loosen the model (e.g. non-parametric estimation of $Q = f(f_c, I_{dB})$).
- Adapt the model so as to reflect time processing of inner ear.
- Asymmetry $\leftrightarrow$ enforce sparsity patterns.
- Still open question: is the frequency selectivity of humans' cochlea different from other mammals ? (and more adapted to speech ?)

Appendix: Redundancy reduction

Redundancy (Shannon/Barlow):

$$1 - \frac{H(Y)}{C}$$

where $H(Y) = -\mathbb{E}(\log p(y))$ is output entropy,
and C is the channel coding capacity.

Appendix: Redundancy reduction

Decomposition of redundancy

$$R = 1 - \frac{H(Y)}{C} = \frac{1}{C}\left(\sum_i H(Y_i) - H(Y)\right) - \frac{1}{C}\left(C - \sum_i H(Y_i)\right)$$

Two associated principles [Atick, 1992]:

- $(\sum_i H(Y_i) - H(Y))$: minimize mutual information between components $\rightarrow$ **Redundancy reduction, minimum-entropy codes**
- $(C - \sum_i H(Y_i))$: maximize information $\rightarrow$ **Infomax**

Appendix: Redundancy reduction

Decomposition of redundancy

$$R = 1 - \frac{H(Y)}{C} = \frac{1}{C}(\sum_i H(Y_i) - H(Y)) - \frac{1}{C}(C - \sum_i H(Y_i))$$

- $I(Y_1, \cdots, Y_m) = (\sum_i H(Y_i) - H(Y))$: minimize mutual information between components → Redundancy reduction, minimum-entropy codes
  Goal: find a set of independent features
- $(C - \sum_i H(Y_i))$: maximize information → Infomax
  also requires a set of independent features!
  [Nadal and Parga, 1994, Bell and Sejnowski, 1995]

Appendix: overcompleteness

In general, W is a rectangular matrix $n \times m$ with $m > n$.
What happens to the $-\log|\det W|$ penalty ?

- Every overcomplete dictionaries have correlated components.
- Minimum entropy/Sparseness gain importance from independence.
- Still, we want the dictionnaries to represent all directions of the space (e.g. filters uniformly distributed in time-freq-phase space)

Appendix: Overcompleteness

- **Solution 1**: enforce sparsity with reconstruction from a few filters

$$\min_{W,Y} ||X - W^{-T}Y||_2 + \gamma \sum ||Y||_1$$

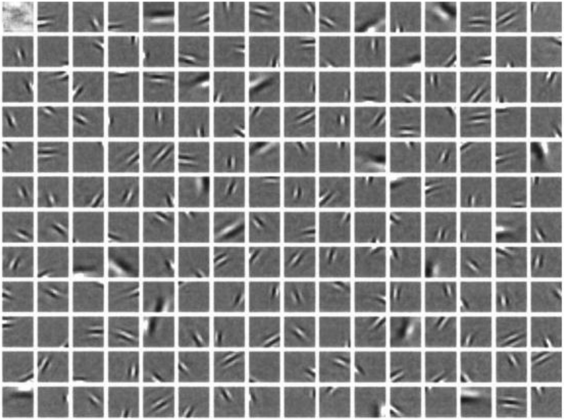
matching pursuit, sparse autoencoders...

- **Solution 2**: Use an appropriate family of dictionnaries and forget the penalty term

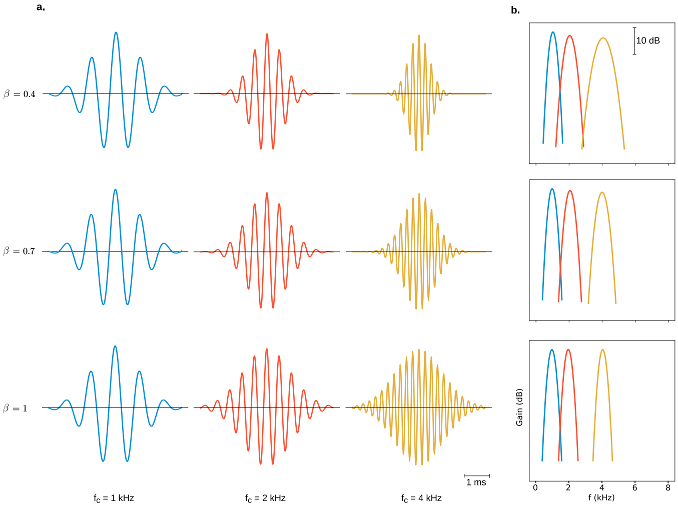
$$h = ||Y||_1$$

Appendix: Sparse coding and V1

A sparse coding algorithm on natural images produces filters that resemble receptive profiles of V1 [Olshausen and Field, 1996].



Appendix: Gabor dictionaries



a. Waveforms of several Gabor dictionnary atoms

b. Associated frequency responses

Appendix: Asymmetry

However: auditory filters are asymmetric. Asymmetric filters are also found with an algorithm of matching pursuit [Smith and Lewicki, 2006].

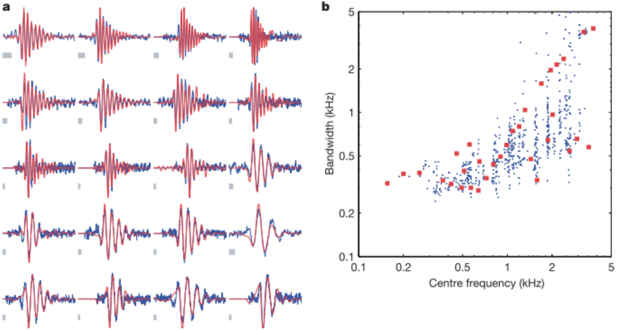


Figure 3 | Human speech is adapted to the mammalian cochlear code. a, As with the kernel functions trained on the natural sounds ensemble, the efficient code for speech consists of asymmetric sinusoids that closely match auditory revcor filters. **b,** The population of speech-trained kernels also matches the population centre bandwidth– frequency relationship of cochlear revcor filters. Details are as in Fig. 2.

References I

Atick, J. J. (1992).
Could information theory provide an ecological theory of sensory processing?
Network: Computation in Neural Systems, 3(2):213–251.

Barlow, H. (1961).
Possible principles underlying the transformations of sensory messages.
In Rosenblith, W. A., editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA.

Barlow, H. (1989).
Finding minimum entropy codes.
Neural Computation, 1(3):412–423.

Barlow, H. (2001).
Redundancy reduction revisited.
Network: Computation in Neural Systems, 12(3):241–253.

References II

Bell, A. J. and Sejnowski, T. J. (1995).
An information-maximisation approach to blind separation and blind deconvolution.
Neural Computation, 7(6):1129–1159.

Gazor, S. and Zhang, W. (2003).
Speech probability distribution.
IEEE Signal Processing Letters, 10(7):204–207.

Hsu, A., Woolley, S. M. N., Fremouw, T. E., and Theunissen, F. E. (2004).
Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons.
The Journal of neuroscience : the official journal of the Society for Neuroscience, 24(41):9201–11.

References III

Jutten, C. and Herault, J. (1988).
Une solution neuromimétique au problème de séparation de sources.
Traitement du signal, 5(6):389–403.

Lewicki, M. S. (2002).
Efficient coding of natural sounds.
Nature neuroscience, 5(4):356–363.

Lieb, E. H. (1990).
Integral bounds for radar ambiguity functions and Wigner distributions.
Journal of Mathematical Physics, pages 625–630.

Linsker, R. (1988).
Self-organization in a perceptual network.
Computer, 21(3):105–117.

References IV

Nadal, J.-P. and Parga, N. (1994).
Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer.
Network: Computation in neural systems, 5(4):565–581.

Olshausen, B. A. and Field, D. (1996).
Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images.
Nature, 381:607–609.

Olshausen, B. A. and Field, D. J. (2004).
Sparse coding of sensory inputs.
Current Opinion in Neurobiology, 14(4):481–487.

References V

Rieke, F., Bodnar, D. A., Bialek, W., and Bialek1, W. (1995).
Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents.
Proceedings of the Royal Society B: Biological Sciences, 262(1365):259–265.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997).
Basilar-membrane responses to tones at the base of the chinchilla cochlea.
The Journal of the Acoustical Society of America, 101(4):2151–2163.

Smith, E. C. and Lewicki, M. S. (2006).
Efficient auditory coding.
Nature, 439(7079):978–82.

References VI

Stilp, C. E. and Lewicki, M. S. (2013).
Statistical structure of speech sound classes is congruent with cochlear nucleus response properties.
In *Proceedings of Meetings on Acoustics 166ASA*, volume 20, page 050001.

Verschooten, E., Robles, L., Kovačić, D., and Joris, P. X. (2012).
Auditory nerve frequency tuning measured with forward-masked compound action potentials.
JARO - Journal of the Association for Research in Otolaryngology, 13(6):799–817.