Efficient Auditory Coding, Smith and Lewicki, Nature 439, 978-982, 2006

Rich Turner (turner@gatsby.ucl.ac.uk)

Gatsby Unit, 21/07/2006

Outline

- **GOAL: represent the complete acoustic waveform on a population spike code, and compare this representation to nature**

- **RESULTS:** Derive a code in which it turns out that **spikes** could represent the **temporal position** and **magnitudes** of **acoustic features**

- Optimise the features, finding they

  1. look like time domain cochlea filter estimates
  2. have a frequency-bandwidth dependence like real cells
  3. have a greater coding efficiency than conventional representations

Generative model

$$p\left[x(t)|\{\tau_{im}, s_{im}\}_{i,m=1}^{I,M}, \theta\right] = \text{Norm}\left[\sum_m \sum_i s_{im}\Phi_m(t-\tau_{im}), \sigma_x^2\right] \quad (1)$$

$$p[s|\theta] = \text{sparse} \quad (2)$$

$$p[\tau|\theta] = \text{sparse} \quad (3)$$

- Notation: $x(t)$ = waveform, $\Phi_m = m^{\text{th}}$ kernel, $\{\tau_{im}, s_{im}\}_{i,m=1}^{I,M}$ = temporal position and 'strength' of the i^{th} occurrence of basis function m.

- Generative model **does not correpond to a physical model of sound production** (same for GSMs)

- But it provides an **efficient representation**.

Learning and Inference

Hacky - zero temp EMish
Would like to **maximise the likelihood**:

$$\log p(x|\theta) = \int d\tau ds p(x|\tau, s, \theta) p(\tau, s|\theta) \tag{4}$$

... can do this if we can **iteratively update the free-energy**

$$\begin{aligned} F[q(s, \tau|x), \theta] &= \log p(x|\theta) - KL[q(\tau, s|x, \theta)||p(\tau, s|x, \theta)] \qquad (5) \\ &= \langle \log p(x, \tau, s|\theta) \rangle_{q(\tau, s|x, \theta)} - H[q(\tau, s|x, \theta)] \qquad (6) \end{aligned}$$

But $p(\tau, s|x, \theta)$ is intractable - have to make some approximations:

$$q(\tau, s|x, \theta) = \delta(\tau - \tau_0)\delta(s - s_0) \tag{7}$$

Inference E-Step: Matching pursuit

- Need to set: τ_0 s_0 to modes of the posterior $q(\tau, s|x, \theta)$, **find these approximately using matching pursuit**:
- Decompose waveform as a projection $\langle x(t), \Phi_m(t-\tau)\rangle = s_m$ and residual $R_x(t)$:

$$x(t) = \langle x(t), \Phi_m(t-\tau)\rangle \Phi_m(t-\tau) + R_x(t) \tag{8}$$

- Find the largest projection: $\arg\max_{\tau,m} \langle x(t), \Phi_m(t-\tau)\rangle$
- Note s_m and τ_m
- Repeat, treating residual as new waveform: $\arg\max_{\tau,m} \langle R_x(t), \Phi_m(t-\tau)\rangle$
- Repeat, stopping when the largest projection is under a threshold.

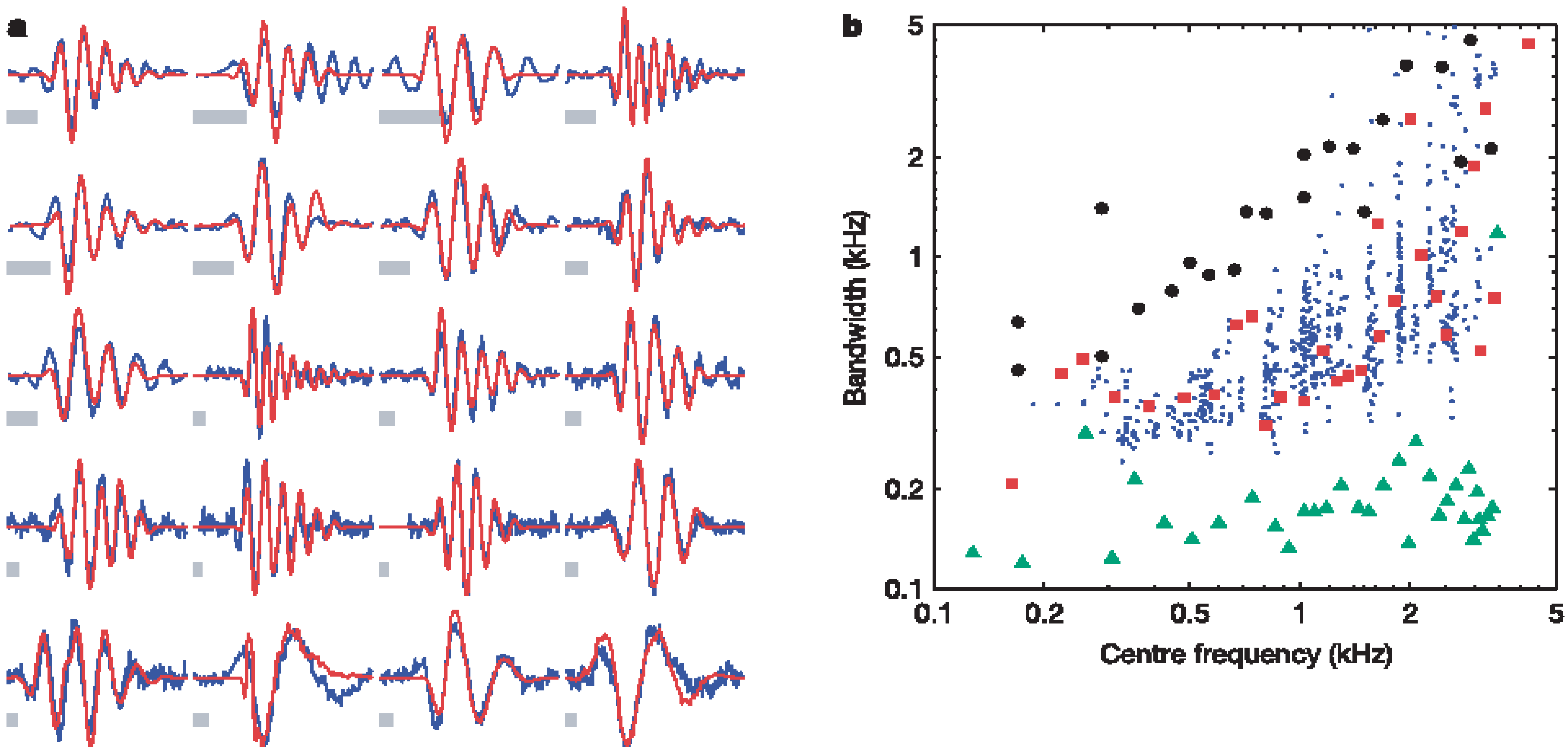
Inference M-Step

$$\arg\max_{\Phi} \log p[x(t), \tau_{MAP}, s_{MAP}|\theta] \tag{9}$$

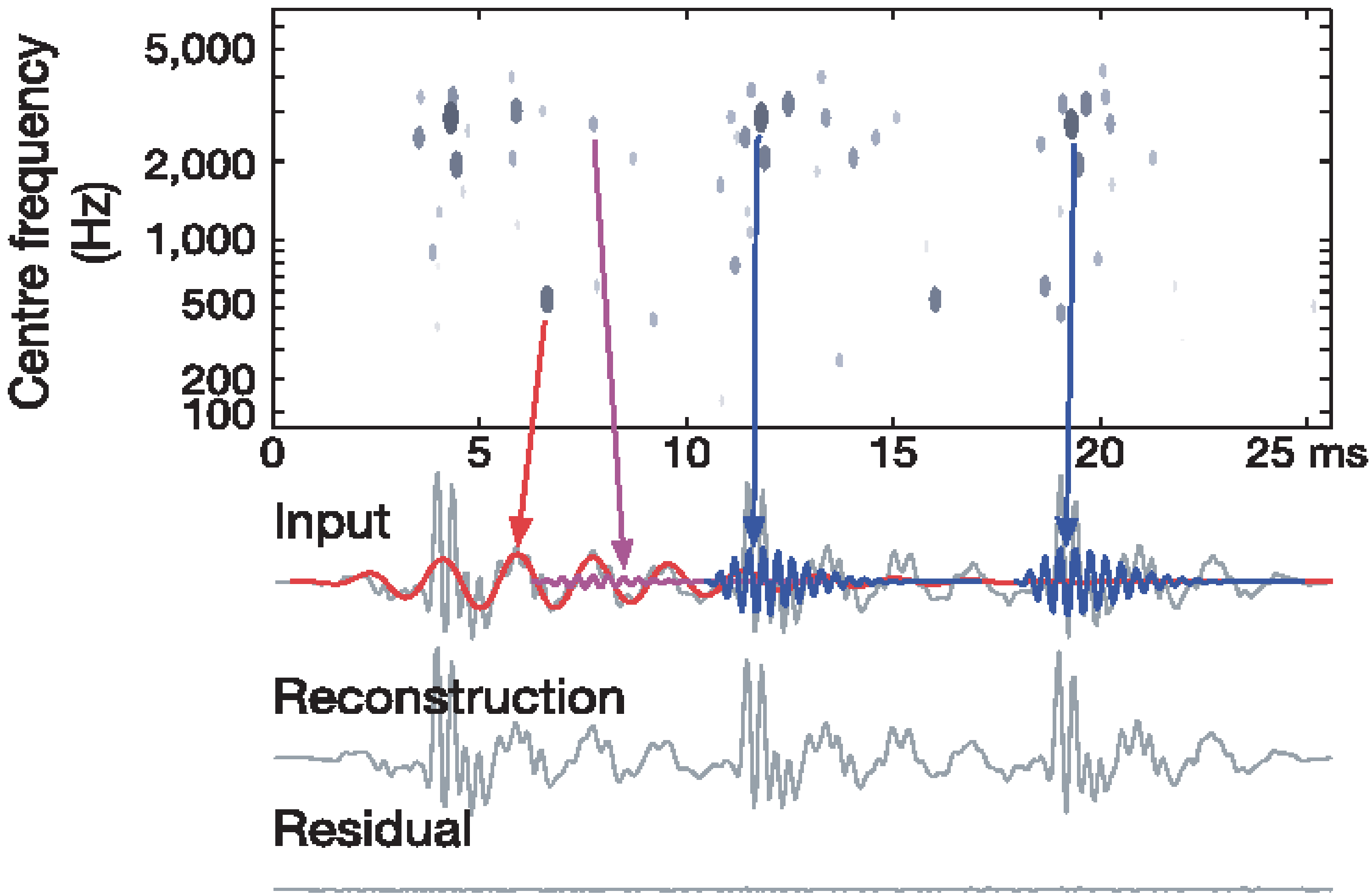
- $\Phi_m =$ a vector of length L_m
- **Optimise each element of Φ and the length**
- Gradient based approach, where more elements are added if zero-padding starts to become non-zero

Run the algorithm on different databases of 'natural' sounds...

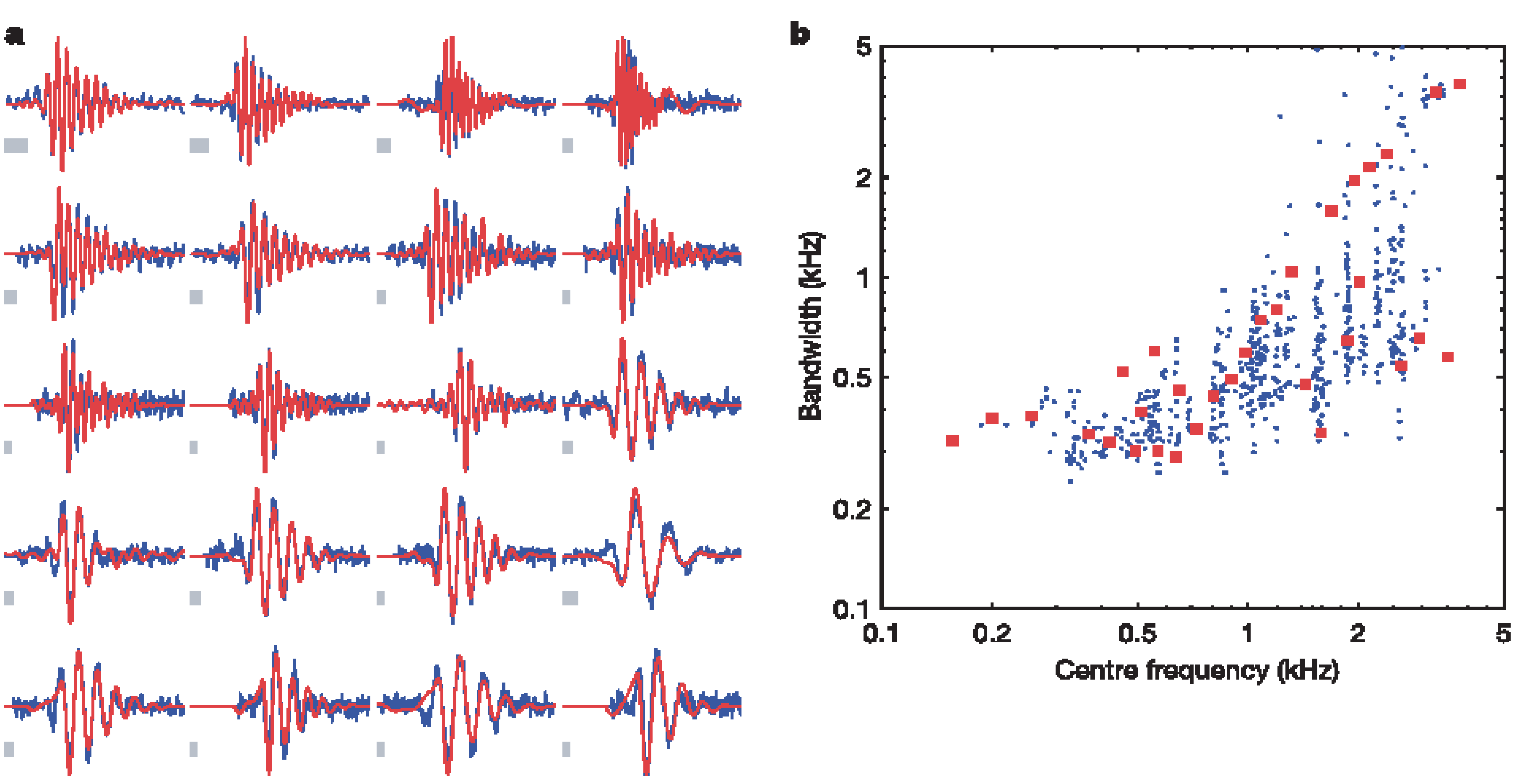
Results



a
b
5
2
1
0.5
0.2
0.1
0.1
0.2
0.5
1
2
5
Bandwidth (kHz)
Centre frequency (kHz)



Centre frequency (Hz)
5,000
2,000
1,000
500
200
100
0
5
10
15
20
25 ms
Input
Reconstruction
Residual


a
b
5
2
1
0.5
0.2
0.1
0.1
0.2
0.5
1
2
5
Bandwidth (kHz)
Centre frequency (kHz)

Results - summary

- $\{s_{MAP}, \tau_{MAP}\}$ are localised, discrete, sparse: **SPIKE LIKE**
- The kernels Φ_m are very like REVCOR filter shapes of the auditory nerve
- Training on a mixture of animal vocalisations (harmonic) and environmental sounds (transient) results in similar bandwith-centre frequency tiling.
- Similar results for a speech corpus show the same, indicating that speech might be optimised for the mammalian cochlear code