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

Gatsby Unit, 21/07/2006

Rich T.
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

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

\[
p[s|\theta] = \text{sparse} \tag{2}
\]

\[
p[\tau|\theta] = \text{sparse} \tag{3}
\]

• Notation: \(x(t)\) = waveform, \(\Phi_m = m^{th}\) 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 correspond 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**:

\[
\log p(x|\theta) = \int d\tau ds p(x|\tau, s, \theta) p(\tau, s|\theta)
\] (4)

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

\[
F[q(s, \tau|x), \theta] = \log p(x|\theta) - KL[q(\tau, s|x, \theta)||p(\tau, s|x, \theta)]
\] (5)

\[
= \langle \log p(x, \tau, s|\theta) \rangle_{q(\tau, s|x, \theta)} - H[q(\tau, s|x, \theta)]
\] (6)

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

\[
q(\tau, s|x, \theta) = \delta(\tau - \tau_0)\delta(s - s_0)
\] (7)

Rich T.
Inference E-Step: Matching pursuit

- Need to set: $\tau_0, s_0$ 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 $\tau_m$

- 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

\[
\arg\max_{\Phi} \log p[x(t), \tau_{MAP}, s_{MAP}|\theta]
\]  \hspace{1cm} (9)

- \(\Phi_m\) = a vector of length \(L_m\)

- Optimise each element of \(\Phi\) 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
Results - summary

- \{s_{MAP}, \tau_{MAP}\} are localised, discrete, sparse: SPIKE LIKE

- The kernels \(\Phi_m\) 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 bandwidth-centre frequency tiling.

- Similar results for a speech corpus show the same, indicating that speech might be optimised for the mammalian cochlear code