Questions

1. (a) In vision, the optical blur that occurs at an image location depends on the accommodation of the lens and on the distance to the point. The blurred image that results is the convolution of the sharp image with the impulse response function (due to the blur).

What is the analogy in audition?

(b) In vision, it sometimes happens that you see two image superimposed on each other. This can happen, for example, when you have a dirty glass surface and you see both the glass surface as well as the image transmitted through the glass.

The visual system sometimes can perceive the two images superimposed images separately, for example, by converging the eyes to bring the glass to zero disparity. What is the analogy in audition?

2. Describe the sequence of transformations of a voiced sound, starting from the glottal pulse train, all the way to its coding in the auditory nerve (leaving the cochlea on its way to the brain).

3. Suppose a ganglion cell in inner ear is tuned to high frequency, say 4 kHz. Suppose the sound source is dominated by a pure tone of frequency 1 kHz. Describe the response (if any) of this ganglion cell.

4. Suppose a ganglion cell in the cochlea can produce spikes at a maximum rate of 500 per second.

Describe the autocorrelation function of the cell’s response to a pure 3 kHz tone. (Assume that the cell has a non-negligible sensitivity to this frequency.)

Hint: the auto-correlation of a signal is defined is the cross-correlation of the signal with itself. (Recall from the linear systems notes (p. 4) how the cross-correlation is defined.)

Hint: For a signal $s(t)$, the autocorrelation function is:

$$ R(\tau) = \sum_{t=0}^{T-1} s(t) \, s(t + \tau) $$

where $s(t)$ is discretized time, with values 0 or 1 depending on whether or not there is a spike in bin $t$. 
Solutions

1. (a) The analogy in audition is that the sound pressure signal that arrives at the ear drum when there is a source from a single direction \((\phi, \theta)\) is the temporal convolution of the source sound \(I(t)\) with the HRIR function \(h(t, \phi, \theta)\) for that direction.

   In the case of vision, the eye doesn’t measure the original sharp image. In the case of audition, the ear doesn’t measure the original source signal.

   (b) In audition, you may have two sounds coming from different directions. These sounds will be superimposed.

   To perceptually separate the sounds, it sometimes helps to reorient the head. This can amplify some sounds and attenuate others. (When I am at a very loud event such as a poster session at a conference, I usually turn my head slightly away from the poster presenter and plug the ear facing away from him/her).

2. Each glottal pulse is transformed by the articulators (jaw, tongue, oral cavity etc) which is modelled by convolving with some \(a(t)\). The sound leaves the mouth and nose and travels through space where it attenuated in strength with distance. The head and ear of the listener then reshape the sound further. This is modelled by convolving with some \(h(t)\), the HRIR. The sound wave then contributes to the oscilations of the basilar membrane (gammatone filter). Assuming this is the only contribution, the various mechanical frequency components of the BM vibrations get transduced into chemical signals (neurotransmitters released). The times of the peaks of these signals are coded by phase locked volleys of spiking ganglion cells, at least for sounds up to 2 kHz.

3. The ganglion cell has relatively weak sensitivity to a 1 kHz sound, but if that sound is loud enough then the 4 kHz sensitive cell will respond. It just won’t spike with nearly as great a frequency as it would if the sound had a 4 kHz component of the same intensity.

   An interesting question is whether the 4 kHz cell would respond in synch with the peaks of the 1 kHz sound. I am not sure.

4. The autocorrelation function has a sharp peak at \(\tau = 0\). Then it has a value of 0 up to \(t = 2\) ms, because of the maximum spiking frequency constraint, that is, two spikes cannot occur less than 2 ms apart.

   Then, because of phase locking, it has a set of small peaks that are at \(t = \frac{i}{3}\) ms (i.e. corresponding to 3 kHz), but only for \(i > 6\) because of the maximum spiking constraint.