# Questions

1. Describe the sequence of transformations of a voiced sound that is spoken by one person and heard by another. Start from the glottal pulse train, and go all the way to its coding in the cochlear nerve.

2. 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. What do you think would be the response (if any) of this ganglion cell?

3. The *auto-correlation* of a signal is defined is the cross-correlation of the signal with itself. For a spike train $s(t)$, 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 $t$.

   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 1 kHz tone, assuming that the cell is sensitive to this frequency.

4. The LSO computes the level differences. We saw that these are computed on both sides of the head, namely the left LSO computes the level difference 'left - right' and the right LSO computes the difference 'right - left'. But isn't this redundant?

5. The maximum time it takes sound to travel the width of a typical head (17 cm) is about $\frac{1}{2}$ ms . For a 2 kHz sound, this distance is one wavelength.

   Cells in the MSO (medial superior olive) are believed to be sensitive to precise timing differences between spike arrivals from the two ears, in particular, for lower frequency sounds. The most fundamental reason for this is that spike timing has limited precision and the low frequency cutoff reflects that limit.

   But suppose that the spike timing precision were *not* a fundamental limitation. Is there any other reason why spike timing information would be less useful for frequencies greater than 2 kHz?

# Solutions

1. 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 causes oscillations of the basilar membrane. Different positions on the BM oscillate based on the amplitude of the signal in a range of frequencies. The various mechanical frequency components of the BM vibrations get transduced into chemical signals (neurotransmitters released by hair cells). The times of the peaks of these signals are coded by phase locked volleys of spiking ganglion cells up to 2 kH and the number of spikes increases with the amplitude of the signal in that frequency band. Beyond 2 kHz, phase locking is not possible and only the amplitude is encoded.

2. The 4 kHz tuned ganglion cell would have a relatively weak sensitivity to a 1 kHz sound since the 1 kHz sound is two octaves away from the peak tuning of the 4 kHz cell. But if a 1 kHz sound were very loud and if there was no sound component near the 4 kHz then the loud 1 kHz component could cause the BM to vibrate a bit (at frequency 1 kHz) at a 4 kHz location, and this could cause the hair cells at the 4 kHz location to release neurotransmitters following the 1 kHz vibration, which could cause the 4 kHz ganglion cell to respond (send spikes) and the spikes could be phase locked to the 1 kHz component. Notice that the 1 kHz sound would have to be very loud to cause the BM at 4 kHz to vibrate even a little bit, and so the 'phase locked' response would be very sparse, i.e. there would be relatively few of the 1 kHz peaks that gets marked with a spike, and so the firing reate would be far far less than 1 kHz. Indeed the maximum firing rate of *any* ganglion cell is 500 per second, so in the case we are considering here, the firing rate of the 4 kHz cell to the very loud 1 kHz sound might be say 20 spikes per second. (I am just pulling these numbers out of the air. The point is that the firing rate would be much lower than the peak of 500 per second.) Note that these spikes that do occur could still be phase locked to the 1 kHz peaks.

   Also notice that this example assumes that there is no sound component near 4 kHz since if there were such a component, then the 4 kHz cell would be much more sensitive to this component and would respond with a rate of several hundred spikes per second, but these spikes would not be phase locked to the 4 kHz component since this is too fast to be coded by the hair cell neurotransmitter release. At 4 kHz vibration, only the amplitude (envelope) of the sound over some time duration can be to coded.

3. The autocorrelation function would have a sharp peak at $\tau = 0$ since if $s(t)$ has a spike at some $t$ then by definition $s(t+\tau)$ will have a spike at exactly $t$ also when $\tau = 0$. But $s(t+\tau)$ will not have a spike at any time $|\tau| < 2$ ms since the cell can spike at most at 500 times per second. So the autocorrelation function will be 0 for $|\tau| < 2ms$.

   Because the cell will phase lock to the 1 kHz sound, the autocorrelation function will have other sharp peaks at $\tau = i$ ms (i.e. corresponding to peaks in the basilar membrane motion at a period of 1 ms, corresponding to the 1 kHz frequency of the band). The time between spikes must be at least 2 ms. So the peaks will be at $\tau = i$ ms where $i \geq 2$. That is, *the autocorrelation function will be missing the peak at $\tau = 1$ ms.*

4. You can think of the left side as computing the level difference $\log \frac{\sum_t I_l^j(t)^2}{\sum_t I_r^j(t)^2}$ and the right side as computing the level difference $\log \frac{\sum_t I_r^j(t)^2}{\sum_t I_l^j(t)^2}$, where the summations are over some small time interval, and for frequency band $j$. Mathematically these are redundant since one is just the negative of the other. However, neurons cannot code negative values. So both of these quantities need to be computed.

   This is analogous to the way positive and negative values are coded retinal ganglion cells, namely ON-center OFF-surround DOG cells and also OFF-center ON-surround cells.

5. The issue I was hoping you would identify is that the head size defines a bound on the timing difference between the ears, and for high frequency components (small wavelength components), if the wavelength is less than the head then there may be other ambiguities introduced.

   Take a sound component of frequency $\omega > 2000$ Hz. The basilar membrane with that center frequency component will vibrate such it has peaks every $\frac{1}{\omega}$ seconds where $\frac{1}{\omega} < 0.5$ ms. Suppose some subset of these peaks would be accurately encoded as spikes in the spike trains coming from the left and right ears, i.e. suppose precision wasn't limited. Let the *true* delay in arrival times (say left - right) of any particular wave of the sound be $\Delta t$ for some source direction. Then there would be more than one possible *estimated* difference in spike arrival times for this source direction, namely any $\Delta t + \frac{j}{\omega}$ would be a valid estimate as long as $|\Delta t + \frac{j}{\omega}| < 0.5$ ms, since the maximum delay is 0.5 due to maximum arrival time difference between the ears. In terms of the coincidence detector model of Jeffress (1948), there would be peaks detected for each of these candidate $j$'s.

   For example, suppose the source were in the medial plane so that the true timing different was 0. Then neighboring peaks in the BM would produce spike arrival times in the MSO that differed by $\frac{1}{\omega}$, which would be less than 0.5 ms in magnitude which would be consistent with a sound that is not in the medial plane.

   That said, the above ambiguity might not be a problem. If the sound has a sudden onset (impact), then you might reduce this problem by matching the "first spike" from the left or right ears. This is not *guarenteed to work*, since the ganglion cells in the cochlea don't spike on every peak of the basilar membrane's motion, so the first spike of a ganglion cell might not corresponds to the first peak of the basilar membrane. That said, there are many high frequency bands and each will have different ambiguities, but all these frequency bands will share the same *true* timing difference. So the ambiguities could be reduced to some extent.

   Bottom line: presumably the 2 kHz limit is a combination of several factors, including the limited spike timing precision, but also the factors just discussed.