Inner Ear (continued from last lecture)

At the end of last lecture, I discussed how the basilar membrane vibrates in response to the sound pressure signal that has been transduced from the air to the fluid inside the cochlea. Today we will examine how these vibrations of the basilar membrane are encoded by the nervous system. Much is known about the detailed anatomy here but we will skip most of the days. We will consider a very simplified model that gives us enough to understand the sequence and location of events, and to describe a computational model of what is happening.

Basilar Membrane and the Tonotopic Map

As mentioned last lecture, difference positions on the basilar membrane move up and down at different peak frequencies with low frequencies at the far end (apex) and high frequencies at the near end (base). In this way the basilar membrane defines a *tonotopic map* with different positions on the BM coding different frequencies of the underlying sound.

The coding is not done by the basilar membrane but rather by sensory nerve cells along the membraine. These nerve cells include both hair cells (which don't spike) and ganglion cells (which do spike). This is analogous to the retina, where the photoreceptors give a continuous response to the signals from the environment and the ganglion cells give spike responses that are sent to the brain. In the cochlea, when fibres of the basilar membrane vibrate at some location, the hair cells and ganglion cells at that location respond in turn. Let's look at the neural coding of sounds in the cochlea in a bit more detail.

The hair cells on the basilar membrane are analogous to the photoreceptors of the eye. The hair cells response to mechanical stimulation by releasing neurotransmitters. Think of these cells as riding the basilar membrane at some location - up, down, up, down. This motion and stretching of the hair cell body releases neurotransmitters (temporary opening of the cell membrane) at the same temporal frequency of this wave. Think of the transmitters being released at the top of the BM wave.

The ganglion cells along the basilar membrane respond to the neurotransmitters that are released by the hair cells. Importantly, the ganglion cells are capable of precise temporal responses, and so if the transmitter level has precise temporal structure then so will the ganglion cells. As I will describe nexxt, hair cells and hence ganglion cells can have detailed timing structure up to about 2 kiloHerz.

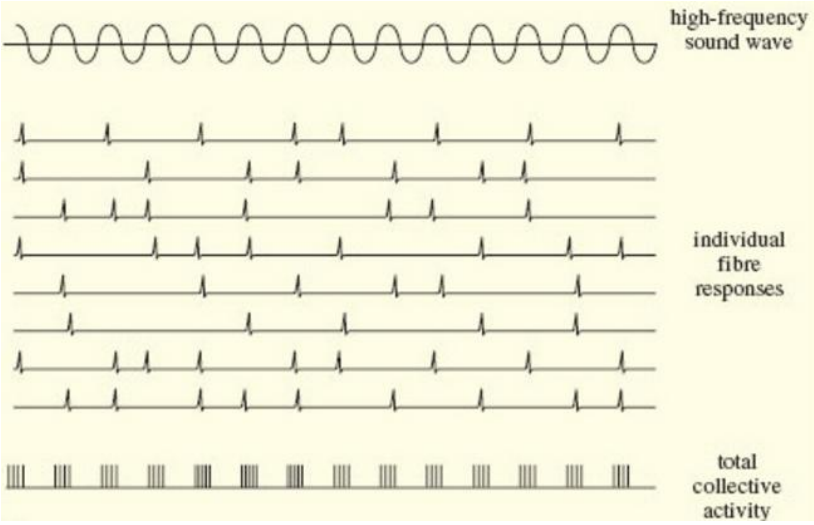
Phase locking and volley code

For each of the two cochleas (left and right), there are only about 3,000 hair cells over the entire basilar membrane, and about 30,000 ganglion cells. So think of each hair cell mapping to about 10 ganglion cells. The reason for this 1-to-many mapping is that the ganglion cells cannot spike at rates of more than a few hundred spikes per second. So, in order to code the exact times of the peak amplitude of the basilar membrane at some position of the BM when the BM peak frequency at that location is more than a few hundred Hz (but less than a few thousand Hz), many ganglion cells are needed at that location.

The spikes for any one ganglion cell thus occur at a subset of the peaks of the basilar membrane (or equivalently, at a subset of the peaks of the hair cell neurotransmitter release). We say that the ganglion cell spikes are *phase locked* with the peaks of the basilar membrane motion at that

location. By having say 10 ganglion cells for each hair cell, this *volley code* [1] allows the group of (say 10) ganglion cells associated with each hair cell to represent the spikes. See the illustration below.

You might ask: If the location on the basilar membrane determines the approximate frequency and if the cell spikes are locked to that frequency, then what information is communicated by the spikes? There are two answers to this, and they are related. First, the exact timing (phase) is important, in particular, for combining the left and right ear signals. Second, when the amplitude of response of the basilar membrane at any position is larger, the probability of any particular ganglion cell at that location having a spike at the peak is also larger. This is important because the reliability of the timing information in the spikes increases when there are more spikes. Also, the amplitude (loudness) information itself is important – as we'll discuss later today.



Phase locking only occurs up to a few kHz. At higher frequencies than that, the exact timing of the neurotransmitter release by the hair cells cannot follow the BM motion exactly. Instead, the amount of neurotransmitter released depends simply on the amplitude of BM motion at that location. This amplitude information is still important, even in source localization since it can be compared between the two ears and this can give information about source direction – as we'll see below.

Auditory Pathway

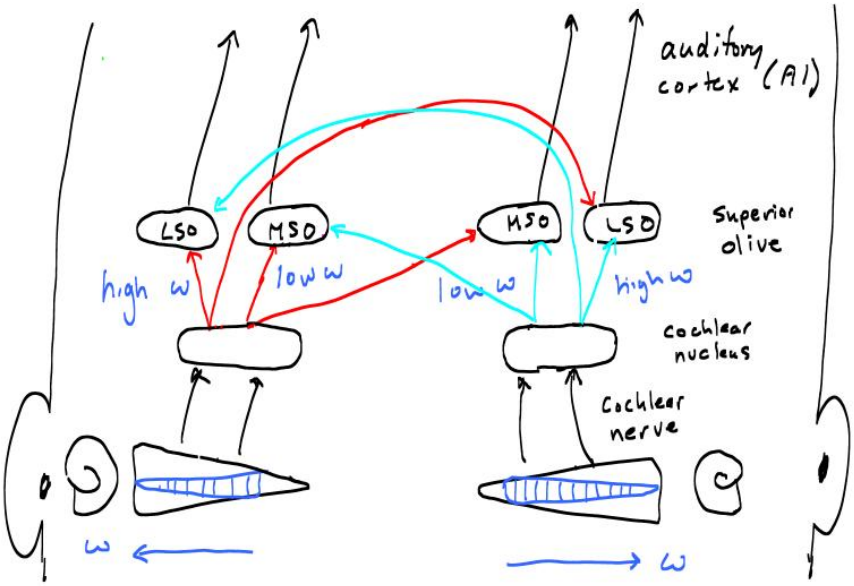
The axons from the ganglion cells in the cochlea are bundled together into the *auditory nerve* (or *cochlear nerve*) which carries spike trains from the cochlea to the brain. (The nerve is often called the vestibulocochlear nerve, since it also carries information from the vestibular body.) The auditory nerve from the ear is analogous to the optic nerve from the eye, which carries the spikes from the retinal ganglion cells to the LGN.

The nerve carries the spike trains from the left and right cochlea to the *cochlear nucleus* (CN) which is in an old part of the brain, specifically in the brainstem. `https://en.wikipedia.org/`

[1] analogy `https://en.wikipedia.org/wiki/Volley_fire`

`wiki/Brainstem`. The mapping is also *tonotopic* namely fibres are arranged spatially according to temporal frequency, just as cells on the basilar membrane are tonotopic and arranged according to temporal frequency.

The cells in the cochlear nucleus then send axons either to the MSO (medial superior olive) or LSO (lateral superior olive) on each side of the brain. "Medial" here means closer to the middle of the brain, and "lateral" means away from the middle of the brain. The MSO receives the low frequency signals and the LSO gets the high frequency signals. The cells in each MSO and LSO receive inputs from both ears, and indeed this is the site in the brain where inputs from the two ears are first combined. Note that, unlike in the visual system where left and right eye images are first combined in the cortex, in the auditory system the left and right ear signals are combined in the brainstem prior to the cortex.



Duplex theory

It is easy to get lost in the names of body parts and so we would like to step back and remind ourselves of a particular computational problem being solved here, namely source localization.[2] Low and high frequency sounds provide different information for solving this problem. Low frequencies carry information from timing differences (delays between the two ears) but not level differences, which are negligible because wavelengths bigger than the size of the head do not undergo significant shadowing and reflection effects. High frequency sounds do carry information about level differences since shadowing of the head and reflections and refractions of the sound wave from the pinna and auditory canal are significantly different between the ears.

Because they carry different information, low and high frequencies are separated by the auditory system and processed separately. As mentioned above, the LSO receives the high frequency

[2]This pathway carries the signals for solving many computuational problems including recognition e.g. speech, music. But these are topics for a different course.

components and computes the level differences between left and right ears. Cells in the LSO are excited by inputs from the CN on the same side of the head and are inhibited by inputs from the opposite side of the head. If the input levels are the same from the two sides, then there is no net excitation or inhibition of an LSO cell. If the input level is greater in the left than the right for some frequency band, then the LSO cells on the left side that encode those frequencies will respond, but the LSO cells on the right side will not (since a cell cannot have a negative response). Similarly, the input level is greater in the right than the left for some frequency band, then the LSO cells on the right side will respond, but the LSO cells on the left side will not.

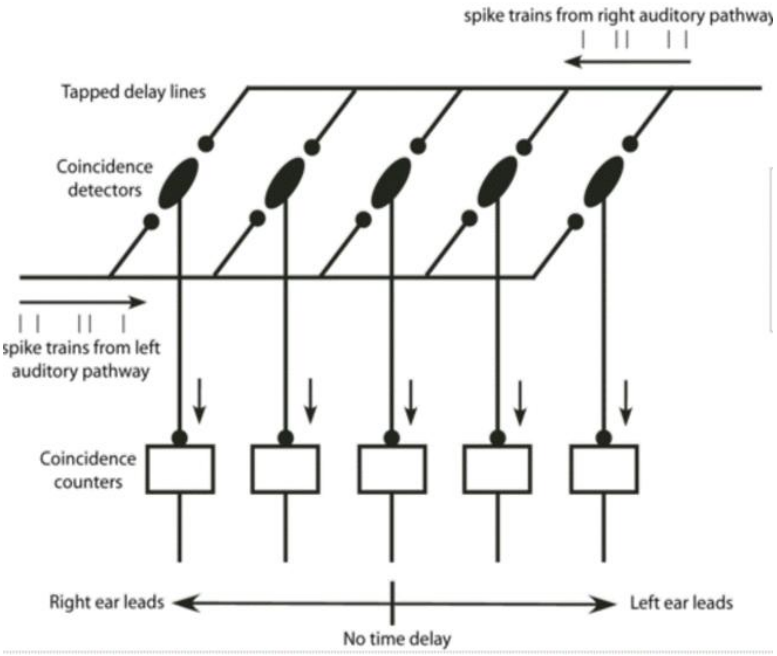
The MSO receives low frequency inputs from the CNs of both sides, and both inputs are excitatory. The MSO compute timing differences but it isn't clear exactly how this is done, and a few different theories have been proposed. The best known theory was proposed by Jeffress (1948) and has become known as the *Jeffress* model.

`http://www.scholarpedia.org/article/Jeffress_model`

Jeffress did not know about the MSO, and it is still controversial whether Jeffress'es model describes the MSO's mechanism for comparing timing in the two ears. There is evidence both for it and against it, and it seems to depend on the animal species e.g. bird versus mammal.

The main idea of the Jeffress model is that there are cells ("coincidence detectors") that each receive input from the same bandpass signal from the two ears, such that the inputs arrive on lines of different lengths. The different lengths give rise to different delays in the signals. The length differences are hardwired, and so each 'coincidence detector' cell in the MSO has a preferred timing difference for arrival in the two ears. To visualize this model, see here:

`https://auditoryneuroscience.com/topics/jeffress-model-animation`.



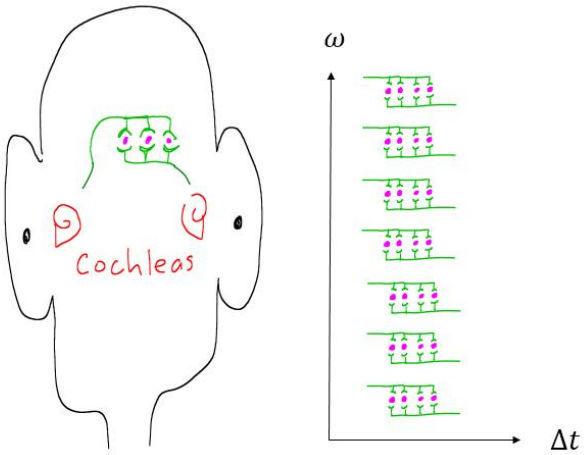
For any sound source in some direction in space and for any frequency band, one of the coincidence detectors for that band will have the greatest response. This greatest response will occur when the signals from the two ears arrive at the coincidence detector at the same time. How exactly this 'coincidence' of arrivals gives rise to the largest response is unspecified by the model, but one obvious scheme is just to add the signals together and look for a sharp peak. This model of adding the signals would be analogous to our model of binocular disparity sensitive cells in vision. (Recall Assignment 1.)

Note that different frequency bands each estimate the delay in arrival times. This provides multiple estimates of the delays. The multiple delay lines for different frequency bands is sketched below.

To get a sense of the time and space scales involved here, consider two binaural cells that sit next to each other in the MSO (coincidence detectors). If corresponding spikes from left and right ear arrive at one of these cells (A) at the same time, then how much of a time difference is required for the cells to arrive at the neighboring cell (B) ? Suppose these cells are $\frac{1}{10}$ mm apart and the spikes travel at a speed of 10 m/s on an axon. Then using $t = d/v$, the signal take $\frac{1}{100}$ ms to travel the distance between the cells. For both signals to arrive at B at the same time instead of A, the sound would need to arrive at the left ear $\frac{1}{100}$ ms earlier and it would need to arrive at the right ear $\frac{1}{100}$ ms later.

This difference in the arrival times corresponds roughly to the difference for a sound in the medial plane versus a sound come from a cone a few degrees away from the medial plane. Amazingly, this is roughly the human sensitivity (threshold, also called "just noticable difference" JND) to sound source azimuth direction in the neighborhood of azimuth $\theta = 0$ degrees and elevation $\phi = 0$ degrees i.e. the straight ahead direction.

It is easy to be skeptical that the auditory system is capable of such high precision. To understand how this is achieved, one should keep in mind that there are many frequency bands and cells involved in this computation. The auditory system doesn't just rely on one cell to do this.[3]



[3] This phenomenon that the performance of a sensory system can be much more precise than its elements is called *hyperacuity*. Examples of visual hyperacuity are well known e.g. Vernier acuity.

Computational model revisited

Recall the timing and level differences were represented by τ and α in the model from last lecture. We set up the problem as one of minimizing the sum of squared differences between one ear's sound and a shifted and scaled sound in the other ear. We found the time delay τ that maximizes the cross correlation

$$\sum_{t=1}^{T} I_l(t)\ I_r(t-\tau)$$

and we solved for α using:

$$\alpha^2 \approx \frac{\sum_{t=1}^{T} I_l(t)^2}{\sum_{t=1}^{T} I_r(t)^2}$$

We now know that sounds are filtered by each ear and so rather than comparing level and timing differences of I_l and I_r, we do these comparisons within each bandpass channel I_l^j and I_r^j. We can find τ_j that maximizes the cross correlation

$$\sum_{t=1}^{T} I_l^j(t)\ I_r^j(t-\tau).$$

For simplicity, let's just assume that the actual timing differences are the same in each frequency band, namely there is a delay between ears that is due to the cone of confusion geometry.[4] In this case, we can estimate τ by combining estimates for τ from the different channels j.

What about level differences? We can estimate the α_j^2 for band j and over some short time interval T by:

$$10 \log_{10} \frac{\sum_{t=1}^{T} I_l^j(t)^2}{\sum_{t=1}^{T} I_r^j(t)^2}$$

which is in dB units. But these level differences for each band will depend on the HRTF and on the source source. How can these two factors be disentangled?

Here is the idea. The signals in band j in the left and right ear are:

$$I_l^j(t;\phi,\theta) = g^j(t) * h_r(t;\phi,\theta) * I_{src}(t;\phi,\theta).$$

$$I_r^j(t;\phi,\theta) = g^j(t) * h_r(t;\phi,\theta) * I_{src}(t;\phi,\theta).$$

Now, use the convolution theorem, and take the Fourier transform of each of the above over some time interval with T samples. Then take the ratio:

$$\frac{\hat{I}_l^j(\omega;\phi,\theta)}{\hat{I}_r^j(\omega;\phi,\theta)} = \frac{\hat{g}^j(\omega)\ \hat{h}_l(\omega;\phi,\theta)\ \hat{I}_{src}(\omega;\phi,\theta)}{\hat{g}^j(\omega)\ \hat{h}_r(\omega;\phi,\theta)\ \hat{I}_{src}(\omega;\phi,\theta)}$$

Cancelling $\hat{g}$ and $\hat{I}_{src}$ terms on the right side (which we can only do when they are non-zero, so this is an assumption) and taking the absolute values gives:

$$\frac{|\hat{I}_l^j(\omega;\phi,\theta)|}{|\hat{I}_r^j(\omega;\phi,\theta)|} = \frac{|\hat{h}_l(\omega;\phi,\theta)\ |}{|\hat{h}_r(\omega;\phi,\theta)\ |}$$

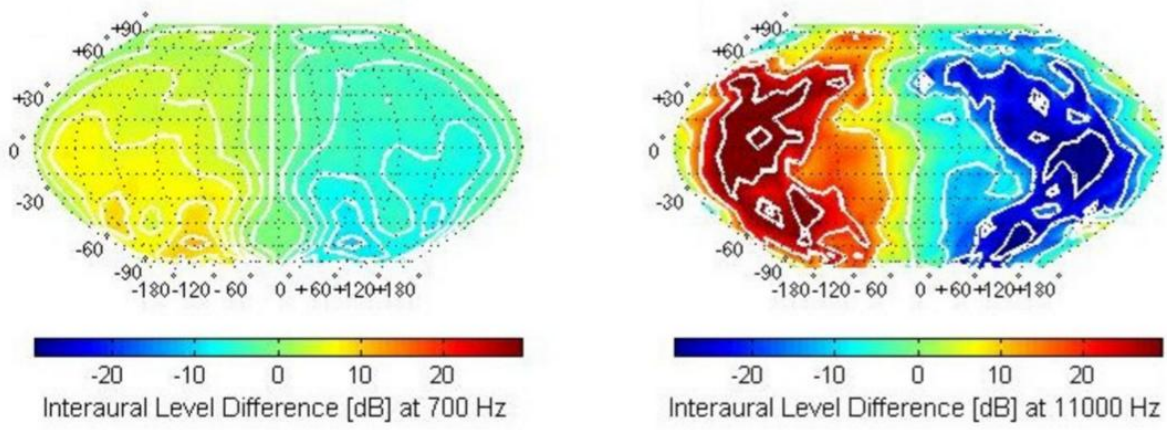
[4] This assumption is an approximation only. Recall the HRIR function from last lecture, where in the medial plane there was some variability in the HRIR over elevation angles ϕ. This suggests that there also *would be* some timing difference in the filtered signals $I_l^j(t,\phi,\theta=0)$ versus $I_r^j(t,\phi,\theta=0)$ within any band j and for any fixed elevation ϕ.

Thus we see that the ratio of the ampitudes of the filtered sound at each frequency only depends on the ratio of the HRTF's at that frequency.

Using a mathematical result known as Parseval's theorem[5] and assuming that, for any band j and for any source direction (θ, ϕ) the HRTFs $\hat{h}_l^j(\omega, \theta, \phi)$ and $\hat{h}_r^j(\omega, \theta, \phi)$ are smooth enough that we can treat them as approximately constant over the frequencies ω *within the band* j, the following approximately holds:

$$\frac{\sum_{t=1}^{T} I_l^j(t; \phi, \theta)^2}{\sum_{t=1}^{T} I_r^j(t; \phi, \theta)^2} = \frac{\hat{h}_l^j(\theta, \phi)^2}{\hat{h}_l^j(\theta, \phi)^2}$$

Taking the $\log_{10}$ of both sides gives that the level differences in band j measured in dB are approximately the same as the level differences in the HRTFs measured in dB.



Thus, to use this information about the level differences in the sound to estimate the source direction (ϕ, θ), the auditory system would need to know how the dB difference of the HRTFs for band j vary as a function of (ϕ, θ). The idea is that *for given value of the dB difference of the HRTFs for band* j*, there would be only a subset of directions* (ϕ, θ) *that such that a source from these directions would produce that level difference.* So for band j, knowing the level difference of the sound in the left and right ear would narrow down the possible source directions. *Combining the contraints from different bands would narrow it down further.*

The figure above is from `https://auditoryneuroscience.com/topics/acoustic-cues-sound-location`. It shows the level differences (left ear - right ear) as a function of (θ, ϕ) for two frequencies: $\omega = 700$ Hz is shown on the left and $11,000$ Hz is shown on the right. These data were obtained by measuring the sound reaching the inside of the ears of a subject, when the sound comes from all different directions (θ, ϕ). Some iso-value (constant value of HRTF) curves are shown. For each (θ, ϕ) direction shown, if we assume that this map is roughly constant over ω within a band j – namely for frequencies near 700 Hz and 11,000 Hz respectively, then we can treat this map as the HRTF differences mentioned above.

[5]Parseval's theorem just says that the Fourier transform is a rotation in an n-D space, and single scaling, and so the L2 norm of a signal is equal to the L2 norm of the Fourier transform of the signal, times a scale factor

Note that this map is different than the HRTF maps shown last lecture. If we think of a function HRTF of variables (ω, θ, ϕ), then the plots above are for ω fixed, whereas the plots last lecture were for θ fixed and ω, ϕ varying, or ϕ fixed and ω, θ varying.

Monaural cues

Our emphasis has been on binaural hearing. However, there is available monoaural information about the direction of the source as well, and people do use it. But how? Consider the Fourier transform of a short duration sound heard in one ear:

$$\hat{I}(\omega) \;=\; \hat{g}(\omega)\; \hat{h}(\omega; \phi, \theta)\; \hat{I}_{src}(\omega; \phi, \theta)$$

The $\hat{h}(\omega; \phi, \theta)$ and $\hat{I}_{src}(\omega; \phi, \theta)$ factors seem to be confounded here. For example, one obtains the same value by multiplying $\hat{h}(\omega; \phi, \theta)$ by some constant c and multiplying $\hat{I}_{src}(\omega; \phi, \theta)$ by $\frac{1}{c}$. This is similar to how in color constancy the illumination spectrum is confounded with the reflectance spectrum.

To avoid this confound, the auditory system needs to make an assumption about the source. Consider a noise source sound $I_{src}(t) = n(t; \phi, \theta)$ coming from direction (ϕ, θ). This noise sound has roughly constant amplitude spectrum in all frequency bands. Or consider an impulse sound that has roughly equal components at all frequencies – e.g. an impact, or an unvoiced stop sound p, k, t or an s sound. What can be concluded in these cases?

The source $I_{src}(t, \phi, \theta)$ is has a flat amplitude spectrum in the different bands then the measured signal $I^j()$ in the different bands will follow the peaks and valleys of the HRTF for that (ϕ, θ). So if there is a peak or notch in the measured $\mid I^j(\phi, \theta) \mid$ for some band j, and the auditory system assumed that the peak or notch was due to the HRTF $\hat{h}(\omega, \phi, \theta)$, then it could identify candidate (ϕ, θ) that would produce the peak or notch.

The *pinnal notch* is an example of how a monaural cue can be used. For sources in the medial plane, there are no binaural cues – no timing or level differences between the ears – but one can perceive the elevation to some extent. Monoaural cues must play some role here. It is believed that the pinnal notch in particular is used. (Recall lecture 21, p. 3-4). If one band of frequencies gives no response but most of the others do, this notch in the response is extremely unlikely to be due to the source. Rather it is mostly likely due to a notch in the HRTF.

[In the lecture, I briefly discussed an experiment done about 20 years ago to show the people could learn a new HRIR over a period of a few weeks. Details are omitted in these lecture notes.]