## Sound impulse

Consider an isolated perturbation of air pressure at 3D point $(X_o, Y_o, Z_o)$ and at time $t = t_0$, for example, due to some impact. Or, you can imagine a digital sound generation system with a speaker that generates a pulse. The idea is that pressure is constant (complete silence) and then suddenly there is an instantaneous jump in pressure at some particular spatial location.

Mathematically, we could model this sound pressure perturbation as an *impulse* function

$$I(X, Y, Z) = \delta(X - X_o, Y - Y_o, Z - Z_o)$$

at $t = t_0$. But how does this impulse evolve for $t > 0$ ? Think of a stone dropped in the water. After the impact, there is an expanding circle. Because sound also obeys a wave equation, the same phenomenon of an expanding circle occurs, except now we are in 3D and so we get an expanding *sphere*. The speed of the wavefront is the speed of sound. After one millisecond, the sphere is of radius 34 cm. After two milliseconds, the sphere is of radius 68 cm, etc.

The expanding sphere will have a finite thickness since the sound impulse will have a very short duration rather than be instantaneous. This thickness will not change as the sphere expands: the leading and trailing edge of the expanding sphere (separated by the small thickness of the sphere) will both travel at the speed $v$ of sound.

How does the level of the sound change as the sphere expands? Obviously there will be a falloff in level as the sphere expands, as we know from experience. Sound sources that are close to the ear are louder than those that are far from the ear, other other things being equal. But what exactly is the falloff rate?

According to physics which I will not explain (since it is subtle and this isn't a physics course), the total energy of a sound in some finite volume and at some instantaneous time $t$ after the impact is proportional to the sum of the squared pressure $I(X, Y, Z)^2$ over that volume. The energy of an expanding impulse is distributed over a thin spherical shell of volume $4\pi r^2 \Delta r$ where

$$r = \sqrt{(X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2}$$

and

$$r = v \cdot t$$

where $v$ is the speed of sound, and $\Delta r$ is the thickness of the shell which is constant over time.

If we ignore for now the loss of energy over time which is due to friction/attenuation in the air (and which in fact can be substantial for high frequencies) then the energy of the sound becomes distributed over a shell whose volume grows as $r^2$. This implies that the energy per unit volume in the shell shrinks like $\frac{1}{r^2}$. This means that the values of $I^2$ shrink like $\frac{1}{r^2}$, which means that $I$ falls off like $\frac{1}{r}$.

So, let $I_{src}$ be a constant that indicates the strength of the impulse which occurs at time $t = 0$. Then at a distance $r$ away from the source, when the impulse reaches that distance, the sound pressure will have fallen to

$$I(t) = \frac{I_{src}(t_0)}{r} \delta(r - vt).$$

In particular, if the point $(X_0, Y_0, Z_0)$ where the impulse occurs is far from the origin, then the impulse will reach the origin at time $t = \frac{v}{r}$ and the sphere can be approximated locally by a plane.

Finally, note that a real sound source won't be just a single impulse, but rather will have a finite time duration. Think of a person talking or shaking keys, etc. Even an impact that seems to have quite a short duration will in fact have a duration over tens of milliseconds. We can model a more general sound source that originates at some 3D position as a sum of impulses, and the sound heard at a distance $r$ from the source has pressure:

$$I(t) = \sum_{t_0} \frac{I_{src}(t_0)}{r} \delta(r - v(t - t_0)).$$

## Interaural Timing Differences

To compare the arrival time difference for the two ears, we begin with a simplified model to relate the pressure signals measured by the left and right ears:

$$I_l(t) \ = \ \alpha I_r(t - \tau) \ + \epsilon(t)$$

where $\tau$ is the time delay, $\alpha$ is a scale factor that accounts for the shadowing of the head, and $\epsilon(t)$ is an error term that is due to factors such as noise and to approximations in the model.

The auditory system is not given $\alpha, \tau$ explicitly, of course. Rather it has to estimate them. We can formulate this estimation problem as finding $\alpha, \tau$ that minimizes the sum of squared errors:

$$\sum_{t=1}^{T} \ (I_l(t) \ - \ \alpha I_r(t - \tau))^2 \ . \tag{1}$$

Intuitively, we wish to shift and scale the right ear's signal by $\tau$ so that it matches the left ear's signal as well as possible. If the signals could be matched perfectly, then the sum of square differences would be zero.

Note that, to find the minimum over $\tau$, the auditory system only needs to consider $\tau$ in the range $[-\frac{1}{2}, \frac{1}{2}]$ ms, which is the time it takes sound to go the distance between the ears.

Minimizing (1) is equivalent to minimizing

$$\sum_{t=1}^{T} I_l(t)^2 + \alpha^2 \sum_{t=1}^{T} I_r(t - \tau)^2 \ - \ 2 \ \alpha \sum_{t=1}^{T} I_l(t) I_r(t - \tau) \quad .$$

The summations in the first two terms are over slightly different intervals because of the shift $\tau$ in the second term. However, if $\tau$ is small relative to $T$, then the second summation will vary little with $\tau$. The third term in the summation is the one that depends heavily on $\tau$, since when the signals line up properly, $I_l(t) \approx \alpha I_r(t - \tau)$ and so $I_l(t) I_r(t - \tau)$ will be positive for all $t$ and the sum will be a large number.

With these assumptions, one can find the $\tau$ that maximizes

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

This summation is essentially the cross-correlation of $I_l(t)$ and $I_r(t)$, so one can find the $\tau$ that maximizes the cross-correlation of sound pressures measured in the two ears over a small time interval.

To estimate $\alpha$, one could use the model:

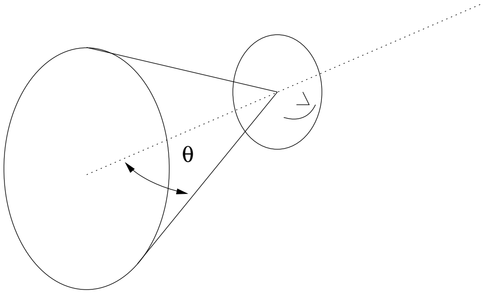
$$I_l(t) \approx \alpha I_r(t-\tau)$$

and so

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

## Cone of confusion

Note that timing differences do not uniquely specify direction. Consider the line passing through the two ears. This line and the center of the head, together with an angle $\theta \in [0, \pi]$, defines a *cone of confusion*. If we treat the head as an isolated sphere floating in space, then all directions along a single cone of confusion produce the same intensity difference and the same timing difference. The reason is that the sphere is symmetric about the line between the ears, so all points that a fixed distance from the cone apex (i.e. any circle shown in the figure) are equivalent. A source at any of those points would produce the same shadowing effects – i.e. level difference – and the same timing difference between the two ears.



Does the cone of confusion provide an ultimate limit on our ability to detect where sounds are coming from? No it doesn't and the reason is that the head is not a sphere floating in space. The head is attached to the body (in particular the shoulders) which reflects sound in an asymmetric way, and the head has ears (the pinna) which shape the sound wave in a manner that depends on the direction from which the sound is coming. As we will see in an upcoming lecture, there is an enormous amount of information available which breaks the cone of confusion.

## Outer Ear

We next turn to how the sound that arrives at the ear is transformed when it enters the ear. Then we'll examine the processing of this sounds within the ear.

Let's clarify what we mean by "ear". We think of our ears as the two fleshy appendages on the side of the head. These appendages are called the *pinnae* (one pinna, two pinnae). Pinnae are not involved in the sensing of the sound waves but they do have a role in hearing, namely in changing the shape of the sound wave.

Each pinna leads to a tube-like cavity called the *auditory canal*. At the end of this canal is the *ear drum (tympanic membrane)* which vibrates in response to air pressure variations. The ear drum marks the boundary between the *outer ear* and the *middle ear.* I will discuss the middle and inner ear later. For now, let's consider how the sound that arrives at the ear gets transformed when it enters the ear.

## Head related impulse response (HRIR)

When we noted the timing difference between the left and right ears, we assumed that space between the ears was empty and that sound travelled freely between the ears without interruption, reflection, etc. This is not the case, however. A person's head transforms an incoming sound and it does so in a direction-dependent way.

For any incoming direction $(\theta, \phi)$ of a sound wave relative to head coordinates, consider the impulse function $\delta(r - v(t - t_0))$ which leaves from a position a distance $r$ away at time $t_0$. The head, ear, shoulders deform this wave of sound. This deformation is a combination of shadowing and reflections. One typically does not model the physics of this. Instead, we one can just measure how an impulse function is transformed (see below). When there is an impulse from direction $(\theta, \phi)$, the sound pressure wave that is measured inside the head is a function $h(t; \theta, \phi)$. This is known as the *head related impulse response* function.

The $\theta$ and $\phi$ define a spherical coordinate system, with the poles being directly above and below the head. The angle $\theta$ is the *azimuth* and goes from 0 to 360 degrees (front, left, behind, right). The angle $\phi$ is the elevation and goes from -90 (below) to 90 degrees (above). Note that this spherical coordinate system is different from the one used in the cone of confusion above, where the poles were directly to the left and right.

Think of the *head* as a filter, which transforms an incoming sound wave. For a general incoming sound wave $I(t; \phi, \theta)$ arriving at the left ear from direction $(\theta, \phi)$, this incoming sound wave would be transformed by the ear by convolving with the head related impluse response function. Letting subscript $l$ stand for left ear:

$$I_l(t; \theta, \phi) \ = \ h_l(t; \theta, \phi) * I_{src}(t; \phi, \theta) \ .$$

To understand why this is a convolution, think of the original source as a sequence of impulses and each of these impulses gets transformed in the same way, and the resulting sound is just the sum.

Similarly, the sound pressure function measured at the right ear would be

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

A few points to note: First, obviously both the left and right ear are at different locations in space, so the $h_l$ and $h_r$ must be suitably shifted in time relative to each other. Second, we are assuming that there is only a single source direction $(\phi, \theta)$. If we had multiple sound sources in different directions, then we would need to sum up the sound pressures $I(t; \phi, \theta)$ from different $\phi, \theta$. Third, the functions $h_l$ and $h_r$ vary from person to person, since they depend on the shape of the person's body (head, ear, shoulders). For any single person, though, the $h_l$ is a typically a mirror reflection of $h_r$, where the mirror reflection is about the (medial) plane of symmetry of the person's body, so

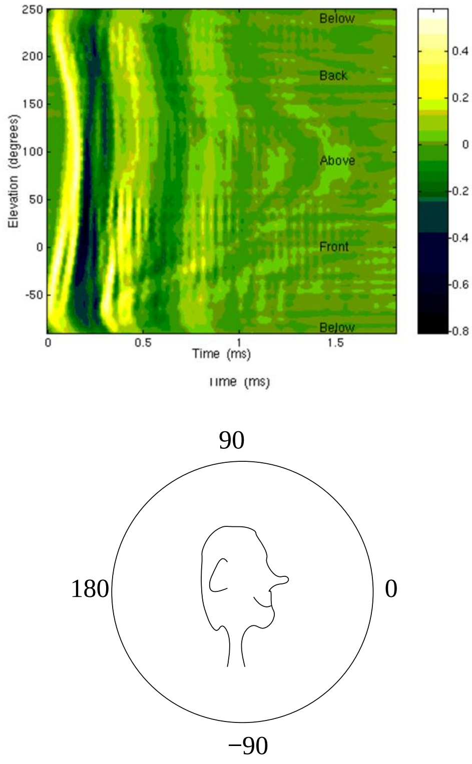
$$h_r(\phi, \theta) = h_r(\phi, -\theta)$$

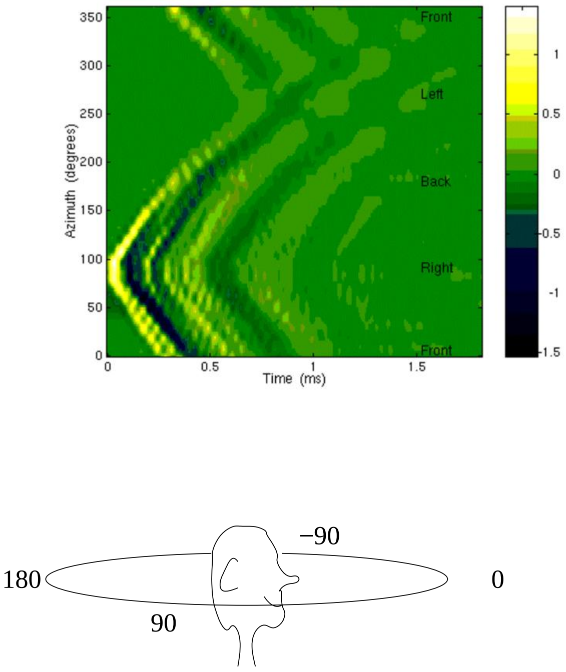
where $\theta = 0$ is the straight ahead azimuth.

To measure the function $h(t;\phi,\theta)$ for a given person, one can place a tiny microphone inside the person's auditory canal and then record the sound produced by an impulse source function at some distance away in direction $(\phi,\theta)$, and repeat the experiment for many different directions $\phi,\theta$. Another approach is to work with a model human, such as a mannequin similar to what you find in a clothing store, but one that has holes in the ears. Such mannequins have been developed for scientific study of HRIR functions: see

`http://kemar.us`

Examples of measurements using a KEMAR mannequin are shown below. Data taken from `http://interface.cipic.ucdavis.edu/`. That web site also has some nice tutorials).





ELEVATION AZIMUTH

On the left is a set of HRIR functions for elevation directions in the medial plane YZ. On the right are HRIR functions for the azimuth directions in the horizontal plane XZ. Time is sampled every 6 $\mu$s (1 $\mu$s $= 10^{-6}$ s), so 100 samples corresponds is 0.6 ms, which is about the time it takes for sound to travel the width of the head.

For the elevation plot, we see that the impulse responses (rows) do vary with elevation. Each impulse becomes a small wave (positive, negative, positive, negative,...) but the exact details vary continuously from row. As we will see later, this provides some information to distinguish elevations. [One minor observation is the presence of a diagonal streak with a long delay. The authors claim this is due to a reflection off of the floor.]

For the azimuth plot on the right, there is a systematic delay in the HRIR with azimuth. The earliest the sound reaches the ear drum is when the sound comes from the right, when $\theta = 90$. The latest that the sound reaches the ear drum is when the sound is coming from $\theta = 270$ deg which is from the left. These systematic delays are qualitatively consistent with the cone of confusion argument earlier. However, notice that the HRIR function has more details, namely an impulse is transformed to a wave with positive and negative values

## Head related transfer function (HRTF)

If we take the Fourier transform of

$$I_l(t;\theta,\phi) \; = \; h_l(t;\theta,\phi) * I(t;\phi,\theta)$$

and apply the convolution theorem, we get:

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

where $\hat{h}_l(\omega;\theta,\phi)$ is called the *head related transfer function.* The term “transfer function” has very general usage. In the context of this course, it refers to the Fourier transform of a filter which is convolved with a signal.

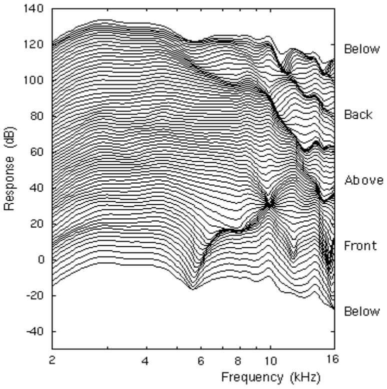
Essentially what we are doing here is decomposing the incoming sound $I_{src}(t;\phi,\theta)$ which is arriving from just one direction into its frequency components and looking at how each frequency component gets transformed by the head and ear. The HRTF $\hat{h}_l(\omega;\theta,\phi)$ is a complex number which specifies a gain (amplitude) and phase shift for each temporal frequency $\omega$ of the incoming wave. You can think of the gain and phase shift as the net effect of shadowing of the head, and reflections off the shoulder and pinna, and any attenuation or amplification inside the auditory canal.

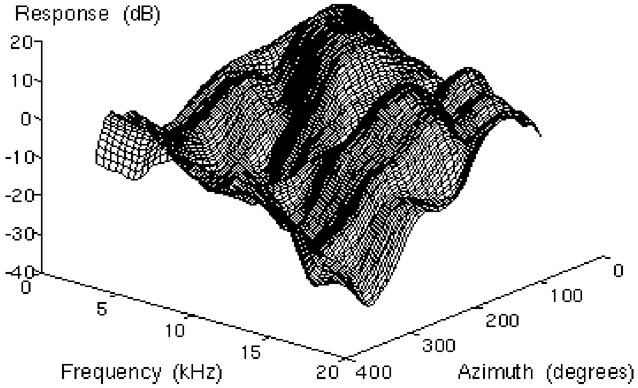
The figure below shows the amplitude spectra $|\hat{h}_l(\omega;\theta,\phi)|$ of HRTF functions, for (left column) the circle of directions in the medial plane dividing the head, and (right column) the horizontal plane i.e. azimuth varying. Only frequencies above 2 kHz are shown. This corresponds to wavelengths of 17 cm or less.

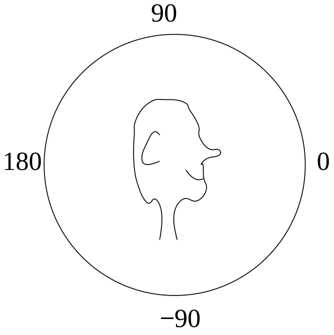
There is a notch (local minimum of the HRTF) at about 6 kHz for sounds coming from the front and below. This is known as the *pinnal notch* because it is believed to be due to the pinna. (The notch disappears when the pinna is removed from the mannequin.) Any energy in the incoming sound from some direction will have severely attenuated energy within the frequencies of the pinna notch. Thus, the *absence* of energy in particular frequency bands is evidence that sound is coming from a certain direction. We will return to this idea next lecture.

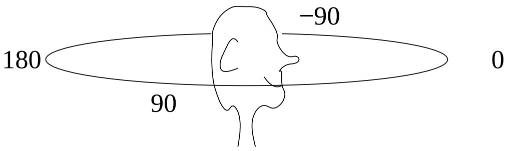
For the azimuth plot on the right, note the general falloff in the transfer function from 90 degrees azimuth down to 270 degrees azimuth. This is due to shadowing of the head. The falloff is pronounced at high frequencies, where the heights at 90 and 270 degrees differ by about 30 dB.

The HRTF is a function of three variables $\omega, \theta, \phi$. The above plots showed 2D slices for a fixed $\theta$ or a fixed $\phi$. You can see examples of a 2D HRTF slice which is a function of $\theta, \phi$ for two different frequencies $\omega$ here: `https://auditoryneuroscience.com/topics/acoustic-cues-sound-location`









ELEVATION

AZIMUTH

## Middle ear

The HRIR function describes how a sound pressurve waves that arrive at the head are transformed by the head and outer ear. The ear drum vibrates in direct response to the sound pressure in its immediate neighborhood in the ear canal. The ear drum marks the end of the outer ear.

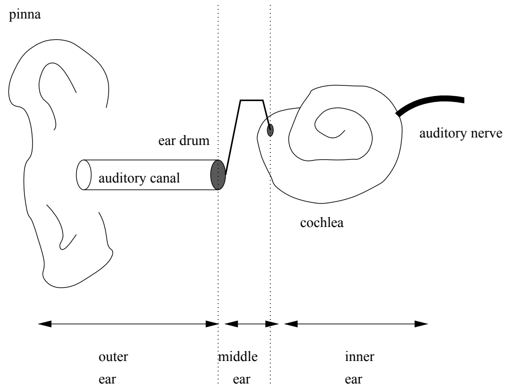
Beyond the ear drum is middle ear. The middle ear is an air filled cavity [1] behind the ear drum. The middle ear contains a rigid chain of three small bones (*ossicles*). One end of the chain is attached to the ear drum and the other end is attached to the *cochlea* which is part of the *inner ear*. The ossicles act as a lever, transferring large amplitude oscillations of the eardrum to small amplitude (but large pressure) oscillations to the base of the cochlea. Next we will examine how these vibrations are encoded by the nervous system.

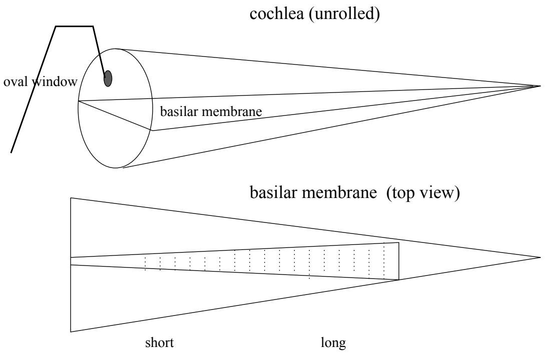
## Inner ear (intro only)

The cochlea is a fluid-filled snail-shaped organ which contains the nerve cells that encode the pressure changes. If we would unwind the cochlea[2], it would have a cone-like shape: the thick end is the *base* and the thin end is the *apex*. The interior of the cochlea is partitioned into two vestibules which are separated by long triangular membrane called the *basilar membrane*. For simplicity, think

[1] This cavity is connected to various other cavities in the head, i.e. mouth and nasal cavities, which is why it can get infected. Children in particular often get middle ear infections.

[2] We cannot unwind it because the shell is hard. Indeed it is said to be the hardest bone in the body!





of the basilar membrane as containing both hair cells (mechanoreceptors) and ganglion cells (which will send spike trains to the brain). Think of these hair and ganglion cells as laying on an inverted triangle of elastic fibres that reach across the membrane. By "inverted", I mean that the fibres are shorter at the base of the cochlea and longer at the apex of the cochlea (see sketch above). In fact the anatomy is more detailed than this, but the details do not concern us here.

Different positions along the length of the basilar membrane contain transverse fibres that oscillate in response to different temporal frequency components of the sound wave. Think of these fibres as piano strings, but only the fundamental vibration occurs, i.e with the fibre length being a half cycle. The basilar membrane responds best to low temporal frequencies (long wavelengths) at the far end (apex) where the fibres are longest, and it responds to high temporal frequencies (short wavelengths) at the near end (base) where the fibres are shortest. By "respond" here, I just mean that it oscillates at these frequencies. If you recall the theory of a vibrating string with $\omega = \frac{c}{L}$, you can think of both $c$ and $L$ varying along the fibres of the basilar membrane. You can get higher frequencies by increasing $c$ (higher tension) and decreasing $L$.

See the nice demo here:
`https://auditoryneuroscience.com/topics/basilar-membrane-motion-0-frequency-modulated-tone`

Next lecture we will discuss how these oscillations in the basilar membrane are coded.